Mathematical modeling and statistical inference to better understand arbovirus dynamics

Clara Champagne

NNT : 2018SACLG006

THÈSE DE DOCTORAT

de

L'UNIVERSITÉ PARIS-SACLAY

École doctorale de mathématiques Hadamard (EDMH, ED 574)

Établissement d'inscription : École nationale de la statistique et de l'administration économique

Établissement d'accueil : Institut de biologie de l'ENS

Laboratoire d'accueil : ENSAE-X Centre d'économie, statistique et sociologie, UMR 9194 CNRS

Spécialité de doctorat : Mathématiques aux interfaces

Clara CHAMPAGNE

Modélisation mathématique et inférence statistique pour une meilleure compréhension des dynamiques des arboviroses

Date de soutenance : 11 Décembre 2018

Après avis des rapporteurs :
HENRIK SALJE (Institut Pasteur)
VIET CHI TRAN (Université de Lille)

Jury de soutenance :

BERNARD CAZELLES	(ENS - UPMC) Codirecteur de thèse
NICOLAS CHOPIN	(CREST) Président du jury
HENRIETTE DE VALK	(Santé Publique France) Examinatrice
HENRIK SALJE	(Institut Pasteur) Rapporteur
VIET CHI TRAN	(Université de Lille) Rapporteur
ALEXANDRE TSYBAKOV	(CREST) Codirecteur de thèse

Résumé

L'importance et l'expansion des arboviroses comme la dengue ou le virus Zika nécessite des modèles pour mieux comprendre et prédire leurs dynamiques. La propagation vectorielle de ces maladies est influencée par de multiples facteurs humains et environnementaux qui rendent complexe la construction de modèles épidémiologiques parcimonieux. Parallèlement, de nombreux outils théoriques et computationnels existent désormais pour confronter ces modèles aux données observées. L'objectif de ce travail de thèse est donc d'apporter l'éclairage des données sur les modèles de propagation des arboviroses. Dans un premier temps, il s'agit d'identifier les éléments les plus importants à incorporer pour modéliser les dynamiques de la dengue en milieu rural, dans la région de Kampong Cham (Cambodge). Différents modèles sont comparés, complexifiant à la fois le détail de l'histoire de la maladie et la prise en compte des formes de stochasticité. Dans le cadre déterministe, on a pu souligner l'importance des interactions entre sérotypes, et le faible intérêt pour la représentation explicite des moustiques vecteurs et des individus asymptomatiques. Par ailleurs, la prise en compte des incertitudes indique qu'une large part de la dynamique est capturée seulement par la stochasticité et non par les éléments du squelette déterministe du modèle. Aussi étudie-t-on dans un second temps d'autres aspects de la transmission de la dengue, comme la saisonnalité et la structure spatiale, grâce à des données d'épidémies à Rio de Janeiro (Brésil). Dans un dernier temps, ces méthodes et modèles sont appliqués à l'étude d'un arbovirus émergent, le virus Zika. A partir de données d'épidémies survenues dans le Pacifique, les paramètres-clé de la propagation du virus sont estimés dans le cadre stochastique, et leur variabilité est envisagée à la fois en termes de contexte géographique et de modèle épidémiologique, par la comparaison de quatre îles et de deux modèles à transmission vectorielle. Par ailleurs, la question des interactions potentielles du virus Zika avec celui de la dengue est explorée.

Abstract

Arboviruses such as the dengue and Zika viruses are expanding worldwide and modeling their dynamics can help to better understand and predict their propagation, as well as experiment control scenarios. These mosquito-borne diseases are influenced by a multiplicity of human and environmental factors that are complex to include in parsimonious epidemiological models. In parallel, statistical and computational tools are nowadays available to confront theoretical models to the observed data. The objective of this PhD work is therefore to study arbovirus propagation models in the light of data. Firstly, in order to identify the most important elements to incorporate in models for dengue dynamics in a rural setting, several dengue models are compared using data from the Kampong Cham region in Cambodia. Models incorporate increasing complexity both in the details of disease life history and in the account for several forms of stochasticity. In the deterministic framework, including serotype interactions proved decisive, whereas explicit modeling of mosquito vectors and asymptomatic infections had limited added value, when seasonality and underreporting are already accounted for. Moreover, including several forms of uncertainties highlighted that a large part of the disease dynamics is only captured by stochasticity and not by the elements of the deterministic skeleton. Therefore, secondly, we explore other aspects of transmission, such as seasonality and spatial structure, in the case of dengue epidemics in Rio de Janeiro (Brazil). Finally, the models and estimation methods are applied to study an emerging arbovirus, the Zika virus. Using data from epidemics in the Pacific, we estimate the key parameters of disease propagation in the stochastic framework and explore their variability in terms of geographic setting and model formulation by comparing four islands and two models with vector-borne transmission. In addition, potential interactions with the dengue virus are explored.

Remerciements

Ce travail n'aurait pas vu le jour sans l'aide précieuse de très nombreuses personnes, que je tiens particulièrement à remercier.

En tout premier lieu, je tiens à remercier Bernard Cazelles pour m'avoir donné le cap, guidée, et tant appris tout au long de ces années de thèse. Un grand merci pour ses relectures en vacances, ses conseils, sa patience, sa disponibilité et son soutien pendant ce long périple.

Je suis profondément reconnaissante à Alexandre Tsybakov pour m'avoir fait confiance afin de mener à bien ce projet interdisciplinaire, pour m'avoir soutenue et orientée durant ces années, ainsi que pour ses perçantes remarques mathématiques.

Je souhaite remercier Henrik Salje et Tran Viet Chi pour avoir accepté d'évaluer ce travail de thèse. Je remercie Nicolas Chopin, pour ses précieux conseils, ainsi que pour avoir accepté de prendre part au jury. Je remercie également Henriette De Valk pour avoir accepté de participer au jury.

Un très grand merci à Joseph Dureau, co-créateur de la librairie Ssm grâce à laquelle les analyses de ce travail ont pu être réalisées, et qui a su m'aider avec patience à apprendre à l'utiliser. Merci aussi à Anton Camacho, pour ses astuces et ses interfaces R pour la librairie Ssm.

Je suis reconnaissante à toute l'équipe du département d'Epidémiologie et Santé publique de l'Institut Pasteur du Cambodge, et l'ensemble de l'équipe impliquée dans le projet DenFREE. Je tiens à remercier en particulier Ly Sowath pour son aide et ses réponses tout comme son accueil à Phnom Penh ; Arnaud Tarantola pour son accueil et ses explications sur la dengue lorsque, alors en stage, j'était en mission dans son unité ; Souv Kimsan, pour l'accès aux données du système de surveillance et pour sa réactivité ; et aussi l'équipe que j'ai accompagnée quelques jours sur les routes de Kampong Cham, Julia Ledien, Sorn Sopheak, et notre chauffeur Touch Saravoan.
Je remercie également Richard Paul pour ses relectures attentives et ses précieuses remarques, épidémiologiques, entomologiques et même linguistiques.

Je remercie l'équipe du Lfa au Crest, auprès de qui j'ai beaucoup appris lors de séminaires et de discussions, à Malakoff comme à Saclay. A l'Ensae, je tiens aussi à remercier toutes les personnes avec qui j'ai interagi lors de mes TD, professeurs, assistants, chargés, étudiants. Je remercie Pierre Alquier pour son suivi et son aide durant ces années avec l'Edmh.
Je souhaite remercier l'équipe Eco-Evolution mathématique, à l'Ibens et au Ceres, pour leurs conseils et leur enthousiasme, de l'organisation des lab-meetings aux discussions du midi, en passant par les déménagements et les explorations culinaires. Merci également à Pierre Vincens et Mathieu Bahin pour m'aider à percer les secrets du cluster. Et un immense merci à David

SALTHOUSE pour m'avoir aidée en tant de circonstances, par ses conseils en mathématiques et informatique et par son amitié (et aussi par de très nombreux cafés, thés, chocolats, ...)

Ces années de thèse n'auraient pas été les mêmes sans les doctorants, post-docs et stagiaires avec qui j'ai partagé des bureaux et des discussions passionnées, grâce à leurs idées et leur bonne humeur, à l'IBENS avec Benjamin N., Benjamin R., Billy, Boris, Célian, Christophe, Elsa, Emmeline, Guilhem, Julia, Mathieu, Orso, Phuong, Pierre-Antoine, Stéphane ; tout comme au LFA avec Alexis, Anna, Jalel, Ophélie, Sébastien. Un merci tout particulier à Alexis, Benjamin N. et Phuong pour leurs relectures !

Une pensée pour les amis de toujours et d'un peu partout qui se reconnaîtront, entre Toulouse et Paris, en passant par Zürich, le Castéra, Tournefeuille, Amiens, Saint-Ouen, Malakoff, Strasbourg, Londres, Sceaux, ... Pour les amis musiciens, aussi, grâce à qui je fais le plein d'énergie et de bonne humeur à chaque répétition.

Enfin, je dois énormément au soutien indéfectible de ma famille, mes parents, ma grand-mère, mon frère, malgré les kilomètres qui séparent Toulouse et Paris, par l'exemple qu'ils me donnent, leurs conseils, leurs encouragements et leur réconfort.
Zu guter Letzt, ein riesiges Dankeschön geht an Lucas, sans qui ce document ne serait même pas compilé, pour son aide de toujours, ses relectures, ses débug, et surtout son soutien et sa patience au quotidien.

Contents

1 Introduction **1**
- 1.1 Eléments de modélisation mathématique des maladies infectieuses 2
- 1.2 Méthodes d'estimation 10
- 1.3 Modéliser les arboviroses 26
- 1.4 Objectif de la thèse (résumé substantiel en langue française - I) 32

2 Comparaison de modèles de dengue **35**
- 2.1 Introduction 35
- 2.2 Dengue modeling in rural Cambodia: statistical performance versus epidemiological relevance 37
- 2.3 Comparison of stochastic and deterministic frameworks in dengue modeling . . . 62

3 Introduire l'aspect spatial **83**
- 3.1 Introduction 83
- 3.2 Material and Methods 84
- 3.3 Results 89
- 3.4 Discussion 104

4 Une autre arbovirose : le virus Zika **109**
- 4.1 Introduction 109
- 4.2 Structure in the variability of the basic reproductive number (R_0) for Zika epidemics in the Pacific islands 110
- 4.3 Interactions between Zika and dengue 138

5 Discussion et perspectives **153**
- 5.1 Synthèse des résultats (résumé substantiel en langue française - II) 153
- 5.2 Discussion et difficultés rencontrées 154
- 5.3 Perspectives 156

A Annexe: Dengue modeling in rural Cambodia: statistical performance versus epidemiological relevance **163**

B Annexe: Comparison of stochastic and deterministic frameworks in dengue modeling **171**

C Accounting for non-stationarity in epidemiology by embedding time-varying parameters in stochastic models **182**

Bibliographie **209**

Chapter 1

Introduction

Alors que les progrès en termes d'hygiène, d'assainissement et de vaccination ont permis une très forte diminution des maladies infectieuses depuis le XIX^{e} siècle, la seconde moitié du XX^{e} siècle a, elle, été marquée par une expansion des maladies émergentes (Weiss and McMichael, 2004; Jones et al., 2008). Les mutations sociales et environnementales, notamment liés au changement climatique, à l'urbanisation ou à l'intensification des échanges, sont des facteurs importants de ce phénomène. Elles favorisent en effet le développement de pathogènes dangereux pour l'homme et leur propagation à grande échelle. Les maladies vectorielles, parmi lesquelles celles transmises par les moustiques, représentent plus de 20% des maladies émergentes depuis les années 1950 (Jones et al., 2008), et constituent un fort enjeu de santé publique.

Transmise par le biais des moustiques de type *Aedes*, la dengue est une maladie virale qui touche l'ensemble des régions tropicales et sub-tropicales du globe (Guzman et al., 2010) et qui connaît une forte expansion jusque dans les régions tempérées (Rezza, 2016). Avec plus de 50 millions d'infections et environ 500 000 hospitalisations chaque année (Guzman et al., 2010), elle représente un enjeu majeur de santé publique au niveau international. Elle a de surcroît de lourdes conséquences économiques à l'échelle familiale autant que nationale, notamment dans les pays où le niveau de vie est faible. Un proche virus, celui du chikungunya, a, lui, été à l'origine de grandes épidémies dans l'océan Indien en 2005-2006 (Renault et al., 2007), puis sur le continent américain à partir de 2013 (Roche et al., 2017; Riou et al., 2017), étant associé à de sévères symptômes de fièvres et de douleurs articulaires, voire à des cas de décès. Plus récemment, le virus Zika, autre virus proche de celui de la dengue, a suscité un émoi international suite aux épidémies de microcéphalie au Brésil en 2015. Ayant été associé à de graves complications neurologiques, chez l'adulte mais aussi sur le développement pré-natal, son étude a été considérée comme une priorité de santé publique (WHO, 2016b).

La dengue, le chikungunya, le Zika, et plus généralement les maladies virales transmises par le biais des arthropodes (dites arboviroses), nécessitent donc d'être mieux comprises, à la fois pour mieux appréhender les mécanismes de leur expansion et pour proposer des solutions de contrôle. Toutefois la propagation de ces maladies est complexe, car elle fait intervenir des facteurs immunologiques, écologiques, entomologiques et humains, au croisement de nombreux champs disciplinaires. Les outils mathématiques et statistiques sont eux aussi mobilisés, les premiers pour modéliser l'interaction entre tous ces facteurs, afin de tester des hypothèses scientifiques ou d'explorer des méthodes de contrôle, et les seconds pour confronter ces modèles aux données observées. Ce travail de thèse s'intéresse particulièrement à l'éclairage porté par les données sur les modèles mathématiques de propagation des arboviroses.

Aussi cette introduction a-t-elle pour but de familiariser le lecteur avec les trois éléments sur lesquels se fondent les chapitres suivants. Dans un premier temps, on présentera des éléments de modélisation mathématique pour représenter de façon mécaniste la propagation des maladies infectieuses chez les humains, dans le cadre déterministe et dans le cadre stochastique. Dans un second temps, on introduira différentes méthodes existantes pour estimer ces modèles à l'aide de données observées. Enfin, dans un dernier temps, on décrira les arboviroses étudiées, la dengue et celle due au virus Zika, et présentera les différents modèles qui ont été proposés pour représenter leurs dynamiques.

1.1 Eléments de modélisation mathématique des maladies infectieuses

L'une des premières tentatives pour représenter le développement d'une épidémie par un modèle mathématique remonte à 1760 et aux travaux de Daniel Bernoulli sur la variole (Bernoulli, 1760). Toutefois, la modélisation des maladies infectieuses a surtout pris son essor au début du XX^e siècle, avec les travaux de Hamer, Soper, Ross et Kermack et McKendrick sur des modèles déterministes et ceux de Reed, Frost, McKendrick et Greenwood qui incorporent une dimension aléatoire dans la propagation des infections (cf. Bailey (1975), chapitre 2, pour une description historique de l'évolution de ces approches). Nous introduisons ici les éléments de modélisation principaux en présentant d'une part les modèles déterministes (section 1.1.1) et d'autre part les modèles stochastiques (section 1.1.2).

1.1.1 Modèles déterministes

Le modèle SIR

Le modèle dit SIR, présenté et formalisé par Kermack and McKendrick (1927), constitue une référence à partir de laquelle de nombreux modèles épidémiologiques sont construits (Keeling and Rohani, 2008). On suppose que l'ensemble de la population, de taille constante N, se répartit en trois classes homogènes : les individus susceptibles S, infectés I et rétablis R (ou retirés, c'est -à-dire qu'ils sont immunisés contre la maladie et ne participent plus à la transmission), ce qui implique $S+I+R = N$. La dynamique entre ces trois compartiments est représentée par le système suivant:

$$\begin{aligned} \frac{dS}{dt} &= -\beta \frac{SI}{N} \\ \frac{dI}{dt} &= \beta \frac{SI}{N} - \gamma I \\ \frac{dR}{dt} &= \gamma I \end{aligned} \tag{1.1}$$

Dans ce cadre, γ correspond au taux de guérison (ou de la fin de la transmissibilité) et β à l'intensité de la transmission, que l'on peut décomposer comme le produit d'un taux de contacts entre les individus et la probabilité qu'un contact conduise à une infection. La formulation selon la loi d'action de masse $\beta \frac{I}{N}$ implique que la force d'infection est proportionnelle à la part d'individus infectés dans la population (alternativement, il est possible de formuler le modèle selon la pseudo action de masse, βI, ces choix n'ayant un implication que lorsque la population varie) (Keeling and Rohani, 2008).

L'analyse de ce système permet de mettre en évidence un propriété fondamentale sur sa stabilité: si $\frac{\beta}{\gamma}\frac{S}{N} > 1$, l'épidémie se poursuit, mais si $\frac{\beta}{\gamma}\frac{S}{N} < 1$, on observe un déclin du nombre d'infectés et l'épidémie s'éteint d'elle-même. Le taux de reproduction de base, noté $R_0 = \beta/\gamma$, s'interprète comme le nombre moyen d'infections secondaires suite à l'introduction d'un infecté dans une population entièrement susceptible. L'étude analytique du système permet de conclure que l'épidémie peut s'éteindre sans épuiser la totalité des susceptibles présents initialement. De surcroît, la vaccination d'une fraction $x = 1 - \frac{1}{R_0}$ de la population permet de maintenir le taux de reproduction effectif $R_e = \frac{\beta}{\gamma}\frac{S}{N}$ inférieur à 1, et ainsi empêcher la transmission de la maladie.

L'étude du R_0 associé à un pathogène permet donc une quantification de sa capacité de propagation, et peut être utile pour orienter des décisions de santé publique. L'obtention de l'expression de R_0 associée à un modèle épidémiologique n'est toutefois pas toujours directe et peut nécessiter des outils comme la matrice Next Generation (Diekmann et al., 2010).

Extensions du modèle SIR déterministe

Le modèle SIR est un modèle très simple mais différentes extensions existent pour pallier les hypothèses sur lesquelles il repose.

Histoire de la maladie La segmentation de la population entre susceptibles, infectés et rétablis peut se révéler simpliste voire inadéquate pour de nombreuses maladies. On peut tout d'abord introduire une dynamique démographique, pour tenir compte des effets de naissance et de mort au sein de la population, notamment pour des modèles sur le temps long.
Par ailleurs, pour plus de réalisme biologique, on peut ajouter des compartiments dans le système pour modéliser d'autres états de la maladie spécifiques au type de pathogène étudié, comme un état d'incubation entre l'état susceptible et l'état infectieux, un état de porteur de la maladie (Keeling and Rohani, 2008), ou encore, par exemple, un état infectieux post-mortem (Legrand et al., 2007). Il est également possible de considérer les états des individus pour plusieurs pathogènes simultanément afin de modéliser d'éventuelles co-infections ou interactions (Opatowski et al., 2018), ou d'introduire plusieurs populations. En particulier, dans ce dernier cas, le modèle classique dit de Ross-Macdonald proposé pour représenter la malaria, et dont la genèse et les différentes versions sont analysées par Smith et al. (2012), comporte deux populations: les hôtes et les vecteurs (moustiques). Dans une des versions de ce modèle (Smith et al., 2012), présentée par Anderson and May (1992), les proportions d'hôtes et de vecteurs infectés sont notées respectivement h_i et v_i, conduisant au modèle suivant :

$$\begin{aligned} \frac{dh_i}{dt} &= mabv_i(1-h_i) - \gamma h_i \\ \frac{dv_i}{dt} &= ach_i(1-v_i) - \mu_V v_i \end{aligned} \tag{1.2}$$

où a correspond au taux de piqûre, b et c aux probabilités de transmission respectivement d'un vecteur à un hôte et vice-versa, m au nombre de moustiques par humain, γ au taux de guérison pour un individu et μ_V au taux de mortalité pour un vecteur. Ce modèle a connu de nombreuses extensions et adaptations à différents pathogènes à transmission vectorielle (Reiner et al., 2013).

Mixage homogène Le modèle SIR (tout comme le modèle de Ross-Macdonald) repose sur l'hypothèse de "mixage homogène". Le taux de transmission $\beta \frac{IS}{N}$ implique une linéarité entre S et I, ce qui signifie que les contacts sont aléatoires et homogènes entre tous les individus de la population. Dans le cadre du modèle de Ross-Macdonald, cela signifie que les vecteurs ont accès et piquent indifféremment tous les hôtes.
Afin de limiter cette hypothèse, une solution est d'introduire des sous-catégories au sein des compartiments et d'utiliser des taux de transmission spécifiques pour les interactions entre ces catégories. Ces catégories peuvent représenter des classes de risque, des classes d'âge, ou des sous-populations spatiales par exemple (Keeling and Rohani, 2008; Dye and Hasibeder, 1986). Ces effets de structure de la population peuvent être pris en compte de plusieurs manières : en considérant des classes discrètes au sein de la population (métapopulations), en supposant un continuum entre les différentes catégories à l'aide d'équations aux dérivées partielles ou dans le cadre de modèles individu-centrés.
Il est également possible de modifier les taux de transmission, pour introduire des relations non linéaires entre S et I, du type $\beta I^{\alpha_1} S^{\alpha_2}$ (Finkenstädt and Grenfell, 2002), ou selon de nombreuses autres expressions (Wang et al., 2017). Des effets de concentration de la population, de saturation de la force d'infection, ou de changement de comportement des individus en réaction à la couverture médiatique peuvent ainsi être pris en compte (Wang et al., 2017).

Paramètres statiques ou variables Dans le modèle (1.1), les paramètres sont statiques, mais on peut considérer des situations où ils fluctuent au cours du temps. Ceci peut permettre de refléter des effets saisonniers sociaux (Soper, 1929; London and Yorke, 1973) ou climatiques (Pascual and Dobson, 2005), ou des modifications des comportements humains (Verelst et al., 2016). Par exemple, le taux de transmission peut varier selon un forçage sinusoïdal, ou selon des relations fonctionnelles prédéfinies avec des variables reflétant le climat (Lourenço and Recker, 2014) ou le calendrier scolaire (Keeling and Rohani, 2008).
D'autres modifications du taux de transmission permettent de représenter l'évolution des comportements humains au cours du temps. Yu et al. (2017) utilisent une relation fonctionnelle avec une nouvelle variable d'état construite à partir de la mortalité observée pour représenter les phénomènes de distanciation sociale. Ainsi, plus la mortalité observée est grande, plus les individus adoptent des comportements qui limitent les contacts infectieux, et plus le taux de transmission baisse. Ils montrent notamment que ces phénomènes ont pour effet de réduire la taille finale de l'épidémie. Chen et al. (2018) proposent, eux, un modèle de l'impact des recommandations par les médias concernant l'hygiène ou les décisions de mise en quarantaine : certains paramètres du modèle changent de valeur lorsque le nombre d'infectés ou le nombre de susceptibles dépasse un certain seuil, reflétant le déclenchement de mesures de contrôle. Ils étudient ensuite les propriétés du système en fonction du choix de ces seuils, afin de guider leur calibration.
Une autre approche consiste à utiliser des modèles semi-mécanistes afin de laisser les variations temporelles des paramètres dépendre des données épidémiologiques observées (Cazelles and Chau, 1997; Ellner et al., 1998), à travers l'utilisation de splines (Laneri et al., 2010) ou de diffusion brownienne (Dureau et al., 2013b) (cf. Appendix C pour plus de détails sur ces méthodes).

Durée exponentielle La formulation du système sous forme d'équations différentielles repose sur l'hypothèse que la durée passée dans chaque compartiment suit une loi exponentielle, ce qui est peu réaliste en raison des propriétés de décroissance rapide et d'absence de mémoire de cette loi (Lloyd, 2001). Il est possible de modifier le système pour prendre en compte des durées suivant une autre distribution, le plus courant étant une loi Gamma ou une loi d'Erlang (somme d'un

nombre entier de variables exponentielles indépendantes de même paramètre). Cette dernière peut se modéliser facilement en décomposant les compartiments concernés en n étapes, par exemple $I = I_1 + \ldots + I_n$ (Lloyd, 2001):

$$\begin{aligned}
\frac{dS}{dt} &= -\beta\frac{SI}{N} \\
\frac{dI_1}{dt} &= \beta\frac{SI}{N} - n\gamma I_1 \\
\frac{dI_2}{dt} &= n\gamma I_1 - n\gamma I_1 \\
&\ldots \\
\frac{dI_n}{dt} &= n\gamma I_{n-1} - n\gamma I_n \\
\frac{dR}{dt} &= n\gamma I_n
\end{aligned}$$

L'utilisation de distributions d'Erlang à la place des distributions exponentielles peut avoir de fortes conséquences sur les propriétés dynamiques du système, comme l'illustre Lloyd (2001) dans le cadre des modèles SIR avec forçage sinusoïdal.

La combinaison de tous ces éléments permet de construire de très nombreux modèles, aux propriétés souvent très riches. Toutefois, malgré l'importance du souci de réalisme, il est décisif de veiller à limiter la complexité des modèles, afin de préserver leur lisibilité et la compréhension de leurs comportements.

1.1.2 Modèles stochastiques

Une autre hypothèse forte des modèles considérés précédemment est leur caractère déterministe. Or, la propagation de maladies infectieuses est influencée par de nombreux facteurs aléatoires.
D'une part, les individus sont tous différents et de la stochasticité naît du fait qu'ils ne naissent, ne meurent et ne réagissent pas tous de la même façon. On appelle cet effet la stochasticité démographique. Il est d'autant plus fort que la population est de petite taille, par exemple lorsque les épidémies sont observées à l'échelles des foyers ou des écoles. Mais cet effet peut aussi être important quand la taille de la population est grande mais par exemple que le nombre d'infectés est faible (il y a alors une possibilité d'extinction).
D'autre part, des effets exogènes viennent perturber l'évolution de l'épidémie (aléas climatiques, changements de comportement, etc.), ce que l'on peut regrouper sous le nom de stochasticité environnementale.

Les différentes formes de stochasticité ont un important effet quantitatif. A partir d'un même jeu de paramètres et de conditions initiales, de nombreuses trajectoires sont possibles, et la solution déterministe, qui néglige cette source d'incertitude, apparaît alors comme trop confiante. Ceci est également important lorsque les modèles sont confrontés à des données et qu'il s'agit d'en estimer les paramètres (cf. section 1.2) : l'incertitude associée aux paramètres est également sous-estimée. De plus, la stochasticité a également un effet qualitatif. Certains phénomènes, comme les phénomènes d'extinction, ne peuvent pas être pris en compte dans le cadre déterministe. De plus, les trajectoires stochastiques peuvent être davantage qu'une simple perturbation des trajectoires déterministes, dans les situations d'amplification stochastique ou de résonance (Alonso et al., 2007;

Keeling and Rohani, 2008).

Formalisation des modèles stochastiques

Stochasticité démographique Classiquement, afin de modéliser la stochasticité démographique, les différents compartiments du modèles sont représentés par des variables aléatoires, prenant des valeurs entières, comprises entre 0 et N (la taille totale de la population). Par exemple, dans l'analogue du modèle SIR (cf. équation 1.1), le système épidémiologique est représenté par un triplet de variables aléatoires (S, I, R), telles que $S + I + R = N$. On peut alors représenter l'évolution du système par un modèle markovien (ce qui, par la propriété d'absence de mémoire, fait implicitement l'hypothèse que les durées passées dans les compartiments ont une distribution exponentielle). Dans l'exemple du modèle SIR, on considère le processus X_t à valeurs $x = (s, i)$ où s et i sont à valeurs entières entre 0 et N (en utilisant la relation $r = N - s - i$) (Allen, 2008). On obtient les équations suivantes:

$$\begin{aligned} P\left[X_{t+dt} = (s-1, i+1) | X_t = (s,i)\right] &= \beta \frac{si}{N} dt + o(dt) \\ P\left[X_{t+dt} = (s, i-1) | X_t = (s,i)\right] &= \gamma i dt + o(dt) \\ P\left[X_{t+dt} = (s,i) | X_t = (s,i)\right] &= 1 - \beta \frac{si}{N} dt - \gamma i dt + o(dt) \end{aligned} \tag{1.3}$$

Le modèle (1.3) est un exemple de processus markovien à sauts densité dépendants, défini plus généralement comme (Andersson and Britton, 2000):

$$\begin{aligned} P\left[X_{t+dt} = x + m_j | X_t = x\right] &= N a_j\left(\frac{x}{N}\right) dt + o(dt), \forall j = 1 \ldots M \\ P\left[X_{t+dt} = x | X_t = x\right] &= 1 - N \sum_j a_j\left(\frac{x}{N}\right) dt + o(dt) \end{aligned}$$

pour lequel on considère l'ensemble de deux réactions possibles $\{m_1, m_2\}$. La réaction $m_1 = (-1, +1)$, l'infection, est associée au taux $a_1\left(\frac{s}{N}, \frac{i}{N}\right) = \beta \frac{s}{N} \frac{i}{N}$ et la réaction $m_2 = (0, -1)$, la guérison, associée au taux $a_2\left(\frac{s}{N}, \frac{i}{N}\right) = \gamma \frac{i}{N}$.

Cependant, ce système est très difficile à manipuler même pour des modèles très simples. Certains résultats probabilistes exacts existent, notamment sur la distribution de la taille finale de l'épidémie (Britton, 2010). Mais souvent, il est nécessaire d'utiliser des approximations du système pour en étudier les propriétés. Par exemple, la phase initiale de l'épidémie peut être approximée par un processus de branchement, lorsque la taille de la population est très grande par rapport au nombre d'infectés, afin d'étudier la probabilité d'observer une épidémie majeure (Britton, 2010). Une autre approximation possible du processus markovien à sauts densité dépendants repose sur l'hypothèse que la taille de la population est grande, et rejoint le formalisme par équations différentielles stochastiques (EDS). Les résultats principaux concernant cette approximation sont dus à Kurtz, et le lecteur peut se référer à Andersson and Britton (2000) ou Camacho (2011) pour plus de détails à ce sujet.

En effet, une autre forme de représentation de la dynamique épidémique fait intervenir les EDS, en supposant que les variables aléatoires correspondant au nombre d'individus dans les différentes classes sont continues (et non plus discrètes). Allen (2008) le présente ainsi : en faisant l'hypothèse

que les incréments $\Delta X = X_{t+\Delta t} - X_t$ sont gaussiens pour de petites valeurs de Δt, leur espérance vaut, dans le cadre du modèle (S, I, R) :

$$\mathbb{E}[\Delta X] = \mathbb{E}\begin{bmatrix}\Delta s \\ \Delta i\end{bmatrix} = \begin{bmatrix}-\beta\frac{si}{N} \\ \beta\frac{si}{N} - \gamma i\end{bmatrix}\Delta t + o(\Delta t) = a(X_t)\Delta t + o(\Delta t)$$

et leur variance peut être approchée à l'ordre 1 par $\sigma^2(X_t)\Delta t + o(\Delta t)$. Dans ce cas, on peut écrire

$$X_{t+\Delta t} = X_t + a(X_t)\Delta t + \sigma(X_t)\sqrt{\Delta t}\eta$$

où $\sigma(X_t) = \sqrt{\sigma^2(X_t)}$ et $\eta \sim N(0, I_2)$, avec I_2 la matrice identité de dimension 2. Cette expression correspond au schéma d'Euler associé au système d'équations différentielles suivant (B_t étant ici un mouvement brownien de dimension 2):

$$dX_t = a(X_t)dt + \sigma(X_t)dB_t \tag{1.4}$$

Comme évoqué précédemment, on peut montrer que le modèle (1.3) converge vers une écriture de la forme (1.4) en grande population sous certaines conditions (pour lesquelles le lecteur peut se référer à Andersson and Britton (2000) ou Camacho (2011)). Toutefois, ce cadre de modélisation peut être vu de façon plus générale, en admettant différents choix pour la fonction $\sigma(X_t)$, ce qui peut permettre d'introduire d'autres formes de stochasticité.

Stochasticité environnementale Parallèlement, il est possible d'incorporer de la stochasticité environnementale dans la dynamique de l'épidémie. Pour ceci, le formalisme (1.4) est particulièrement adapté, car il permet par exemple d'inclure une composante aléatoire dans certaines réactions du système (par exemple la réaction d'infection). Selon la façon dont est paramétrée cette composante aléatoire, elle peut être vue comme un bruit additif ou comme un bruit affectant certains paramètres de la réaction (Keeling and Rohani, 2008).

L'interprétation comme un bruit affectant les taux de réaction est utilisée par Bretó et al. (2009) pour incorporer la stochasticité environnementale dans le modèle markovien (1.3) de stochasticité démographique. Il s'agit de considérer un processus à incréments indépendants, positifs et stationnaires noté $\Gamma(t)$ et de définir le bruit comme la dérivée de ce processus $\xi = \frac{d\Gamma(t)}{dt}$. Ainsi, le taux a_j de la réaction m_j est alors multiplié par un bruit ξ_j pour introduire la stochasticité environnementale. Bretó et al. (2009) utilisent l'exemple d'un bruit Gamma, dans lequel $\Gamma_j(t+dt) - \Gamma_j(t) \sim Gamma(dt/s_j^2, s_j^2)$, ayant ainsi une moyenne dt et une variance s_j^2. Les taux stochastiques $a_j\xi_j$ peuvent alors être incorporés dans un schéma de simulation du système stochastique, présenté dans la section suivante.

Méthodes de simulation

Une alternative à l'étude analytique des modèle stochastique est l'étude de simulations de leurs trajectoires. De nombreuses méthodes permettent en effet de simuler ces modèles.
Le modèle markovien pour la stochasticité démographique (1.3) peut être simulé de façon exacte à l'aide de l'algorithme de Gillespie. Supposons que le modèle représente l'état x d'un système X à l'instant t et comporte M réactions m_j, $j = 1...M$, auxquelles sont associés les taux $a_j(x)$. On note $a_0 = \sum_{j=1}^{M} a_j$. On s'intéresse à la probabilité $P(\tau, j|x, t)$, qui correspond à la probabilité que la prochaine réaction se produise à l'instant $t + \tau$ et que cette réaction soit la réaction m_j.

Cette probabilité peut s'écrire comme le produit de la probabilité de réalisation de la réaction m_j à l'instant $t + \tau$ et de la probabilité de l'absence de réaction entre t et $t + \tau$:

$$p(\tau, j|x, t) = a_j(x)exp(-a_0(x)\tau) = p_1(\tau|x, t)p_2(j|\tau, t, x)$$

Par la formule de Bayes, elle peut se décomposer entre la probabilité $p_1(\tau|x, t) = a_0(x)exp(-a_0(x)\tau)$, ce qui correspond à une distribution exponentielle de paramètre $a_0(x)$, et $p_2(j|\tau, t, x) = a_j(x)/a_0(x)$, ce qui correspond à un tirage dans l'espace discret des réactions m_j, $j = 1...M$ avec probabilités $a_j(x)/a_0(x)$. Une méthode de simulation découle alors naturellement : l'instant τ est simulé selon une loi exponentielle par la méthode de la fonction de répartition inverse, et la réaction m_j est choisie avec probabilité $a_j(x)/a_0(x)$ (cf. Algorithme 1).

Algorithm 1 Algorithme de Gillespie, méthode directe (Gillespie, 1976, 2001)

Soit x_t l'état du système à l'instant t.

1: Echantillonner indépendamment $u_1 \sim U(0, 1)$ et $u_2 \sim U(0, 1)$
2: $\tau = -\frac{1}{a_0(x_t)} log\left(\frac{1}{u_1}\right)$
3: $j = \underset{i=1...M}{inf} \sum_{k=1}^{i} a_k(x_t) < u_2 a_0(x_t)$
4: La réaction m_j se produit à l'instant $t + \tau$, conduisant à la valeur $x_{t+\tau} = x + m_j$.
5: Retourner à l'étape 1, avec $x_t := x_{t+\tau}$ et $t := t + \tau$

Cette méthode a en revanche le défaut de devenir très coûteuse d'un point de vue computationnel dès que la taille de la population grandit. En effet, chaque réaction est modélisée individuellement, ce qui devient très fastidieux dans une grande population où un très grand nombre de réactions se produisent de façon presque simultanée. Une approximation, proposée par Gillespie (2001), permet, plutôt que de tracer chaque réaction, de modéliser l'ensemble des événements se produisant pendant un petit intervalle de temps noté τ (d'où l'appellation τ-leap de cette méthode). L'erreur d'approximation vient du fait que, sur cet intervalle de temps, on fait l'hypothèse que les taux de réactions sont constants $a_j(x_t)$, alors qu'en réalité, ils se modifient à chaque fois qu'une réaction se produit. Il faut pour cela que l'intervalle de temps τ soit suffisamment petit pour que cette hypothèse soit valable ("leap condition").
Dans ce cas, le nombre de réactions de chaque type r_j suivent des lois de Poisson indépendantes de paramètres respectifs $a_j(x_t)\tau$, ce qui conduit à l'Algorithme 2.

Algorithm 2 Algorithme τ-leap, (Gillespie, 2001)

Soit x_t l'état du système à l'instant t. Soit τ un pas de temps vérifiant la "leap condition".

1: $\forall j = 1...M$, $k_j \sim P(a_j(x_t\tau))$
2: $x_{t+\tau} = x_t + \sum_j k_j m_j$
3: Retourner à l'étape 1, avec $x_t := x_{t+\tau}$ et $t := t + \tau$

La difficulté pour appliquer cet algorithme vient du choix de l'intervalle τ. En particulier, comme la loi de Poisson n'est pas bornée, il peut arriver qu'un trop grand nombre de réactions se produisent sur le pas de temps, conduisant à des effectifs négatifs dans certains compartiments. Parmi les extensions de ce schéma, on peut citer l'utilisation d'autres distributions, comme la distribution multinomiale (qui permet d'éviter les effectifs négatifs).
En particulier, la distribution multinomiale est utilisée par Bretó et al. (2009) dans un cadre qui permet également d'introduire de la stochasticité environnementale à travers des taux aléatoires (cf. Algorithme 3).

Algorithm 3 Schéma d'Euler multinomial avec taux stochastiques, (Bretó et al., 2009; He et al., 2010)

c = nombre de compartiments de la population
$x_t = (x_t(1), ..., x_t(c))$ représente l'état du système à l'instant t.
τ est un pas de temps.
La réaction $m_{i,j}$ correspond à un transfert d'individus du compartiment i au compartiment j.
On note $a_{ij} = a_{ij}(x_t)$ le taux qui lui est associé, et $\Delta\Gamma_{ij}$ le bruit associé à ce taux, de variance infinitésimale s_{ij}. Le nombre de transferts de i vers j ayant lieu entre t et $t+\tau$ est noté $\Delta N_{i,j}$.
On note R_i le nombre d'individus qui restent dans le compartiment i entre t et $t+\tau$.

1: $\forall i, j = 1...c$, générer les incréments du bruit $\Delta\Gamma_{ij} \sim Gamma(\tau/s_{ij}^2, s_{ij}^2)$
2: $\forall i, j = 1...c$, calculer $p_{i,j} = (1 - exp(-\sum_k a_{i,k}\Delta\Gamma_{ik}))\ \frac{a_{i,j}\Delta\Gamma_{ij}}{\sum_k a_{i,k}\Delta\Gamma_{ik}}$
3: $\forall i = 1...c$, échantillonner

$$(\Delta N_{i,1}, ..., \Delta N_{i,i-1}, \Delta N_{i,i+1}, \Delta N_{i,c}, R_i) \sim Multinom\left(x_t(i), (p_{i,j}, 1 \le j \le c, j \ne c), 1 - \sum_{k \ne i} p_{i,k}\right)$$

4: $\forall i = 1...c$, $x_{t+\tau}(i) = R_i + \sum_{k \ne i} \Delta N_{j,i}$,
5: Retourner à l'étape 1, avec $x_t = x_{t+\tau}$ et $t = t + \tau$

Quelques propriétés du modèle stochastique

Le formalisme stochastique permet d'étudier des questions spécifiques. En particulier, la probabilité qu'une épidémie majeure se produise peut être obtenue en approchant la dynamique initiale du système par un processus de branchement. Dans le modèle (1.3), l'infection finira par s'éteindre de la population dans le temps long, toutefois, cet événement peut soit survenir très rapidement après l'introduction de quelques individus infectés soit survenir après une épidémie de grande ampleur (épidémie majeure), lorsque le nombre d'infecté redevient faible. On obtient alors le théorème du seuil suivant pour la probabilité d'observer une épidémie majeure $p_{\text{épidémie}}$, en introduisant m infectés dans une population entièrement susceptible (Allen, 2008):

$$\text{Si } R_0 < 1, p_{\text{épidémie}} = 0$$
$$\text{Si } R_0 > 1, p_{\text{épidémie}} = 1 - (1/R_0)^m$$

Ce résultat souligne le rôle de seuil de R_0 identifié dans le cadre déterministe, en ajoutant la propriété que plus la valeur de R_0 est faible, plus les extinctions sont fréquentes. Il est également possible de calculer la distribution de la taille finale de l'épidémie (nombre total d'infectés au cours de l'épidémie). Si $R_0 < 1$, cette distribution est concentrée autour de faibles valeurs proches de zéro. Si $R_0 > 1$, cette distribution est bimodale, reflétant les deux types de situations possibles (épidémies majeures et extinctions rapides).

Dans le cadre des maladies endémiques, se pose également la question de la persistence de la maladie, en raison des extinctions probables lorsque le nombre d'infectés devient faible. Bartlett introduit dès les années 1950 la notion de taille de communauté critique (en anglais "critical community size", CCS), comme la taille de population au-dessus de laquelle on n'observe pas d'extinctions de la maladie. Il observe en effet, à partir de séries de cas de rougeoles dans différentes villes au Royaume-Uni (Bartlett, 1957) et aux Etats-Unis (Bartlett, 1960), une relation entre la durée inter-épidémique et la taille de la population, un constat qu'il corrobore par des

simulations par ordinateur. La persistence de la maladie dépend aussi de la fréquence des imports de cas infectés depuis d'autres villes, permettant de relancer une épidémie déjà éteinte. Les imports peuvent être modélisés explicitement en représentant les déplacements entre plusieurs aires géographiques (par exemple dans un modèle spatial), ou en ajoutant un terme constant à la force d'infection. En complément de la taille de communauté critique, plusieurs indicateurs existent afin de quantifier la persistence d'une maladie, selon si les imports sont pris en compte ou si l'on applique un conditionnement par l'absence d'extinction (Keeling and Rohani, 2008).

Enfin, le modèle stochastique peut également conduire à des dynamiques qualitativement différentes du modèle déterministe, selon plusieurs mécanismes. Par exemple, Keeling and Rohani (2008) évoquent les phénomènes de résonance, par lesquels la moyenne du processus stochastique peut être différente de sa contre-partie déterministe. Les mécanismes d'amplification stochastique introduits par Alonso et al. (2007) indiquent l'importance qualitative de la stochasticité dans les propriétés dynamiques des systèmes étudiés (Camacho, 2011).

1.2 Méthodes d'estimation

Même si on peut tirer des résultats théoriques de leurs analyse, pour la plupart des rôles que l'on veut faire jouer aux modèles en santé publique, il est crucial que ceux-ci soient paramétrés de façon appropriée. De nombreuses méthodes d'estimation existent pour confronter les modèles aux données observées, afin d'en inférer les paramètres ou d'évaluer leur qualité.

1.2.1 Quel type de données

Les données parfaites pour étudier un modèle épidémiologique seraient l'observation de tous les événements de l'épidémie (toutes les infections, guérisons, naissances, etc.). Evidemment, en pratique, la plupart de ces événements ne sont pas observables, et les données disponibles sont à la fois incomplètes et incertaines. D'une part, seuls certains compartiments du système sont observés : par exemple, les systèmes de santé ne recueillent le plus souvent que les informations sur les individus malades (et non pas sur les individus sains), ainsi seuls les individus infectés sont observés, et non pas les individus susceptibles, immunisés ou en phase d'incubation. D'autre part, toute méthode de collecte de données est elle-même sujette à des bruits d'observation ou différents types de biais.
On évoquera dans un premier temps les données médicales ou épidémiologiques, directement en lien avec l'étude de la maladie, et dans un second temps les données extérieures, qui peuvent informer sur d'autres aspects de la population modélisée.

Données médicales ou épidémiologiques Deux grandeurs épidémiologiques sont utilisées pour représenter la présence d'une maladie dans une population donnée. La prévalence renseigne sur le nombre de personnes malades à une date particulière. L'incidence renseigne sur le nombre de nouveaux cas survenus sur une unité de temps prédéfinie. On présentera ici quelques sources d'observation de ces quantités, ainsi que des biais qui leur sont associés.

En premier lieu, une épidémie peut être observée à travers les cas d'infection. Très souvent, on dispose de séries temporelles du nombre de nouveaux cas observés par intervalle de temps. Ces données informent sur la dynamique de l'épidémie à travers une mesure de l'incidence. Elles peuvent provenir directement du système de santé (hôpitaux, centres de soin), à des fins de surveillance. Ce mode de recueil a l'avantage de couvrir de longues périodes de temps et de larges

zones géographiques ; en revanche, les données sont agrégées par unité de temps (par exemple par mois ou par semaine, beaucoup plus rarement par jour), et sur l'ensemble de la population (ou sur un certain nombre de sous-populations). Les séries de cas peuvent également provenir d'enquêtes épidémiologiques spécifiques. Plusieurs méthodes d'enquête existent, comme les enquêtes par cluster (où l'entourage de cas préalablement détectés est dépisté pour la maladie, ou suivi de manière intensive), ou par cohortes (un nombre d'individus est suivi durant une longue période de temps). Ces enquêtes permettent souvent des données plus détaillées : par exemple les données individuelles peuvent permettre une description plus fine de chaque infection, ou de modéliser des sous-groupes par âge ou par foyer.
De manière générale, les données de cas présentent deux types de biais. D'une part, un certain nombre de cas ne sont pas observés, car ils échappent au périmètre de l'enquête ou au système de soins en raison de leurs faibles symptômes, de leurs caractéristiques socio-économiques, ou d'autres facteurs. D'autre part, des erreurs de diagnostic peuvent survenir (faux positifs et faux négatifs). Ces deux types de biais dépendent fortement du mode de collecte des données (diagnostic clinique ou biologique par exemple), ainsi que du pathogène étudié (proportion de cas graves par exemple).

En second lieu, des données de prévalence ou de séroprévalence sont parfois disponibles. A une date donnée, une mesure est faite de la proportion de la population portant une maladie, ou portant des anticorps attestant qu'il a déjà connu la maladie au cours de sa vie (séroprévalence). Ces données sont informatives sur l'état du système épidémiologique à une date donnée (par exemple, elles peuvent renseigner sur ses conditions initiales). De surcroît, lorsqu'elles sont stratifiées par âge, elles informent également sur la dynamique de la maladie, en faisant l'hypothèse que les individus plus âgés peuvent témoigner de l'incidence à des périodes antérieures. En ce qui concerne leurs biais, ces données sont également sujettes aux erreurs de diagnostic. De plus, comme les mesures ne peuvent généralement pas être réalisées sur la population entière, elles comportent également les biais et aléas associés aux méthodes de sondage (taille et représentativité de l'échantillon).

Enfin, les méthodes de séquençage de l'ADN des virus ont permis l'accès à une autre forme de données épidémiologiques, qui a commencé récemment à être utilisée pour paramétrer les modèles. La reconstruction de la phylogénie des virus à l'aide des séquences observées permet d'informer sur la dynamique de l'épidémie, en faisant l'hypothèse que les branchements de l'arbre phylogénétique correspondent à des événements d'infection. Ces données sont néanmoins rares et coûteuses et elles nécessitent des méthodes statistiques spécifiques non abordées dans ce travail de thèse.

Autres sources de données Parallèlement, les modèles épidémiologiques peuvent être enrichis à l'aide de diverses sources de données non médicales, dont nous citons ici quelques exemples.
D'une part, les recensements de population peuvent informer sur la taille de la population, sa structure d'âge et parfois ses caractéristiques socio-économiques. D'autres informations topologiques ou environnementales (végétation, urbanisation, routes) peuvent éclairer le contexte dans lequel le modèle est conçu (notamment lorsque celui-ci comporte une dimension spatiale).
Dans le cadre des maladies à transmission vectorielle, comme par exemple la dengue qui est transmises par les moustiques de type *Aedes*, des mesures entomologiques peuvent également être incorporées dans les modèles (ces données reflètent souvent la taille de la population de moustiques, plus rarement les moustiques infectieux).

Enfin, certaines données peuvent informer sur les comportement humains. Il peut s'agir de données de déplacements, observés directement, comme les flux aériens, ou indirectement par exemple à l'aide de données téléphoniques (Wesolowski et al., 2015). Il peut également s'agir de données de contact entre sous-groupes d'individus, par exemple acquises à partir d'enquêtes spécifiques. Enfin, les données provenant d'Internet (nombre de requêtes sur des moteurs de recherche, mots-clés sur les réseaux sociaux par exemple) peuvent également éclairer sur les attitudes par rapport à une maladie (Roche et al., 2017) ou la présence de cas sous-cliniques.

Chaque type de données a ses propres enjeux et peut donner lieu à des analyses statistiques spécifiques. Par la suite, on se concentrera surtout sur les séries temporelles de cas.

1.2.2 Modèles statistiques et modèles d'observation

La prise en compte des données fait apparaître une nouvelle source de stochasticité, liée au bruit d'observation. Les observations sont effectivement une représentation incomplète et bruitée de la réalisation de la trajectoire épidémiologique, dont le processus de génération peut lui-même être modélisé. Aussi les données observées y_k sont-elles liées aux solutions du modèle épidémiologique à travers un modèle d'observation. Par exemple, dans le cadre d'un modèle SIR comme défini dans le système d'équations (1.1), supposons que l'on observe l'incidence $\frac{dC_t}{dt} = \beta \frac{S_t I_t}{N}$ cumulée par semaine $C_k = \int_{t_k}^{t_k+7} \frac{dC_t}{dt} dt$ (les t_k étant comptés en jours), avec un certain taux de report ρ chaque semaine de façon indépendante.
Plusieurs modèles d'observation peuvent être utilisées, comme le modèle gaussien, ou le modèle binomial. Le modèle gaussien est utilisé implicitement dans le cas de l'estimation par moindres carrés, mais il est parfois également employé en spécifiant des variances différentes le long de la trajectoire (comme dans les modèles de Yang et al. (2014) par exemple).
Un modèle fréquemment utilisé est le modèle d'observation de Poisson. On suppose que les observations y_k sont indépendantes conditionnellement aux C_k et suivent la loi :

$$y_k | C_k \sim \mathcal{P}(\rho C_k) \tag{1.5}$$

Etant donné que la loi de Poisson possède un support non borné, ce modèle attribue une probabilité non nulle aux cas où l'on observe davantage de cas que réellement simulés, permettant une prise en compte des faux-positifs (Bretó, 2018).
Une autre hypothèse couramment adoptée est le modèle binomial négatif. On suppose dans ce cas que les taux de report ρ_k sont des variables indépendantes de loi Gamma$(1/\phi, \rho\phi)$ et que, conditionnellement aux ρ_k les observations suivent des lois de Poisson indépendantes $y_k | \rho_k, C_k \sim \mathcal{P}(\rho_k C_k)$ (Bretó et al., 2009). On a ainsi

$$y_k | C_k \sim \mathcal{NB}(\frac{1}{\phi}, \frac{1}{1 + \phi\rho C_k})$$

où l'on note $\mathcal{NB}$ la loi négative binomiale de paramètres $\frac{1}{\phi} \in \mathbb{R}_+^*$ et $\frac{1}{1+\phi\rho C_k} \in [0, 1]$. Ceci implique $\mathbb{E}[y_k | C_k] = \rho C_k$ et $Var[y_k | C_k] = \rho C_k + \phi(\rho C_k)^2$, et peut ainsi s'interpréter également comme une distribution de Poisson "surdispersée" (la variance étant plus élevée que la moyenne).

Bretó (2018) généralise l'idée (présente dans le modèle binomial négatif) d'utiliser des taux de reports eux-même aléatoires, comme un parallèle à l'introduction de taux aléatoires dans les réactions du système stochastique pour représenter la stochasticité environnementale (cf. section 1.1.2). Cette formalisation, permettant de séparer d'une part le processus d'observation et d'autre

part l'aléatoire sur les paramètres qui le composent permet selon lui d'ouvrir la voie à des améliorations de ces modèles d'observation. Par exemple, il devient possible de prendre en compte de la dépendance potentielle entre les bruits associés aux paramètres. Il devient également possible de prendre en compte des erreurs de diagnostic en incluant une modélisation des cas faux-positifs.

D'un point de vue empirique, Knape et al. (2011) comparent différents modèles d'observation sur un même jeu de données en écologie et concluent à une influence notable de ces choix sur les résultats obtenus.

Dans les modèles considérés ici, on peut poser le modèle à espace d'état suivant:

$$\begin{cases} x_{1:T} = f(\theta', x_0, u_t) \\ y_{1:K} \sim g(h(x_{1:T}), \theta) \end{cases}$$

où les y_k sont les observations, x_t les états du système aux temps t, et $\theta = (\theta', x_0)$ les paramètres du système (incluant les paramètres et les conditions initiales x_0). f correspond aux équations définissant le modèle épidémiologique, qui peuvent également dépendre d'un bruit u_t. g correspond à la distribution du modèle d'observation, appliquée à une transformée h des états $x_{1:T}$ du système.

1.2.3 Estimations de modèles déterministes

Dans le cadre déterministe ($u_t = 0$), les états du système $x_{1:T}$ sont déterminés entièrement par le paramètre $\theta = (\theta', x_0)$, souvent de façon numérique. La seule source de stochasticité provient du modèle d'observation.

Cadre fréquentiste

Le modèle d'observation permet de définir la vraisemblance, même si celle-ci n'est souvent calculable que numériquement, via la résolution du système d'équations différentielles du modèle épidémiologique.

Dans l'exemple du modèle de Poisson donné ci-dessus (cf. equation 1.5), on a $\theta = (x_0, \beta, \gamma, \rho)$, la vraisemblance du modèle s'écrit comme:

$$L(y_{1:K}, \theta) = \prod_{k=1}^{K} \frac{(\rho C_k)^{y_k} e^{-\rho C_k}}{y_k!}$$

où les C_k sont obtenus en résolvant le système d'équations (1.1), pour (x_0, β, γ).

La vraisemblance définie peut ensuite être maximisée selon des méthodes numériques adaptées, comme l'algorithme de Nelder-Mead (simplexe) par exemple (cf. (Eisenberg et al., 2013) pour un exemple d'application).

Cadre bayésien et Monte Carlo à chaîne de Markov (MCMC)

Dans le cadre bayésien, la fonction cible est la distribution a posteriori

$$p(\theta|y) \propto L(y, \theta)\pi(\theta)$$

où $\pi(\theta)$ est une distribution a priori pour les paramètres, et la vraisemblance marginale $L(y, \theta) = L(x, y, \theta) = g(h(x_{1:T}), \theta)$ est évaluée à travers la résolution numérique du modèle épidémiologique pour le paramètre θ.

La distribution a posteriori étant connue à une constante près et de façon numérique, les méthodes simulées de Monte Carlo à chaîne de Markov (MCMC) sont particulièrement adaptées à son étude. Il s'agit de construire une chaîne de Markov ergodique ayant pour distribution invariante la distribution a posteriori: la chaîne constitue alors un échantillon d'observations dépendantes tirées selon la distribution voulue (Robert and Casella, 2004). Elle peut être utilisée pour calculer les distributions marginales pour les différents paramètres, et des indicateurs tels que la moyenne a posteriori ou des intervalles de crédibilité. Elle peut également être utilisée pour re-simuler les trajectoires du modèle en incorporant l'incertitude a posteriori sur les paramètres.

Par exemple, l'algorithme de Metropolis-Hastings (cf. algorithme 4) permet de générer une telle chaîne, $(\theta_0, \theta_1, ..., \theta_N)$, à partir d'un point d'initialisation quelconque θ_0 et d'une densité de proposition $q(.|\theta_i)$: il faut pour cela que le support de la fonction cible soit connecté et inclus dans le support de la densité de proposition q, et de faibles conditions sur la distribution q [1] (Robert and Casella, 2004). La distribution $q(.|\theta_i)$ peut être par exemple centrée sur le point θ_i (situation de marche aléatoire). La chaîne (une fois éliminée la phase préconvergence, dite de *burn-in*) constitue un échantillon d'observations dépendantes tirées selon la distribution a posteriori jointe. Cet algorithme est également valable lorsque la fonction-cible est connue à une constante près.

Algorithm 4 METROPOLIS-HASTINGS

La fonction cible est ici la distribution a posteriori $p(\theta|y)$.

1: Initialiser θ_0.
2: **for** $i = 1...N$ **do**
3: Echantillonner θ^* avec $q(.|\theta_i)$
4: Accepter θ^* avec probabilité $\alpha = 1 \wedge \frac{p(\theta^*|y)q(\theta_i|\theta^*)}{p(\theta_i|y)q(\theta^*|\theta_i)}$
5: Si θ^* est accepté, $\theta_{i+1} = \theta^*$. Sinon $\theta_{i+1} = \theta_i$
6: **end for**
7: **return** $(\theta_0, \theta_1, ..., \theta_N)$

Les méthodes de MCMC ont été appliquées à des séries épidémiologiques depuis la fin des années 1990 et le début des années 2000 (cf. Putter et al. (2002) par exemple), et l'algorithme de Metropolis-Hastings est fréquemment utilisé pour les modèles déterministes (Pandey et al., 2013; Lourenço and Recker, 2014; Kucharski et al., 2016; Funk et al., 2016). Le formalisme bayésien permet notamment d'incorporer sous forme de distributions a priori les informations disponibles sur certains paramètres (comme les paramètres mesurables cliniquement par exemple).

[1] La condition sur q pour garantir l'ergodicité de la chaîne est la suivante (Roberts and Tweedie, 1996) : la densité cible est minorée par une constante positive et majorée sur tout compact, et $\exists \delta > 0$ et $\epsilon > 0$ tels que $\forall x$

$$|x - y| < \delta \Rightarrow q(y|x) > \epsilon$$

Les conditions garantissant l'ergodicité géométrique sont, elles, plus restrictives (Robert and Casella, 2004).

La convergence du MCMC

Diagnostiquer la convergence Si en théorie la chaîne de Markov converge vers la distribution a posteriori, il n'est pas aisé de s'assurer en pratique que la loi invariante a bien été atteinte. Même s'il n'est jamais possible d'être complètement certain de la convergence, de nombreux diagnostics existent (Robert and Casella, 2004). Il est tout d'abord nécessaire de réaliser certaines vérifications graphiques, comme l'appréciation visuelle de la stationnarité de la chaîne, de la convergence de la moyenne cumulée ou l'étude d'autocorrélogrammes ou de corrélations croisées. Ensuite, certains outils diagnostics quantitatifs peuvent être utilisés.
Certaines méthodes analysent des propriétés de la chaîne comme la stationnarité, ou la convergence de la moyenne empirique (Robert and Casella, 2004). Par exemple, afin d'évaluer la stationnarité de la chaîne, des tests non-paramétriques d'homogénéité (comme Kolmogorov-Smirnov) entre le début et la fin de la chaîne peuvent être utilisés (à noter qu'il est approprié de sous-échantillonner la chaîne au préalable pour s'approche du cadre indépendant et identiquement distribué) (Robert and Casella, 2004). Le test de Geweke (Geweke, 1992), autre test fréquemment utilisé, compare des estimations de la moyenne du début et de la fin de la chaîne, à l'aide d'estimations non paramétriques de la densité spectrale.
Un autre indicateur est la taille effective de l'échantillon (Effective Sample Size, ou ESS en anglais). Il s'agit de calculer la taille qu'aurait un échantillon indépendant de même variance que celle observée dans la chaîne, afin de mesurer la perte d'efficacité liée à l'utilisation d'un échantillon dépendant (Robert and Casella, 2004). L'ESS se calcule comme le ratio entre le nombre d'itérations considérées de la chaîne N et son temps d'autocorrélation, qui nécessite d'être approximé numériquement.
D'autres méthodes comparent plusieurs chaînes initialisées en des points différents et s'assurent qu'elles convergent vers la même distribution. Par exemple, dans le cadre du test de Gelman-Rubin (Gelman and Rubin, 1992), il s'agit tout d'abord de générer plusieurs points d'initialisation, à partir d'une estimation surdispersée de la densité cible, à partir desquels des chaînes de MCMC sont simulées. Deux estimateurs de la variance du posterior, fondés sur les variances inter et intra-chaîne, sont ensuite comparés à travers un indicateur dont la distribution peut être approximée à partir d'une loi de Student. Il est néanmoins important de bien échantillonner l'espace des paramètres lors du choix des points d'initialisation pour que le diagnostic soit valide (Robert and Casella, 2004).
Ces diagnostics ont toutefois de nombreuses limites, et il est important de combiner plusieurs formes de tests (graphiques et quantitatifs, univariés et multivariés, par exemple) afin de détecter les éventuels problèmes de convergence (Cowles and P. Carlin, 1995).

Schémas adaptatifs Par ailleurs, si en théorie l'algorithme de Metropolis-Hastings converge pour toute densité de proposition q, les propriété d'ergodicité forte ne sont pas toujours satisfaites et la convergence de la chaîne peut être lente (Robert and Casella, 2004). En pratique, le choix de q a donc de fortes conséquences sur la chaîne: de façon heuristique, dans l'exemple d'une marche aléatoire gaussienne, on peut dire que si la variance de q est trop élevée, les nouveaux paramètres échantillonnés sont trop souvent rejetés et la chaîne reste longtemps fixée sur une même valeur. Au contraire, si la variance est trop faible, seul un voisinage du point d'initialisation est exploré, ne permettant pas la convergence vers la loi cible, qui peut en être éloignée. Parmi les solutions possibles, il est possible d'utiliser des méthodes adaptatives, utilisant des propriétés du début de la chaîne (taux d'acceptation, covariance) pour mieux choisir la densité de proposition q. L'utilisation de valeurs passées de la chaîne conduit à une perte de la propriété de Markov, et, pour de nombreux schémas adaptatifs, les propriétés de convergence ne sont prouvés que sous des hypothèses parfois

difficiles à vérifier en pratique (Robert and Casella, 2004; Yang and Rosenthal, 2017).
Malgré ces réserves, l'adaptation des algorithmes de MCMC peut se révéler très utile et de nombreuses méthodes existent (Roberts and Rosenthal, 2009). Par exemple, dans le cas d'une marche aléatoire gaussienne, la matrice de covariance de la densité de proposition est redimensionnée afin d'obtenir un taux d'acceptation proche des résultats d'optimalité prouvés dans de nombreux contextes (Roberts and Rosenthal, 2001). Cette matrice de covariance est parfois également calculée à partir de la covariance passée de la chaîne (Haario et al., 2001; Roberts and Rosenthal, 2009), afin de s'approcher des situations optimales dans le cas gaussien. Plusieurs schémas adaptatifs sont parfois enchaînés les uns aux autres (Yang and Rosenthal, 2017; Dureau et al., 2013b).

1.2.4 Estimation de modèles stochastiques

Dans les situations où, en plus du bruit d'observation, le processus dynamique de propagation de l'épidémie est lui-même stochastique, l'estimation requiert des méthodes plus complexes. O'Neill (2010) sépare deux types de situations: celles où la vraisemblance est tractable (ce qui dépend à la fois du modèle choisi et des données disponibles), et celles où elle ne l'est pas. On se concentre ici sur les cas les plus fréquents, ceux où la vraisemblance est intractable ; pour les autres situations, le lecteur peut se référer à O'Neill (2010) ou Andersson and Britton (2000).

Plus particulièrement, dans de nombreuses situations traitées ici, on considère que le modèle n'est disponible que sous la forme de simulations, comme une boîte noire (il est possible de calculer ponctuellement la vraisemblance d'une trajectoire réalisée, mais pas de l'ensemble des trajectoires possibles pour un jeu de paramètres donné). Les estimations réalisables dans ce cadre sont parfois dites avoir la propriété "plug-and-play" (Bretó et al., 2009). King et al. (2016) propose également une typologie des méthodes d'estimation selon leur caractère "plug-and-play" ou non, l'approche retenue (bayésienne ou fréquentiste), et le recours à la vraisemblance ou à des statistiques-résumés.

On détaillera ici deux types de stratégies : estimer la vraisemblance par des méthodes de filtrage, ou utiliser des méthodes sans vraisemblance fondées sur des statistiques-résumés.

Méthodes de filtrage

On suppose dans cette section que l'on dispose d'une trajectoire inobservée $\{X_k\}$ pour k=*0:K* modélisée comme un processus markovien dont la distribution initiale est notée $\mu_\theta(x_0)$ (cette trajectoire peut avoir un processus sous-jacent $\{X_t\}$ pouvant être en temps continu ou discret). On dispose également d'observations $y_{1:K} = (y_1, ...y_K)$ indépendantes conditionnellement aux $x_{0:k} = (x_1, ...x_K)$, et d'un jeu de paramètres θ. Ceci conduit au modèle suivant (Doucet et al., 2001b):

$$\begin{cases} f_\theta(x_k|x_{k-1}) & k = 1:K \\ g_\theta(y_k|x_k) & k = 1:K \end{cases} \tag{1.6}$$

On s'intéresse alors à la vraisemblance

$$L(y,\theta) = p_\theta(y_{1:K}) = \int \mu_\theta(x_0) \prod_{k=1}^{K} g_\theta(y_k|x_k) f_\theta(x_k|x_{k-1}) dx_{0:K} \tag{1.7}$$

ainsi qu'à la distribution $p_\theta(x_{0:K}|y_{1:K})$, qui peut s'écrire de façon récursive

$$p_\theta(x_{0:K}|y_{1:K}) = \frac{1}{p_\theta(y_{1:K})}\mu_\theta(x_0)\prod_{k=1}^{K} g_\theta(y_k|x_k)f_\theta(x_k|x_{k-1}) \tag{1.8}$$

Il n'est possible de calculer analytiquement ces quantités que dans de rares situations, notamment le cadre linéaire gaussien (utilisation du filtre de Kalman), ou lorsque les x_k ne peuvent prendre qu'un nombre fini de valeurs (Doucet et al., 2001b), ce qui n'est généralement pas le cas des modèles épidémiologiques. Il existe toutefois des méthodes simulées pour traiter ces modèles.

Filtre particulaire (ou Monte Carlo Séquentiel) Le Monte Carlo séquentiel permet d'estimer la distribution (1.8) et d'estimer la vraisemblance (1.7), en supposant connu le jeu de paramètres θ. L'algorithme 5 présente sa version dite "bootstrap".

Intuitivement, l'état du système à l'instant k, noté X_k, est représenté par un échantillon de particules $\{x_k^{(j)}, j \in [1; J]\}$. A chaque itération de l'algorithme, les particules sont projetées au temps $k+1$ selon la dynamique du modèle épidémiologique $f_\theta(x_k|x_{k-1})$ et un poids leur est attribué en fonction de leur adéquation à l'observation y_k. Avant l'itération suivante, les particules (et leur trajectoire depuis le temps $t = 0$) sont ré-échantillonnées selon ces poids, afin d'éliminer les particules ayant des poids trop faibles et de concentrer l'effort de computation sur les régions les plus vraisemblables. La vraisemblance est également estimée de façon séquentielle.
Il s'agit d'une méthode d'*importance sampling* dans laquelle les observations sont incorporées de façon séquentielle, la fonction d'importance permettant de calculer les poids de façon séquentielle (étapes 4 à 8 dans l'algorithme 5), accompagnée d'une étape de ré-échantillonnage (étape 9 dans l'algorithme 5) (Doucet et al., 2001b). L'algorithme fournit également un estimateur non biaisé de la vraisemblance, cf. équation (1.7).

Algorithm 5 Monte Carlo séquentiel (SMC) - Filtre particulaire Bootstrap

L est la vraisemblance $p_\theta(y_{1:K})$. $W_k^{(j)}$ est le poids et $x_k^{(j)}$ est l'état associés à la particule j à la date k.

1: Fixer $L = 1$, $W_0^{(j)} = 1/J$.
2: Échantillonner $(x_0^{(j)})_{j=1:J}$ selon $\mu_\theta(x_0)$.
3: **for** $k = 1 : K$ **do**
4: **for** $j = 1 : J$ **do**
5: Échantillonner $x_k^{(j)}$ selon $f_\theta(x_k|x_{k-1}^{(j)})$
6: Fixer $\alpha^{(j)} = g_\theta(y_k|x_k^{(j)})$
7: **end for**
8: Fixer $W_k^{(j)} = \frac{\alpha^{(j)}}{\sum_{l=1}^{J}\alpha^{(l)}}$ et $L = L\frac{1}{J}\sum_l \alpha^{(l)}$
9: Re-échantillonner $(x_{0:k}^{(j)})_{j=1:J}$ selon $W_k^{(j)}$
10: **end for**

Le filtre bootstrap a l'inconvénient que les particules sont projetées de $k-1$ à k sans tenir compte des observations y. Cela peut conduire à de nombreuses particules avec un faible poids, et à une disparition de nombreuses particules à mesure que de nouvelles observations sont incorporées. Ce phénomène est particulièrement fort lorsque la distribution $g_\theta(y_k|x_k)$ est très informative. Afin de pallier ce problème, de nombreuses extensions du filtre bootstrap existent (filtre particulaire guidé, filtre particulaire auxiliaire, par exemple). Toutefois, elles nécessitent souvent de calculer la densité $f_\theta(x_k|x_{k-1})$, qui n'est pas tractable dans de nombreux modèles épidémiologiques, et elles

ne possèdent ainsi souvent pas la propriété "plug-and-play".

Afin d'estimer à son tour le jeu de paramètres θ, le filtre SMC (algorithme 5) peut être incorporé dans un schéma d'estimation, par exemple dans le cadre bayésien selon des méthodes de MCMC (Particle Markov Chain Monte Carlo, PMCMC (Andrieu et al., 2010)) ou de SMC (SMC^2 (Chopin et al., 2013)), ou dans le cadre fréquentiste en maximisant la vraisemblance (Maximum Likelihood via Iterated Filtering, (Ionides et al., 2006)). On introduira brièvement ces différentes approches, mais dans le cadre de cette thèse, c'est la méthode du PMCMC (dans sa version particle marginal Metropolis Hastings, PMMH) qui a été utilisée.

Cadre bayésien et méthodes de Monte Carlo Dans le cadre bayésien, on suppose une distribution a priori $\pi(\theta)$ et l'objectif est désormais d'estimer la loi a posteriori :

$$\begin{aligned} p(x_{0:K}, \theta | y_{1:K}) &= p(\theta | y_{1:K}) p_\theta(x_{0:K} | y_{1:K}) \\ &\propto p_\theta(y_{1:K}) \pi(\theta) p_\theta(x_{0:K} | y_{1:K}) \end{aligned} \tag{1.9}$$

Le Particle Marginal Metropolis Hastings (Andrieu et al., 2010) permet d'échantillonner selon cette loi a posteriori. Il s'agit de construire un algorithme de Metropolis-Hastings dont la fonction cible est la loi (1.9) et dont la densité de proposition s'écrit comme :

$$q\left((\theta^*, x^*_{0:K}) | (\theta, x_{0:K})\right) = q(\theta^* | \theta) p_{\theta^*}(x^*_{0:K} | y_{1:K})$$

Le ratio de Metropolis-Hastings s'écrit alors de la sorte (Andrieu et al., 2010):

$$\begin{aligned} \frac{p(x^*_{0:K}, \theta^* | y_{1:K}) q\left((\theta, x_{0:K}) | (\theta^*, x^*_{0:K})\right)}{p(x_{0:K}, \theta | y_{1:K}) q\left((\theta^*, x^*_{0:K}) | (\theta, x_{0:K})\right)} &= \frac{p(\theta^* | y_{1:K}) p_{\theta^*}(x^*_{0:K} | y_{1:K}) q(\theta | \theta^*) p_\theta(x_{0:K} | y_{1:K})}{p(\theta | y_{1:K}) p_\theta(x_{0:K} | y_{1:K}) q(\theta^* | \theta) p_{\theta^*}(x^*_{0:K} | y_{1:K})} \\ &= \frac{p(\theta^* | y_{1:K}) q(\theta | \theta^*)}{p(\theta | y_{1:K}) q(\theta^* | \theta)} \\ &= \frac{p_{\theta^*}(y_{1:K}) \pi(\theta^*) q(\theta | \theta^*)}{p_\theta(y_{1:K}) \pi(\theta) q(\theta^* | \theta)} \end{aligned}$$

Les quantités inconnues $p_\theta(y_{1:K})$ et $p_\theta(x_{0:K} | y_{1:K})$ sont approximées par leurs estimations à l'aide du filtre SMC, ce qui conduit à l'algorithme 6, dont la validité et les propriétés de convergence ont été établies par Andrieu et al. (2010).

Algorithm 6 PARTICLE MARGINAL METROPOLIS-HASTINGS (Andrieu et al., 2010)

Dans un modèle avec K observations et J particules.
$q(.|\theta^{(i)})$ est la densité de proposition.

1: Initialiser $\theta^{(0)}$.
2: Avec SMC, calculer $\hat{p}_{\theta^{(0)}}(y_{1:K})$ et échantillonner $x^{(0)}_{0:n}$ selon $\hat{p}_{\theta^{(0)}}(x_{0:K}|y_{1:K})$.
3: **for** $i = 1...N-1$ **do**
4: Echantillonner θ^* selon $q(.|\theta^{(i)})$
5: Avec SMC, calculer $L(\theta^*) = \hat{p}_{\theta^*}(y_{1:K})$ et échantillonner $x^*_{0:n}$ selon $\hat{p}_{\theta^*}(x_{0:K}|y_{1:K})$
6: Accepter θ^* (et $x^*_{0:K}$) avec probabilité

$$1 \wedge \frac{L(\theta^*)\pi(\theta^*)q(\theta^{(i)}|\theta^*)}{L(\theta^{(i)})\pi(\theta^{(i)})q(\theta^*|\theta^{(i)})}$$

7: Si θ^* est accepté, $\theta^{(i+1)} = \theta^*$ et $x^{(i+1)}_{0:K} = x^*_{0:K}$. Sinon $\theta^{(i+1)} = \theta^{(i)}$ et $x^{(i+1)}_{0:K} = x^{(i)}_{0:K}$.
8: **end for**
9: **return** $(\theta_0, \theta_1, ..., \theta_N)$

L'utilisation du PMCMC en épidémiologie a permis notamment l'étude de paramètres variables dans le temps (Dureau et al., 2013b; Camacho et al., 2015; Cazelles et al., 2018), ou la prise en compte de données phylogénétiques (Rasmussen et al., 2011, 2014).

Dans le cadre du SMC^2 (Chopin et al., 2013), le filtre particulaire (algorithme 5) est incorporé dans un autre algorithme SMC (nommé IBIS) pour lequel c'est la distribution de θ qui est représentée par un échantillon de particules. Dans l'algorithme IBIS, les observations $y_{1:K}$ sont prises en compte de façon séquentielle, en utilisant les incréments de la vraisemblance $p(y_k|y_{1:k-1}, \theta)$. L'algorithme SMC^2 permet, entre autres, de traiter les situations où ces incréments ne sont pas tractables, en utilisant leur estimation non biaisée par le filtre particulaire (algorithme 5). À chaque itération k de l'algorithme, une nouvelle observation y_k est prise en compte, afin d'actualiser à la fois la distribution de $p(x_k|y_{1:k}, \theta)$ et celle de $p(\theta|y_{1:k})$, en utilisant N_θ particules $(\theta^m)_{m=1...N_\theta}$, et N_x particules pour $\left(x^{m,n}_{0:k}\right)_{n=1...N_x}$ pour chaque θ^m.

Maximum Likelihood via Iterated Filtering (MIF) L'approche de Ionides et al. (2006) consiste à chercher le jeu de paramètres θ qui maximise la vraisemblance $p_\theta(y_{1:K})$, en construisant un algorithme itératif (Maximum Likelihood via Iterated Filtering, ou MIF).
Pour cela, ils considèrent le modèle étendu dans lequel θ est remplacé par un processus variant dans le temps θ_k suivant une marche aléatoire d'espérance $\mathbb{E}[\theta_k|\theta_{k-1}] = \theta_{k-1}$ et de variance $Var[\theta_k|\theta_{k-1}] = \sigma^2\Sigma$ (le modèle initial correspondant au cas $\sigma^2 = 0$). Par itérations successives, σ^2 est approché de zéro de façon exponentielle $\sigma_i = \alpha^{i-1}$, où α est un facteur de "refroidissement" strictement compris entre 0 et 1.
A chaque itération i de l'algorithme, un filtre particulaire de type SMC est appliqué pour estimer $p(x_{0:K}, \theta_{0:K}|y_{1:K})$, ce qui permet d'obtenir une séquence de $\hat{\theta}^{(i)}_{0:K}$ et de leurs variances $V^{(i)}_{0:K}$ en utilisant la variabilité des particules. L'estimateur $\hat{\theta}^{(i+1)}$est alors estimé comme une moyenne pondérée des $\hat{\theta}^{(i)}_{0:K}$, soit $\hat{\theta}^{(i+1)} = \hat{\theta}^{(i)} + V^{(i)}_1 \sum_{k=1}^{K}(V^{(i)}_k)^{-1}\left(\hat{\theta}^{(i)}_k - \hat{\theta}^{(i)}_{k-1}\right)$.
Le lecteur peut se référer à Ionides et al. (2006) et Ionides et al. (2011) où les propriétés de convergence de cet algorithme sont détaillées et prouvées, à Ionides et al. (2015) pour une version améliorée de l'algorithme et à King et al. (2016) pour les détails d'implémentation.

Afin d'obtenir des écarts-types ou des intervalles de confiance pour $\hat{\theta}$, des profils de vraisemblance sont utilisés.
L'algorithme MIF a été fréquemment utilisé en épidémiologie, sur des données de cholera (King et al., 2008), rougeole (He et al., 2010), malaria (Laneri et al., 2010), grippe (Camacho et al., 2011), ou encore rotavirus (Stocks et al., 2017).

Il existe plusieurs librairies pour implémenter ces algorithmes coûteux, comme SSM (Dureau et al., 2013a), Libbi (Murray, 2013) ou POMP (King et al., 2016). C'est SSM qui a été utilisée dans le cadre de cette thèse.

Filtre de Kalman Dans le cadre de modèles linéaires et gaussiens, le problème de filtrage peut être résolu de façon analytique à l'aide du filtre de Kalman. Dans ce cadre, les processus (x_k) et (y_k) sont gaussiens et ainsi entièrement définis par leur espérance et leur matrice de covariance, qui peuvent se calculer de manière récursive. Si l'hypothèse de normalité n'est pas respectée, le filtre de Kalman permet tout de même d'obtenir la meilleure estimation linéaire pour le problème de filtrage (Sarkka, 2013).
Toutefois, les modèles épidémiologiques sont très souvent non linéaires, et le filtre de Kalman ne peut pas être appliqué de façon directe. Le filtre de Kalman étendu (Extended Kalman filter, EKF, cf. algorithme 7) consiste à réaliser un filtre de Kalman sur une linéarisation du système par la formule de Taylor (Sarkka, 2013). Par exemple, en supposant des bruits additifs dans les équations de transition et d'observation, l'équation (1.6) devient :

$$\begin{cases} x_k = f_{\theta,k-1}(x_{k-1}) + u_{k-1} \\ y_k = g_{\theta,k}(x_k) + v_k \end{cases} \tag{1.10}$$

avec $\{u_k\}, \{v_k\}$ deux bruits blancs gaussiens indépendants de variances respectives U_k et V_k. La loi de filtrage $p_\theta(x_k|y_{1:k})$ est alors approximée par une gaussienne de moyenne m_k et de covariance P_k, qui sont calculées récursivement (Sarkka, 2013).

Algorithm 7 Extended Kalman Filter

On note $m_{k|k-1}$ l'approximation de $\mathbb{E}[x_k|y_{1:k-1}]$
On note $F_{\theta,k}$ la matrice jacobienne de $f_{\theta,k}$ et $G_{\theta,k}$ la matrice jacobienne de $g_{\theta,k}$.

Fixer m_0 et P_0.
for $k = 1 : K$ **do**
  Prédiction :
  $m_{k|k-1} = f_{\theta,k-1}(m_{k-1})$
  $P_{k|k-1} = F_{\theta,k-1}(m_{k-1})P_{k-1}F^T_{\theta,k-1}(m_{k-1}) + U_{k-1}$
  Mise à jour :
  $m_k = m_{k|k-1} + P_{k|k-1}.G^T_{\theta,k}(m_{k|k-1}).S^{-1}_{k|k-1}.(y_k - g_{\theta,k}(m_{k|k-1}))$
  $P_k = P_{k|k-1} - P_{k|k-1}.G^T_{\theta,k}(m_{k|k-1}).S^{-1}_{k|k-1}.G_{\theta,k}(m_{k|k-1}).P_{k|k-1}$
  où $S_{k|k-1} = G_{\theta,k}(m_{k|k-1}).P_{k|k-1}.G^T_{\theta,k}(m_{k|k-1}) + V_k$
end for

Par exemple, le filtre de Kalman étendu a été utilisé pour estimer les paramètres variables d'un modèle épidémiologique (Cazelles and Chau, 1997) (les paramètres variables étant considérés comme des variables d'état).

Afin d'estimer le jeu de paramètres fixes θ, le filtre de Kalman peut être introduit dans une procédure d'optimisation de la vraisemblance ou d'échantillonnage de la distribution a posteriori. Toutefois, étant donné qu'il repose sur une approximation du modèle à espace d'état, les estimations sont biaisées et l'incertitude peut être sous-estimée (Sarkka, 2013). En raison de son moindre coût computationnel par rapport au filtre particulaire, le filtre de Kalman étendu est utilisé au sein d'un algorithme de Metropolis-Hastings adaptatif par Dureau et al. (2013b) comme algorithme d'initialisation, avant d'utiliser le PMCMC (notamment pour initialiser la matrice de covariance de la densité de proposition gaussienne).

Une alternative au filtre de Kalman étendu pour les modèles non-linéaires est le filtre de Kalman d'Ensemble. Dans cet algorithme, le système n'est pas approximé par linéarisation, mais par un ensemble d'états simulés du système (on peut voir l'analogie avec le filtre particulaire). Cet ensemble constitue un échantillon sur lequel peut être calculée la covariance qui intervient dans les équations du filtre de Kalman (Evensen, 2003). Cette méthode a été utilisée en épidémiologie, pour étudier des pathogènes comme la grippe ou encore le virus du Nil occidental (Yang et al., 2014; DeFelice et al., 2017), notamment à des fins de prédiction (Hickmann et al., 2015).

Utilisation de statistiques-résumés : l'exemple des algorithmes ABC

Une autre approche pour l'estimation des paramètres du modèle épidémiologique lorsque la vraisemblance est intractable est l'utilisation d'algorithmes comme l'ABC (Approximate Bayesian Computation), dont un exemple simple est donné par l'algorithme 8 (Marin et al., 2012). Pour cet algorithme, il est seulement nécessaire de pouvoir simuler selon le modèle à espace d'état, et même le calcul de la vraisemblance n'est pas nécessaire, ce qui le rend très adapté pour les modèles complexes, notamment lorsque des données de séquences génétiques sont utilisées (Saulnier et al., 2017).

Algorithm 8 ABC (version algorithme de rejet)

Soit $S(y_{1:K})$ un ensemble de statistiques-résumés de $y_{1:K}$.
Soit $d(.,.) > 0$ une distance.
Soit $\epsilon > 0$ un seuil de tolérance.

1: **for** $i = 1 : N$ **do**
2: Générer $\tilde{\theta} \sim \pi(\theta)$
3: Simuler $\tilde{x}_{1:K} \sim p(x_{1:K}|\tilde{\theta})$ et $\tilde{y}_{1:K} \sim p(y_{1:K}|\tilde{x}_{1:K}, \tilde{\theta})$
4: Accepter $\theta_i = \tilde{\theta}$ si $d\left(S(\tilde{y}_{1:K}), S(y_{1:K})\right) \leq \epsilon$
5: **end for**

L'algorithme 8 présente une version simple de l'algorithme, pour laquelle les paramètres sont proposés de façon non informative, et acceptés ou rejetés uniquement selon un seuil, ce qui le rend peu performant en pratique. De nombreuses extensions faisant appel aux méthodes de MCMC existent, notamment détaillées par Marin et al. (2012), comme par exemple ABC-MCMC ou ABC-SMC (Toni et al., 2009). Des méthodes de régression sont également utilisées pour ajuster les paramètres estimés afin de mieux représenter la distribution a posteriori (cf. par exemple Saulnier et al. (2017)).

En contrepartie de la simplicité d'utilisation, l'ABC repose sur le choix du seuil de tolérance, de la distance et surtout des statistiques-résumés. Certaines méthodes existent pour tenter de calibrer

au mieux ces quantités (Marin et al., 2012). Une autre approche consiste à conserver un grand nombre de statistiques-résumés et les étudier par des méthodes de machine learning comme les forêts aléatoires (Raynal et al., 2017).

Parmi les applications en épidémiologie, McKinley et al. (2009) utilisent des méthodes d'ABC pour estimer les paramètres d'un modèle de transmission de la fièvre Ebola, à partir de données d'incidence et les simulations du modèle, en introduisant également une métrique sur la taille finale de l'épidémie et la date de fin de la dernière infection. Ils comparent leurs résultats à ceux d'un modèle d'augmentation de données, et à des situations où davantage d'informations sont disponibles (nombre d'infectés et non incidence par exemple). Saulnier et al. (2017), eux, utilisent les méthodes d'ABC pour calculer les paramètres épidémiologiques à partir de données phylogénétiques, pour lesquelles l'expression ou le calcul d'une vraisemblance sont parfois complexes.

Comparaison de ces méthodes

Comparaison des différentes méthodes fondées sur le filtrage Bhadra (2010) réalise une comparaison du PMCMC avec le MIF sur les exemples de l'article de Andrieu et al. (2010): sur cet exemple, le MIF est plus rapide d'un point de vue computationnel mais nécessite ensuite de calculer des profils de vraisemblance. Il indique également que le choix entre ces deux algorithmes dépend si le cadre bayésien (et donc l'utilisation de lois a priori) ou le cadre fréquentiste est privilégié dans l'analyse.
Yang et al. (2014) comparent, eux, sur des modèles de grippe, différentes méthodes de filtrage, fondées d'une part sur des filtres particulaires (filtre SMC, MIF et PMCMC) et d'autre part sur des filtres d'ensemble. Dans leur situation, les filtres d'ensemble et le filtre SMC reproduisent mieux les données passées, notamment en présence de pics multimodaux. Ceci est du au fait que ces filtres actualisent les paramètres du système tout au long de la série temporelle, tandis que les filtres particulaires n'actualisent que les variables d'état, les paramètres étant fixes au long de la série. Cependant, les meilleures méthodes de filtrages ne s'avèrent pas les meilleures en prédiction selon leurs analyses (Yang et al., 2014).

Comparaison des approches par filtrage et par réduction d'information Owen et al. (2015a) proposent une comparaison des approches par PMCMC et ABC sur des données simulées. D'après leurs analyses, lorsque le bruit d'observation est élevé, le PMCMC permet des estimations plus fines, notamment en présence de nombreuses observations, et l'ABC ne permet pas de retrouver les paramètres du modèle d'observation. Lorsque le bruit d'observation est faible, le PMCMC présente des problèmes de convergence, la chaîne restant bloquée de longues périodes sur les mêmes valeurs de paramètres. Par ailleurs, le PMCMC présente de grandes difficultés de calibration, et une mauvaise initialisation du point de départ de la chaîne et de la matrice de covariance de la densité de proposition peut conduire à des situations de non-convergence. Une suggestion de leur étude serait d'utiliser l'ABC comme initialisation du PMCMC (Owen et al., 2015b).

Fasiolo et al. (2016) proposent, eux, une comparaison détaillée de plusieurs méthodes, dont certaines sont fondées sur le filtrage particulaire (PMMH et MIF) et d'autres sur la réduction d'information par statistiques-résumés (Synthetic Likelihood (SL) et ABC), à la fois sur des modèles simulés et des exemples en écologie et en épidémiologie (PMMH et SL). D'après leurs travaux, les méthodes de filtrage conduisent à des estimations plus précises, notamment quand le modèle est faiblement identifiable, même si elles peuvent nécessiter des phases de calibration préalable à l'aide

de méthodes plus synthétiques. En revanche, ces méthodes ont de réels désavantages lorsque la vraisemblance est fortement multimodale, ce qui arrive notamment en présence d'un faible bruit de processus et de dynamiques chaotiques. Dans ces situations, le PMMH surestime la stochasticité du système, tandis que les méthodes de réduction d'information semblent plus robustes. Les méthodes par réduction d'information semblent donc plus adaptées pour l'exploration de nouveaux modèles, même si leur calibration (choix de statistiques-résumé, par exemple) peut être contraignante, tandis que les méthodes par filtrage permettent de finaliser les estimations d'un modèle préalablement établi.

Autres méthodes d'inférence

On a détaillé ici des méthodes reposant sur l'écriture du modèle à espace d'état et les méthodes sans vraisemblance, notamment en raison de leur caratère "plug-and-play". Il existe cependant d'autres approches pour traiter les situations où la vraisemblance est intractable, pour lesquelles le lecteur peut se référer à l'introduction par Britton and Giardina (2016).

Parmi celles-ci, on peut citer les méthodes par augmentation des données, dans lesquelles des variables latentes (comme les dates et types des événements non observés) sont introduites dans le modèle (Gibson and Renshaw, 1998; O'Neill and Roberts, 1999). La vraisemblance devient tractable pour le modèle complété par les variables latentes, et l'inférence menée à l'aide de l'algorithme E-M ou de méthodes de MCMC. Ces méthodes sont aussi utilisées pour étudier des données individuelles (Auranen et al., 2000; Cauchemez et al., 2004), afin de retrouver les lieux principaux de la transmission (foyer, école, par exemple).

Une autre approche consiste à approximer le modèle en temps continu par un modèle en temps discret (Finkenstädt and Grenfell, 2002). Selon ce modèle, nommé TSIR (time series susceptible infected recovered), les données sont agrégées selon un pas de temps t correspondant à la durée caractéristique de la maladie (dans le cas de la rougeole, la durée entre l'infection et la guérison), et les infections ayant lieu à la période t sont liées aux nombres d'individus infectés et de susceptibles à la date $t-1$. Cette écriture récursive conduit à des expressions tractables de la vraisemblance, permettant l'estimation par moindres carrés ou par MCMC par exemple.

1.2.5 Comparaison et évaluation des modèles

La complexité des modèles épidémiologiques et la multiplicité des effets qui peuvent être représentés nécessite des méthodes pour comparer les différents modèles. Ceci peut permettre de comparer la plausibilité de différentes hypothèses biologiques (King et al., 2008; Camacho et al., 2011), ou de mettre en évidence les mécanismes les plus importants dans la dynamique épidémique observée, afin de conserver des modèles parsimonieux (Pandey et al., 2013). La comparaison de modèles épidémiologiques est une question non triviale (Gibson et al., 2018), les méthodes existantes pour certains contextes sont parfois inadaptées à d'autres, par leur difficulté d'implémentation ou leur manque de fiabilité. Toutefois, de nombreuses stratégies ont été proposées, dont nous présentons quelques unes ici. La revue de Gibson et al. (2018) synthétise beaucoup d'entre elles, dans une perspective en lien avec l'estimation fondée sur la vraisemblance complétée par augmentation de données via MCMC.

Critères d'information

Critère d'information d'Akaike (AIC) (Akaike, 1974) Le critère AIC se calcule très simplement à partir de la log-vraisemblance $\log L(y, \theta)$ calculée en son maximum $\hat{\theta}$, pénalisée par le nombre de paramètres du modèle (noté p):

$$AIC = -2\log(L(y, \hat{\theta})) + 2p$$

Cette quantité est une approximation de l'espérance de l'information de Kullback-Leibler relative, la pénalisation par le nombre de paramètres étant une correction de biais. Pour les échantillons de petite taille n, une version de l'AIC dite "AIC corrigé" (AICc) existe, prenant en compte la correction du biais à l'ordre 2 (Burnham and Anderson, 2004):

$$AICc = AIC + \frac{2p(p+1)}{n-p-1}$$

Le critère AIC a été fréquemment utilisé en épidémiologie, par exemple par King et al. (2008); Camacho et al. (2011); Pandey et al. (2013).

Critère d'information DIC (Deviance Information Criterion) (Spiegelhalter et al., 2002) Le DIC proposé par Spiegelhalter et al. (2002) repose sur la notion de déviance :

$$D(\theta) = -2\log(L(y, \theta)) + 2\log(h(y))$$

où $h(y)$ est un terme de normalisation qui ne dépend que des données y, et qui est fixé à $h(y) = 1$ dans le cadre de la comparaison de modèles. Le DIC repose également sur la notion de "nombre effectif de paramètres", définie comme

$$p_D = \mathbb{E}_{\theta|y}[D(\theta)] - D(\tilde{\theta}(y))$$

où $\tilde{\theta}(Y)$ est un choix d'estimateur ponctuel du paramètre d'intérêt θ, comme par exemple la moyenne a posteriori $\mathbb{E}_{\theta|y}[\theta] = \bar{\theta}$. Le DIC se définit alors comme (Celeux et al., 2006):

$$\begin{aligned} DIC &= \mathbb{E}_{\theta|y}\left[D(\theta)\right] + p_D && (1.11)\\ &= D(\tilde{\theta}(y)) + 2p_D && (1.12)\\ &= 2\,\mathbb{E}_{\theta|y}\left[D(\theta)\right] - D(\tilde{\theta}(y)) && (1.13)\\ &= -4\,\mathbb{E}_{\theta|y}\left[\log(L(y, \theta))\right] + 2\log(L(y, \tilde{\theta}(y)) && (1.14) \end{aligned}$$

L'expression (1.12) fait apparaître la similitude avec le critère AIC, et Spiegelhalter et al. (2002) motivent le choix de p_D comme mesure du nombre de paramètres en utilisant plusieurs arguments, par exemple à travers une approximation gaussienne du posterior et lorsque $\tilde{\theta}(y) = \bar{\theta}$. Leur raisonnement sur ce point est le suivant : en développant $D(\theta)$ au voisinage de $\bar{\theta}$, on a

$$
\begin{aligned}
\mathbb{E}_{\theta|y}[D(\theta)] &\approx \mathbb{E}_{\theta|y}\left[D(\bar{\theta}) + (\theta-\bar{\theta})^T \frac{\partial D}{\partial \theta}(\bar{\theta}) + \frac{1}{2}(\theta-\bar{\theta})^T \frac{\partial^2 D}{\partial \theta^2}(\bar{\theta})(\theta-\bar{\theta})\right] \\
&= D(\bar{\theta}) + \mathbb{E}_{\theta|y}\left[(\theta-\bar{\theta})^T \frac{\partial^2 \log L}{\partial \theta^2}(y,\bar{\theta})(\theta-\bar{\theta})\right] \\
&= D(\bar{\theta}) - \mathbb{E}_{\theta|y}\left[tr\left(\frac{\partial^2 \log L}{\partial \theta^2}(y,\bar{\theta})(\theta-\bar{\theta})(\theta-\bar{\theta})^T\right)\right] \\
&= D(\bar{\theta}) - tr\left(\frac{\partial^2 \log L}{\partial \theta^2}(y,\bar{\theta})\mathbb{E}_{\theta|y}\left[(\theta-\bar{\theta})(\theta-\bar{\theta})^T\right]\right) \\
&= D(\bar{\theta}) + tr\left(-\frac{\partial^2 \log L}{\partial \theta^2}(y,\bar{\theta})V\right) \\
&= D(\bar{\theta}) + tr\left(-\frac{\partial^2 p(\theta|y)}{\partial \theta^2}(\bar{\theta})V\right) - tr\left(-\frac{\partial^2 p(\theta)}{\partial \theta^2}(\bar{\theta})V\right) \\
&= D(\bar{\theta}) + tr\left(-Q_{\bar{\theta}}V\right) - tr\left(-P_{\bar{\theta}}V\right) \\
&\approx D(\bar{\theta}) + tr\left(V^{-1}V\right) - tr\left(-P_{\bar{\theta}}V\right) \text{ en utilisant l'approximation gaussienne} \\
&= D(\bar{\theta}) + p - tr\left(-P_{\bar{\theta}}V\right)
\end{aligned}
$$

On a ainsi $p_D \approx p - tr\left(-P_{\bar{\theta}}V\right)$, dont le second terme tend vers zéro lorsque l'information a priori est négligeable.

Le DIC peut également s'interpréter en (1.11) comme une mesure de l'adéquation du modèle (déviance moyenne a posteriori : $\mathbb{E}_{\theta|y}[D(\theta)]$) pénalisée par une mesure de sa complexité (p_D), et le meilleur modèle est celui pour lequel le DIC est le plus faible (Spiegelhalter et al., 2002). Ce critère s'adapte notamment aux situations pour lesquelles la définition de la complexité du modèle n'est pas directe (par exemple quand des priors informatifs sont utilisés), et son expression permet de le calculer facilement à partir des résultats d'une chaîne de MCMC.

Bien que souvent utilisé en pratique, ce critère a été l'objet de fort débats théoriques (cf. par exemple la discussion de Spiegelhalter et al. (2002)). De plus, certains aspects d'utilisation ne sont pas toujours évidents, comme le choix de l'estimateur $\tilde{\theta}(Y)$ (moyenne a posteriori, mais aussi médiane ou mode) qui est laissé à l'utilisateur et peut avoir une influence sur la pertinence du DIC (la quantité p_D pouvant s'avérer négative). Celeux et al. (2006) formulent plusieurs "DIC observés", correspondant à des choix différents pour $\tilde{\theta}(Y)$, ainsi que plusieurs versions de l'indicateurs dans le cadre des modèles à données manquantes.

Autres difficultés Ces méthodes fondées sur la vraisemblance dépendent donc du jeu de données utilisé. Dans les situations où l'on compare des modèles estimés à partir d'observations différentes (par exemple, un modèle agrégé et un modèle dans lequel les données sont sub-divisées par aire géographique, ou sérotype de virus), ces méthodes ne sont pas appropriées.

Vérifications "posterior predictive"

Un des avantages du contexte bayésien est la possibilité de re-simuler le modèle préalablement estimé, et, sous l'hypothèse que le modèle soit valable, il est un pré-requis désirable que les simulations du modèle (notées y^{rep}) ressemblent aux vraies données (y) (Gelman et al., 2013).

Les vérifications "posterior predictive" reposent sur cette idée, en comparant des mesures de "discrepancy" calculées sur les données simulées $T(y^{rep}, \theta)$ à celles calculées sur les vraies observations $T(y, \theta)$. Il est alors possible de calculer une "p-valeur posterior predictive" (PPP-valeur) qui correspond à la proportion de simulations pour lesquelles $T(y^{rep}, \theta) > T(y, \theta)$. Des valeurs proches de 0 ou de 1 pour cette proportion indiquent des doutes sur la validité du modèle. On peut également tracer le graphique des valeurs de $T(y^{rep}, \theta)$ en fonction de celles de $T(y, \theta)$ afin d'évaluer dans quelle mesure elles diffèrent (notamment en les comparant au cas idéal dans lequel on aurait une égalité parfaite, c'est-à-dire $T(y^{rep}, \theta) = T(y, \theta)$). Ces éléments de diagnostic peuvent éclairer sur les défauts du modèle, afin de comprendre ses limites et de proposer des améliorations.

Gibson et al. (2018) citent plusieurs travaux qui utilisent des PPP-valeurs ou des vérifications "posterior predictives", notamment utilisant des indicateurs d'autocorrélation spatiale comme mesure de "discrepancy". Camacho et al. (2011) utilisent également des graphiques de "discrepancies" en complément de l'indicateur AIC dans leur comparaison de modèles de grippe.

Une limite à l'utilisation des PPP-valeurs vient du fait que celles-ci n'ont pas la propriété des p-valeurs fréquentistes d'uniformité sous l'hypothèse que le modèle est le bon. Elles ont notamment tendance à être plus conservatrices car les données sont utilisées deux fois (pour estimer les paramètres et pour évaluer le modèle). Sans connaître leur distribution, il est difficile d'interpréter leur valeur afin de savoir si celle-ci est ou non extrême. Cette critique a conduit à des formulations alternatives (parfois appelées "u-valeurs"). Gelman et al. (2013) insistent toutefois sur le fait que cette propriété ne constitue pas une limitation à l'utilisation des diagnostics graphiques et à l'utilisation des "discrepancies" pour mettre en évidence les faiblesses d'un modèle.
Gibson et al. (2018) évoquent, eux, la difficulté du choix des mesures de "discrepancy", qu'ils comparent à la difficulté du choix des statistiques-résumés dans l'estimation par ABC. Ils insistent sur l'importance de choisir des indicateurs en lien avec les utilisations futures du modèle. Ils introduisent également des extensions de ces méthodes, fondées sur les résidus latents (Gibson et al., 2018; Lau et al., 2014).

1.3 Modéliser les arboviroses

1.3.1 Les arboviroses

Les arboviroses sont des maladies causées par des virus ayant pour vecteurs des arthropodes, comme les moustiques ou les tiques ("**ar**thropod-**bo**rne viruses" ou "**arbo**viruses" en anglais). Le virus se transmet entre l'hôte et le vecteur lorsque ce dernier prélève le sang d'un hôte pour se nourrir. Il existe de nombreuses arboviroses, affectant les humains comme les animaux: nous nous concentrerons ici sur les arboviroses humaines. Certaines sont connues et étudiées depuis plusieurs décennies comme la fièvre jaune, la dengue, le chikungunya, l'encéphalite japonaise ou l'encéphalite à tiques ; d'autres ont été découvertes ou réétudiées plus récemment comme celles causées par les virus Zika, Mayaro ou Spondweni (Gould et al., 2017). Les symptômes associés à ces maladies peuvent être très divers, allant de manifestations très légères à de sévères séquelles neurologiques ou des fièvres hémorragiques parfois mortelles. Toutefois, ces virus sont proches d'un point de vue antigénique, ce qui peut compliquer leur diagnostic sérologique, et induire des réactions immunitaires croisées au sein de l'hôte lors d'infections successives.

La présence de ces maladies à l'échelle du globe est liée à celle de leurs vecteurs qui dépend fortement des conditions climatiques et environnementales (Kraemer et al., 2015). Toutefois, les échanges internationaux jouent aussi un rôle décisif dans l'introduction des virus et des vecteurs dans de nouvelles régions, comme ce fut probablement le cas autour du XVIIe siècle sur le continent américain pour le virus de la fièvre jaune (Gould et al., 2017) ou le moustique *Aedes aegypti* (Brown et al., 2014). Les changements climatiques, l'intensification des échanges et les modifications de l'environnement liées à la déforestation ou l'urbanisation contribuent à l'expansion de ces maladies vers de nouvelles aires géographiques (Messina et al., 2014; Gould et al., 2017), comme en témoignent par exemple les premières épidémies de chikungunya dans sud de l'Europe (Rezza et al., 2007; Delisle et al., 2015). La compréhension des mécanismes de leur propagation est donc d'autant plus importante.

Dans le cadre de ce travail, deux arboviroses ont été étudiées en particulier: la dengue et le virus Zika.

La dengue

La dengue est causée par un flavivirus transmis par les moustiques *Aedes*. Elle se manifeste par un type grippal sévère, pouvant conduire à de graves complications, comme la fièvre hémorragique ou le syndrome de choc. Néanmoins, de nombreuses infections par la dengue ne présentent que des symptômes faibles, voire inexistants, qui ne nécessitent pas le recours aux services de santé (Grange et al., 2014). Ces cas dits "inapparents" pouvant tout de même transmettre le virus (Duong et al., 2015), ils jouent un rôle important dans la propagation de la maladie. On estime environ 390 millions d'infections par la dengue chaque année dans le monde, dont 50 à 100 millions de cas symptomatiques ce qui constitue un problème majeur de santé publique, qui a de surcroît de lourdes conséquences économiques (Bhatt et al., 2013; Shepard et al., 2016).

L'épidémiologie de la dengue est très liée à l'écologie de ses vecteurs, les deux principaux étant *Aedes aegyti* et *Aedes albopictus*. Le premier est le principal vecteur de la maladie, et il est présent dans l'ensemble des régions tropicales du globe (Kraemer et al., 2015). Le second est un vecteur moins compétent (Lambrechts et al., 2010) mais il est plus répandu à l'échelle mondiale, notamment dans les régions tempérées (Kraemer et al., 2015). Les conditions climatiques et environnementales influencent fortement la transmission du virus (Naish et al., 2014), par leur influence sur la durée de vie des vecteurs, leur reproduction, ou leur capacité à transmettre le virus (durée d'incubation, probabilité de transmission) (Liu-Helmersson et al., 2014). La dengue est endémique dans la plupart des régions tropicales et subtropicales du globe, et connaît parfois des dynamiques saisonnières, marquée par la résurgence annuelle d'épidémies d'ampleur variable, comme par exemple en Asie du Sud-Est (Teurlai et al., 2012; Nisalak et al., 2016; Yap et al., 2016). La survie du pathogène lors des saisons peu propices à la transmission peut s'expliquer par des mécanismes de transmission verticale au sein du vecteur (Lequime and Lambrechts, 2014), ou par la présence de cas asymptomatiques permettant une circulation à bas bruit tout au long de l'année (Adams and Boots, 2010).

Si la dengue et son vecteur sont fortement influencés par les facteurs climatiques, ils dépendent aussi de facteurs humains. Le vecteur *Aedes aegyti* est particulièrement anthropophile et les petits réservoirs d'eau autour des habitats humains constituent de nombreux lieux de ponte. Les facteurs socio-économiques ont donc une forte influence sur la propagation de la maladie, car des éléments comme la climatisation par exemple peuvent avoir un rôle pour limiter la présence de moustiques

et ainsi la transmission (Reiter et al., 2003). Toutefois, ces relations sont complexes et changeantes, la dengue affectant des zones aux caractéristiques diverses, selon des schémas différents d'année en année (Telle et al., 2016). Le fait que les moustiques *Aedes* soient actifs le jour et se déplacent dans un faible rayon donne une importance prépondérante aux lieux fréquentés en journée en plus des lieux d'habitation, et ainsi aux déplacements humains. Les humains sont sûrement les acteurs de la dispersion de la maladie à l'échelle d'une zone urbaine (Stoddard et al., 2013) comme à l'échelle d'un pays (Teurlai et al., 2012). Historiquement la dengue a été surtout décrite dans les milieux urbains et associée au phénomène d'urbanisation mais de plus en plus de travaux décrivent l'importance de la dengue en zone rurale. De manière générale, les cas de dengue présentent une signature spatiale très spécifique, caractérisée par un fort regroupement à très fine échelle (Salje et al., 2012, 2017).

Le virus est présent sous la forme de quatre sérotypes, qui coexistent et interagissent entre eux. Si l'infection par un sérotype confère généralement une immunité définitive contre ce sérotype, elle ne protège pas d'une infection par un autre, sauf de façon temporaire (Reich et al., 2013). Les secondes infections conduisent même plus souvent que les premières aux complications graves de la maladie (Sangkawibha et al., 1984; Kouri et al., 1989; Graham et al., 1999; Hammond et al., 2005). L'ADE (antibody-dependent enhancement) est un mécanisme fréquemment invoqué pour expliquer ce phénomène: les anticorps liés à une infection antérieure auraient un rôle contre-productif et contribueraient à renforcer une infection secondaire (Halstead, 2007). Cet effet qui n'est pas observé systématiquement pourrait être lié à des valeurs spécifiques de concentration pour les anticorps (Katzelnick et al., 2017).

Bien que ces mécanismes de renforcement croisé rendent complexe la mise au point d'un vaccin, le premier vaccin contre la dengue a été commercialisé en 2016 (WHO, 2016a). Toutefois, les résultats sont à ce jour contrastés quant à son efficacité, ses effets à long terme et ses potentiels dangers (Flasche et al., 2016), et la campagne de vaccination initiée aux Philippines a été interrompue pour raisons de sécurité (Dyer, 2017). Comme il n'existe pour l'heure pas de traitement spécifique contre la maladie, le contrôle vectoriel est la première stratégie de lutte. De nombreuses méthodes existent (combinaisons d'insecticides, utilisation de moustiques génétiquement modifiés, de moustiques irradiés et stériles, ou de moustiques infectés par la bactérie *Wolbachia*), mais malgré leurs effets importants sur la population de moustiques, leur efficacité à réduire l'incidence de la dengue n'est pas fermement établie (Bowman et al., 2016).

Le Zika

Découvert en 1947 dans la forêt Zika en Ouganda, le virus Zika était considéré jusqu'au milieu des années 2010 comme un virus essentiellement bénin, pouvant causer des rougeurs cutanées, des douleurs articulaires, des conjonctivites et des fièvres légères, et très souvent des symptômes presque imperceptibles (Duffy et al., 2009). Toutefois, suites aux épidémies en Polynésie Française et au Brésil, ce virus a été associé à de graves complications neurologiques : le syndrôme de Guillain-Barré (Cao-Lormeau et al., 2016) et des malformations congénitales comme la microcéphalie (Schuler-Faccini, 2016; Rasmussen et al., 2016). Ces complications sont rares, mais, associées aux forts taux d'attaque du virus, elles ont conduit à un nombre de cas important, et l'Organisation mondiale de la Santé (OMS) a déclaré en 2016 l'épidémie de microcéphalies au Brésil comme une urgence de santé publique de niveau international (WHO, 2016b). A l'heure actuelle, on ne connaît pas l'étendue des conséquences d'une infection congénitale sur le développement infantile, mais des effets non décelables à la naissance peuvent également exister

(Broussard et al., 2018).

Avant les années 2010, le virus a circulé en Afrique, en Asie et dans le Pacifique (Petersen et al., 2016), avant que la première épidémie de grande ampleur soit documentée en Micronésie sur l'île de Yap en 2007 (Duffy et al., 2009). Le virus a ensuite causé des épidémies en Polynésie française (2013-2014), en Nouvelle Calédonie (2014), sur l'île de Pâques (2014), et les îles Cook (2015), puis sur le continent latino-américain à partir de 2015 (Petersen et al., 2016). Toutefois, des études phylogénétiques (Faria et al., 2016) ainsi que l'analyse retrospective d'échantillons sanguins (Passos et al., 2017) indiquent la présence du virus au Brésil dès l'année 2013.

Le virus Zika est essentiellement transmis par les moustiques *Aedes*, mais d'autres modes de transmission, notamment sexuelle, maternofoetale, ou via transfusion sanguine ont été mis en évidence (Musso and Gubler, 2016). Le virus possède une structure très proche de celle de la dengue, ce qui induit des difficultés de diagnostic sérologique chez les patients séropositifs pour la dengue (Petersen et al., 2016). Des possibilité d'interaction entre Zika et d'autres arbovirus, via des mécanismes comme l'ADE ont également été mis en évidence (Martín-Acebes et al., 2018; Rey et al., 2018).

Les dynamiques des arboviroses dépendent donc de l'entremêlement complexe de multiples facteurs humains, environnementaux et immunologiques. Face à cette complexité, les modèles mathématiques constituent un outil important pour mieux appréhender ces systèmes et éclairer les décisions de santé publique. Ils permettent en effet de formuler des prédictions, de tester et comparer différentes mesures de prévention ou de contrôle, ou d'évaluer la plausibilité de plusieurs hypothèses biologiques en comparant des modèles reposant sur différents mécanismes. On présentera donc dans la section 1.3.2 différents types de modèles mathématiques utilisés en épidémiologie pour représenter la dengue et le Zika.

1.3.2 Quels modèles pour la dengue et le Zika ?

Modélisation de la dengue

L'étude des dynamiques de la dengue a conduit à de nombreux travaux de modélisation, notamment synthétisés dans les revues d'Andraud et al. (2012) sur les modèles déterministes, de Johansson et al. (2011) qui insiste sur les conséquences pour la vaccination et de Perkins et al. (2014). En raison de la multiplicité de facteurs influençant la transmission de la maladie, les éléments pris en compte dans les modèles sont très divers, incluant tantôt l'écologie du vecteur, la co-existence des sérotypes, ou encore la structure de la population.

Prise en compte du vecteur de transmission Certains auteurs privilégient une modélisation explicite de la population de moustiques, avec différents niveaux de précision. Souvent, un modèle de type Ross-MacDonald est utilisé (Reiner et al., 2013), la population de moustiques étant représentée par une dynamique SI ou SEI (Wearing and Rohani, 2006; Lourenço and Recker, 2010). Dans certains cas, le cycle de vie des vecteurs est détaillé de façon très fine, en représentant les stades aquatiques d'oeuf ou de larve (Burattini et al., 2008; Erickson et al., 2010), ou tenant compte de la transmission verticale (Burattini et al., 2008; Adams and Boots, 2010). Souvent, dans les modèles hôtes-vecteurs, certains paramètres liés au cycle de vie du moustiques sont supposés variables au cours du temps, dépendant par exemple de façon plus ou moins complexe de variables climatiques (Erickson et al., 2010; Lourenço and Recker, 2014).

D'autres modélisent directement les variations saisonnières, sans avoir recours à la formulation explicite d'une population de moustiques, par exemple en utilisant un taux de transmission variable selon une fonction périodique (Adams et al., 2006; Aguiar et al., 2011).

Sérotypes et interactions Bien qu'on dénombre quatre sérotypes de virus, dans certaines situations, un seul sérotype est représenté. C'est le cas de l'article d'Erickson et al. (2010), parce qu'il étudie les possibilités d'introduction de la maladie dans une zone où elle n'a jamais sévi (Texas). C'est aussi le cas du modèle de Burattini et al. (2008), plutôt tourné vers l'étude du vecteur de transmission que vers les interactions entre sérotypes. A l'inverse, certains travaux incorporent les quatre sérotypes (Chikaki and Ishikawa, 2009; Nagao and Koelle, 2008; Recker et al., 2009; Coudeville and Garnett, 2012; Reich et al., 2013), leurs interactions pouvant être à l'origine des fluctuations pluriannuelles de la maladie. Dans certains cas, l'immunité est totale après deux infections, dans d'autres cas il faut attendre la quatrième infection. Beaucoup de modèles, enfin, retiennent seulement deux sérotypes de virus, afin de prendre en compte ces effets tout en modérant la complexité du modèle (Ferguson et al., 1999; Aguiar et al., 2008). Les sérotypes peuvent être supposés identiques (Ferguson et al., 1999; Aguiar et al., 2011) ou différer selon certaines caractéristiques comme leur transmissibilité par exemple (Adams and Boots, 2006).

Lorsqu'un modèle à plusieurs sérotypes est retenu, différents mécanismes d'interaction entre eux peuvent être introduits. En particulier, deux types d'interactions ont été souvent modélisés: le renforcement croisé (cross enhancement) et l'immunité croisée (cross immunity).
Plusieurs formes de renforcement croisé ont été pris en compte dans les modèles. En premier lieu, c'est la transmissibilité du virus qui peut être modifiée lors de la seconde infection. Dans la plupart des modèles, on parle de renforcement de transmissibilité, phénomène soutenu par l'hypothèse d'une plus forte charge virale des individus secondo-infectés (Ferguson et al., 1999; Chikaki and Ishikawa, 2009). Aguiar et al. (2008) soutiennent à l'inverse une baisse de la transmissibilité en raison de la plus forte propension à être hospitalisé. D'autres types de renforcements croisés peuvent exister, notamment le renforcement de la susceptibilité (les individus ayant déjà connu une première infection sont plus souvent infectés) (Adams and Boots, 2006; Recker et al., 2009) ou le renforcement de la mortalité lors de la deuxième infection (Adams and Boots, 2006)
L'immunité croisée a également été modélisée sous plusieurs formes. Dans certains cas, il s'agit d'une immunité partielle, réduisant la probabilité de réinfection (Adams and Boots, 2006). D'autres travaux considèrent une immunité croisée temporaire, au terme de laquelle l'individu peut à nouveau être infecté par les autres sérotypes. Wearing and Rohani (2006) montrent par exemple que la prise en compte d'une période d'immunité croisée temporaire permet de reproduire les fluctuations pluriannuelles de la maladie, et sa durée a été évaluée entre six mois et trois ans (Coudeville and Garnett, 2012; Reich et al., 2013). Des auteurs modèlisent également une immunité temporaire contre les symptômes mais non contre l'infection (clinical cross protection): certains individus pourraient ainsi être réinfectés sans aucune manifestation clinique apparente et ainsi acquérir l'immunité contre un autre sérotype (Wearing and Rohani, 2006; Chikaki and Ishikawa, 2009).

Structure de la population et hétérogénéités Certains modèles prennent en compte des phénomènes de structure de la population, comme la structure par âge (Chikaki and Ishikawa, 2009; Pongsumpun and Tang, 2003), qui est particulièrement utile dans l'étude des stratégies de vaccination (Coudeville and Garnett, 2012).
La structure spatiale a été très largement étudiée, souvent à l'aide de modèles individus centrés (Barmak et al., 2011; Lourenço and Recker, 2013; Karl et al., 2014; Reiner et al., 2014; Kang

and Aldstadt, 2017), mais aussi de modèles à métapopulations (Sarzynska et al., 2013; Amaku et al., 2016). Elle permet de générer des dynamiques spécifiques, non reproductibles à une échelle agrégée (Lourenço and Recker, 2013; Amaku et al., 2016). En particulier, Lourenço and Recker (2013) parviennent à reproduire les dynamiques pluriannuelles de la maladie sans interactions entre les quatre sérotypes représentés mais en introduisant des phénomènes spatiaux, d'hétérogénéité et de stochasticité. Les contacts entre les différentes unités spatiales peuvent également dépendre d'une structure en réseau (Stolerman et al., 2015; Nevai and Soewono, 2014), d'hypothèses paramétriques sur la mobilité (Sarzynska et al., 2013) ou d'informations externes comme les données téléphoniques (Wesolowski et al., 2015). L'échelle spatiale considérée peut aller du foyer (Favier et al., 2005; Reiner et al., 2014) ou des quartiers ou blocs au sein d'une ville (Stolerman et al., 2015; Kang and Aldstadt, 2017), à des échanges entre villes (Nevai and Soewono, 2014) ou entre provinces (Sarzynska et al., 2013). Par ailleurs, certains modèles non spatiaux incluent toutefois la possibilité de rencontrer des individus infectés provenant d'une autre population, représentée par une constante dans le taux de transmission (Nagao and Koelle, 2008; Aguiar et al., 2011), notamment pour éviter les phénomènes d'extinction du virus.

Les individus peuvent également être structurés en termes d'intensité de la maladie. Dans certains modèles, les cas asymptomatiques ou sévères sont explicitement modélisés (Coudeville and Garnett, 2012), et peuvent jouer des rôles distincts dans la transmission (Chikaki and Ishikawa, 2009). Sinon, les cas inapparents peuvent être modélisés implicitement en appliquant un taux de report entre le nombre de cas prédits par le modèle et le nombre de cas observés (Reich et al., 2013). Afin de prendre en compte la plus grande probabilité d'hospitalisation des infections secondaires, Aguiar et al. (2011) supposent que seuls les cas secondaires sont reportés parmi les cas graves.

Modélisation du Zika

Les modèles sur les dynamiques du Zika sont beaucoup plus récents, débutant en 2016 avec l'irruption de la maladie parmi les urgences de santé publique. En raison du peu d'information sur le virus Zika à cette période, il a tout d'abord été modélisé de façon similaire à la dengue, avant que des modèles plus spécifiques à Zika soient développés. Suivant la chronologie des épidémies et donc des données disponibles, les premiers modèles ont été adaptés aux épidémies du Pacifique (Kucharski et al., 2016; Funk et al., 2016; Nishiura et al., 2016a), puis aux épidémies du continent américain (Ferguson et al., 2016; Villela et al., 2017; Towers et al., 2016a), à des échelles allant d'une île ou d'une ville (Rojas et al., 2016; Lourenço et al., 2017) jusqu'à l'ensemble d'un continent (Zhang et al., 2017).

Eléments de modélisation inspirés par d'autres arboviroses Les premiers modèles de Zika ont transposé des modèles conçus pour d'autres arboviroses, notamment la dengue, afin de permettre de premières estimations et prédictions (Kucharski et al., 2016; Nishiura et al., 2016a; Towers et al., 2016a). L'hypothèse de la proximité entre la transmission des différents virus a elle-même été étudiée, à travers des modèles qui comparent les caractéristiques de la transmission du Zika à celles du chikungunya (Riou et al., 2017) ou de la dengue (Funk et al., 2016). En particulier, selon Funk et al. (2016), il semble que les différences entre contextes géographiques soient plus grandes que les différences entre pathogènes. De fait, de nombreux travaux ont également souligné les différences dans la propagation du virus Zika en différents lieux (Rojas et al., 2016; He et al., 2017), et certains ont incorporé des éléments écologiques ou climatiques dans les modèles pour mieux les adapter aux contextes étudiés (Lourenço et al., 2017; He et al., 2017). La propagation spatiale du virus a également été explicitement modélisée, à travers des modèles individus-centrés (Zhang et al., 2017) ou des modèles d'équations aux dérivées partielles (Fitzgibbon et al., 2017).

Eléments de modélisation spécifiques à Zika Certains modèles ont également incorporé des éléments spécifiques de la transmission du Zika, comme la présence de deux modes de transmission, vectorielle et sexuelle (Gao et al., 2016; Towers et al., 2016a). Malgré des méthodes différentes, notamment pour dériver l'expression du nombre de reproduction de base, les deux études concluent à un faible impact de la transmission sexuelle, insuffisante pour maintenir seule l'épidémie, mais pouvant avoir des conséquences sur certaines classes d'âge de la population.
En raison des complications liées au virus, l'âge des individus infectés ou les facteurs de risques ont également été incorporés dans les modèles. Aussi les prédictions des modèles ont-elles été adaptées afin de mettre en évidence les risques pour les femmes enceintes ou en âge d'avoir des enfants (Perkins et al., 2016b), parfois à travers une modélisation fine de la structure d'âge de la population (Zhang et al., 2017). Andronico et al. (2017) modélisent, eux, les cas sévères de Guillain-Barré afin de planifier les besoins hospitaliers en Martinique.

1.4 Objectif de la thèse (résumé substantiel en langue française - I)

L'importance et l'expansion des arboviroses nécessite des modèles pour comprendre et prédire leurs dynamiques ou expérimenter des scénarios de vaccination et de contrôle. Comme ces maladies dépendent de multiples facteurs humains, environnementaux et immunologiques, de très nombreux éléments peuvent être incorporés dans la construction de modèles réalistes d'un point de vue biologique et épidémiologique. Dans le cadre de ce travail de thèse, différents modèles représentant la dynamique des arboviroses sont confrontés à des données observées, à l'aide de méthodes d'estimation bayésiennes. Il s'agit de mettre en évidence les modèles les plus adaptés aux données disponibles ainsi que les hypothèses qui s'avèrent trop restrictives, et d'identifier les données supplémentaires permettant de lever les incertitudes encore présentes.

Afin d'estimer les modèles à partir des données observées, qui sont incomplètes et bruitées, tous les éléments caractéristiques de la propagation des arboviroses ne peuvent pas être pris en compte simultanément. Il est nécessaire d'identifier les éléments de modélisation les plus importants à incorporer et ceux qui peuvent être simplifiés dans un souci de parcimonie. Le chapitre 2 s'intéresse à cette question, dans le cas des dynamiques de la dengue en milieu rural, contexte dans lequel encore peu de travaux de modélisation ont été réalisés. Il s'agit de comparer des modèles à la complexité croissante en les confrontant à un double jeu de données, récolté dans la région de Kampong Cham, au Cambodge. L'augmentation de la complexité des modèles est étudiée selon deux axes. Dans un premier temps, c'est la complexité en termes de mécanismes épidémiologiques qui est explorée : six modèles déterministes sont comparés, incorporant tour à tour la dynamique vectorielle, les cas asymptomatiques et les différents sérotypes de virus. Dans un second temps, c'est la complexité en termes d'implémentation du modèle qui est explorée, en incorporant, en plus du bruit d'observation, de la stochasticité démographique et de la stochasticité environnementale.

Un autre élément important de la propagation de la dengue qui n'a pas pu être exploré avec les données précédentes est la dynamique spatiale. Aussi le chapitre 3 s'intéresse-t-il à cet aspect, à travers les séries de cas de dengue observées à Rio de Janeiro, au Brésil. Des modèles incorporant les phénomènes de saisonnalité et de structure spatiale entre les quartiers de la ville sont alors explorés.

Enfin, le chapitre 4 applique les méthodes et les modèles utilisés précédemment à l'étude d'une arbovirose émergente encore méconnue, le virus Zika. Dans un premier temps, il s'agit

d'utiliser les données d'épidémies survenues dans le Pacifique afin d'estimer les paramètres-clé de la propagation du virus et d'explorer leur variabilité. Ce travail prolonge les travaux déjà publiés à ce sujet en étendant les modèles au cadre stochastique, et en étudiant la variabilité des résultats sous l'angle de la variabilité géographique, par la comparaison de quatre îles de tailles différentes, et sous l'angle de la variabilité du modèle épidémiologique, par la comparaison de deux modèles à transmission vectorielle. Afin de concilier les données d'incidence et de séroprévalence, un nouveau paramètre est inclus dans le modèle, représentant la part de la population participant à l'épidémie. Dans un second temps, la question des interactions potentielles du virus Zika avec celui de la dengue, qui a circulé à la même période, est explorée à l'aide de données sur l'île de Tahiti (Polynésie française) et la ville de Rio de Janeiro (Brésil).

Le chapitre 5 présente une synthèse et une discussion de ces travaux. Des perspectives de prolongement sont discutées, en lien avec les questions d'identifiabilité, d'utilisation de sources de données diverses et de comparaison de modèles.

Les chapitres qui vont suivre reprennent les travaux suivants :

- Champagne C., Salthouse D.G., Paul R., Cao-Lormeau V.-M., Roche B., Cazelles B., Structure in the variability of the basic reproductive number (R_0) for Zika epidemics in the Pacific islands. *eLife*, 2016 — **présenté au chapitre 4, section 4.2**
- Champagne C., Paul R., Ly S., Duong V., Leang R., Cazelles B., Dengue modeling in rural Cambodia: statistical performance versus epidemiological relevance, *Epidemics*, 2018 — **présenté au chapitre 2, section 2.2**
- Champagne C., Cazelles B., Comparison of stochastic and deterministic frameworks in dengue modeling, *preprint en révision* — **présenté au chapitre 2, section 2.3**

ainsi que

- Cazelles B., Champagne C., Dureau J., Accounting for non-stationarity in epidemiology by embedding time-varying parameters in stochastic models. *PLoS Computational Biology*, 2018 — **présenté dans l'annexe C**

Chapter 2

Comparaison de modèles de dengue

2.1 Introduction

Dengue is challenging to model due to the multiplicity of factors involved, that include the interaction between four virus serotypes, the ecology of the mosquito vector and the high proportion of inapparent infections. In contexts of limited data where model complexity needs to be restricted, all these elements cannot be included together. Moreover, observation and transmission of the pathogen are both influenced by uncertainties which are crucial to account for, but which induce a cost in computation and interpretation. Therefore, it is difficult to find the appropriate level of complexity and stochasticity needed to reproduce the dynamics of pathogen propagation.

Long considered as an urban disease, dengue was more rarely studied in rural settings, even though there is increasing evidence of its high burden in these areas. For instance, annual outbreaks have been recorded for more than ten years in the Kampong Cham province in Cambodia, and this highly populated and connected region might have a role in national disease propagation (Teurlai et al., 2012). For this study, we take advantage of a small scale cluster survey that was conducted in the Kampong Cham region and combine it with national surveillance data for this area to model the rural dengue dynamics. This double dataset, which includes asymptomatic infections, virus serotypes but also long term time series, enables the exploration and comparison of several modeling formulations.

In section 2.2, relying on Bayesian MCMC estimation in a state space framework, we estimate and compare six different models of dengue transmission according to several statistical and epidemiological criteria. By simulating the models with the estimated parameters, we also study their behaviour regarding hidden states of the systems, in order to assess their epidemiological coherence. This step is important in the evaluation of the models' quality, especially when inference relies only on incidence data and when the models are to be used for prediction or testing of control scenarios. The model including two interacting strains reproduces best the dengue dynamics. On the contrary, the explicit modeling of the mosquito vector does not enrich the dynamics and is subject to more uncertainties in the absence of entomological data.

In section 2.3, we extend the previous analysis by taking into account uncertainties in the epidemiological transmission process. We introduce sequentially stochasticity due to observation, demographic variability and environmental effects, relying on a framework formulated by Bretó et al. (2009). In a Bayesian setting, using MCMC and particle MCMC methods (Andrieu

et al., 2010), we estimate five epidemiological models in one deterministic and two stochastic implementations for which we contrast and compare the estimated epidemic trajectories and parameter values. In our context, deterministic models provide a good approximation of the mean behaviour, and important epidemiological parameters such as the reproduction number are estimated in the same range of values as in more computationally intensive stochastic implementations. However, the uncertainty associated with parameter estimates is underestimated, and a large amount of observation noise is required to reproduce the variability in the data. Stochastic models give a more realistic view of the model uncertainties but are sometimes so flexible that they are not informative on the disease trajectories, even when increasing the details in compartmental models. These aspects highlight the complexity of epidemiological models estimation, a difficult but crucial step towards understanding disease dynamics and testing the impacts of potential control scenarios.

2.2 Dengue modeling in rural Cambodia: statistical performance versus epidemiological relevance [1]

Abstract

Dengue dynamics are shaped by the complex interplay between several factors, including vector seasonality, interaction between four virus serotypes, and inapparent infections. However, paucity or quality of data do not allow for all of these to be taken into account in mathematical models. In order to explore separately the importance of these factors in models, we combined surveillance data with a local-scale cluster study in the rural province of Kampong Cham (Cambodia), in which serotypes and asymptomatic infections were documented. We formulate several mechanistic models, each one relying on a different set of hypotheses, such as explicit vector dynamics, transmission via asymptomatic infections and coexistence of several virus serotypes. Models are confronted with the observed time series using Bayesian inference, through Markov chain Monte Carlo. Model selection is then performed using statistical information criteria, and the coherence of epidemiological characteristics (reproduction numbers, incidence proportion, dynamics of the susceptible classes) is assessed in each model. Our analyses on transmission dynamics in a rural endemic setting highlight that two-strain models with interacting effects better reproduce the long term data, but they are difficult to parameterize when relying on incidence cases only. On the other hand, considering the available data, incorporating vector and asymptomatic components seems of limited added-value when seasonality and underreporting are already accounted for.

Introduction

Dengue is a vector-borne viral disease transmitted by *Aedes* spp. caused by any of four dengue virus (DENV) serotypes. Infection can result in a flu-like illness, and sometimes potentially lethal complications called Dengue Hemorrhagic Fever (DHF) and Dengue Shock Syndrome (DSS), although a significant proportion are subclinical or asymptomatic, causing insufficient discomfort for clinical presentation (Grange et al., 2014). Dengue is ubiquitous in the tropics and the subtropics, particularly in Southeast Asia, the Pacific and the Americas (Guzman et al., 2010). The World Health Organization (WHO) considers that dengue is a major public health issue worldwide, with four billion people in 128 countries exposed to the dengue virus (Messina et al., 2014; WHO, 2017), an estimated 390 million infections every year and about 50-100 million symptomatic cases worldwide and a high disease burden (Bhatt et al., 2013; Shepard et al., 2016). Nowadays, there are more cases of dengue worldwide than any other arboviral disease (Halstead, 2007; Messina et al., 2015; Sharp et al., 2017).

The value of mathematical models and associated statistical tools for investigating public health policy questions has long been recognized and has provided insights into their transmission and control for more than one hundred years (Heesterbeek et al., 2015; Reiner et al., 2013). It is important, however, to adapt them as much as possible to a specific setting, in order to derive appropriate public health recommendations and accurately generate the key parameters using estimation tools, so that they can produce realistic conclusions, in accordance with the observed data.

[1] Champagne C., Paul R., Ly S., Duong V., Leang R., Cazelles B., Dengue modeling in rural Cambodia: statistical performance versus epidemiological relevance, *Epidemics*, 2018

Dengue dynamics are shaped by the complex interplay between many factors associated with the mosquito vector and human hosts and their interactions with the virus. Hitherto, the exploration of dengue dynamics has focused on the urban setting, where the incidence of dengue is highest (Clapham et al., 2015; Nisalak et al., 2016; Reich et al., 2013; Salje et al., 2012). Few studies have been carried out in rural settings (Aldstadt et al., 2012; Mammen Jr et al., 2008; Strickman et al., 2000), despite growing evidence that rural dengue is an increasing problem. Guha-Sapir and Schimmer (2005) observed shifts in modal age, rural spread, and social determinants of dengue susceptibility, with major implications for health services. Muhammad Azami et al. (2011) observed similar dengue seroprevalence rates between urban and rural samples, showing that dengue is not confined to urban areas in Malaysia. Chareonsook et al. (1999) showed that DHF in Thailand, which was originally thought to be an urban disease, has spread to most areas of Thailand, and is now more common in rural than urban areas and studies suggest that rural dengue incidence can surpass urban and semi-urban communities within the same region (Reller et al., 2012; Vong et al., 2010). In addition, several studies have stressed that rural settings play an important role in the timing of dengue epidemics in Southeast Asia, with the seasonal dengue waves typically arriving later in major urban centers (Cazelles and Cazelles, 2014; Cuong et al., 2013; Teurlai et al., 2012).

In this study, we combine two datasets from rural Cambodia that provide information on different key factors. We contrast and compare several mechanistic models, incorporating differing levels of complexity with respect to vector dynamics, coexistence of several virus strains, and transmission via asymptomatic infections. Models are adapted to the observed time series using Bayesian inference, through Markov Chain Monte Carlo (MCMC) and compared in light of the data, using statistical indicators to identify the best model (Camacho et al., 2011; King et al., 2008; Pandey et al., 2013; Reich et al., 2013). In addition, we also analyze the epidemiological coherence of the estimated models in simulations. Critically, we do not merely focus on the observed infected individuals but also on other compartments, such as the susceptible class of individuals. By comparing these models, we try to find a realistic but parsimonious way of modeling dengue epidemics in rural Cambodia. The best model may then be used in the study of intervention scenarios or in comparative analyses with other settings. For instance, it could be readily expanded to understand the potential impact of different vaccination strategies in rural settings.

Methods

Data

Study area

Kampong Cham province is a densely populated rural province 120km northeast from the capital Phnom Penh. Dengue is endemic and strongly seasonal, with outbreaks occurring every year from June to September, during the rainy season. The four virus serotypes co-circulate, even though one usually dominates the three others for about 3 to 5 years. We used two different datasets reporting dengue cases in the province: the results of a punctual study conducted in a 30km radius around the city of Kampong Cham (DENFREE data), and the national surveillance data (NDSS data) in the four districts comprising the DENFREE study area (Kampong Cham, Kampong Siem, Prey Chhor and Tboung Khmum, with the administrative divisions of 2012-2013).

DENFREE Data

The DENFREE study took place in the Kampong Cham region during the 2012 and 2013 outbreaks. Patients with acute dengue-like illness were enrolled in three hospitals in the Kampong Cham province. Positive DENV cases were considered as index cases, and an outbreak investigation was initiated in their neighbourhood, in order to detect additional symptomatic cases but also asymptomatic or mildly symptomatic cases. For both index and outbreak investigation cases, DENV infection was confirmed by qRT-PCR. The study protocol is extensively detailed in Duong et al. (2015).

We used the series of the total number of cases per week (index cases and outbreak investigation cases) between 18th Jun - 5th Nov 2012 and 3rd Jun - 30th Sep 2013. We also restricted the study to children under 15 years old for two major reasons : most of the reported dengue cases were in this age class (89.6% and 90.8% in 2012 and 2013 respectively, cf. Appendix A.2) , and it allowed a comparison with other dengue reporting systems in Cambodia, which are mainly done at paedriatric hospitals. Information on the serotype responsible for infection, and symptomatic/asymptomatic status of the patients were available.

Children under 15 years old	**Total number of observed cases**		**Theoretical quantity**	**Observation rate**	**Model**
DENFREE data	**year 2012**	**year 2013**			
Confirmed symptomatic cases	236	574	C_I	r_D	All
Denv-1	226	451	C_{I1}	r_D	SEIR2, SEIR2psi
Denv-2, Denv-3, Denv-4	5	122	C_{I2}	r_D	SEIR2, SEIR2psi
Unknown serotype	5	0			
Confirmed asymptomatic cases	5	28	C_A	r_A	SEIAR
Denv-1	5	21			
Denv-2, Denv-3, Denv-4	0	7			
NDSS data	2002-2015				
Surveillance cases	10,780 (10,096 in 2002-2013)		C_I	r_N	All except SEIAR
			C_H	r_H	SEIAR

Table 2.1 – **Number of cases under 15 years old in the DENFREE study (index cases and community cases) and in the surveillance system NDSS, and associated theoretical quantities in models.** One case in 2013 was coinfected with DENV-1 and DENV-2 and was not included in models with two strains (serotypes).

NDSS Data

Because the DENFREE data covers only a relatively short period of time, surveillance data were added to improve the estimations. Surveillance of dengue is conducted at the national level in Cambodia, through the National Dengue Surveillance system (NDSS)(Huy et al., 2010; Teurlai et al., 2012), involving the paediatric departments of several hospitals throughout the country. Since surveillance is hospital based, mostly severe cases are observed. Diagnosis is done clinically and only a small fraction of the cases are confirmed serologically. Because of the co-circulation of other flaviviruses (Chikungunya, Japanese Encephalitis) and the relative non-specificity of symptoms, clinical mis-diagnosis may be frequent. Since surveillance is carried out in paediatric departments, only cases among children under 16 years old are reported.

We selected all the cases under 15 years old in the four districts involved in the DENFREE

study between January 2002 and December 2015 and aggregated them per week (cases under 15 years old represent 96.2% of reported cases over the period, cf. Appendix A.2). In this area, on average, 770 cases under 15 years old are reported per year (maximum 1985 cases in 2007, minimum 209 cases in 2014). We used data from 2002 to 2013 for estimations, and data for 2014 and 2015 as the test set.

	DENFREE	**NDSS**
Time window	Observations in 2012-2013	Observations in 2002-2015
	No observation during inter-epidemic period, dataset starts during the epidemic peak	Observations all year round
Area	30km radius around Kampong Cham city	4 districts comprising the DENFREE study
	Clustered collecting process	Stable reporting process over time
	An observation rate can be calculated	Unknown observation rate
Diagnosis	Laboratory confirmation of all cases (qRT-PCR)	Clinical diagnosis
	Observation of symptomatic infections (index cases and outbreak investigation) and mildly symptomatic or asymptomatic infections (outbreak investigation)	Symptoms ranging from dengue fever to DHF and DSS (mostly hospitalized)
	Known serotype for both index and outbreak investigation cases	Unknown serotype
	Primary and secondary infections are not distinguished	Primary and secondary infections are not distinguished

Table 2.2 – **Comparison of both datasets**

Population

We take as the reference population (N=161391) the number of children below 15 years old in four districts of the Kampong Cham province (Kampong Cham, Kampong Siem, Prey Chhor and Tboung Khmum, with the administrative divisions of 2012-2013) according to 2008 National Census (National Institute of Statistics, 2009). Since the DENFREE study was conducted in a subpart of this area, we calculated the total population for the DENFREE study (n=65208) as the sum of the population of children under 15 years old in all the villages investigated in either 2012 or 2013 (National Institute of Statistics, 2009).

Models

All model parameters are defined in the figures captions and in Table 2.3.

One-strain models

We take a Susceptible-Exposed-Infected-Recovered (SEIR) model as the simplest model (cf. Figure 2.1). C_I represents the count of new cases and it is aggregated weekly to be compared with data on both NDSS and DENFREE symptomatic cases. In this model, the basic reproduction number, i.e. the number of secondary infections resulting from the introduction of a single infected in an entirely susceptible population, is $R_0^{SEIR}(t) = \frac{\beta(t)\sigma}{(\gamma+\mu_H)(\sigma+\mu_H)}$.

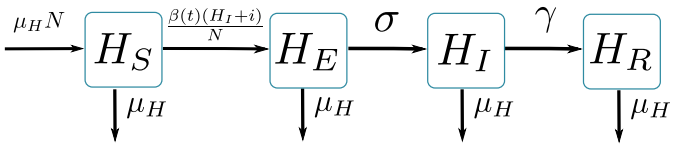


Figure 2.1 – **Graphical representation of SEIR model.** H_S susceptible individuals; H_E infected (not yet infectious) individuals; H_I infectious individuals; H_R recovered individuals; $\beta(t)$ is the transmission parameter; σ is the rate at which H_E-individuals move to the infectious class H_I; infectious individuals (H_I) then recover at rate γ; individuals leave the children population at rate μ_H. $H_S + H_E + H_I + H_R = N$.

$$
\begin{aligned}
\frac{dH_S}{dt} &= \mu_H N - \beta(t)\frac{(H_I + i)H_S}{N} - \mu_H H_S \\
\frac{dH_E}{dt} &= \beta(t)\frac{(H_I + i)H_S}{N} - \sigma H_E - \mu_H H_E \\
\frac{dH_I}{dt} &= \sigma H_E - \gamma H_I - \mu_H H_I \\
\frac{dH_R}{dt} &= \gamma H_I - \mu_H H_R \\
\frac{dC_I}{dt} &= \sigma H_E
\end{aligned}
\tag{2.1}
$$

This model is compared with two other models that include the mosquito vector transmission components. In the first one, which is a Ross-Mcdonald type model derived from Pandey et al. (2013), the vector is modelled explicitly with three compartments (Susceptible-Exposed-Infected) (cf. Figure 2.2). In the second one, derived from Laneri et al. (2010), the vector is modelled implicitly as an external force of infection including two stages, latent (κ) and current (λ) (cf. Figure 2.3). We derived R_0 for each model as $R_0^{Pandey}(t) = \frac{\beta_H \beta_V(t)\sigma\tau}{(\gamma+\mu_H)(\sigma+\mu_H)\mu_V(\mu_V+\tau)}$ and the estimation $R_0^{Laneri}(t) = \frac{\beta(t)\sigma}{(\gamma+\mu_H)(\sigma+\mu_H)}$ (Champagne et al., 2016). In order to compare these models with the non-vector models, we considered the same definition (i.e. the number of secondary human infections resulting from the introduction of a single infected human in a entirely susceptible population), and not the reproduction ratio per generation provided through the use of the next generation matrix.

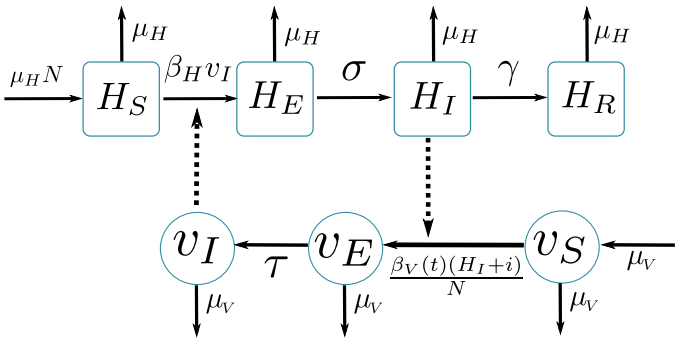


Figure 2.2 – **Graphical representation of Pandey model (Pandey et al., 2013).** Squared boxes and circles correspond respectively to human and vector compartments. Plain arrows represent transitions from one state to the next. Dashed arrows indicate interactions between humans and vectors. H_S susceptible individuals; H_E infected (not yet infectious) individuals; H_I infectious individuals; H_R recovered individuals; β_H is the transmission parameter from vector to human; σ is the rate at which H_E-individuals move to the infectious class H_I; infectious individuals (H_I) then recover at rate γ; individuals leave the children population at rate μ_H; $H_S + H_E + H_I + H_R = N$; v_S proportion of susceptible vectors; v_E proportion of infected (not yet infectious) vectors; v_I proportion of infectious vectors; $\beta_V(t)$ is the transmission parameter from human to vector; τ is the rate at which v_E-vectors move to the infectious class v_I; vectors die at rate μ_V.

The equations describing the Pandey model are:

$$\begin{aligned}
\frac{dH_S}{dt} &= \mu_H N - \beta_H v_I H_S - \mu_H H_S \\
\frac{dH_E}{dt} &= \beta_H v_I H_S - \sigma H_E - \mu_H H_E \\
\frac{dH_I}{dt} &= \sigma H_E - \gamma H_I - \mu_H H_I \\
\frac{dH_R}{dt} &= \gamma H_I - \mu_H H_R \\
\frac{dv_S}{dt} &= \mu_V - \beta_V(t)\frac{(H_I+i)}{N} v_S - \mu_V v_S \\
\frac{dv_E}{dt} &= \beta_V(t)\frac{(H_I+i)}{N} v_S - \tau v_E - \mu_V v_E \\
\frac{dv_I}{dt} &= \tau v_E - \mu_V v_I \\
\frac{dC_I}{dt} &= \sigma H_E
\end{aligned} \tag{2.2}$$

where v_s is the proportion of susceptible mosquitoes, v_E the proportion of exposed mosquitoes, and v_I the proportion of infected mosquitoes.

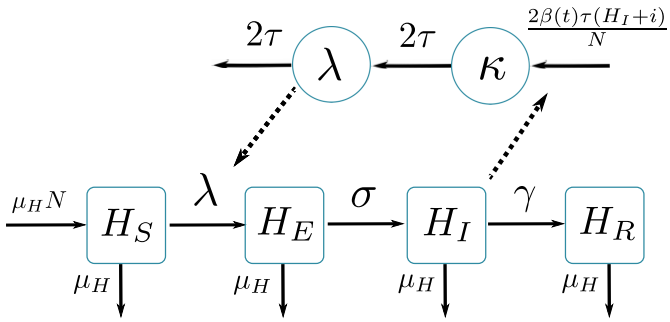


Figure 2.3 – **Graphical representation of Laneri model (Laneri et al., 2010).** Squared boxes and circles correspond respectively to human and vector compartments. Plain arrows represent transitions from one state to the next. Dashed arrows indicate interactions between humans and vectors. H_S susceptible individuals; H_E infected (not yet infectious) individuals; H_I infectious individuals; H_R recovered individuals; σ is the rate at which H_E-individuals move to the infectious class H_I; infectious individuals (H_I) then recover at rate γ; individuals leave the children population at rate μ_H; $H_S + H_E + H_I + H_R = N$; implicit vector-borne transmission is modelled with the compartments κ and λ; λ current force of infection; κ latent force of infection reflecting the exposed state for mosquitoes during the extrinsic incubation period; $\beta(t)$ is the transmission parameter; τ is the transition rate associated with the extrinsic incubation period.

The equations describing the Laneri model are:

$$\begin{aligned}
\frac{dH_S}{dt} &= \mu_H N - \lambda H_S - \mu_H H_S \\
\frac{dH_E}{dt} &= \lambda H_S - \sigma H_E - \mu_H H_E \\
\frac{dH_I}{dt} &= \sigma H_E - \gamma H_I - \mu_H H_I \\
\frac{dH_R}{dt} &= \gamma H_I - \mu_H H_R \\
\frac{d\kappa}{dt} &= \beta(t)\frac{2(H_I + i)\tau}{N} - 2\tau\kappa \\
\frac{d\lambda}{dt} &= 2\tau\kappa - 2\tau\lambda \\
\frac{dC_I}{dt} &= \sigma H_E
\end{aligned} \tag{2.3}$$

Model with explicit asymptomatic individuals (SEIAR)

We also consider a model in which asymptomatic infections are explicitly taken into account in the transmission process (cf. Figure 2.4). In this model, we assume that, after the incubation period, there are three possible manifestations of the disease: asymptomatic (H_A), mildly symptomatic not requiring hospitalization (H_I) and hospitalized cases (H_H). Asymptomatic cases are defined in the dengue study as asymptomatic or pauci-symptomatic (presence of other symptoms not being sufficient to classify as symptomatic). Hospital cases are defined as NDSS cases (reported by the surveillance system in hospitals). We assume that symptomatic DENFREE cases are either (H_I) or (H_H). We also assume that asymptomatic cases transmit the disease as much as symptomatic cases, as recently shown (Duong et al., 2015), and therefore, $R_0^{SEIAR}(t) = \frac{\beta(t)\sigma}{(\gamma+\mu_H)(\sigma+\mu_H)}$. C_H, C_I, C_A represent respectively the count of new hospitalized, symptomatic and asymptomatic cases,

and each is aggregated weekly to be compared respectively with NDSS data, DENFREE data on symptomatic cases, and DENFREE data on asymptomatic cases.

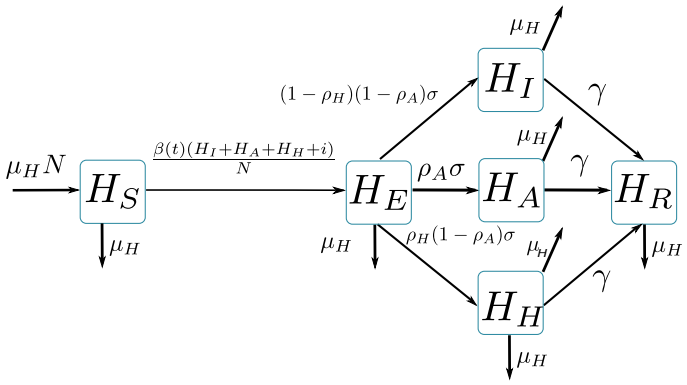


Figure 2.4 – **Graphical representation of SEIAR model.** H_S susceptible individuals; H_E infected (not yet infectious) individuals; H_A asymptomatic infectious individuals; H_I mildly symptomatic infectious individuals; H_H hospitalized infectious individuals; H_R recovered individuals; $\beta(t)$ is the transmission parameter; σ is the rate at which H_E-individuals move to the infectious classes H_I, H_A and H_H; a proportion ρ_A of H_E-individuals do not show symptoms during the infectious period; a proportion ρ_H of symptomatic individuals go to hospital; infectious individuals (H_I,H_A,H_H) then recover at rate γ; individuals leave the children population at rate μ_H. $H_S + H_E + H_A + H_I + H_H + H_R = N$.

$$
\begin{aligned}
\frac{dH_S}{dt} &= \mu_H N - \beta(t)\frac{(H_I + H_A + H_H + i)H_S}{N} - \mu_H H_S \\
\frac{dH_E}{dt} &= \beta(t)\frac{(H_I + H_A + H_H + i)H_S}{N} - \sigma H_E - \mu_H H_E \\
\frac{dH_A}{dt} &= \rho_A \sigma H_E - \gamma H_A - \mu_H H_A \\
\frac{dH_H}{dt} &= \rho_H(1-\rho_A)\sigma H_E - \gamma H_H - \mu_H H_H \\
\frac{dH_I}{dt} &= (1-\rho_H)(1-\rho_A)\sigma H_E - \gamma H_I - \mu_H H_I \\
\frac{dH_R}{dt} &= \gamma(H_I + H_A + H_H) - \mu_H H_R \\
\frac{dC_I}{dt} &= (1-\rho_A)\sigma H_E \\
\frac{dC_A}{dt} &= \rho_A \sigma H_E \\
\frac{dC_H}{dt} &= \rho_H(1-\rho_A)\sigma H_E
\end{aligned}
\tag{2.4}
$$

Model with two virus serotypes

In the 2012 and 2013 epidemics, DENV-1 was highly dominant : the three other serotypes represented about 2% of the cases reported in the DENFREE study in 2012 and about 21% in 2013 (cf. Table 2.1). Therefore, a two-strain model is also studied, in which we separate DENV-1 cases from DENV-2, DENV-3 and DENV-4 combined (cf. Figure 2.5). For simplicity and parsimony in the number of parameters, the two strains share the same parameter values. We first assume

both strains to be independent ($\psi = 1$ in equation 2.5, called SEIR2 model). In this context, the reproduction numbers for each strain are equal, $R_0^{SEIR2_1}(t) = R_0^{SEIR2_2}(t) = \frac{\beta(t)\sigma}{(\gamma+\mu_H)(\sigma+\mu_H)}$.

$$
\begin{aligned}
\frac{dH_S}{dt} &= \mu_H N - \beta(t)\frac{(H_{I1} + \psi H_{I21} + i)H_S}{N} - \beta(t)\frac{(H_{I2} + \psi H_{I12} + i)H_S}{N} - \mu_H H_S \\
\frac{dH_{E1}}{dt} &= \beta(t)\frac{(H_{I1} + \psi H_{I21} + i)H_S}{N} - \sigma H_{E1} - \mu_H H_{E1} \\
\frac{dH_{I1}}{dt} &= \sigma H_{E1} - \gamma H_{I1} - \mu_H H_{I1} \\
\frac{dH_{S1}}{dt} &= \gamma H_{I1} - \beta(t)\frac{(H_{I2} + \psi H_{I12} + i)H_{S1}}{N} - \mu_H H_{S1} \\
\frac{dH_{E12}}{dt} &= \beta(t)\frac{(H_{I2} + \psi H_{I12} + i)H_{S1}}{N} - \sigma H_{E12} - \mu_H H_{E12} \\
\frac{dH_{I12}}{dt} &= \sigma H_{E12} - \gamma H_{I12} - \mu_H H_{I12} \\
\frac{dH_{E2}}{dt} &= \beta(t)\frac{(H_{I2} + \psi H_{I12} + i)H_S}{N} - \sigma H_{E2} - \mu_H H_{E2} \\
\frac{dH_{I2}}{dt} &= \sigma H_{E2} - \gamma H_{I2} - \mu_H H_{I2} \\
\frac{dH_{S2}}{dt} &= \gamma H_{I2} - \beta(t)\frac{(H_{I1} + \psi H_{I21} + i)H_{S2}}{N} - \mu_H H_{S2} \\
\frac{dH_{E21}}{dt} &= \beta(t)\frac{(H_{I1} + \psi H_{I21} + i)H_{S2}}{N} - \sigma H_{E21} - \mu_H H_{E21} \\
\frac{dH_{I21}}{dt} &= \sigma H_{E21} - \gamma H_{I21} - \mu_H H_{I21} \\
\frac{dH_R}{dt} &= \gamma(H_{I12} + H_{I21}) - \mu_H H_R \\
\frac{dC_I}{dt} &= \sigma(H_{E1} + H_{E2} + H_{E21} + H_{E12}) \\
\frac{dC_{I1}}{dt} &= \sigma(H_{E1} + H_{E21}) \\
\frac{dC_{I2}}{dt} &= \sigma(H_{E2} + H_{E12})
\end{aligned}
\tag{2.5}
$$

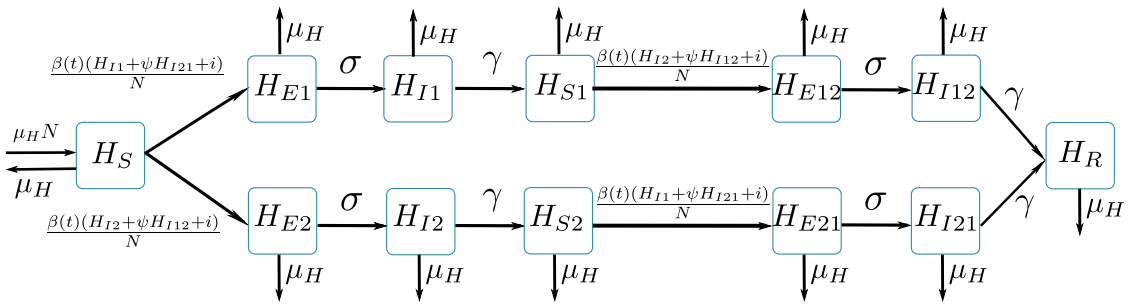


Figure 2.5 – **Graphical representation of SEIR2 models.** H_S individuals susceptible to both strains; H_{E1} (resp. H_{E2}) individuals infected (not yet infectious) to strain 1 (resp. strain 2); H_{I1} (resp. H_{I2}) individuals infectious to strain 1 (resp. strain 2); H_{S1} (resp. H_{S2}) individuals immune to strain 1 only (resp. strain 2); H_{E12} (resp. H_{E21}) individuals (not yet infectious) with a secondary infection to strain 2 (resp. strain 1); H_{I12} (resp. H_{I21}) infectious individuals with a secondary infection to strain 2 (resp. strain 1); H_R individuals immune to both strains; $\beta(t)$ is the transmission parameter; σ is the rate at which exposed individuals move to the infectious class; infectious individuals then recover at rate γ; ψ is the change in infectivity for secondary infected individuals in SEIR2psi model (in SEIR2 model, $\psi = 1$); individuals leave the children population at rate μ_H. $H_S + H_{E1} + H_{E2} + H_{I1} + H_{I2} + H_{S1} + H_{S2} + H_{E12} + H_{E21} + H_{I12} + H_{I21} + H_R = N$.

We also considered another version of the model including interaction between strains (ten Bosch et al., 2016), in order to reflect the fact that secondary infection with a heterologous serotype leads more often than primary infection to severe manifestations of the disease (Halstead, 2007). In our model (called SEIR2psi model), primary and secondary infections differ in infectiousness, through a parameter ψ (Ferguson et al., 1999). This parameter is estimated between 0.5 and 3: values superior to 1 correspond to transmission cross-enhancement (because of higher virus titers during secondary infections (Ferguson et al., 1999)) and values inferior to 1 suggest a lower infectivity for secondary infected individuals (for example because they are hospitalized and less in contact with the population (Aguiar et al., 2011)). As in Ferguson et al. (1999), we define $R_0^{SEIR2_1}(t) = R_0^{SEIR2_2}(t) = \frac{\beta(t)\sigma}{(\gamma+\mu_H)(\sigma+\mu_H)}$ the basic reproduction number for each strain.

C_I represents the count of new cases for both serotypes and it is aggregated weekly to be compared with NDSS data. C_{I1} and C_{I2} represent the count of new cases for strain 1 and strain 2, and are aggregated weekly to be compared with symptomatic cases from the DENFREE study for DENV-1 and DENV-2/ DENV-3/ DENV-4 respectively.

Seasonality

All models include seasonality through the use of a time-varying transmission parameter $\beta(t) = \beta[1 + b.sin(2\pi(\frac{t}{365} + p)]$, according to a sinusoidal function whose phase p and amplitude b are estimated.
We also assume that a constant number of cases i are imported.

Prior distributions

The prior distributions are listed in Table 2.3.

Parameter		Prior distribution	Reference	Models
Infectiousness, incubation and mortality rates				
γ^{-1}	infectious period (days)	4.5	WHO (2017)	All
σ^{-1}	intrinsic incubation period (days)	5.9	Chan and Johansson (2012)	Pandey, Laneri
τ^{-1}	extrinsic incubation period (days)	10	Chan and Johansson (2012)	Pandey, Laneri
σ^{-1}	both incubation periods (days)	15.9	Chan and Johansson (2012)	SEIR, SEIAR, SEIR2, SEIR2psi
μ_H^{-1}	age duration (years)	15	assumed	All
μ_V^{-1}	mosquito lifespan (days)	15	Liu-Helmersson et al. (2014)	Pandey
ρ_A	Proportion of asymptomatic cases	Uniform[0, 1]	assumed	SEIAR
ρ_H	Proportion of hospitalized cases	Uniform[0, 1]	assumed	SEIAR
Transmission parameters				
R_0	average basic reproduction number	Uniform[0, 20]	assumed	All
β_V	transmission from human to mosquito	Uniform[0.1, 2]	Pandey et al. (2013)	Pandey
ψ	inh./enh. of infectiousness	Uniform[0.5, 3]	assumed	SEIR2psi
Initial conditions				
$H_I(0)$	Initial number of infected individuals	Uniform[0, 100]	assumed	SEIR, Pandey, Laneri
$H_E(0)$	Initial number of exposed individuals	$H_I(0)$	assumed	SEIR, Pandey, Laneri, ($3H_I(0)$ in SEIAR)
$H_S(0)$	Initial number of susceptible individuals	N*Normal(0.44, 0.05) in [0.2, 1]	Thai et al. (2005)	All
$v_I(0)$ or $\lambda(0)$	Initial number of infected mosquitoes	0	assumed	Pandey or Laneri
$v_E(0)$ or $\kappa(0)$	Initial number of exposed mosquitoes	0	assumed	Pandey or Laneri
$H_H(0)$, $H_A(0)$	Initial number of asymptomatic and hospitalized individuals	$H_I(0)$	assumed	SEIAR
$H_{S1}(0)$, $H_{S2}(0)$	Initial number of individuals susceptibles to 1 strain	N*Uniform[0.01, 0.5]	assumed	SEIR2, SEIR2psi
$H_{I1}(0)$, $H_{I2}(0)$	Initial number of infected individuals	Uniform[0, 100]	assumed	SEIR2, SEIR2psi
$H_{E1}(0)$, $H_{I21}(0)$, $H_{E21}(0)$	Initial number of infected and exposed individuals	$H_{I1}(0)$	assumed	SEIR2, SEIR2psi
$H_{E2}(0)$, $H_{I12}(0)$, $H_{E12}(0)$	Initial number of infected and exposed individuals	$H_{I2}(0)$	assumed	SEIR2, SEIR2psi
Observation process				
r_N	observation rate for NDSS data	Uniform[0, 1]	assumed	SEIR, Laneri, Pandey, SEIR2, SEIR2psi
r_H	observation rate for NDSS data	1	assumed	SEIAR
r_D	observation rate for DENFREE data (symptomatic)	fixed	cf. Table 2.4	All
r_A	observation rate for DENFREE data (asymptomatic)	fixed	cf. Table 2.4	SEIAR
ϕ	overdispersion	Uniform[0, 1]	assumed	All
Seasonality parameters				
b	amplitude of the sinusoidal forcing	Uniform[0, 1]	assumed	All
p	phase of the sinusoidal forcing	Uniform[−0.5, 0.5]	assumed	All
i	import parameter	Uniform[0, 10]	assumed	All
Total population				
N	total population	161391		All

Table 2.3 – **Prior distributions of parameters**. "Uniform[0,20]" indicates a uniform distribution in the range [0,20]. "Normal$(0.44, 0.05)$ in [0.2, 1]" indicates a a truncated normal distribution with mean 0.44 and standard deviation 0.05, restricted to the range [0.2,1].

Dirac priors based on the literature were used for the durations of infectiousness and incubation, as well as the mortality rates. In models without vectorial transmission, the incubation period is assumed to be the sum of the extrinsic (in mosquito) and intrinsic (in human) incubation periods, to reflect the generation time of the disease. For transmission parameters, we used wide weakly informative priors.

Initial conditions

The initial number of infected individuals is assumed to be equal to the number of exposed individuals and to be lower than 100, as the model starts in January, during the epidemic trough. Except for the initial proportion of susceptibles, all priors on initial conditions are uniform distributions.

The initial proportion of susceptibles is an influential parameter on the model outputs. It is highly correlated to the transmission parameter β (and therefore to the basic reproductive number), which makes it difficult to estimate them both. An informative gaussian prior was therefore used on $H_S(0)$. To date, no large scale seroprevalence study is available for Cambodia, and we relied on a study conducted among schoolchildren in rural Vietnam (Thai et al., 2005), which we considered as the closest setting to be compared with Kampong Cham. We extrapolated their results on schoolchildren (7 to 14 years old) to a 1-15 years old population as follows, where S_λ is the proportion of susceptibles among 1-15 year-old children (using their estimation $\lambda = 0.117$):

$$S_\lambda = \frac{1}{15}\sum_{a=1}^{15} exp(-\lambda a) = 0.44$$

S_λ is used as the mean of the gaussian prior, and the standard deviation is fixed at 0.05.

Estimation

Observation model

The propagation models are related to the observed data using a negative binomial observation model (Bretó et al., 2009). The observed number of cases during week k in each dataset, $C_{obs}^{(k)}$, is assumed to be drawn from a a negative binomial distribution with mean $rC^{(k)}$ and variance $rC^{(k)} + (rC^{(k)})^2\phi$, where $C^{(k)}$ is the total number of new cases simulated by the model during week k, r is the observation rate quantifying the amount of non reported cases and ϕ is an overdispersion parameter. The quantity C and observation rate r associated to each dataset are indicated in Table 2.1.
For each epidemiological model, estimations are performed simultaneously on NDSS and DENFREE data (DENFREE data being decomposed into two time series in SEIAR, SEIR2 and SEIR2psi models). The observations of all time series are assumed independent conditional on the underlying disease process, and the model likelihood is the product of the likelihoods calculated on each series.

The observation rate is estimated for NDSS data but fixed for DENFREE data, using informations from the sampling scheme (cf. Table 2.4), in order to account for the difference in coverage during outbreak investigation between 2012 and 2013. We assume that index cases are all reported and that the observation rate for community cases equals the ratio of people tested over the population of the area. We then extrapolate this observation rate to the total population of the four districts.
Different observation rates are used in the SEIAR model. As NDSS data are interpreted as hospitalized cases, the observation rate is fixed to 1 (which assumes that all hospitalized cases go through surveillance and neglects the presence of private hospitals or non reports from hospitals), but the proportion of hospitalized cases is estimated. For DENFREE cases, the observation rate for symptomatic and asymptomatic individuals is fixed as indicated in Table 2.4.

Children < 15 years old	**2012**		**2013**	
	Symptomatic	**Asymptomatic**	**Symptomatic**	**Asymptomatic**
Population in investigated villages (n)	65,208		65,208	
Population in the 4 districts (N)	161,391		161,391	
Index cases (a)	151	0	376	0
Observation rate for index cases	1	.	1	.
Children tested in communities (b)	1722		4119	
DENV positive in communities (c)	85	5	198	28
Observation rate for community cases (d=b/n)	0.0264		0.0632	
Extrapolated number of community cases (e=c/d)	3219	189	3135	443
Extrapolated total number of cases (f=a+e)	3370	189	3511	443
Observed total number of cases (g=a+c)	236	5	574	28
Observation rate (r_D=(g/f)*(n/N) and r_A=b/N)	**0.0283**	**0.0107**	**0.0661**	**0.0255**

Table 2.4 – **Calculation of the observation rate for DENFREE data.**

The overdispersion parameter ϕ is estimated along with the other parameters.

Estimation with Markov Chain Monte Carlo

Models are considered in deterministic framework and estimations are made using random walk Metropolis Hastings. SSM software (Dureau et al., 2013a) is used for simulations and calculations. The covariance matrix of the proposal distribution was initialized using adaptative MCMC as in Dureau et al. (2013a).
Due to unknown initial conditions, the posterior distribution is multimodal, especially in two strain models. This implies that the adaptive MCMC remains trapped in local maxima and global exploration of the parameter space was performed with latin hypercube sampling. We ran a simplex algorithm on 10,000 parameter sets sampled with latin hypercube sampling (with *lhs* R package (Carnell, 2016)) and chose the one with the highest posterior value as initialization of the adaptive MCMC algorithm. The results associated to this best fitting equilibrium are displayed in the main text, but other examples of good fitting equilibria for two-strain models are indicated in Appendix A.5.

Model comparison

Statistical indicators

In order to identify the best model, the Deviance Information Criterion (DIC) (Spiegelhalter et al., 2002) is used. DIC is an indicator that combines a measure of model fit and a penalty on model complexity, commonly used with MCMC estimations. The best model is the one with the smallest DIC. As it does not enable comparison of models with differing number of observations, we also calculate the RMSE (root mean square error) between observations $C_{obs}^{(k)}$ and simulations of the model with parameters sampled in the posterior distribution $C^{(k,i)}$, for weeks $k = k_0, ..., K$ and simulation i :

$$\text{RMSE}_i = \sqrt{\frac{1}{K-k_0}\sum_{k=k_0}^{K}(C^{(k,i)} - C_{obs}^{(k)})^2}$$

We then compute the mean and the 2.5% and 97.5% quantiles of this indicator among the simulated trajectories. We first calculated them on the data used for estimation, separating NDSS data and DENFREE data. For DENFREE data we considered the total number of symptomatic cases to enable comparison of all models: therefore, for this metric, asymptomatic DENFREE cases were not taken into account in SEIAR model, and cases of all serotypes were aggregated in two-strain models. Then we used these indicators to assess the predictive performance of the models, comparing projections of the model with NDSS observations for 2014 and 2015.

Epidemiological indicators

Models are also compared according to several indicators to describe their epidemiological behaviour. The basic reproduction number, the observation rate and the initial proportion of susceptibles are estimated using the MCMC chain. In the model with asymptomatic infections we report the estimated proportions of asymptomatic and hospitalized cases.
With parameters sampled in the posterior distribution, we can also re-simulate the model to study hidden states, such as the susceptible and infected classes. The effective reproduction number (R_e) is calculated as the seasonal basic reproduction number multiplied by the proportion of susceptibles at each time step, as indicated in Table 2.5. We then calculate the annual incidence proportion as the total number of infections over one year divided by the total population of susceptibles at the beginning of the year. In models with two strains, we separate the annual incidence of primary infections (as the total number of primary infections over one year divided by the total population of naive individuals at the beginning of the year) and secondary infections (as the total number of secondary infections over one year divided by the total population of susceptibles to one strain only at the beginning of the year).

	SEIR, Laneri, SEIAR	**Pandey**	**SEIR2**	**SEIR2psi**
$R_0(t)$	$\frac{\beta(t)\sigma}{(\gamma+\mu_H)(\sigma+\mu_H)}$	$\frac{\beta_H\beta_V(t)\sigma\tau}{(\gamma+\mu_H)(\sigma+\mu_H)\mu_V(\mu_V+\tau)}$	$\frac{\beta(t)\sigma}{(\gamma+\mu_H)(\sigma+\mu_H)}$	$\frac{\beta(t)\sigma}{(\gamma+\mu_H)(\sigma+\mu_H)}$
$R_e(t)$ (or R_e^i, $i=1,2$)	$R_0(t)\frac{H_S(t)}{N}$	$R_0(t)\frac{H_S(t)}{N}v_s(t)$	$R_0(t)\frac{H_S(t)+H_{Sj}(t)}{N}$	$R_0(t)\frac{H_S(t)+\psi H_{Sj}(t)}{N}$

Table 2.5 – **Reproduction numbers calculation in each model.** Calculations are based on the next generation matrix method (Diekmann et al., 2010).

Calculations are made using R version 3.2.2 (Team, 2015), and graphics using *ggplot2* (Wickham, 2009).

Results

The calculations are based on the MCMC chain associated with the highest posterior after LHS exploration. In multistrain models, the posterior distribution is multimodal, and examples of other local posterior modes are displayed in Appendix A.5.

Statistical comparison

Model	SEIR	Laneri	Pandey	SEIAR	SEIR2	SEIR2psi
nb parameters	8	8	9	9	11	12
nb observations	665	665	665	704	704	704
ESTIMATION SET						
DIC	4343	4360	4364	4472	4492	4446
RMSE NDSS	25 (22-28)	25 (22-29)	26 (23-29)	26 (23-29)	25 (22-29)	21 (19-24)
RMSE NDSS 2002	43 (28-59)	42 (27-58)	45 (29-63)	43 (27-61)	40 (24-61)	17 (12-25)
RMSE NDSS 2003-2013	23 (20-25)	23 (21-26)	23 (21-25)	23 (21-26)	24 (21-26)	22 (19-24)
RMSE DENFREE	24 (19-29)	24 (19-29)	24 (19-29)	24 (19-29)	23 (19-27)	25 (19-31)
TEST SET						
RMSE 2014-2015	15 (11-19)	15 (11-19)	16 (12-20)	15 (11-20)	17 (13-22)	13 (10-16)
RMSE 2014	16 (11-23)	16 (11-23)	17 (11-24)	17 (11-24)	17 (11-24)	14 (10-19)
RMSE 2015	13 (9-17)	13 (9-17)	13 (10-18)	13 (10-17)	17 (12-23)	12 (9-16)

Table 2.6 – **Information criteria in deterministic models.** DIC is the Deviance Information Criterion (Spiegelhalter et al., 2002). RMSE is the root mean square error (mean and 2.5% and 97.5% quantiles) between simulations and observations: it is calculated separately on the datasets used for estimations (NDSS data for 2002-2013 and DENFREE data for 2012-2013) and on the test set (NDSS data for 2014-2015). It is also computed on separated years in order to highlight well or badly estimated years: for example, for each simulation i, $\frac{1}{12}(\text{RMSE}_i^{\text{NDSS 2002}})^2 + \frac{11}{12}(\text{RMSE}_i^{\text{NDSS 2003-2013}})^2 = (\text{RMSE}_i^{\text{NDSS}})^2$. Calculations are based on the MCMC chain associated with the highest posterior.

For single strain models, the SEIR model proved the best with the DIC criterion (cf. Table 2.6). For two strain models, the SEIR2psi proved best. As regards simulation-based indicators on the 2002-2014 data, SEIR, Laneri, Pandey models have RMSE values in the same order of magnitude. Indeed, they produce a similar dynamic with respect to the 2002-2014 data (cf. Figures 2.6-2.7), with a period of approximately six years, and a large overestimation of the 2002 outbreak (the mean RMSE for 2002 is far higher (> 40) than the average for the other years (<25)). Due to the small number of observed asymptomatic cases, SEIAR model also produces a similar dynamic and the SEIR2 model, despite a larger number of parameters, does not improve over single-strain models. The model with two interacting strains (SEIR2psi) outperforms the other models in terms of RMSE, but the difference is mainly explained by the first year of simulation (2002). However, the RMSE indicator does not penalize for the complexity of this more flexible model, and the differences are not very pronounced given the variability in RMSE values. When visualizing the simulations compared to the data (cf. Figures 2.6-2.7), all models underestimate a large number of epidemic peaks and all models but the one with two interacting strains overpredict the first epidemic peak (as pointed out in Table 2.6). SEIR2psi is the model in which the large epidemic in 2007 is best reproduced and overall, SEIR2psi is the model that reproduces most accurately the observed data. Moreover, the estimated overdispersion is lower in this model (cf. Table 2.7), indicating that less observation noise is required to explain the data and that the model explains a larger part of the variability in the data.

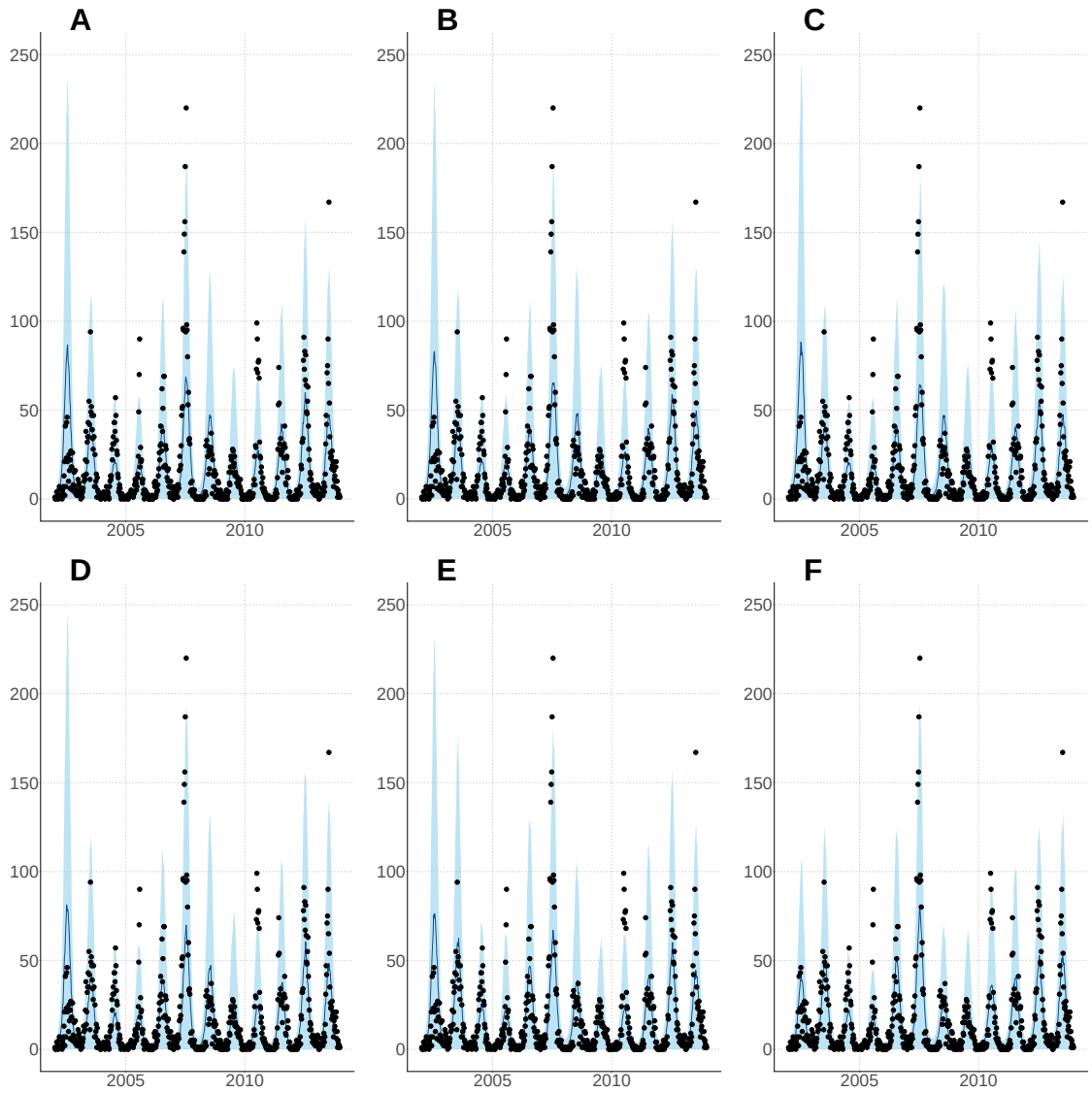


Figure 2.6 – **Number of observed cases per week and NDSS data, 2002-2013.** Simulations with negative binomial noise using parameters from the MCMC chain associated with the highest posterior, calculated using both NDSS and DENFREE datasets. Posterior median (solid line), 95% credible intervals (shaded blue area) and NDSS data points (black dots). A. SEIR model. B. Laneri model. C. Pandey model. D. SEIAR model. E. SEIR2 model. F. SEIR2psi model.

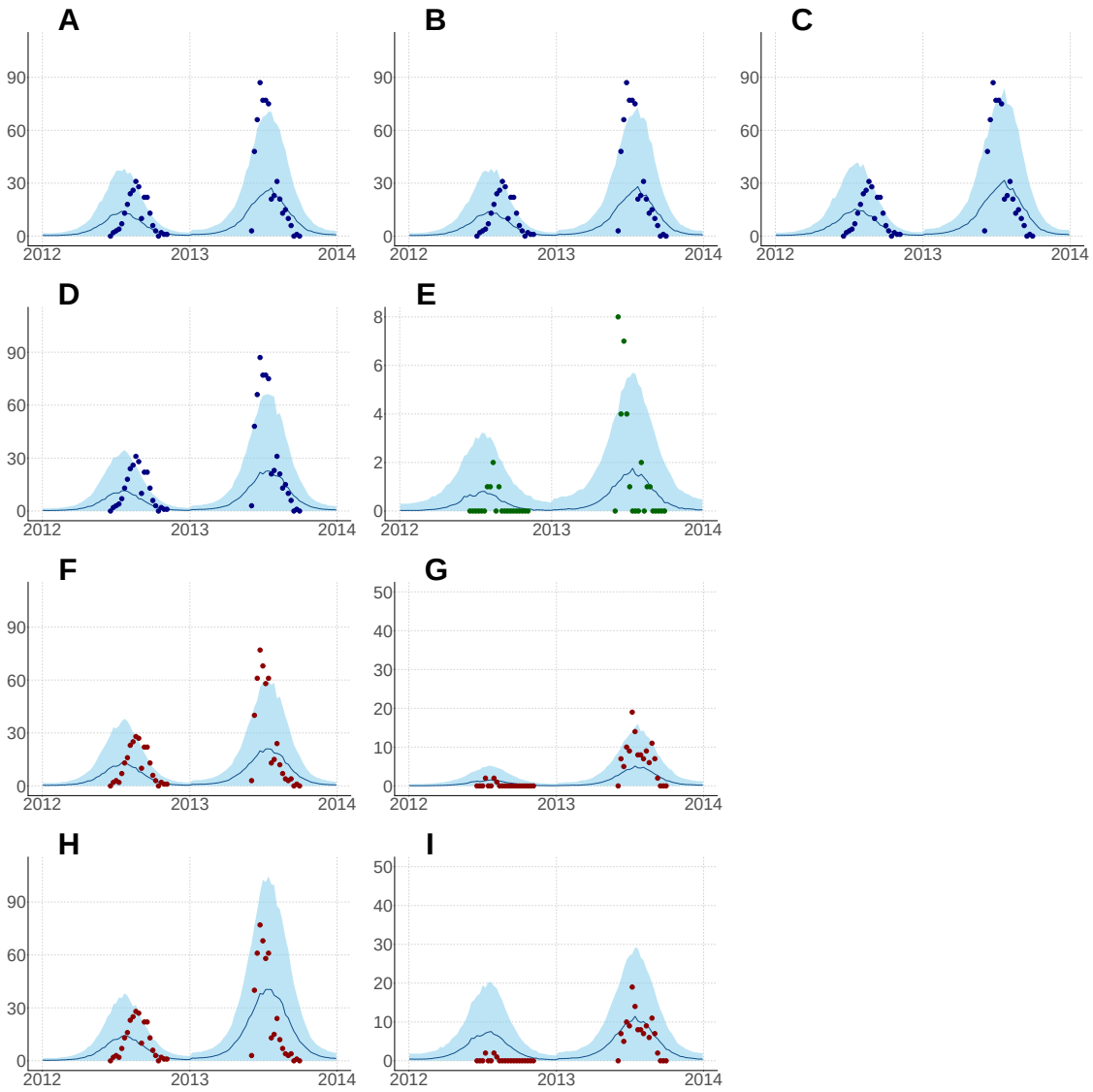


Figure 2.7 – **Number of observed cases per week and DENFREE data, 2012-2013.** Simulations with negative binomial noise using parameters from the MCMC chain associated with the highest posterior, calculated using both NDSS and DENFREE datasets. Posterior median (solid line), 95% credible intervals (shaded blue area) and DENFREE data points (blue dots: symptomatic cases, green dots: asymptomatic cases, red dots: serotype specific symptomatic cases). A. SEIR model. B. Laneri model. C. Pandey model. D. SEIAR model, symptomatic cases. E. SEIAR model, asymptomatic cases. F. SEIR2 model, cases due to strain 1 and DENV1 observations. G. SEIR2 model, cases due to strain 2 and DENV2/DENV3/DENV4 observations. H. SEIR2 model with interactions, cases due to strain 1 and DENV1 observations. I. SEIR2 model with interactions, cases due to strain 2 and DENV2/DENV3/DENV4 observations.

Considering the predictive capacity of the models, all models overestimate the 2014 small epidemic, and to a lesser extent the one in 2015 (cf. Figure 2.8). In the SEIR2psi model, this overestimation is less pronounced, and the associated RMSE is also smaller (cf. Table 2.6).

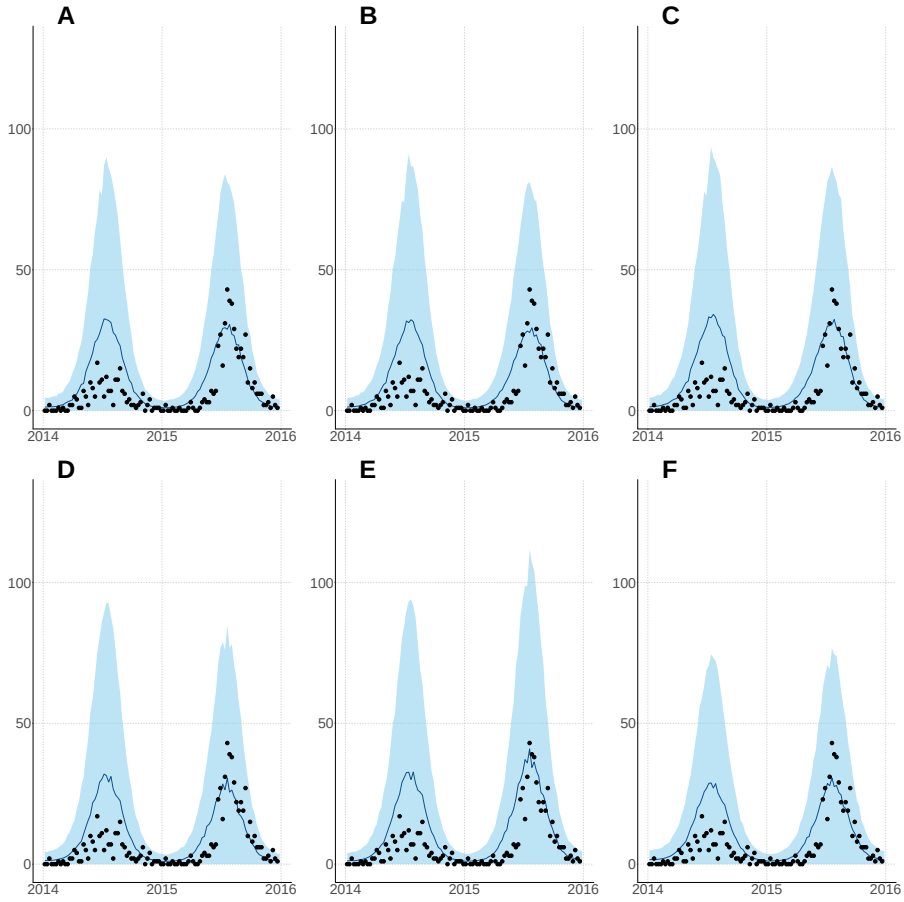


Figure 2.8 – **Projections of the number of observed cases per week and NDSS data, 2014-2015.** Simulations with negative binomial noise using parameters from the MCMC chain associated with the highest posterior, calculated using both NDSS and DENFREE datasets. Posterior median (solid line), 95% credible intervals (shaded blue area) and NDSS data points (black dots). A. SEIR model. B. Laneri model. C. Pandey model. D. SEIAR model. E. SEIR2 model. F. SEIR2psi model.

Epidemiological comparison

Model		SEIR	Laneri	Pandey	SEIAR	SEIR2	SEIR2 psi
mean R_0	median (95%CI)	2.32 (2.25-2.41)	2.38 (2.3-2.46)	3.32 (3.16-3.47)	2.3 (2.23-2.37)	1.45 (1.42-1.48)	2.47 (2.35-2.57)
max R_0	median (95%CI)	3.64 (3.51-3.77)	3.66 (3.53-3.8)	5.87 (5.57-6.17)	3.6 (3.48-3.72)	2.29 (2.23-2.35)	3.78 (3.59-3.94)
ψ	median (95%CI)						0.67 (0.62-0.73)
$H_S(0)/N$ (%)	median (95%CI)	46 (45-47)	46 (45-47)	37 (36-38)	47 (46-47)	49 (45-56)	35 (30-38)
$H_{S1}(0)/N$ (%)	median (95%CI)					22 (14-25)	14 (10-21)
$H_{S2}(0)/N$ (%)	median (95%CI)					27 (19-30)	4 (1-11)
Observation rate (%)	median (95%CI)	12 (11-13)	12 (11-13)	10 (9-11)		11 (10-12)	7 (6-7)
Hospitalized (%)	median (95%CI)				14 (13-15)		
Asymptomatic (%)	median (95%CI)				15 (13-19)		
Over-dispersion	median (95%CI)	0.73 (0.59-0.91)	0.75 (0.61-0.93)	0.7 (0.57-0.87)	0.9 (0.74-0.99)	0.73 (0.59-0.91)	0.56 (0.45-0.76)
Median annual incidence proportion							
primary infection (%)	median 2002-2015 (min-max)	8 (5-16)	8 (4-16)	12 (7-25)	8 (4-16)	6 (3-10)	14 (7-27)
secondary infection (%)	median 2002-2015 (min-max)					3 (1-5)	7 (3-14)
mean Re	median (95%CI)	1.03 (1.03-1.03)	1.06 (1.06-1.07)	1.14 (1.13-1.15)	1.03 (1.03-1.03)		
mean Re strain 1	median (95%CI)					1.03 (1.02-1.03)	1.03 (1.02-1.03)
mean Re strain 2	median (95%CI)					1.04 (1.03-1.04)	1.02 (1.02-1.03)
max Re	median (95%CI)	1.7 (1.66-1.74)	1.72 (1.69-1.77)	2.18 (2.12-2.27)	1.7 (1.67-1.74)		
max Re strain 1	median (95%CI)					1.73 (1.69-1.77)	1.74 (1.71-1.77)
max Re strain 2	median (95%CI)					1.72 (1.69-1.75)	1.69 (1.65-1.73)

Table 2.7 – **Epidemiological criteria among children under 15 years old.** Estimated parameters and indicators based on simulations over 2002-2015 (posterior median and 95% credible intervals). Calculations are based on the MCMC chain associated with the highest posterior.

The average R_0 is estimated to be between 2 and 3 in most of the models and the maximum value between 3 and 4 (cf. Table 2.7), except with the Pandey model, in which it is higher (mean value above 3 and maximum value above 6). The estimated values are very close in the SEIR, Laneri and SEIAR models. These values relate to transmission among children only and we expect therefore the R_0 in the whole population to be higher, since children represent approximately one third of the total population, but more than 90% of the total cases observed. Estimates for R_0 in South East Asia in the literature range from 1 to 5 (Imai et al., 2015, 2016), with sometimes values above 10 (Johansson et al., 2011), displaying large variations in time and space, and in particular, the estimates for R_0 in Cambodia based on age-stratified case-notification data are between 2 and 7 (Imai et al., 2016), so that our estimates are in the same order of magnitude. Nevertheless, our estimation of R_0 strongly depends on the estimation of the initial proportion of susceptibles ($H_S(0)$), which is unknown in the case of Cambodia and can bias the estimates. It is indeed very difficult to evaluate R_0 via incidence data (Imai et al., 2016), especially for endemic diseases without information on seroprevalence, and it may be more informative to study the effective reproductive number (R_e), which accounts for the proportion of immune individuals in the population. The estimated R_e has a mean value around 1 and a maximum value around 1.7 in most models, but it is higher in the Pandey model (the time-varying behaviour of R_e is displayed in Appendix A.4).
In SEIR2psi model, the parameter ψ quantifying the interaction between strains is inferior to 1, suggesting a reduced infectivity of secondary infections on average, as in Aguiar et al. (2011), or in Coudeville and Garnett (2012). This suggests that, as far as infectiousness is concerned, cross-protection among serotypes is more important than cross-enhancement to explain the results observed in the field.

It is estimated in the models that approximately half of the children are susceptible to the disease, which is close to the informative prior used, highlighting the difficulty to estimate this

quantity due to identifiability issues. This proportion is lower in the Pandey model. In the models with two strains, the number of susceptibles differs between models, but as a whole, at least 50% of children are not totally immune to the disease and can still suffer infection. The proportion of children who are susceptible to one or both strains are however correlated in the MCMC chain (cf. Appendix A.3), indicating that their relative proportions are not well identified. These values are in the range of the measures of seroprevalence in several Asian countries (L'Azou et al., 2016, 2018). As the measures reveal large differences between countries (L'Azou et al., 2016, 2018; Prayitno et al., 2017; Thai et al., 2005; Vongpunsawad et al., 2017), a seroprevalence survey in Cambodia would be particularly useful to evaluate which scenario is more plausible.

The observation rate for NDSS data varies between models, from 6 to 13%. These values indicate that a large proportion of dengue cases are not reported in national surveillance, likely reflecting mild symptoms that do not require hospitalization or misdiagnosis or misreporting.

In the SEIR2psi model, we also plotted the current number of infected individuals for each strain (cf. Figure 2.9). In our simulations, the first strain is responsible for large explosive outbreaks, whereas the second one has a more regular pattern over the years. Moreover, the two strains are asynchronous and each one dominates for two or three years. It is also qualitatively close to the dynamics observed in Thailand (Reich et al., 2013) or Singapore (Ong et al., 2017).
The proportion of susceptible individuals also displays asynchronous dynamics between strains, as they reflect the history of past epidemics. Despite the seasonality and the year-to-year variations, the total number of susceptibles remains high (cf. Figure 2.9 for SEIR2psi and SEIR models), allowing the possibility for large outbreaks to occur in the future. However, multi-strain models are difficult to parameterize due to the uncertainty in the initial conditions, and models with differing trajectories and parameters may have close performances in terms of model fitting (cf. Appendix A.5).

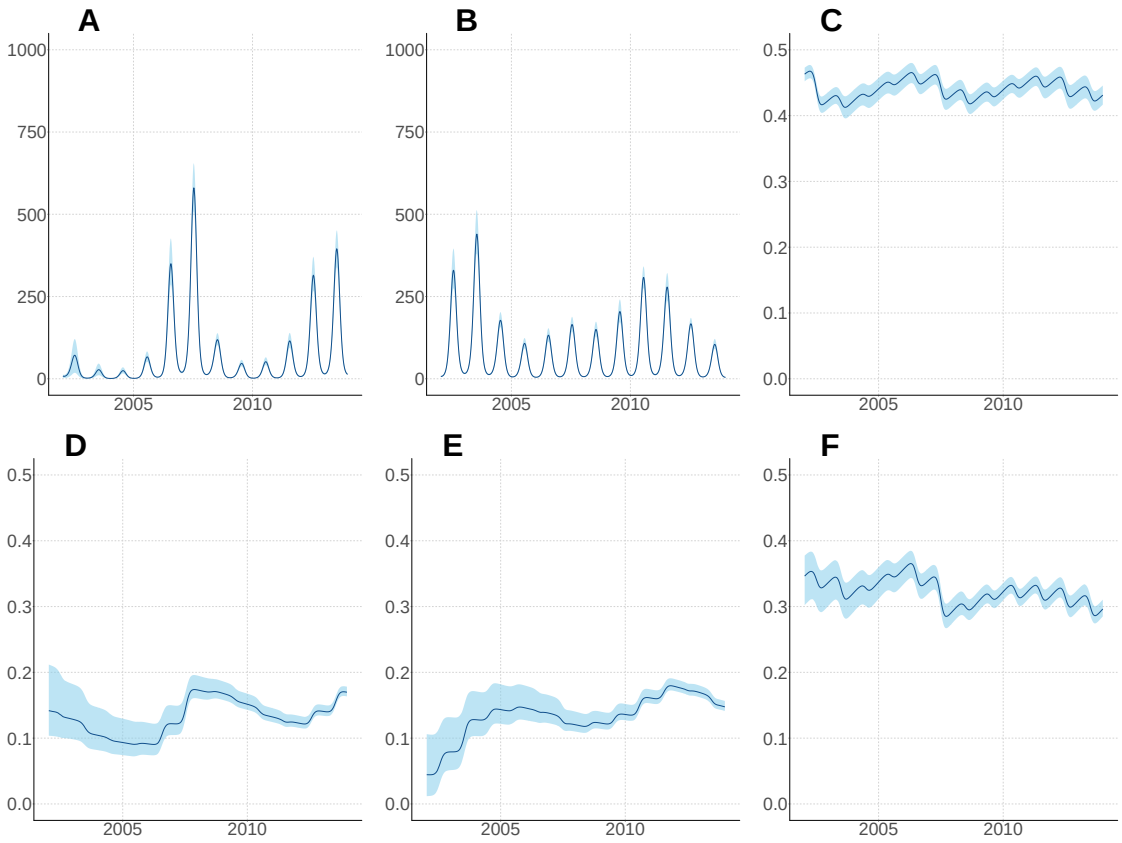


Figure 2.9 – **Number of infected and proportion of susceptible individuals.** Simulations without negative binomial noise using parameters from the MCMC chain associated with the highest posterior, calculated using both NDSS and DENFREE datasets. Posterior median (solid line) and 95% credible intervals (shaded blue area). A. Individuals infected with strain 1 in SEIR2psi model ($H_{I1} + H_{I21}$). B. Individuals infected with strain 2 in SEIR2psi model ($H_{I2} + H_{I12}$). C. Proportion of susceptible individuals in SEIR model (H_S/N). D. Proportion of susceptible individuals with immunity to strain 1 in SEIR2psi model (H_{S1}/N). E. Proportion of susceptible individuals with immunity to strain 2 in SEIR2psi model (H_{S2}/N). F. Proportion of individuals susceptible to both strains in SEIR2psi model (H_S/N).

We calculated the annual incidence proportion as the proportion of new infections over one year among the susceptibles at the beginning of that year (cf. Figure 2.10). The values for primary infections are coherent with other studies in Vietnam or Indonesia who analyzed seroprevalence data or seroconversion data (Graham et al., 1999; Prayitno et al., 2017; Thai et al., 2005, 2007; Tien et al., 2010). The incidence proportion is highly variable from one year to the next, especially in models with two interacting strains.

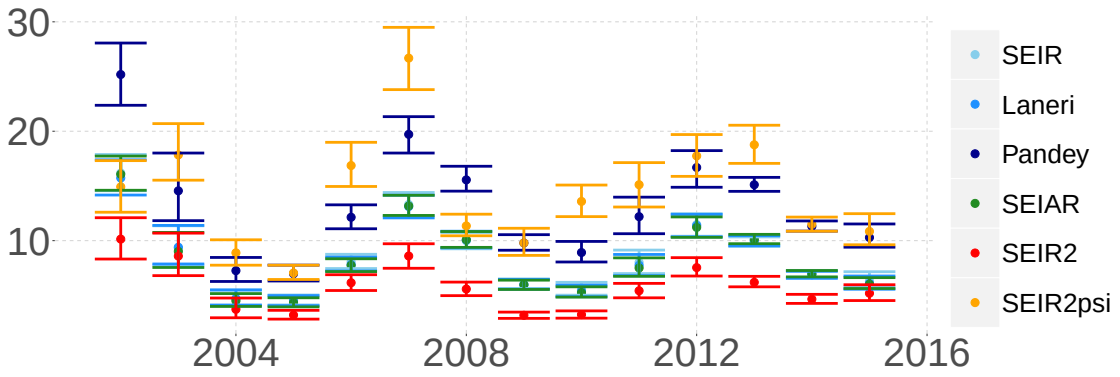


Figure 2.10 – **Annual incidence proportion of first dengue infection (%).** Median and credible intervals per year, over 2002-2015, based on simulations without negative binomial noise using parameters from the MCMC chain associated with the highest posterior, calculated using both NDSS and DENFREE datasets.

Discussion

With two datasets reporting dengue cases in the Kampong Cham region in Cambodia, we compared several models to represent dengue transmission dynamics in a rural setting. In order to assess the quality of the models, we compared their statistical properties and analysed their epidemiological features.

The best model describing the dengue trend over 14 years of data was the two strain model, with reduced infectivity for secondary infected individuals. Secondary infections being more prone to severe dengue (Halstead, 2007), these individuals may stay at home or at the hospital, and be less involved in the spread of the disease than the ones whose mild symptoms do not alter what Stoddard et al. (2013) call "house-to-house movement" (Yoon et al., 2012; Perkins et al., 2016a). There is also some indication from studies on the infectiousness of infected people to mosquitoes, that symptomatic infections are less infectious than asymptomatic individuals (Duong et al., 2015). Likewise there was a small negative impact of IgG on transmission rates that would be indicative of secondary and potentially more symptomatic cases (Nguyen et al., 2013). This feature was previously analytically studied and coherent with dengue incidence time series in Thailand (Aguiar et al., 2011, 2014).
On the contrary, including vectorial transmission or a compartment for asymptomatic infections did not seem to improve the model fit despite the additional complexity. In our models, the role of mosquitoes in transmission seems well captured by the seasonal forcing implemented in the model. The limited added-value of including vectorial components has been observed by several authors previously, when studying one dengue season only (Pandey et al., 2013) or when a seasonal forcing was included in the model (Rocha et al., 2016). Mathematical analyses have suggested that because the time scale of the mosquito epidemiology is so fast compared to that in humans, it will be slaved by the slower human epidemiology (Rocha et al., 2013). Thus, for understanding human disease epidemiology, mainly the dynamics of the human time scale are essential and inclusion of mosquito dynamics results in an unnecessary increase in model complexity when vector data is not available (Dye and Williams, 1995; Rocha et al., 2013, 2016), and when potential seasonal variations are already accounted for.
Our models account for underreporting, and the lack of additional improvement when including explicitly the asymptomatic class is likely due to the very few asymptomatic infections observed,

which may be due to a very strict definition of asymptomatic infections in the DENFREE protocol. Yoon et al. (2012) also observe many inapparent cases in their cohort but few strictly asymptomatic cases in their cluster study in Thailand. Therefore, in our model, including a compartment for asymptomatic individuals had only little influence on the overall transmission due to their small number.

However, even though the two strain model with interactions reproduces the observed data more adequately, it is difficult to parameterize. Differing model parameterizations may lead to differing dynamics that relate closely to the observed data, as shown in Appendix A.5, indicating identifiability issues (Miao et al., 2011). Identifiability is nonetheless a central problem in designing and estimating epidemiological models, that also affects models with vector-borne transmission (Kao and Eisenberg, 2018) and even simple models (Evans et al., 2002). In our case, this problem was related to two phenomena. On the one hand, some parameters values were correlated, and therefore difficult to distinguish from one another. On the other hand, in some models, the posterior distribution was multimodal, and therefore both hard to explore with many inference methods and hard to interpret given the presence of several possible trajectories. In order to limit these difficulties in all models, we fixed some parameters (in particular the durations of infection and incubation, or the observation rate in DENFREE data), and relied on informative priors (in particular for the initial proportion of susceptibles). In addition, it is possible to rule out some trajectories that correctly fit the data but display epidemiological incoherences in the qualitative dynamics of their hidden states (such as the depletion of the susceptible population or the absence of serotype asynchrony), in the spirit of pattern matching methods (ten Bosch et al., 2016). We therefore studied the epidemiological characteristics of the selected models to ensure such coherence. Other trajectories may, nevertheless, remain equally plausible given the available information and require additional data to be further explored. These issues are even more important when evaluating public health measures, as models with similar fit may have different responses to control scenarios (Kao and Eisenberg, 2018; ten Bosch et al., 2016). Therefore, care is required when choosing the best model, and the uncertainty in the model structure must also be taken into account, especially when the observed data are scarce.

Over the six models that were explored, we also obtained some insight on the parameters describing transmission. The average annual R_0 in our paediatric population is estimated between 1 and 4 over all models. These values are within the range observed in urban settings, suggesting that, despite very different population densities, the rural dynamics of dengue are not that dissimilar, or that dissimilarities are hidden by the variations between countries and populations and the uncertainties due to diverse estimation procedures, as this quantity is difficult to evaluate using incidence data. The median annual incidence at primary infection over the period is between 5 and 14, with large year-to-year variations. The estimated observation rate on surveillance data varies between models (9-13% in most models and 6-7% in the model with two interacting strains), indicating in both cases a high proportion of unreported infections, so that most of the transmission is due to unobserved infections. These values are in line with the large underrecognition highlighted in Wichmann et al. (2011), and also with other studies in South-East Asia (Graham et al., 1999; Thai et al., 2007; Yoon et al., 2012). These results are also coherent with a recent study emphasizing that more that 80% of dengue transmission is due to inapparent infections (ten Bosch et al., 2018).

This work has however several limitations. Firstly, model comparison was not straightforward between one strain and two strain models, both statistically and epidemiologically. From the statistical point of view, the differing number of observations between models led us to use

simulation based-indicators. From the epidemiological point of view, single strain models, two strain models and observations on four serotypes may be hard to compare because some indicators cover different interpretations. For example, in single strain models, there is no distinction in the susceptible class between individuals immune to one strain only and naive ones, and there is no strain specific reproduction number or incidence.
Secondly, the selected model formulations were restricted due to data availability. In particular, despite the endemicity of the four serotypes in rural Cambodia, we did not consider more than two dengue serotypes. This was done to limit model complexity, especially in the number of unknown initial conditions, but has also been previously shown to adequately describe dengue dynamics (Aguiar et al., 2013). When two serotypes were considered, we tested only interactions in terms of enhancement or restriction of infectiousness. We did not include models with (temporary) cross immunity, because of the too large increase in the number of parameters with respect to the data. We also did not include models with a finer spatial scale, even if small scale transmission plays a decisive role in dengue dynamics (Salje et al., 2012, 2017). On the one hand, NDSS data were only available at the district level, which was too large to follow transmission chains and too small for observing a sufficient number of cases. On the other hand, the clustered sampling protocol in the DENFREE study abnegated the interpretation of the spatial distribution of community cases. We also restricted the analysis to children under 15 years old and did not study the role of adults in transmission.
Thirdly, the projections are not completely able to describe the observed data, as most models overestimate the dengue epidemic in 2014. Nevertheless, 2014 was a particular year, with the lowest number of cases in the whole time series, potentially due to particular climatic conditions. In many countries in South-East Asia, except Malaysia, the reported incidence was lower than in 2013 (Cheng et al., 2017). Many provinces of Thailand also reported fewer cases than usual in 2014 (Reich et al., 2016). Our models are deterministic and do not take into account variations due to demographic stochasticity or environmental hazards such as climate.

Despite these limitations, combining two datasets permitted us to overcome some observation biases, such as the fact that surveillance data did not report serotype and DENFREE data did not reflect the long term dynamics. Nevertheless, some information is lacking in both datasets, in particular that on seroprevalence. Clearly the parameter estimations depend on the immunological status at the beginning of the simulations, especially in complex multi-strain models. As in our previous work (Champagne et al. (2016), chapter 4), our modeling study stresses the importance of seroprevalence data in order to more accurately estimate the initial conditions of our simulation and reduce identifiability problems. A seroprevalence survey in Cambodia would be of great value to evaluate the dengue burden, transmissibility potential and consider vaccination scenarios.

Conclusion

In conclusion, our analyses highlight that two-strain models with interacting effects better reproduce the long term dengue dynamics, but they are also difficult to parameterize relying on incidence data only. On the other hand, incorporating vector and asymptomatic components seems of limited added-value in this case when seasonality and underreporting are already accounted for. Although model complexity is framed by the scientific objectives, it must also be driven by the available data. The unavailability of mosquito data and the difficulty of observing asymptomatic infections questions their incorporation explicitly in the models. Another important aspect is

related to the comparison of models considering the available data. In addition to model selection based on goodness of fit (Pandey et al., 2013), assessing the validity of model's outputs in terms of epidemiological features (such as reproduction numbers, annual incidence, dynamics of the susceptible classes) is an important step when data are scarce and identifiability issues are present.

Acknowledgments

We thank the members of the Epidemiology and Public Health unit and the Virology unit in Institut Pasteur in Cambodia for their work during the DENFREE project. All the authors have been supported by a grant from Agence Nationale de la Recherche for the PANIC project (Targeting PAthogen's NICche: a new approach for infectious diseases control in low-income countries: ANR-14-CE02-0015-01). The research leading to these results has also received funding from the European Commission Seventh Framework Program FP7 for the DENFREE project under Grant Agreement 282 378. CC and BC are partially supported by the "Pépiniere interdisciplinaire Eco-Evo-Devo" from the Centre National de la Recherche Scientifique (CNRS). CC is supported by the PhD studentship of the Groupe des Ecoles Nationales d'Economie et Statistique (GENES). The funders played no role in the study design, data collection, analysis, or preparation of the manuscript.

2.3 Comparison of stochastic and deterministic frameworks in dengue modeling [2]

Abstract

We perform estimations of compartment models for dengue transmission in rural Cambodia with increasing complexity regarding both model structure and the account for stochasticity. On the one hand, we successively account for three embedded sources of stochasticity: observation noise, demographic variability and environmental hazard. On the other hand, complexity in the model structure is increased by introducing vector-borne transmission, explicit asymptomatic infections and interacting virus serotypes. Using two sources of case data from dengue epidemics in Kampong Cham (Cambodia), models are estimated in the bayesian framework, with Markov Chain Monte Carlo and Particle Markov Chain Monte Carlo. We highlight the advantages and drawbacks of the different formulations in a practical setting. Although in this case the deterministic models provide a good approximation of the mean trajectory for a low computational cost, the stochastic frameworks better reflect and account for parameter and simulation uncertainty.

Introduction

Stochasticity is a key element in infectious disease dynamics, as transmission of pathogens depends on individual specificities, randomness of contacts and environmental context. These effects are crucial in small populations, for instance at the household or school level, but also in larger populations, when the number of infected individuals is small and the extinction of the disease likely to occur (Bailey, 1975; Britton, 2010). Therefore, demographic stochasticity plays a major role in the study of invasion and persistence of the pathogen (Bartlett, 1957, 1964). First, it tackles questions that do not arise in the deterministic case, such as the probability of a major outbreak or the time to extinction (Allen, 2008; Lloyd et al., 2007). Second, it may lead to specific model behaviours that are qualitatively different from the deterministic ones (Nasell, 2002; Rohani et al., 2002). The uncertainties of the environment also shape disease spreading patterns, regardless of population size (He et al., 2010). Moreover, when inference is performed, deterministic models may lead to biased results (Britton and Lindenstrand, 2009) or underestimate the uncertainty in parameters and predictions, as they account only for noise in the observation process and not in the disease dynamic itself (King et al., 2015). Stochasticity can then compensate and sometimes highlight model misspecification (King et al., 2015), and in some contexts, simple stochastic models may be competitive with more complex deterministic ones (Stocks et al., 2018). However, stochastic models induce a cost in computation, estimation and interpretation of the results, and an analysis of their deterministic counterpart can prove highly informative.

Dengue is a vector-borne viral disease whose dynamics are highly influenced by both demographic and environmental stochasticity. Ubiquitous in the tropics and the subtropics, particularly in Southeast Asia, the Pacific and the Americas (Guzman et al., 2010), it is caused by any of four dengue virus (DENV) serotypes and transmitted by *Aedes* spp. Dengue dynamics are therefore shaped by the complex system involving mosquito vectors, human hosts and virus specificities, all of which are also influenced by random phenomena in populations and environment. The small spatial scale of transmission, due to the short mosquito flying rate, leads to a special geographical signature (Salje et al., 2012, 2017), in which demographic stochasticity plays a major role. Climatic

[2] Champagne C., Cazelles B., Comparison of stochastic and deterministic frameworks in dengue modeling, *in revision*

variations affect the mosquito life cycle and the transmission process (Liu-Helmersson et al., 2014; Cazelles et al., 2005), and human activities also impact the disease's spread, through movement (Teurlai et al., 2012; Stoddard et al., 2009) or behaviour (Reiter et al., 2003).

In this work, we compare several mechanistic models for dengue dynamics in rural Cambodia, that incorporate differing levels of complexity with respect to vector dynamics, coexistence of several virus strains, and transmission via asymptomatic infections. We consider these models both in deterministic and stochastic formulations, including step by step demographic and environmental stochasticity, in the framework by Bretó et al. (2009), as in Stocks et al. (2018). Whereas demographic stochasticity results from the random births, deaths and interactions between individuals, environmental stochasticity refers to the uncertainty from external sources and is accounted for by multiplying some transition rates with a white noise (Bretó et al., 2009). Models are adapted to observed time series of dengue cases in the Kampong Cham province (Cambodia) through Markov Chain Monte Carlo (MCMC) or Particle Markov Chain Monte Carlo (PMCMC) (Andrieu et al., 2010) using the SSM library (Dureau et al., 2013a). Models are compared in light of the data, relying on statistical indicators (King et al., 2008; Camacho et al., 2011; Pandey et al., 2013), and we also analyze their epidemiological coherence. We can therefore explore the increase in complexity both in the model structure and in the account for uncertainties and identify the advantages and drawbacks of the different formulations in a practical estimation setting.

Methods

Data

Kampong Cham province is a densely populated rural province in Cambodia, 120km northeast from the capital Phnom Penh, where dengue is endemic and strongly seasonal (cf. Figure 2.11). We used two different datasets reporting dengue cases in the province: the results of a punctual study conducted in a 30km radius around the city of Kampong Cham (DENFREE data), and the national surveillance data (NDSS data) in the four districts comprising the DENFREE study area (Kampong Cham, Kampong Siem, Prey Chhor and Tboung Khmum, with the administrative divisions of 2012-2013).

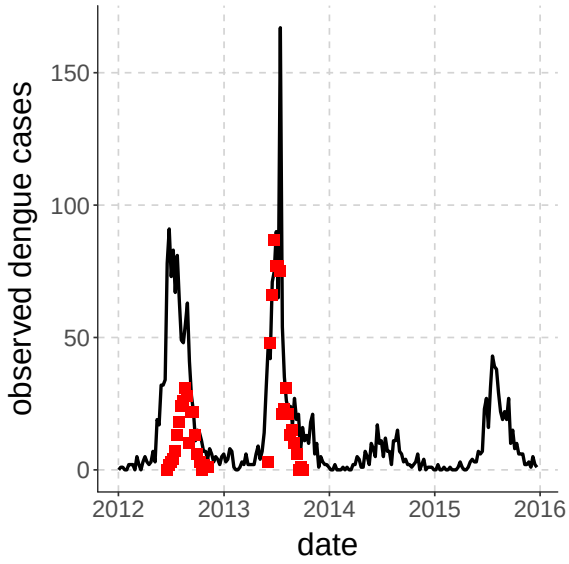


Figure 2.11 – **NDSS and DENFREE data.** The black line the series of NDSS cases per week in four districts of the Kampong Cham province. The red dots are the DENFREE cases per week in the DENFREE study area.

DENFREE data

The DENFREE study took place in the Kampong Cham region during the 2012 and 2013 outbreaks. Patients with acute dengue-like illness were enrolled in three hospitals in the Kampong Cham province and DENV infection was confirmed by qRT-PCR. Positive DENV cases were considered as index cases, and an outbreak investigation was initiated in their neighbourhood, in order to detect asymptomatic or mildly symptomatic cases. The study protocol is extensively detailed in (Duong et al., 2015). Information on the serotype responsible for infection, and symptomatic/asymptomatic status of the patients are available (cf. Table 2.8).

Children under 15 years old		
DENFREE data	**2012**	**2013**
Observed symptomatic cases $\left(C_{DS}^{\text{obs}}\right)$	236	574
Denv-1 $\left(C_{D1}^{\text{obs}}\right)$	226	451
Denv-2, Denv-3, Denv-4 $\left(C_{D2}^{\text{obs}}\right)$	5	122
Observed asymptomatic cases $\left(C_{DA}^{\text{obs}}\right)$	5	28
Children tested in the community	1722	4119
NDSS data	**2012 to 2015**	
Observed NDSS cases $\left(C_{N}^{\text{obs}}\right)$	2853	

Table 2.8 – **Number of cases under 15 years old, in the DENFREE study (index cases and community cases) and in the NDSS dataset.** Serotype was unknown for 5 cases in 2012. One case in 2013 was coinfected by Denv-1 and Denv-2 and was not included in models with two strains (serotypes).

NDSS data

Because the DENFREE data covers only a very short period of time, surveillance data was added to improve the estimations. Surveillance of dengue is conducted at the national level in

Cambodia, through the National Dengue Surveillance system (NDSS) (Huy et al., 2010), involving the paediatric departments of several hospitals throughout the country. Diagnosis is done clinically and only a small fraction of the cases are confirmed serologically. Because of the co-circulation of other flaviviruses (Chikungunya, Japanese Encephalitis) and the relative non-specificity of symptoms, clinical mis-diagnosis may be frequent.

We selected all the cases under 15 years old in the four districts involved in the DENFREE study between January 2012 and December 2015 and aggregated them per week (cf. Figure 2.11). In this area, on average, 713 cases under 15 years old were reported per year (1126, 1043, 209 and 475 in 2012, 2013, 2014 and 2015 respectively). As displayed on Figure 2.11, NDSS and DENFREE data have similar dynamics.

Models description

Five epidemiological models are considered, following Champagne et al. (2018). The reactions describing their stochastic counterparts are detailed in Appendix B.1.

One-strain models

We take a Susceptible-Exposed-Infected-Recovered (SEIR) model as the simplest model (cf. equation 2.6), with H_S susceptible individuals; H_E infected (not yet infectious) individuals; H_I infectious individuals and H_R recovered individuals. The parameter σ is the rate at which H_E-individuals move to the infectious class H_I and infectious individuals (H_I) then recover at rate γ. C_N and C_{DS} represent the count of new symptomatic cases and they are aggregated weekly to be compared respectively with NDSS cases (C_N with reporting rate r_N) and DENFREE symptomatic cases (C_{DS} with reporting rate r_{DS}). We also assume that a constant number of cases δ are imported, in order to prevent extinction of the disease.

$$\begin{aligned}
\frac{dH_S}{dt} &= \mu_H N - \beta(t)\frac{(H_I+\delta)H_S}{N} - \mu_H H_S \\
\frac{dH_E}{dt} &= \beta(t)\frac{(H_I+\delta)H_S}{N} - \sigma H_E - \mu_H H_E \\
\frac{dH_I}{dt} &= \sigma H_E - \gamma H_I - \mu_H H_I \\
\frac{dH_R}{dt} &= \gamma H_I - \mu_H H_R \\
\frac{dC_N}{dt} &= \frac{dC_{DS}}{dt} = \sigma H_E
\end{aligned} \tag{2.6}$$

with $H_S + H_E + H_I + H_R = N$. In this model, the basic reproduction number, i.e. the number of secondary human infections resulting from the introduction of a single infected individual in a entirely susceptible population, is $R_0^{SEIR}(t) = \frac{\beta(t)\sigma}{(\gamma+\mu_H)(\sigma+\mu_H)}$.

This model is compared with two other models that include the mosquito vector transmission components. The first one, called SEIR-vector, is a Ross-Macdonald type model derived from Pandey et al. (2013), in which the vector is modelled explicitly with three compartments (Susceptible-Exposed-Infected) (cf. equation 2.7).

$$
\begin{aligned}
\frac{dH_S}{dt} &= \mu_H N - m\, c_H \frac{V_I}{V} H_S - \mu_H H_S \\
\frac{dH_E}{dt} &= m\, c_H \frac{V_I}{V} H_S - \sigma H_E - \mu_H H_E \\
\frac{dH_I}{dt} &= \sigma H_E - \gamma H_I - \mu_H H_I \\
\frac{dH_R}{dt} &= \gamma H_I - \mu_H H_R \\
\frac{dV_S}{dt} &= \mu_V V - \beta_V(t) \frac{(H_I + \delta)}{N} V_S - \mu_V V_S \\
\frac{dV_E}{dt} &= \beta_V(t) \frac{(H_I + \delta)}{N} V_S - \tau V_E - \mu_V V_E \\
\frac{dV_I}{dt} &= \tau V_E - \mu_V V_I \\
\frac{dC_N}{dt} &= \frac{dC_{DS}}{dt} = \sigma H_E
\end{aligned}
\tag{2.7}
$$

where V_S, V_E and V_I are respectively the number of susceptible, exposed and infected mosquitoes and $V_S + V_E + V_I = V = mN$, m being the unknown number of mosquitoes per human. The parameter τ is the rate at which V_E-vectors move to the infectious class V_I and vectors die at rate μ_V. The parameters c_H and $\beta_V(t)$ represent respectively the transmission rate from vector to human and from human to vector, as the product of the biting rate and the probability of transmission per bite. As noted by Pandey et al. (2013), only the product $m\, c_H$ appears in the equations, so that both m and c_H cannot be estimated separately for identifiability reasons. Therefore, we estimate instead the parameter $\beta_H = m\, c_H$, and the total number of mosquitoes V remains unknown. This is not an issue in the deterministic framework, because both human and vector populations are continuous state variables that can be treated as proportions. However, in the stochastic framework, because the model's compartments are discrete, the assumption on the magnitude of V may influence the amount of demographic stochasticity in the model. In our case, we assumed $V' = N$.
In this model, R_0 is derived using the next-generation matrix (Diekmann et al., 2010) as $R_0^{vect}(t) = \sqrt{\frac{m c_H \beta_V(t) \sigma \tau}{(\gamma+\mu_H)(\sigma+\mu_H)\mu_V(\mu_V+\tau)}} = \sqrt{\frac{\beta_H \beta_V(t) \sigma \tau}{(\gamma+\mu_H)(\sigma+\mu_H)\mu_V(\mu_V+\tau)}}$. This value corresponds to the number of secondary infections per generation. However, as detailed in Roberts and Heesterbeek (2003), the "expected number of secondarily infected humans that result from a single infected human" is $T_1 = (R_0^{vect})^2$, and it quantifies the "effort required to control the disease".

In the second one, derived from Laneri et al. (2010) (cf. equation 2.8), the vector is modelled implicitly as an external force of infection including two stages, latent (κ) and current (λ).

$$
\begin{aligned}
\frac{dH_S}{dt} &= \mu_H N - \lambda H_S - \mu_H H_S \\
\frac{dH_E}{dt} &= \lambda H_S - \sigma H_E - \mu_H H_E \\
\frac{dH_I}{dt} &= \sigma H_E - \gamma H_I - \mu_H H_I \\
\frac{dH_R}{dt} &= \gamma H_I - \mu_H H_R \\
\frac{d\kappa}{dt} &= \beta(t)\frac{2(H_I+\delta)\tau}{N} - 2\tau\kappa \\
\frac{d\lambda}{dt} &= 2\tau\kappa - 2\tau\lambda \\
\frac{dC_N}{dt} &= \frac{dC_{DS}}{dt} = \sigma H_E
\end{aligned}
\tag{2.8}
$$

where the parameter τ is the transition rate associated to the extrinsic incubation period. Because the compartments λ and κ refer to the delayed force of infection and not to real mosquito populations, they should not be subject to demographic stochasticity in the same way as the other compartments, and should be treated as continuous state variables. However, the considered software does not allow to combine discrete and continuous compartments within the same model. Therefore, we consider the two compartment as discrete but rescaled by a factor α, such that $\lambda' = \alpha\lambda$ and $\kappa' = \alpha\kappa$, with $\alpha = N$. We derived R_0 for this model as the estimation $R_0^{Laneri}(t) = \frac{\beta(t)\sigma}{(\gamma+\mu_H)(\sigma+\mu_H)}$ (Champagne et al., 2016).

Model with explicit asymptomatic individuals (SEIAR)

We also consider a model in which asymptomatic infections are explicitly taken into account in the transmission process (cf. equation 2.9). In this model, we assume that, after the incubation period, there are three possible manifestations of the disease: asymptomatic (H_A), mildly symptomatic not requiring hospitalization (H_I) and hospitalized cases (H_H). C_{DA} represents the count of new asymptomatic cases, and it is aggregated weekly to be compared with DENFREE data on asymptomatic cases (with reporting rate r_{DA}).

$$
\begin{aligned}
\frac{dH_S}{dt} &= \mu_H N - \beta(t)\frac{(H_I + H_A + H_H + \delta)H_S}{N} - \mu_H H_S \\
\frac{dH_E}{dt} &= \beta(t)\frac{(H_I + H_A + H_H + \delta)H_S}{N} - \sigma H_E - \mu_H H_E \\
\frac{dH_A}{dt} &= \rho_A \sigma H_E - \gamma H_A - \mu_H H_A \\
\frac{dH_H}{dt} &= \rho_H(1-\rho_A)\sigma H_E - \gamma H_H - \mu_H H_H \\
\frac{dH_I}{dt} &= (1-\rho_H)(1-\rho_A)\sigma H_E - \gamma H_I - \mu_H H_I \\
\frac{dH_R}{dt} &= \gamma(H_I + H_A + H_H) - \mu_H H_R \\
\frac{dC_N}{dt} &= \rho_H(1-\rho_A)\sigma H_E \\
\frac{dC_{DS}}{dt} &= (1-\rho_A)\sigma H_E \\
\frac{dC_{DA}}{dt} &= \rho_A \sigma H_E
\end{aligned}
\tag{2.9}
$$

where a proportion ρ_A of H_E-individuals do not show symptoms during the infectious period; a proportion ρ_H of symptomatic individuals go to hospital and $H_S + H_E + H_A + H_I + H_H + H_R = N$.

Asymptomatic cases are defined in the DENFREE study as asymptomatic or pauci-symptomatic (presence of other symptoms not being sufficient to classify as symptomatic). Hospital cases are defined as NDSS cases (reported by the surveillance system in hospitals). We assume that symptomatic DENFREE cases are $H_I + H_H$. We also assume that asymptomatic cases transmit the disease as much as symptomatic cases, as recently shown (Duong et al., 2015), and therefore, $R_0^{SEIAR}(t) = \frac{\beta(t)\sigma}{(\gamma+\mu_H)(\sigma+\mu_H)}$.

Model with two virus serotypes

Finally, a two-strain model is also studied, in which we separate DENV-1 cases from DENV-2, DENV-3 and DENV-4 combined (cf. equation 2.10). In this model, which was previously found the best for reproducing long term Kampong Cham series in the deterministic framework (Champagne et al., 2018), primary and secondary infections differ in infectiousness, through a parameter ψ (Ferguson et al., 1999). This parameter was previously estimated to 0.67 (Champagne et al., 2018) suggesting a lower infectivity for secondary infected individuals (for example because they are hospitalized and less in contact with the population (Aguiar et al., 2011)). For simplicity and parsimony in the number of parameters, the two strains share the same parameter values. In this model, there are H_S individuals susceptible to both strains; H_{Ei} individuals infected (not yet infectious) to strain i; H_{Ii} individuals infectious to strain i ; H_{Si} individuals immune to strain i only; H_{Eij} individuals (not yet infectious) with a secondary infection to strain j ; H_{Iij} infectious individuals with a secondary infection to strain j and H_R individuals immune to both strains and δ_i imported cases from strain i, with $(i, j) = (1, 2)$ or $(i, j) = (2, 1)$.

$$
\begin{aligned}
\frac{dH_S}{dt} &= \mu_H N - \beta(t)\frac{(H_{I1} + \psi H_{I21} + \delta_1)H_S}{N} - \beta(t)\frac{(H_{I2} + \psi H_{I12} + \delta_2)H_S}{N} - \mu_H H_S \\
\frac{dH_{E1}}{dt} &= \beta(t)\frac{(H_{I1} + \psi H_{I21} + \delta_1)H_S}{N} - \sigma H_{E1} - \mu_H H_{E1} \\
\frac{dH_{I1}}{dt} &= \sigma H_{E1} - \gamma H_{I1} - \mu_H H_{I1} \\
\frac{dH_{S1}}{dt} &= \gamma H_{I1} - \beta(t)\frac{(H_{I2} + \psi H_{I12} + \delta_2)H_{S1}}{N} - \mu_H H_{S1} \\
\frac{dH_{E12}}{dt} &= \beta(t)\frac{(H_{I2} + \psi H_{I12} + \delta_2)H_{S1}}{N} - \sigma H_{E12} - \mu_H H_{E12} \\
\frac{dH_{I12}}{dt} &= \sigma H_{E12} - \gamma H_{I12} - \mu_H H_{I12} \\
\frac{dH_{E2}}{dt} &= \beta(t)\frac{(H_{I2} + \psi H_{I12} + \delta_2)H_S}{N} - \sigma H_{E2} - \mu_H H_{E2} \\
\frac{dH_{I2}}{dt} &= \sigma H_{E2} - \gamma H_{I2} - \mu_H H_{I2} \\
\frac{dH_{S2}}{dt} &= \gamma H_{I2} - \beta(t)\frac{(H_{I1} + \psi H_{I21} + \delta_1)H_{S2}}{N} - \mu_H H_{S2} \\
\frac{dH_{E21}}{dt} &= \beta(t)\frac{(H_{I1} + \psi H_{I21} + \delta_1)H_{S2}}{N} - \sigma H_{E21} - \mu_H H_{E21} \\
\frac{dH_{I21}}{dt} &= \sigma H_{E21} - \gamma H_{I21} - \mu_H H_{I21} \\
\frac{dH_R}{dt} &= \gamma(H_{I12} + H_{I21}) - \mu_H H_R \\
\frac{dC_N}{dt} &= \frac{dC_{DS}}{dt} = \sigma\,(H_{E1} + H_{E21} + H_{E2} + H_{E12}) \\
\frac{dC_{D1}}{dt} &= \sigma\,(H_{E1} + H_{E21}) \\
\frac{dC_{D2}}{dt} &= \sigma\,(H_{E2} + H_{E12})
\end{aligned}
\tag{2.10}
$$

where $H_S + H_{E1} + H_{E2} + H_{I1} + H_{I2} + H_{S1} + H_{S2} + H_{E12} + H_{E21} + H_{I12} + H_{I21} + H_R = N$. For parsimony in the number of parameters, we also assume $\delta_1 = \delta_2 = \delta'$. In this context, the reproduction numbers for each strain are equal, and as in Ferguson et al. (1999), we define $R_0^{SEIR2_1}(t) = R_0^{SEIR2_2}(t) = \frac{\beta(t)\sigma}{(\gamma+\mu_H)(\sigma+\mu_H)}$.

C_{D1} and C_{D2} represent the count of new cases for strain 1 and strain 2, and are aggregated weekly to be compared with symptomatic cases from the DENFREE study for DENV-1 and DENV-2/DENV-3/ DENV-4 respectively (with reporting rate r_{DS}).

Seasonality

All models include seasonality through the use of a time-varying transmission parameter $\beta(t) = \beta[1 + b.sin(2\pi(\frac{t}{365} + p))]$, according to a sinusoidal function whose phase p and amplitude b are estimated.

Parameter values and prior distributions

The parameter values and prior distributions are listed in Table 2.9. Informative priors based on the literature are used for the durations of infectiousness and incubation, as well as the mortality rates. In models without vectorial transmission, the incubation period is assumed to be the sum of the extrinsic (in mosquito) and intrinsic (in human) incubation periods, to reflect the generation time of the disease. For transmission parameters, we use wide weakly informative priors. The observation rate for NDSS data, r_N, is estimated in all models except the SEIAR model. In the SEIAR model, the parameter r_N is equal to 1 by definition of H_H-individuals; instead, the proportion of hospitalized cases, ρ_H, is estimated. The observation rate for DENFREE data is calculated using information on the sampling scheme (Champagne et al., 2018); in the SEIAR model, it differs for symptomatic and asymptomatic cases.

Parameter		Prior distribution	Reference	Models
Infectiousness, incubation and mortality rates				
γ^{-1}	infectious period (days)	Normal(4.5, 0.1) in [4, 5]	(WHO, 2017)	All
σ^{-1}	intrinsic incubation period (days)	Normal(5.9, 0.1) in [3, 10]	(Chan and Johansson, 2012)	SEIR-vector, Laneri
τ^{-1}	extrinsic incubation period (days)	Normal(10, 1) in [5, 20]	(Chan and Johansson, 2012)	SEIR-vector, Laneri
σ^{-1}	both incubation periods (days)	Normal(15.9, 0.1) in [8, 30]	(Chan and Johansson, 2012)	SEIR, SEIAR, SEIR2
μ_H^{-1}	age duration (years)	15	assumed	All
μ_V^{-1}	mosquito lifespan(days)	Normal(15, 1) in [4, 30]	(Liu-Helmersson et al., 2014)	SEIR-vector
Transmission parameters				
R_0	basic reproduction number	Uniform[0, 20]	assumed	SEIR, Laneri, SEIAR, SEIR2
T_1	type reproduction number	Uniform[0, 20]	assumed	SEIR-vector
β_V	transmission from human to mosquito	Uniform[0, 10]	assumed	SEIR-vector
Initial conditions				
$V_I(0)$ or $L(0)$	Initial number of infected mosquitoes	N*Uniform[10^{-6},10^{-3}]	assumed	SEIR-vector or Laneri
$V_E(0)$ or $K(0)$	Initial number of exposed mosquitoes	$V_I(0)$ or $L(0)$	assumed	SEIR-vector or Laneri
Observation process				
r_N	observation rate for NDSS data	Uniform[0, 1]	assumed	SEIR, Laneri, SEIR-vector, SEIR2, SEIR2psi
r_N	observation rate for NDSS data	1	assumed	SEIAR
$r_{DS} = r_{D1} = r_{D2}$ (2012 // 2013)	observation rate for DENFREE data (symptomatic) (%)	2.83 // 6.61	(Champagne et al., 2018)	All
r_{DA} (2012 // 2013)	observation rate for DENFREE data (asymptomatic) (%)	1.07 // 2.55	(Champagne et al., 2018)	SEIAR
ρ_a	Proportion of asymptomatic cases	Uniform[0, 1]	assumed	SEIAR
ρ_h	Proportion of hospitalized cases	Uniform[0, 1]	assumed	SEIAR
Seasonality parameters				
b	amplitude of the sinusoidal forcing	Uniform[0, 1]	assumed	All
p	phase of the sinusoidal forcing	Uniform[−0.5, 0.5]	assumed	All
δ, δ'	import parameter	Uniform[0, 10]	assumed	All
Total population				
N	total population under 5 years old	161,391	(National Institute of Statistics, 2009; Champagne et al., 2018)	All

Table 2.9 – **Values and prior distributions for model parameters**. "Uniform[0,20]" indicates a uniform distribution in the range [0,20]. "Normal(4.5,0.1) in [4,5]" indicates a normal distribution with mean 4.5 and standard deviation 0.1, restricted to the range [4,5].

In order to reduce the parameter space, the initial conditions for the host compartments are fixed to the values obtained in a deterministic model, fitted on 14 years of NDSS data and the two-year

DENFREE data (precisely, we take the median value of 1000 simulations at the corresponding time point; the values are given in Table B.6) (Champagne et al., 2018). Initial conditions on the vector in Laneri (Laneri et al., 2010) and SEIR-vector models are estimated using uniform priors.

Implementation and estimation

Models are considered both in the deterministic and stochastic frameworks.

In the deterministic framework, the transmission process is modelled with ordinary differential equations, and the only source of uncertainty is in the observation process. In the stochastic frameworks, demographic stochasticity is included following the Euler-multinomial scheme as in Bretó et al. (2009). Two versions of the stochastic models are considered, as in Stocks et al. (2018). In the first one, only demographic stochasticity is taken into account. In the second one, following the framework by Bretó et al. (2009), environmental stochasticity is added to the demographic stochasticity, by introducing a white noise on the infection reaction. The infection rate at time t is therefore multiplied by a gamma noise $\xi(t) = \frac{d\Gamma(t)}{dt}$, where $\Gamma(t+\delta) - \Gamma(t) \sim \Gamma(\frac{\delta}{s^2}, s^2)$ and the square root of the infinitesimal variance s^2 is estimated (Bretó et al., 2009). In the Pandey and Laneri models, the white noise is included in the rate of vector to human infection. In the SEIR2 model, the two strains have independent white noises that share the same infinitesimal variance, as in Bretó et al. (2009).

In every case, an observation model is added, to link the epidemiological dynamics to the data in a state space framework and therefore represent the uncertainty in the observation process. The observed number of cases during week k in dataset d, $C_d^{\text{obs}}(k)$, is assumed to follow a discretized gaussian distribution, noted $p_{d,k}$, with mean $r_d C_d(k)$ and variance $r_d C_d(k) + \phi(r_d C_d(k))^2$, where $C_d(k)$ is the number of new cases during week k simulated by the model. The parameter r_d is the observation rate that quantifies underreporting in dataset d and ϕ is an overdispersion parameter. This formulation of the discretized gaussian distribution was motivated to approximate a negative binomial distribution, which is commonly used in observation models (Bretó et al., 2009). Assuming that the observations of the different datasets are independent conditional on the underlying disease process, the observation probability at week k is $\prod_{d \in D} p_{d,k}\left(C_d^{\text{obs}}(k)\right)$ with $D = \{N, DS\}$ in the SEIR, Laneri and SEIR-vector models, $D = \{N, DS, DA\}$ in the SEIAR model, and $D = \{N, D1, D2\}$ in the SEIR2 model.

We consider the bayesian perspective and aim to estimate the posterior distribution of the model parameters given the observed data, which is proportional to the product of the likelihood and the prior distribution. In the deterministic version, this is done using random walk Metropolis-Hastings (MCMC). When process stochasticity is accounted for, however, the model likelihood becomes intractable and this algorithm cannot be used. Therefore, in both stochastic versions, estimations are performed using particle Markov chain Monte Carlo (PMCMC)(Andrieu et al., 2010), an algorithm which relies on a particle filter to estimate the model likelihood within a Metropolis-Hastings scheme.

Open source SSM software (Dureau et al., 2013a) is used for calculations and simulations in all cases. In order to initialize the starting point of the chains, we run a simplex algorithm on 1000 randomly sampled parameter sets and choose the one with the highest posterior value at the end of the chain. The covariance matrix of the proposal distribution is initialized using adaptative MCMC as in (Dureau et al., 2013a) in the deterministic case and adaptative Kalman MCMC and adaptative PMCMC in the stochastic case (Dureau et al., 2013a,b). Convergence is assessed visually and trace

plots are displayed in Appendix B.3, in Figures B.1, B.2 and B.3.

Model comparison

In order to compare the statistical performance of the models, the Deviance Information Criterion (DIC) (Spiegelhalter et al., 2002) is used. DIC is an indicator that combines a measure of model fit and a penalty on model complexity, commonly used with MCMC estimations. The best model is the one with the smallest DIC. However, this indicator does not enable comparison of models estimated on differing sets of observations (such as asymptomatic infections or separate serotype data). Therefore, we take advantage of the possibility to re-simulate these mechanistic models to calculate a metric that is common to all models. In particular, we calculated the root mean square error (RMSE) on the total number of symptomatic cases (with K observations weeks, and S simulated trajectories):

$$\text{mean RMSE NDSS} = \frac{1}{M}\sum_{s=1}^{S}\text{RMSE NDSS}_s = \frac{1}{M}\sum_{s=1}^{S}\sqrt{\frac{1}{K}\sum_{k=1}^{K}\left(C_N^{\text{sim}_s}(k) - C_N^{\text{obs}}(k)\right)^2}$$

$$\text{mean RMSE DENFREE} = \frac{1}{M}\sum_{s=1}^{S}\text{RMSE DENFREE}_s = \frac{1}{M}\sum_{s=1}^{S}\sqrt{\frac{1}{K}\sum_{k=1}^{K}\left(C_{DS}^{\text{sim}_s}(k) - C_{DS}^{\text{obs}}(k)\right)^2}$$

The RMSE indicator does neither penalize models' complexity, nor use a different dataset for estimation and testing. It is therefore sensitive to overfitting and cannot evaluate the quality of model fit. To the very least, it provides a comparison tool between models estimated on differing time series. This indicator penalizes simulation envelopes with low data coverage and also simulation envelopes with large variance.
Moreover, when re-simulating stochastic models, one needs to set simulation horizons (because uncertainties propagate through time) and starting dates (as predictions are easier when the epidemic peak has been passed). In our case, we decided to simulate the models from the starting date until the end of the considered time series, in order to explore the range of possible outcomes over the whole studied period, given a set of initial conditions and a posterior distribution for the model parameters.

Models are also compared in terms of their epidemiological behavior. The basic reproduction number, the import parameter and the observation rate are estimated using the MCMC chain. Filtered trajectories are plotted to study the dynamics of hidden states, such as the susceptible individuals or infected vector classes.

Results

The information criteria for all models in the three frameworks are displayed in Table 2.10. In the deterministic framework, all single-strain models have very close log-likelihood values. Regarding RMSE values, the SEIR-vector model seems the best for reproducing NDSS data, and SEIR and SEIAR models for DENFREE data, but in general, RMSE values are close to one another. Indeed, all models seem to reproduce the data similarly, as the simulated trajectories have a similar qualitative behavior (cf. Figures 2.12 and 2.13). The deterministic credible intervals due to observation noise comprise almost all data points, except for the epidemic peak of 2013, which is

consistently underestimated in both NDSS and DENFREE datasets.

Model	**SEIR**	**Laneri**	**SEIR-vector**	**SEIAR**	**SEIR2**
nb observations	247	247	247	286	286
DETERMINISTIC					
nb parameters	8	10	12	9	8
log-likelihood	-758	-756	-751	-817	-848
DIC	1529	1528	1529	1647	1712
mean RMSE NDSS	18.7	18.9	17.8	19	18.2
mean RMSE DENFREE	24.9	26	26.9	24.9	29.7
STOCHASTIC (demographic)					
nb parameters	8	10	12	9	8
log-likelihood	-733	-731	-719	-789	-800
DIC	1481	1480	1460	1598	1620
mean RMSE NDSS	21.2	25.7	33.5	20.9	22.8
mean RMSE DENFREE	26.4	29.8	33.2	26.1	25.5
STOCHASTIC (demographic + environmental)					
nb parameters	9	11	13	10	9
log-likelihood	-706	-708	-709	-760	-771
DIC	1427	1435	1440	1536	1558
mean RMSE NDSS	39.2	32.9	25.1	36.8	49.3
mean RMSE DENFREE	37.8	34.8	27.5	35	36.1
White noise SD s (median + %95 CI)	2.5 (1.8-3.4)	1.7 (1.2-2.1)	1.9 (1.5-2.4)	2.2 (1.5-3.1)	3.6 (2.6-4.8)

Table 2.10 – **Information criteria.** Information criteria for the five models in the three frameworks. When environmental stochasticity is included, the median and credible intervals for the infinitesimal standard deviation are indicated. Similar results are obtained using absolute errors instead of squared errors (cf. Appendix B.5).

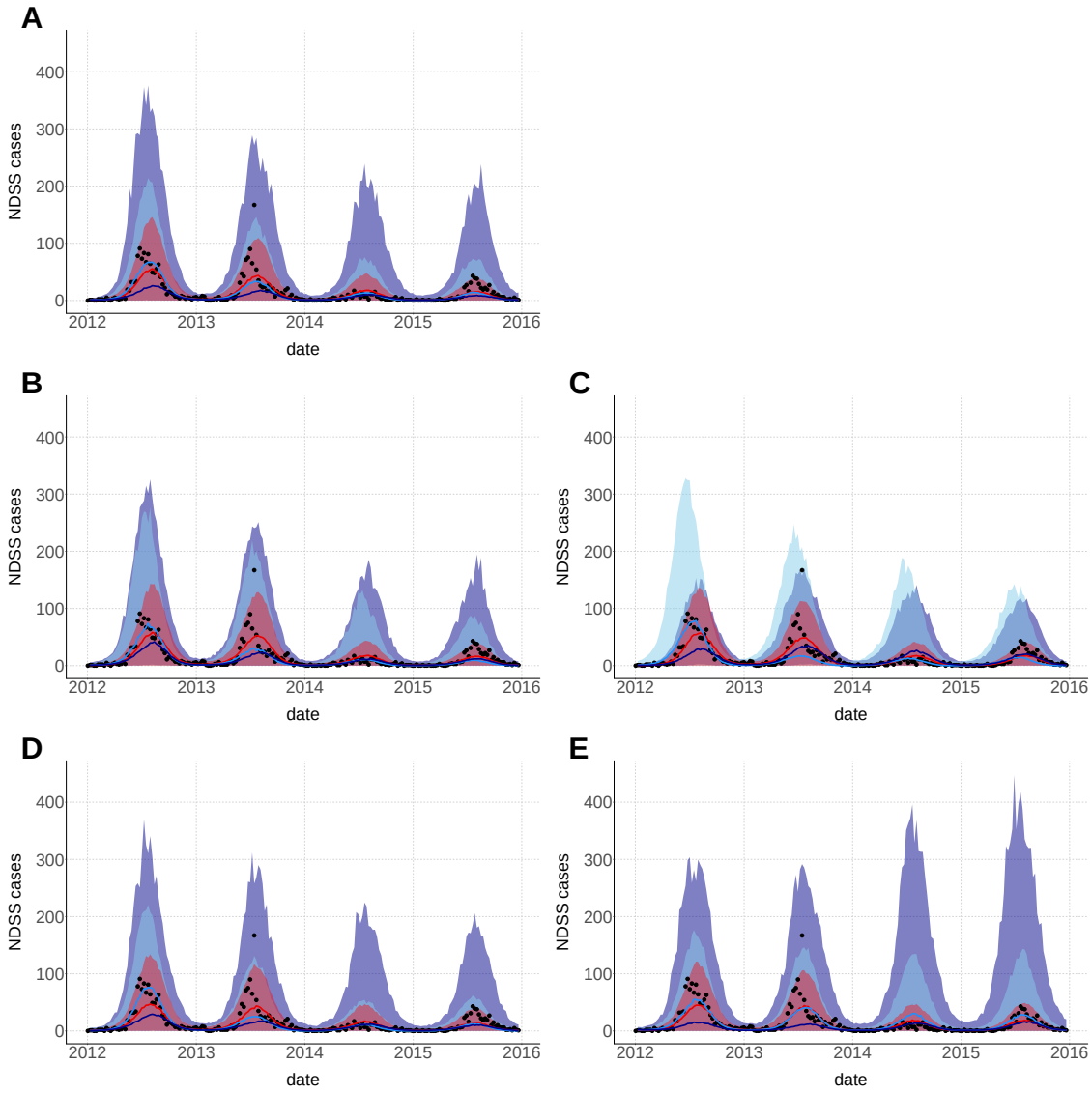


Figure 2.12 – **Number of observed cases per week with deterministic and stochastic models and NDSS data.** Simulations with negative binomial noise using parameters from the posterior distribution. Median (solid line), 95% credible intervals (shaded area) and NDSS data points (black dots). Red: deterministic case. Light blue: stochastic case with demographic stochasticity only. Dark blue: stochastic case with demographic and environmental stochasticity. A. SEIR model. B. Laneri model. C. SEIR-vector model. D. SEIAR model. E. SEIR2 model.

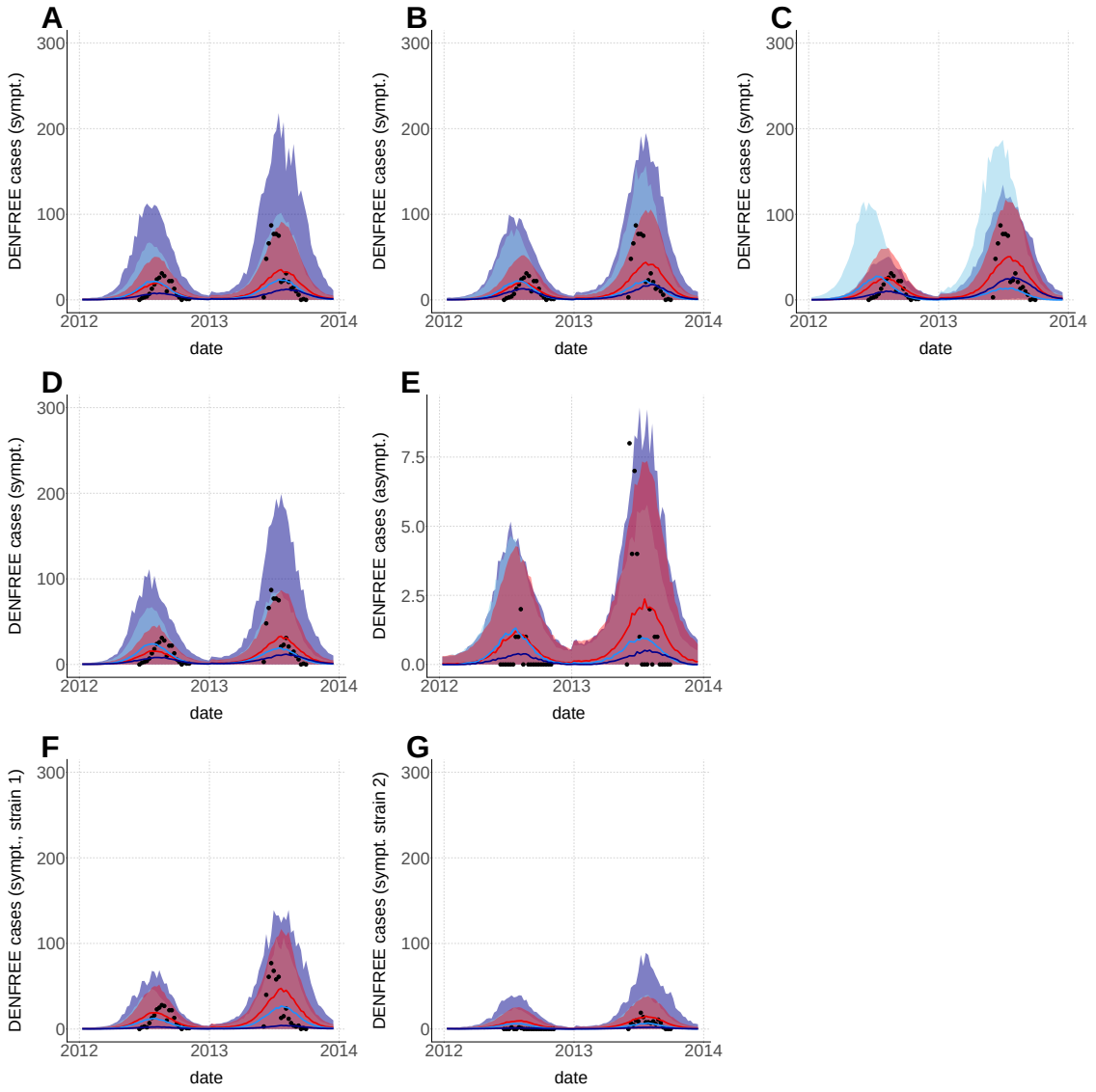


Figure 2.13 – **Number of observed cases per week with deterministic and stochastic models and DENFREE data.** Simulations with negative binomial noise using parameters from the posterior distribution. Median (solid line), 95% credible intervals (shaded area) and DENFREE data points (black dots). Red: deterministic case. Light blue: stochastic case with demographic stochasticity only. Dark blue: stochastic case with demographic and environmental stochasticity. A. SEIR model. B. Laneri model. C. SEIR-vector model. D. SEIAR, symptomatic cases. E. SEIAR, asymptomatic cases. F. SEIR2 model, strain 1 cases (DENV1). G. SEIR2 model, strain 2 cases (DENV2, DENV3, DENV4).

In the framework with demographic stochasticity only, the SEIR-vector model has the highest log-likelihood but also the highest RMSE value. Log-likelihood values are smaller than in the deterministic case, but RMSE values are larger, due to the much greater credible intervals in the stochastic simulations (cf. Figures 2.12 and 2.13). Uncertainty is particularly large in vector-borne models (Laneri and SEIR-vector models), due to the additional compartments for the vector which are subject to demographic stochasticity and uncertainty on both their initial values and parameters. This phenomenon is particularly strong in the SEIR-vector model because a very low absolute number of infected vectors is estimated (cf. Figure 2.14B1 and B2), leading to a greater

influence of stochasticity in the vector population. On the contrary, the filtered trajectories of the external force of infection λ in the Laneri model are much more comparable across stochasticity frameworks.
In all cases, the overdispersion parameter is lower than in the deterministic case (cf. Figure 2.15), indicating that less observation noise is needed to explain the data, when process noise is included.

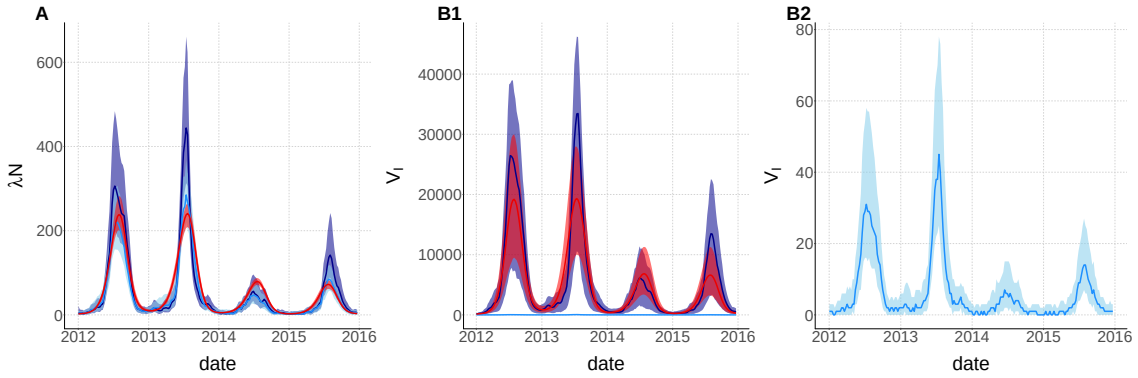


Figure 2.14 – **Number of infected vectors over time.** Median of the filtered trajectories and credible intervals without observation noise. Red: deterministic case. Light blue: stochastic case with demographic stochasticity only. Dark blue: stochastic case with demographic and environmental stochasticity. A. Laneri model: rescaled effective force of infection (λN). B1. SEIR-vector model: number of infected mosquitoes (V_i). B2. SEIR-vector model: number of infected mosquitoes (V_i) in the stochastic case with demographic stochasticity only.

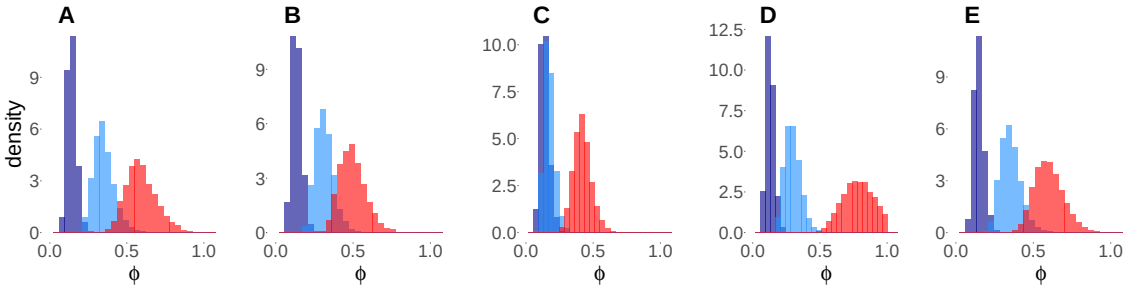


Figure 2.15 – **Posterior distributions for the overdispersion parameter ϕ in the five models.** Red: deterministic case. Light blue: stochastic case with demographic stochasticity only. Dark blue: stochastic case with demographic and environmental stochasticity. A. SEIR model. B. Laneri model. C. SEIR-vector model. D. SEIAR model. E. SEIR2 model.

When environmental stochasticity is added, the log-likelihood values are higher than in both previous cases (cf. Table 2.10), reflecting that these more flexible models manage to better reproduce the observed data by filtering. The SEIR model with environmental stochasticity is identified as the best model by the DIC criterion, probably as it is the most parsimonious one. On the contrary, RMSE values are higher than in both previous cases in non-vector models (cf. Table 2.10), mainly because of the higher variability in simulations with increased stochasticity (cf. Figures 2.12 and 2.13). They reflect the diversity in the simulated trajectories: in some simulations, the 2012 outbreak is very large, and all the subsequent ones much smaller, in other simulations, the 2012 is moderate, and large outbreaks occur in the subsequent years. In the SEIR-vector model, however, the credible intervals are smaller than in the stochastic case with demographic stochasticity only. The absolute number of infected vectors is higher (cf. Figure 2.14), indicating

that extra-demographic stochasticity manages to capture the large variability of the data without increasing demographic stochasticity via the reduction of the compartments' size. In all models, the standard deviation of the environmental noise is high, always higher than one (cf. Table 2.10). On the contrary, the overdispersion parameter is lower than in both previous cases (cf. Figure 2.15): in these estimations with large environmental noise, less observation noise is needed to explain the variability in the data. Indeed, the week-to-week variability is smaller than in both previous cases (except for the SEIR-vector model) and more in adequation with the data, when quantified by the autocorrelation at lag one week of each simulated series as in King et al. (2008) (results not shown).

The dynamics of the proportion of individuals in the susceptible class are very similar between models: it is relatively stable about 40%, and about 30% in the SEIR-vector and SEIR2 models (cf. Figure 2.16). The credible intervals are larger in the stochastic cases, but overall, the dynamics are similar.

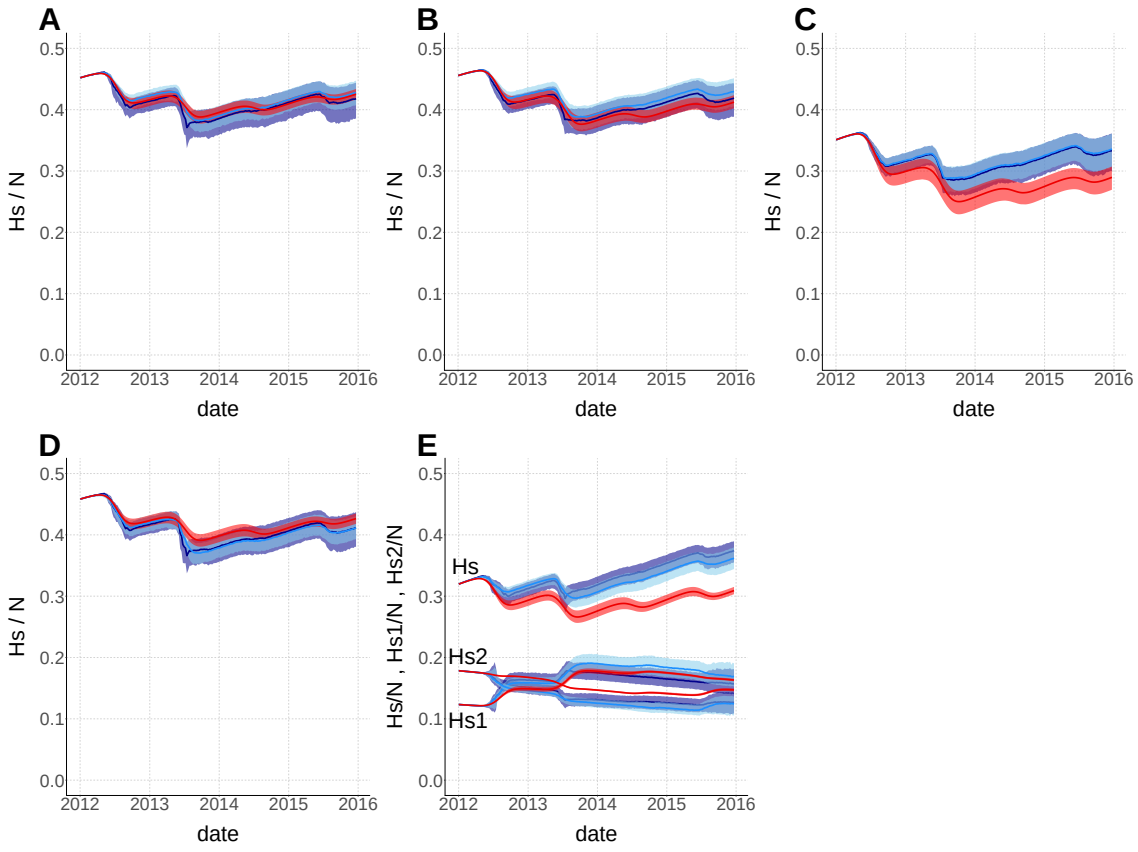


Figure 2.16 – **Proportion of susceptible individuals over time.** Median of the filtered trajectories and credible intervals without observation noise. Red: deterministic case. Light blue: stochastic case with demographic stochasticity only. Dark blue: stochastic case with extra-demographic stochasticity. A. SEIR model. B. Laneri model. C. SEIR-vector model. D. SEIAR model. E. SEIR2 model.

The posterior distributions of the main parameters are displayed in Figure 2.17. In the deterministic framework, in most models, the average annual R_0 is estimated between 2 and 3 and the maximum value between 3 and 4. In the SEIR-vector model, the similar R_0 values correspond to a higher proportion of the population to vaccinate in order to reach herd immunity than in the other models (cf. parameter T_1 in Figure B.4). In the stochastic case with demographic

stochasticity only, the estimated parameters are very close to the deterministic ones (except for the SEIR-vector model), even though the credible intervals are slightly larger. This may be due to the fact that the population size is big enough to limit the influence of demographic stochasticity and that the import parameter prevents from the extinction of the disease. This is not observed in the SEIR-vector model, because the very small absolute number of vectors estimated leads to a large influence of demographic stochasticity. In the stochastic framework with additional environmental stochasticity, the basic reproduction numbers have similar median values, but the credible intervals are much larger. The import parameter is higher (displaying almost a uniform distribution between 0 and 1) which may compensate the increased extinction probability in this highly variable framework. The observation rate differs in the three frameworks, indicating that this parameter is hard to estimate, even though credible intervals overlap in most cases. Nevertheless, despite the large differences in the account for stochasticity in the three framework, the parameters of interest are estimated to similar values.

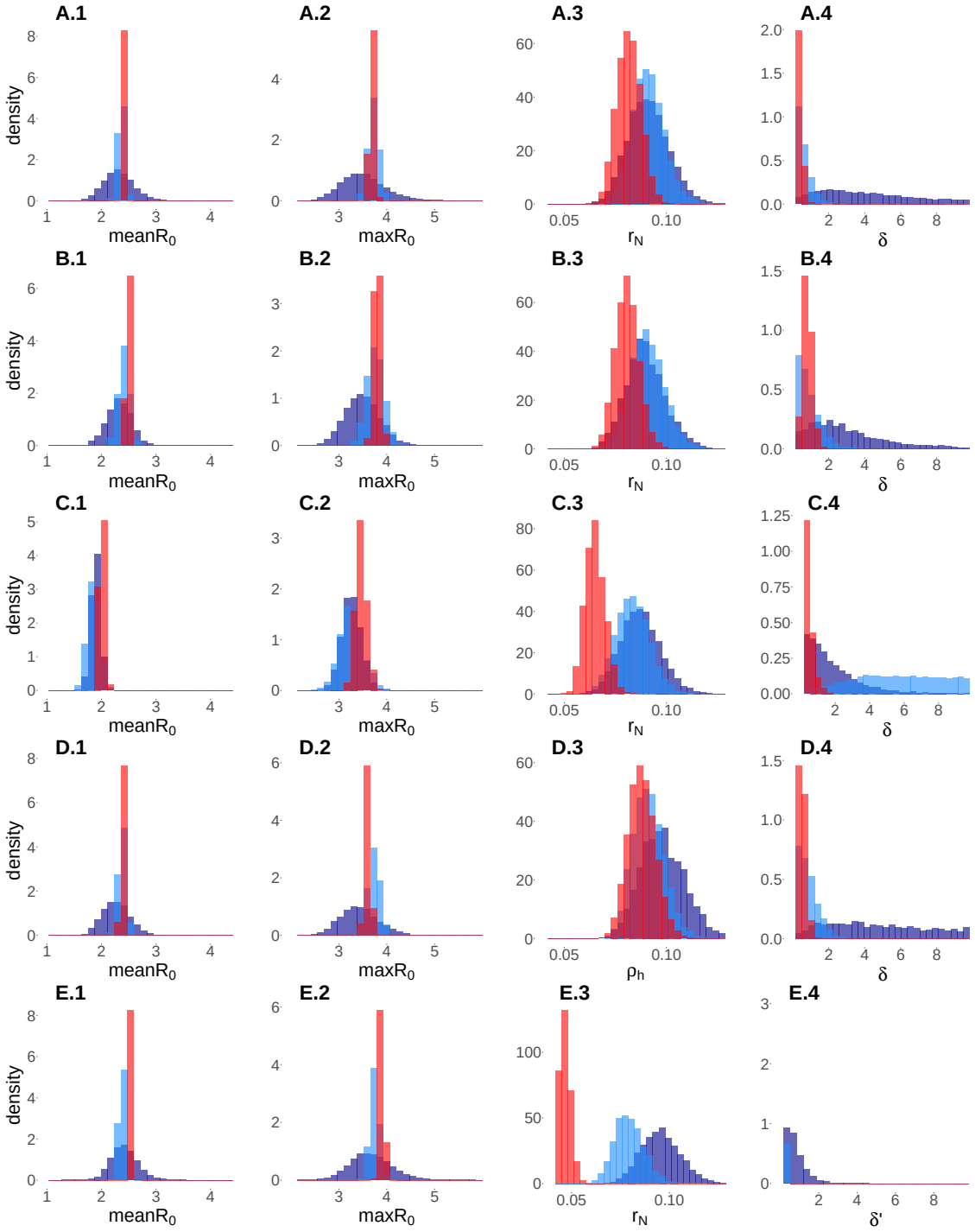


Figure 2.17 – **Posterior distributions.** Posterior distributions in the five models for the mean R_0, the maximum R_0, the observation rate r and the import parameter δ. Red: deterministic case. Light blue: stochastic case with demographic stochasticity only. Dark blue: stochastic case with demographic and environmental stochasticity. Posterior distributions for the other parameters are displayed in Appendix B.4, Figure B.4.A. SEIR model. B. Laneri model. C. SEIR-vector model. D. SEIAR model. E. SEIR2 model.

Discussion

We compared five compartmental models and three frameworks for stochasticity, in order to reproduce dengue data from rural Cambodia. Regarding the differences between compartmental models, the simple ones perform similarly to the more complex ones, indicating that increasing model complexity is not supported by the available information in these four-year datasets. Regarding implementations, stochasticity was introduced in three steps: first we considered only observation noise in a deterministic transmission model, then we allowed for process noise by including demographic stochasticity in the transmission dynamics and finally an environmental white noise was added to the infection reaction (Bretó et al., 2009).

The deterministic framework appears to provide a good approximation of the trajectory, for a low computational cost. The results are not too dissimilar from the ones including demographic stochasticity, maybe because the population is large enough to maintain the epidemic although it is on the lower side of estimated values for the critical community size for dengue (Kuno, 1995; Wearing and Rohani, 2006; Chowell et al., 2008) and because it is not an isolated area and the import parameter prevents extinction events (Bartlett, 1957; Stoddard et al., 2013). Nevertheless, models with vector-borne transmission react more to demographic stochasticity, because low absolute numbers of infected mosquitoes are estimated. This phenomenon is no longer observed when environmental stochasticity is added, so that these low numbers might be a way to compensate the lack of variability of the model by increasing demographic stochasticity, as observed in He et al. (2010) with the reduction of incubation and infectious periods. The observation noise also compensates for the lack of variability in the model, as the overdispersion parameter is estimated to higher values in the absence of process noise: when stochasticity in the propagation model is included and increased, less observation noise is needed to reproduce the data.

However, parameter estimations in the deterministic case may be overconfident, as larger credible intervals are estimated in the stochastic cases, especially with environmental stochasticity. In general, the estimated values of the parameters, especially R_0, are close to previous deterministic estimations for the same setting with a longer time series (Champagne et al. (2018), section 2.2), mainly because the initial conditions were fixed to the same values. This reduction of the parameter space permitted the feasibility of the estimation procedure, but variability may then be underestimated, as uncertainty on the initial state of the system is not taken into account. Nonetheless, despite very different formulations of stochasticity, R_0 is estimated to similar values in all three frameworks, and our estimated R_0 are in the same range as values from the literature (Imai et al., 2015, 2016; Johansson et al., 2011).

Models including demographic and extra-demographic stochasticity better fit the data: Stocks et al. (2018) reached a similar conclusion in their model comparison study across stochasticity frameworks for modelling rotavirus epidemics. In our estimations, due to their increased flexibility, models allowing for both sources of process noise better reproduce extreme observations and pluri-annual variations in incidence. Their associated filtered trajectory can be informative on hidden states of the system such as mosquito compartments, or susceptibles individuals. Yet, when re-simulated, they lead to unpredictable dynamics, with simulations ranging from no outbreak to more than the largest epidemic ever observed in the region. In the framework with extra-demographic stochasticity, the estimated parameter quantifying the amount of environmental noise is very high and the observational over-dispersion very low, so it seems that, with our inference procedure, variability in the data is better explained by process noise than by observation noise,

as far as both sources of noise are properly identified. This large amount of process noise may be an indication of model misspecification (King et al., 2015; He et al., 2010). As increased model complexity did not improve notably the estimations, there may be some non-modeled elements having a strong impact on the dynamic, such as spatial heterogeneity (Lourenço and Recker, 2013), inter-annual climate variations (Lowe et al., 2015), changes in human behaviour (Funk et al., 2010; Verelst et al., 2016), or reporting (Funk et al., 2015). Time-varying parameters can be used to explore these sources of variability (Cazelles and Chau, 1997; Cazelles et al., 2018).

Quality and quantity of data are also important issues. Observing dengue is challenging due to the high proportion of mild infections unseen by the health system and the relative non-specificity of symptoms leading to possible misdiagnosis (Grange et al., 2014) . This feature was taken into account via the overdispersed observation process, and also by combining two datasets with differing collecting strategies (Huy et al., 2010; Duong et al., 2015). In spite of that, some identifiability issues remained, as parameters related to durations of infectiousness, incubation and mosquito lifespan strongly relied on their informative prior distribution. Precise field measures of these quantities are therefore needed for model predictions to be reliable. Moreover, model distinguishability issues (Lee et al., 2017; Walter and Pronzato, 1996) were also observed, as some models of differing complexity performed equally well regarding our datasets, in part because we restricted the analysis to a four-year time series. In addition to case data, entomological or seroprevalence data would be highly interesting to include, as differing sources of information are sometimes difficult to reconcile (e.g. with incidence and seroprevalence data in Kucharski et al. (2016, 2018); Champagne et al. (2016), chapter 4) and may be more effective in discriminating models.

Therefore, all these flexible models with many hidden states require a large amount of data from several sources to be realistically parameterized. In our context, the large uncertainty at large horizons of the stochastic models indicates that they do not seem appropriate for long-term forecasting or for testing counterfactual control scenarios, for which their deterministic counterpart may give in practice a good approximation of the mean dynamic. However, the stochastic frameworks better evaluate uncertainties in parameter estimates and highlight the remaining gaps in understanding and prediction of dengue dynamics.

Acknowledgements

The authors thank the Pasteur Institute in Cambodia for providing the data used in the estimations. The authors have been supported by a grant from Agence Nationale de la Recherche for the PANIC project (Targeting PAthogen's NIChe: a new approach for infectious diseases control in low-income countries: ANR-14-CE02-0015-01). The research leading to these results has also received funding from the European Commission Seventh Framework Program FP7 for the DENFREE project under Grant Agreement 282 378. CC is supported by the PhD studentship of the Groupe des Ecoles Nationales d'Economie et Statistique (GENES). The funders played no role in the study design, data collection, analysis, or preparation of the manuscript.

Chapter 3

Introduire l'aspect spatial

3.1 Introduction

Dengue is caused by a mosquito-borne virus, that circulates in the tropical and subtropical areas of the globe. Modeling dengue is complex due to the multiplicity of factors involved, especially because of the role of mosquito vectors in transmission. The Ross-Macdonald model, initially developed for studying malaria (Smith et al., 2012), is commonly used in order to represent vector-host disease propagation, and it has been frequently applied to dengue (Reiner et al., 2013). However, it has revealed several drawbacks. First of all, it includes many parameters and is sensitive to some of them which are difficult to measure on the field and to identify based on disease data (Luz et al., 2003; Kao and Eisenberg, 2018). Moreover, in some modeling works, it has been outperformed by simpler models that do not include explicit vectors (Pandey et al. (2013), and chapter 2). In addition, it has been observed to generate a higher number of infections than actually observed (Amaku et al. (2016), Champagne et al. (2016), chapter 4). The Ross-Macdonald model relies in particular on the homogeneous mixing assumption, which may be too simplistic to account for the complexity of vector-host encounters (Smith et al., 2014).

To alleviate some of these limitations, Amaku et al. (2016) modify the Ross-Macdonal model by introducing seasonality and spatial structure to reproduce what they call the "dengue oscillator". The epidemics propagates spatially and reaches some areas of a city but is stopped by the unfavorable climatic season for vector-borne transmission; it re-starts the following year, infecting entirely the previously reached areas and spreading to other areas. This may explain the diversity in dengue epidemic patterns observed across cities in Brazil: some cities experience 2 to 3 years of large epidemics followed by 2 to 3 years without epidemics, some experience a single epidemic peak whereas others experience recurrent annual epidemics. By changing the number of sub-populations and the timing between seasonality and disease introduction in their model, Amaku et al. (2016) manage to reproduce qualitatively these different dengue epidemic patterns. The objective of the present chapter is to confront this theory to data from dengue epidemics in Rio de Janeiro, which are available at three spatial scales.

Rio de Janeiro has experienced a resurgence of dengue since the 1980's, with the sequential re-introduction of all four serotypes in 1986, 1990, 2000 and 2010 (Nogueira et al., 2005; Nogueira and Eppinghaus, 2011) leading to devastating epidemics, especially in 2002 and 2008. This mega-city is also characterized by a strong heterogeneity in wealth, infrastructure, urbanization, and land cover, described by Rosa-Freitas et al. (2010) as a "patchy mosaic of urban ecosystems".

Rosa-Freitas et al. (2010) develop a classification of these urban ecosystems, and note that areas with differing characteristic might be equally affected by high dengue incidence. Moreover, results concerning the influence of socio-economic and ecological factors on dengue risk are heterogeneous and contrasted (Teixeira et al., 2013), potentially indicating a role for spatio-temporal dynamics in shaping the city scale herd immunity (Xavier et al., 2017).

Rio de Janeiro is therefore an interesting setting to explore further the theory of Amaku et al. (2016) in the light of data. After a descriptive analysis of the data series in order to identify specific patterns coherent with the theory, the model is confronted to the spatial epidemic data, using MCMC estimation tools.

3.2 Material and Methods

3.2.1 Data

Dengue data

The number of dengue cases in Rio de Janeiro, Brazil are available on Rio's goverment website (SMS, 2016a). In particular, we used weekly data from 2000 to 2010 (consulted on 2017-08-31) and from 2015 to 2017 (consulted on 26 2017-09-26) and monthly data from 2011 to 2014 (consulted on 2017-08-31). The population size for every sub-population are also indicated.

Socio-demographic data

Socio-demographic indicators were available on the `http://www.data.rio` website. In particular, we use the household income (in minimum salaries, based on 2010 census) (data.rio, 2017a), the total number of employees (data.rio, 2017b), the vegetation cover map (PrefeituraRio, 2010) and the administrative map (PrefeituraRio, 2016). From the vegetation cover map, we calculated the mean surface covered with urban areas per administrative region. The distribution of these variables in the area is presented in section 3.4.2.

Climate data

Climatic data from the Climatic Research Unit were used (Harris et al., 2014; Harris and Jones, 2017). In particular, we used the series of average temperature (TMP) and precipitation (PRE), for the grid point associated with the coordinates (lon=-43.25, lat=-22.75), from 2000 to 2003. This data is based on interpolation from other data sources. In order to assess the validity at this spatial scale, we checked to coherence with another dataset for temperature from the Santos Dumont Airport (WeatherUnderground, 2016), (this dataset was not available for the entire studied period and could not be used instead). The comparison of both datasets for temperature is displayed in section 3.4.1.

3.2.2 Model

The model (Amaku et al., 2016)

We rely on the following epidemiological model proposed by Amaku et al. (2016), with some modifications : we assume that the human population is constant, that there is no disease induced mortality and that the duration in latent state for mosquitoes is exponential.

$$\frac{dH_{Si}}{dt} = -abH_{Si}\sum_j \beta_{ij}^H \frac{M_{Ij}}{N_{Hj}} - \mu_H H_{Si} + \mu_H N_{Hi} \tag{3.1}$$

$$\frac{dH_{Ii}}{dt} = abH_{Si}\sum_j \beta_{ij}^H \frac{M_{Ij}}{N_{Hj}} - (\mu_H + \gamma_H)H_{Ii} \tag{3.2}$$

$$\frac{dH_{Ri}}{dt} = \gamma_H H_{Ii} - \mu_H H_{Ri} \tag{3.3}$$

$$\frac{dM_{Si}}{dt} = -ac\frac{M_{Si}}{N_{Hi}}\sum_j \beta_{ij}^M H_{Ij} - \mu_M M_{Si} + \mu_M\,(M_{Si} + M_{Li} + M_{Ii}) \tag{3.4}$$

$$\frac{dM_{Li}}{dt} = ac\frac{M_{Si}}{N_{Hi}}\sum_j \beta_{ij}^M H_{Ij} - \tau M_{Li} - \mu_M M_{Li} \tag{3.5}$$

$$\frac{dM_{Ii}}{dt} = \tau M_{Li} - \mu_M M_{Ii} \tag{3.6}$$

The population is divided into J sub-populations, so that in each sub-population i the individuals can be in the susceptible, infectious or recovered state (respectively H_{Si}, H_{Ii}, H_{Ri}) and the vectors in the susceptible, latent or infected state (respectively M_{Si}, M_{Li}, M_{Ii}). Movements between populations are determined by parameters β_{ij}^H (when a susceptible individual from population i is infected by a mosquito from population j) and β_{ij}^M (when a mosquito from population i is infected by an infected individual from j). The human population size is constant in each area ($H_{Si}+H_{Ii}+H_{Ri} = N_{Hi}$) and the mosquito population size is assumed to follow seasonal variations, based on a cosine forcing. The other model parameters are detailed in Table 3.1.

Seasonality

In order to model the influence of seasonal climate on the disease dynamics, Amaku et al. (2016) introduce a cosine forcing on the mosquito recruitment rate, so that, with our notations, equation (3.4) becomes:

$$\frac{dM_{Si}}{dt} = -ac\frac{M_{Si}}{N_{Hi}}\sum_j \beta_{ij}^M H_{Ij} - \mu_M M_{Si} + \mu_M\,(M_{Si} + M_{Li} + M_{Ii}) * \left(1 + S\,cos\left[2\pi\left(\frac{t}{365} + p\right)\right]\right)$$

where S and p are the amplitude and phase of the cosine forcing.

We also consider an alternative formulation of seasonality that includes local data on climate. In this version of the model (called climate model 1), seasonality in the mosquito population is not driven by a cosine function in the recruitment rate but by a function of standardized average temperature (T) and precipitation (P):

$$\frac{dM_{Si}}{dt} = -ac\frac{M_{Si}}{N_{Hi}}\sum_j \beta_{ij}^M H_{Ij} - \mu_M M_{Si} + V\,N_{Hi}\,(1 + \alpha_t\,T + \alpha_p\,P)^+$$

connection where V, α_t and α_p are constants to be estimated, and we note $(x)^+ = \max(x, 0)$. Another formulation is also explored (called climate model 2), in which the mosquito infection rate is seasonal instead of the mosquito recruitment rate :

$$\frac{dM_{Si}}{dt} = -ac(1 + \alpha_t\,T + \alpha_p\,P)^+\frac{M_{Si}}{N_{Hi}}\sum_j \beta_{ij}^M H_{Ij} - \mu_M M_{Si} + \mu_M\,(M_{Si} + M_{Li} + M_{Ii})$$

where α_t and α_p are constants to be estimated (equation (3.5) is also modified accordingly).

Spatial structure

The spatial structure is considered regarding three separate effects. Firstly, the sub-populations might be heterogeneous in their transmission potential, because they do not have the same mosquito population size. Secondly, sub-populations are connected according to specific patterns (people do move between sub-populations, but not homogeneously). Thirdly, the disease is not introduced uniformly in all sub-populations.

Heterogeneity among sub-populations

We assume that the initial population of mosquitoes is fully susceptible and equal to $M_{Si}(0) = m_i \times N_{Hi}$, so that m_i quantifies the abundance of mosquitoes per human in each sub-population. In a model with J sub-populations, there are therefore J unknown parameters m_i for which several assumptions are explored:

- **Homogeneous mosquito density:** $\forall i, m_i = m$. The abundance of mosquitoes relative to human population size is the same in all sub-populations.
- **Heterogeneous mosquito density:** $\forall i, m_i$. The abundance of mosquitoes relative to human population size is specific to each sub-population.
- **Heterogeneity with covariates:** The abundance of mosquitoes relative to human population size in area i depends linearly on some characteristics of sub-population i. In particular, we considered $\forall i, m_i = m + c_x X_i + c_u U_i$, where X_i and U_i are respectively the average income and the proportion of urbanized areas in sub-population i.

Connections between sub-populations

Regarding the connections between locations, in a model with J sub-populations, there are $2 * (J-1)^2$ unknown parameters β_{ij}^H and β_{ij}^M, that cannot be all estimated together given the type of available data. Therefore, we introduce some assumptions on the structure of movements. We first assume that vectors do not move between patches, due to their short flying range. Regarding hosts, we assume that in each sub-population i, a proportion $1 - \eta_i$ stays in the sub-population and a proportion η_i travels to other sub-populations, and that infectious individuals have the same behaviour as the others. In consequence, we consider that:

(i) the movement of susceptible individuals is described by parameters β_{ij}^H such that $\sum_{j \neq i} \beta_{ij}^H = \eta_i$ and $\beta_{ii}^H = 1 - \eta_i = 1 - \sum_{j \neq i} \beta_{ij}^H$.

(ii) the infection of vectors by infected hosts from different areas is described by parameters β_{ij}^M. Because we assume that vectors do not move between areas, this occurs because of the movement of infected hosts. Therefore, vectors from population i can be infected either by individuals from population i ($\beta_{ii}^M H_{Ii}$) or by incoming individuals from other populations ($\beta_{ij}^M H_{Ij}$, $j \neq i$). Because we also assume that susceptible and infectious individuals have the same mobility patterns, there are $\beta_{ii}^H H_{Ii}$ infectious individuals staying in i and $\beta_{ji}^H H_{Ij}$ infectious individuals coming from j. Therefore, we assume that $\beta_{ii}^M = \beta_{ii}^H$ and $\beta_{ij}^M = \beta_{ji}^H$. In this case, we may have $\sum_j \beta_{ij}^M \neq 1$.

Quantifying human mobility patterns is a difficult task in the absence of specific mobility data, however, several parametric models for connections between areas exist.

The gravity model was introduced in epidemiology by Xia et al. (2004): the term "gravity" refers to the analogy with Newton's law of gravitation, because it is assumed that the intensity of contacts between two locations depends on the product of their population sizes divided by the squared distance between them, or, more generally, on these three quantities (the two population sizes and distance) with specific exponents. This parametric structure, which simplifies greatly the complexity of the contact matrix, has proven powerful to reproduce the dynamics of measles in the United Kingdom (Xia et al., 2004), and influenza in the United States (Viboud et al., 2006). In the latter, data on workflows are used to parameterize the exponent parameters of the models via regression tools and the obtained gravity model is then applied to simulate epidemic scenarios at the country scale (Viboud et al., 2006). The gravity model has sometimes been extended by including other variables in addition to population sizes, such as the population who work in a certain area (Truscott and Ferguson, 2012), vegetation coverage or indicators of human exposure risk (Barrios et al., 2012), or economic indices (Stefanouli and Polyzos, 2017).
Another parametric framework for contact structure was introduced by Simini et al. (2012) and called the radiation model. In this model, the flow between two locations depends positively on their population sizes, but, unlike the gravity model, it also depends negatively on the population in a radius around the source location, reflecting the fact that people tend to move on longer distances when there are fewer opportunities near the area where they live. Recently, Sallah et al. (2017) proposed another parameter-free model, the impedance model, and applied it to simulated and real epidemic data.
Whereas many works used these parametric models at the country or the global scales, some also studied the city scale. Masucci et al. (2013) observed that both the gravity and the radiation models perform badly to reproduce mobility patterns at the city scale. They emphasize the fact that, at this spatial scale, employment areas are not well approximated by residential areas, potentially introducing a bias if the models are parameterized using population sizes. Yan et al. (2014) proposed another parameter-free model, the population weighted opportunity model (PWO), inspired by the radiation model but more adapted to the city scale, in which individuals are more prone to travel long distances and are more mobile. Their work emphasize the capacity of the PWO model to reflect the mobility patterns in four cities (in China, Ivory Coast and the USA), especially when only population data is available and the calibration of additional parameters is not possible.

In this work, we explored three hypothesis on β^H_{ij}:

- **Independent populations:** $\forall i, j, \beta^H_{ij} = 0$. All sub-populations are independent.

- **Gravity model:** $\forall i \neq j, \beta^H_{ij} = \beta \frac{N_i^{\tau_1} N_j^{\tau_2}}{d_{ij}^{\rho}} \in [0; 1]$ and $\forall i, \beta^H_{ii} = 1 - \sum_{j \neq i} \beta^H_{ij}$. The flux from i to j depends positively on the population sizes of i and j ($\tau_1 \geqslant 0$,$\tau_2 \geqslant 0$) and negatively on the distance d_{ij} between i and$j$ ($\rho \geqslant 0$). β is considered as a normalizing constant.
  In order to limit the number of parameters to estimate, we assume that $\tau_1 = 0$, indicating that the movement depends on the attractiveness of the destination population and not on the source population.

- **Employee-based connections:** $\forall i \neq j, \beta^H_{ij} = \beta \frac{N_i^{\tau_1} N_j^{\tau_2} R_j^{\delta}}{d_{ij}^{\rho}} \in [0; 1]$ and $\forall i, \beta^H_{ii} = 1 - \sum_{j \neq i} \beta^H_{ij}$. The flux from i to j depends also positively on some attributes R_j of j reflecting

attractiveness of the destination ($\delta \geqslant 0$): as an indicator, we take the number of employees in the area. β is considered as a normalizing constant.
For parsimony in the number of parameters, we assume that $\tau_2 = \tau_1 = 0$, indicating that the movement depends on the attractiveness of the destination sub-population based on the employment availability and not on the source sub-population.

In order to ensure that $\sum_{j \neq i} \beta_{ij}^H < 1$, we fix $\beta = \frac{\min_{i \neq j}(d_{ij})^\rho}{(J-1)\max_i(N_i)^{\tau_1+\tau_2}\max_i(R_i)^\delta}$.

Disease introduction

We explore two hypotheses for disease introduction:

- **Homogeneous introduction:** $\forall i, H_{Ii}(0) = h_I(0) \times N_i$ and $\forall i, M_{Ii}(0) = 0$. The disease is introduced homogeneously in all sub-populations.
- **Heterogeneous introduction:** $\forall i, H_{Ii}(0) = h_{Ii}(0) \times N_i$ and $\forall i, M_{Ii}(0) = 0$. The initial number of infected individuals is specific to each sub-population.

Parameter		Prior	Reference
a	mosquito biting rate	1 day^{-1}	Amaku et al. (2016)
b	probability infection from mosquito to human	0.6	Amaku et al. (2016)
c	probability infection from human to mosquito	0.5	Amaku et al. (2016)
γ_H^{-1}	infectious period (human)	7 days	Amaku et al. (2016)
τ^{-1}	EIP	7 days	Amaku et al. (2016)
μ_M^{-1}	mosquito lifespan	21 days	Amaku et al. (2016)
μ_H^{-1}	human lifespan	70 years	Amaku et al. (2016)
S	cosine seasonality amplitude	$N(0.47, 0.02^2)$ in [0,1]	Preliminary estimation on aggregated data
p	cosine seasonality phase	$N(0.28, 0.006^2)$ in [0,0.5]	Preliminary estimation on aggregated data
α_t	temperature coeff. (model 1)	$N(0.26, 0.05^2)$ in [0,1]	Preliminary estimation on aggregated data
α_p	precipitation coeff. (model 1)	$N(0.44, 0.05^2)$ in [0,1]	Preliminary estimation on aggregated data
V	scaling coeff. (model 1)	$N(2.3 \times 10^{-5}, (3 \times 10^{-5})^2)$ in $[10^{-6}, 10^{-2}]$	Preliminary estimation on aggregated data
α_t	temperature coeff. (model 2)	$N(0.40, 0.05^2)$ in [0,1]	Preliminary estimation on aggregated data
α_p	precipitation coeff. (model 2)	$N(0.35, 0.05^2)$ in [0,1]	Preliminary estimation on aggregated data
m or m_i	mosquitoes per human	U[0.0001;0.25]	
$H_{Ri}(0)$	initial number of recovered hosts	0	
$H_{Ii}(0)$	initial number of infected hosts	$U[10^{-6}; 10^{-3}] * N_i$	
$M_{Si}(0)$	initial number of susceptible vectors	$m_i \times N_{Hi}$	
$M_{Li}(0)$	initial number of latent vectors	0	
$M_{Ii}(0)$	initial number of infected vectors	0	
ρ	gravity coeff. (distance)	U[0;5]	
τ_1	gravity coeff. (source population)	0	
τ_2	gravity coeff. (destination population, gravity model)	U[0;1]	
τ_2	gravity coeff. (destination population, employee model)	0	
δ	gravity coeff. (employee)	U[0;1]	
c_x	coefficient for income	U[-0.1;0.1]	
c_u	coefficient for urbanization	U[-0.1;0.1]	
r	observation rate	U[0;1]	
ϕ	overdispersion	U[0;10]	

Table 3.1 – **Parameters and priors**

3.2.3 Estimation method

The number of cases observed during week k in sub-population i, noted $C_{k,i}^{\text{obs}}$, is assumed to be drawn from a negative binomial distribution with mean $rC_{k,i}$ and variance $rC_{k,i} + \phi(rC_{k,i})^2$, where

$C_{k,i}$ is the number of new cases simulated by the model during week k in sub-population i, r is an observation rate and ϕ is an overdispersion parameter (Bretó et al., 2009). We note $p_{i,k}\left(C_{k,i}^{\text{obs}}\right)$ its density. The observation in differing weeks $k = 1...K$ and sub-populations $i = 1...J$ are assumed independent conditional on the underlying disease process, and therefore the total likelihood is the product $L = \prod_{i=1}^{J}\prod_{k=1}^{K} p_{i,k}\left(C_{k,i}^{\text{obs}}\right)$.

Estimations of the deterministic models are performed with Markov Chain Monte Carlo (MCMC), with random walk Metropolis-Hastings algorithm, using SSM software (Dureau et al., 2013a). The covariance of the proposal distribution is initially adapted as specified in (Dureau et al., 2013a). In a preliminary initialization phase, the parameter space is explored by running a simplex algorithm on parameter sets sampled via latin hypercube sampling (using the *lhs* package (Carnell, 2016)): the parameter set with the highest posterior is used as initial point for the MCMC chain. In the model with employee-based connections and heterogeneity in mosquito populations, due to the large dimension of the parameter set, the MCMC chain was initialized to the best fitting parameters for employee-based connections and for heterogeneity in mosquito populations. Prior distributions are specified in Table 3.1.
In order to restrict the parameter space and improve convergence, we use informative priors for the seasonality parameters. We first estimated a simple model (J=1) on the aggregated data. The empirical mean and standard deviation of the marginal posterior distribution for seasonality parameters is used to calibrate the truncated gaussian prior used in the spatial models.

The maximum a posteriori and 95% credible intervals of model parameters based on the MCMC chain are displayed. In addition, we calculate the seroprevalence rate as the proportion of individuals in the H_R compartment at the end of the period, in a deterministic simulation using the parameter set with the highest posterior. We compute the total seroprevalence $H_R/N = \sum_{i=1}^{J} H_{Ri} / \sum_{i=1}^{J} N_i$ and of the seroprevalence per sub-population H_{R_i}/N_i. In models including connections, we also compute the proportion of individuals travelling out of each sub-population i, $\eta_i = \sum_{j\neq i} \beta_{ij}^{H}$, using parameters ρ, τ_2 and δ associated with the highest posterior value; we then compute the median, minimum and maximum of this indicator among the considered sub-populations. Models are also compared using the DIC criterion (Spiegelhalter et al., 2002), the best model being the one with the smallest DIC.

3.3 Results

3.3.1 Descriptive data analysis

Disease dynamics at the city scale

The dynamics of dengue incidence over the period 2000-2017 are presented in Figure 3.1. Epidemics are highly seasonal, and there were four separate outbreaks over the period, in 2001-2002, 2006-2008, 2011-2013 and 2015-2016.

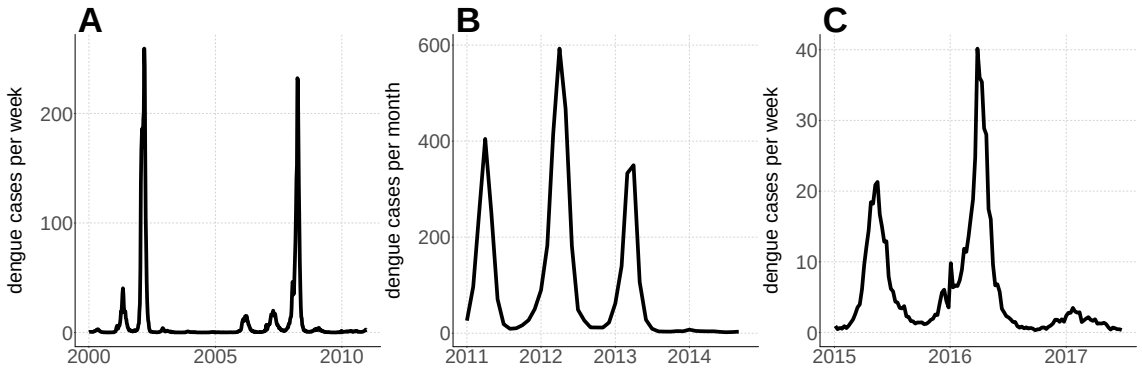


Figure 3.1 – **Dengue incidence in Rio de Janeiro (number of cases per unit of time, per 100,000 inh.).** 2000 to 2010 data are based on 2000 Census. 2014 to 2017 data are based on 2010 Census. **A.** Dengue incidence per week from 2000 to 2010. **B.** Dengue incidence per month from 2011 to 2014. **C.** Dengue incidence per week from 2015 to 2017. (SMS, 2016a)

The first two epidemics (2001-2002 and 2006-2008) follow similar patterns, in which small peaks are followed by a very large outbreak. It is considered that these outbreaks were dominated by serotype DENV-3 and DENV-2 respectively (Fares et al., 2015). Serotype DENV-3 was first detected in the Rio de Janeiro state in December 2000 (De Simone et al., 2004), and serotypes DENV-1, DENV-2 and DENV-3 were co-circulating between 2000 and 2001. However, it seems that the large 2002 epidemic was mainly due to the newly introduced DENV-3 serotype (Nogueira et al., 2005). The 2007-2008 epidemic is mainly attributed to DENV-2 (Fares et al., 2015; Rosa-Freitas et al., 2010). There is a different pattern in 2011-2013 with three epidemics of similar magnitude. DENV-4 was introduced in Rio de Janeiro in 2010 (Nogueira and Eppinghaus, 2011), and co-circulated with other serotypes during the following outbreaks (DENV-4 represented 44% of cases sampled between 2011 and 2013 by Heringer et al. (2017)). By 2015 the most frequent serotypes were DENV-1 and DENV-4 (Fares et al., 2015). In 2015-2016, the magnitude of epidemics is much lower. It is important to note that, by the end of the period, there may be a larger immunity in the population since all serotypes have already been introduced. For instance, Honório et al. (2009) reported age-standardized seroprevalence proportions above 50% in three neighbourhoods sampled between 2007 and 2008. Seroprevalence studies in others cities in Brazil also report high seroprevalence rates (Braga et al., 2010), and Rodriguez-Barraquer et al. (2011) hypothesize that, in the 2000's, the country may be experiencing a shift in transmission from re-emergence to hyperendemicity.
In addition, other arboviruses have circulated in the region in the second decade: the chikungunya virus was introduced in Brazil in 2014 (Nunes et al., 2015) and the Zika virus caused a large epidemic in Rio de Janeiro in 2015-2016. Reporting of Zika cases became mandatory in November 2015 (Villela et al., 2017), but retrospective analysis show evidence of circulation as early as 2013 (Passos et al., 2017). Due to similarities in clinical manifestations and serological cross-reactions in diagnostic tests between arboviruses, these co-occurrences and changes in reporting might also affect dengue reporting after 2013.

Disease dynamics at several spatial scales

Surveillance data are available at three regional scales: 10 planification areas (Areas de Planejamento, AP, cf. Figure 3.2A), 33 administrative regions (AR), and 167 neighbourhoods.

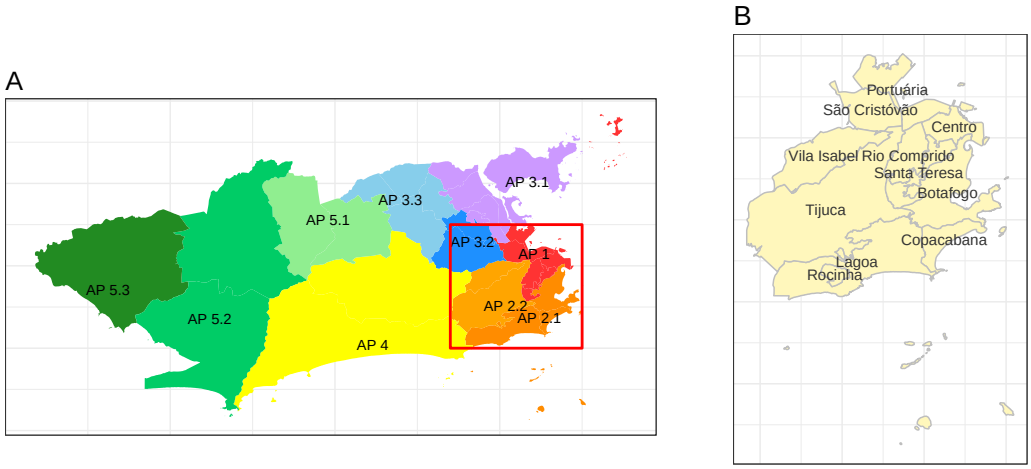


Figure 3.2 – **Administrative divisions in the city of Rio de Janeiro. A.** Planification areas, the red rectangle refers to panel B. **B.** Sub-area of Rio de Janeiro for estimations (refers to the red rectangle in panel A).

Figure 3.3 displays the incidence curves in the 10 AP. There is a very strong synchronization of dengue epidemics among the AP in 2000-2002 and to a lesser extent in 2006-2008. On the contrary, in 2011-2013 and 2015-2016, the patterns are more heterogeneous: for instance, in 2011-2013, the largest peak is the first one in AP5.2, the second one in AP5.1 and the last one AP2.1 and AP2.2, whereas in AP1.0, AP4.0 and AP5.3 three small peaks of similar magnitude are observed.

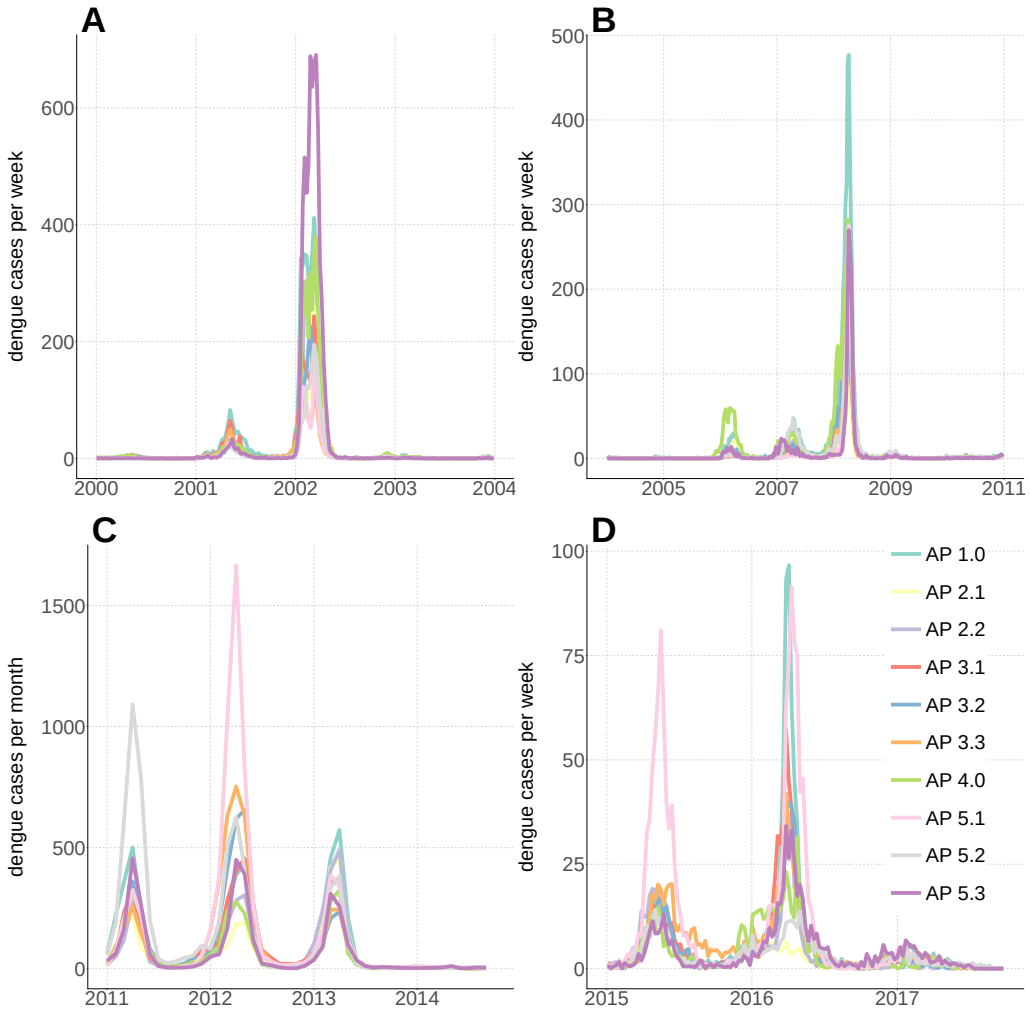


Figure 3.3 – **Dengue incidence in Rio de Janeiro per planification areas (number of cases per unit of time, per 100,000 inh.).** 2000 to 2010 data are based on 2000 Census. 2014 to 2017 data are based on 2010 Census. **A.** Dengue incidence per week from 2000 to 2004. **B.** Dengue incidence per week from 2005 to 2010. **C.** Dengue incidence per month from 2011 to 2014. **D.** Dengue incidence per week from 2015 to 2017. (SMS, 2016a)

As an example of dynamics at a finer scale, we focus on AP2.1 and AP2.2 (cf. Figure 3.4) and AP3.1 (cf. Figure 3.5) in 2000-2003. When AP2.1 and AP2.2 are considered together (cf. Figure 3.4A), the 2002 outbreak is bimodal, but the two sub-areas experience a single peak with a small delay. This pattern is also observed when focusing on two administrative regions of comparable size (Lagoa and Vila Isabel, cf. Figure 3.4B) and it is coherent with an epidemic travelling from one area to the next. However, in 2001, the dynamics in both regions are much more synchronized, indicating that the disease was introduced in both areas before 2002. Within the Tijuca administrative region, the neighbourhood of Alto da Boa Vista also displays a delayed epidemic in 2002 (cf. Figure 3.4C).

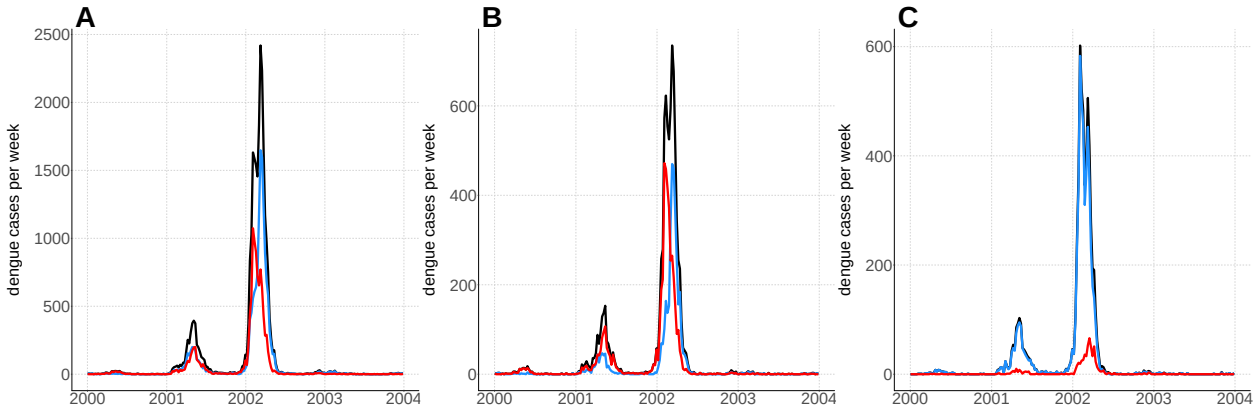


Figure 3.4 – **Number of dengue cases per week from 2000 to 2004 in AP2.1 and AP2.2. A.** Dengue cases in AP2.1 (blue line), in AP2.2 (red line) and in the sum of both (black line). **B.** Dengue cases in Lagoa administrative region (blue line), in Vila Isabel administrative region (red line) and in the sum of both (black line). **C.** Dengue cases in Tijuca and Praça da Bandeira neighbourhoods (blue line), in Alto da Boa Vista neighbourhood (red line) and in the sum of both (black line).(SMS, 2016a)

On the contrary, in AP3.1 (cf. Figure 3.5), the bimodal outbreak in 2002 does not seem to be due to delay between spatial areas, since this pattern is also observed at the regional scale (in the Ramos administrative region), and the neighbourhood scale (three neighbourhoods within the Ramos administrative region).

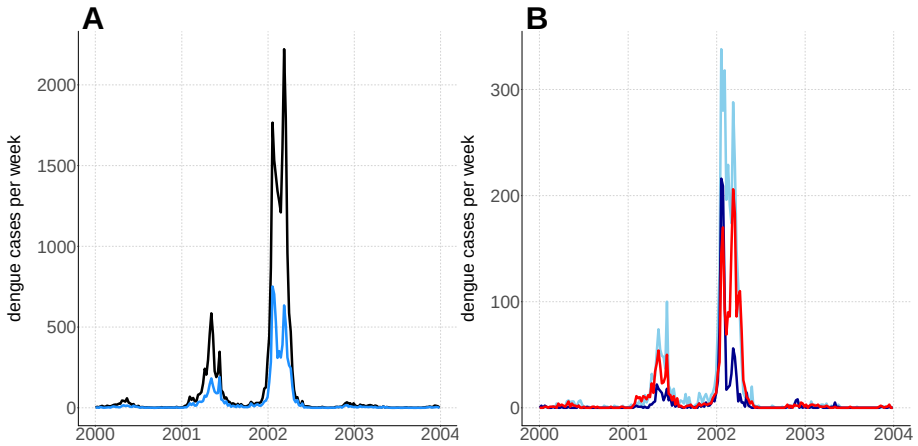


Figure 3.5 – **Number of dengue cases per week from 2000 to 2004 in AP3.1. A.** Dengue cases in AP3.1 (black line), and in the Ramos administrative region (blue line). **B.** Dengue cases in Ramos neighbourhood (red line), in Bonsucesso neighbourhood (light blue line) and in Manguinhos neighbourhood (dark blue line). (SMS, 2016a)

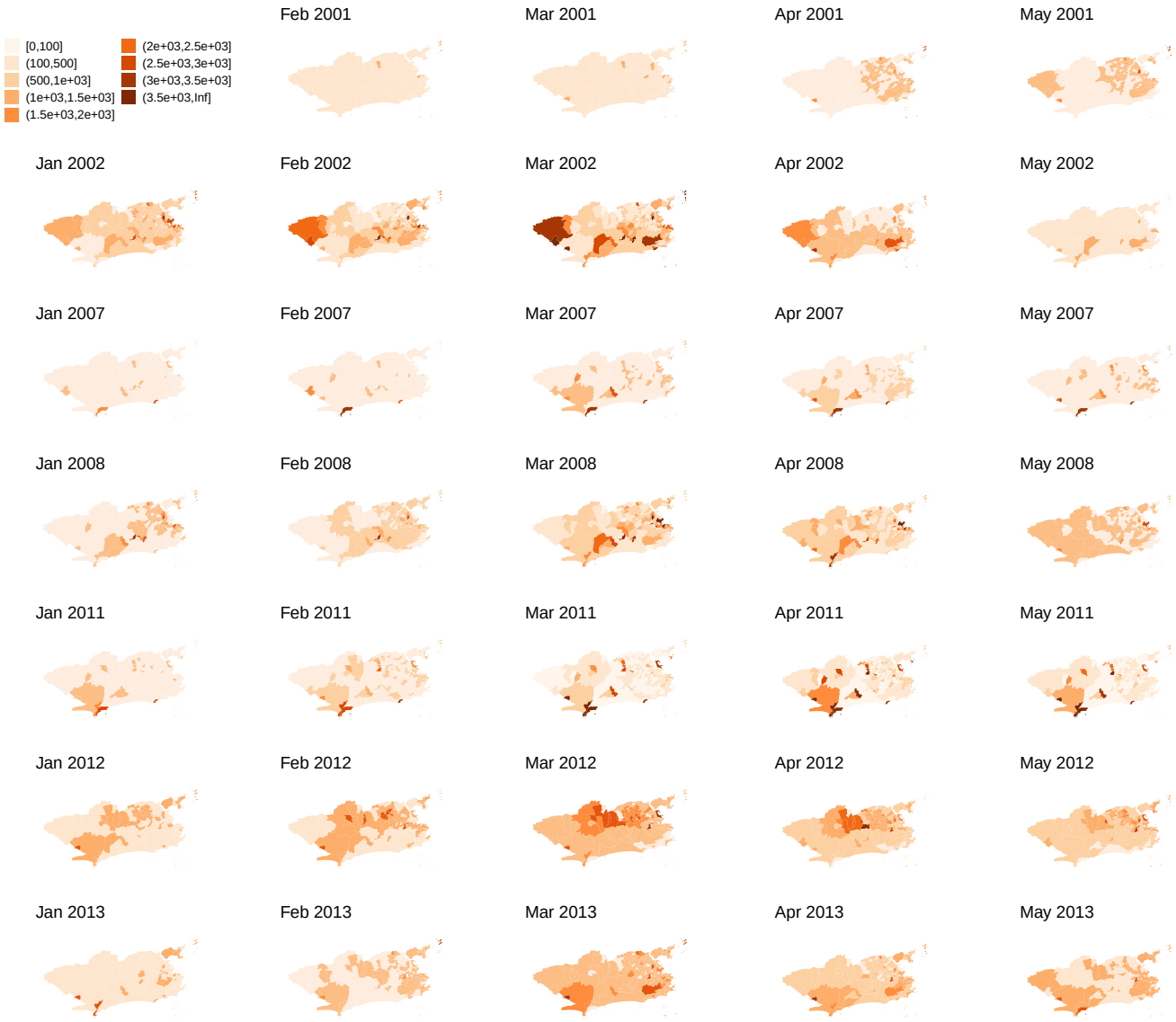


Figure 3.6 – **Dengue incidence per month per neighbourhood in Rio de Janeiro (number of cases per month, per 100,000 inh.)** Only epidemic years and the preceding year are displayed. (SMS, 2016a)

The incidence per neighbourhood (cf. Figure 3.6) is highly heterogeneous across years and areas. For instance, the north-easter part of the city is affected by the end of the 2001 season, and the epidemic spreads to the rest of the city the following year. In 2007-2008, the south-western part seems to be affected before the other areas. In the third epidemic wave between 2011 and 2013, as observed by Xavier et al. (2017), the most affected regions are not the same from one year to the next, in particular the southern neighbourhoods are more affected in 2013, whereas the northern neighbourhoods are more affected in 2012. The fact that some regions are more affected during some epidemics and others during other epidemics is coherent with a spreading of the epidemic from one patch to the next. Large epidemics usually start earlier in the year, whereas small ones

preceding a large outbreak occur later in the season (for example, in 2002 as opposed to 2001), indicating a potentially crucial role for seasonality as in the framework by Amaku et al. (2016).

3.3.2 Simulations of the model

We then performed some simulations of the model with cosine seasonal forcing to explore its behaviour with differing parameter sets. As indicated in Amaku et al. (2016), the framework is very flexible and a wide range of trajectories can be simulated (cf. Figures 3.7 and 3.8). With two sub-populations that share the same mosquito parameters, by varying the intensity of transmission and the strength and asymmetry of connections, epidemic patterns can differ strongly (cf. Figure 3.7).

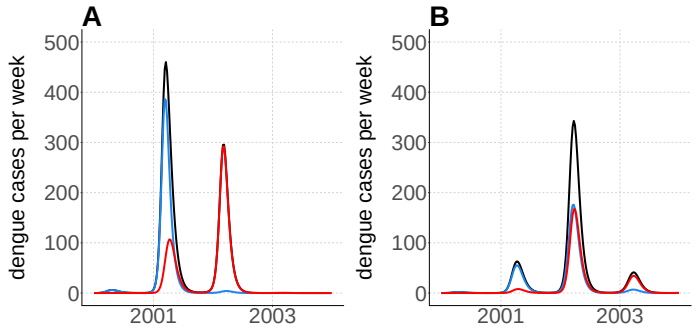


Figure 3.7 – **Examples of deterministic simulations of the model with 2 sub-populations.** Median number of observed dengue cases per week. Total population: black. Sub-population 1: blue. Sub-population 2: red. $N_1 = N_2 = 180,000$, $H_{I1}(0) = 1$, $H_{I2}(0) = 0$, $r = 0.033$, $p = 0.21$, $S = 0.59$. Other parameters are as stated in Table 3.1. **A.** $\beta_{12} = \beta_{21} = 0.002$, $m_1 = m_2 = 0.18$. **B.** $\beta_{12} = 0.006$, $\beta_{21} = 0$, $m_1 = m_2 = 0.15$

We also simulate the model with nine sub-populations of equal size, considering two simple hypothesis on the connections between them: connections along a chain ($\beta_{i,i+1} = 0.0002$, $\forall j \neq i+1, j \neq i \ \beta_{i,j} = 0$, cf. Figure 3.8A) and equal connections ($\forall j \neq i \ \beta_{i,j} = 0.0002$, cf. Figures 3.8C). In the chained model, the epidemic affects the sub-regions one after the other, creating successive peaks, in a pattern that depends on the intensity of connections. However, this model is less credible at this spatial scale in a city (all sub-regions are more connected) and less coherent with the data (the epidemics last 2 to 3 years). In the equally connected model, a difference appears between the first and all the other populations (the intensity of connections influences the speed at which the other patches are infected). The subsequent patches are very synchronized. By introducing heterogeneity in population sizes (cf. Figures 3.8B and D), the magnitude of the subsequent peaks can be modulated.

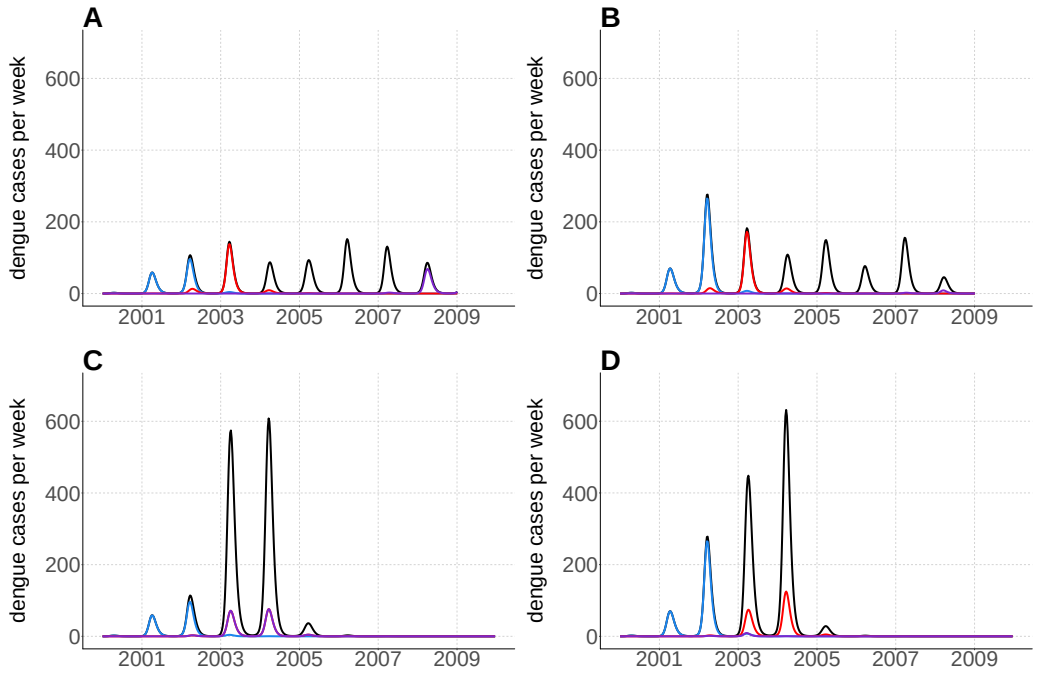


Figure 3.8 – **Examples of deterministic simulations of the model with 9 sub-populations.** Median number of observed dengue cases per week. Total population: black. Sub-population 1: blue. Sub-population 2: red. Sub-population 6: purple. Populations 4, 5, 7, 8 and 9 are not represented. $r = 0.033$, $\forall i, m_i = 0.15$, $p = 0.21$, $S = 0.59$. Other parameters are as stated in Table 3.1.
A. Chained populations. $H_{I1}(0) = 1$, $\forall i > 1$, $H_{Ii}(0) = 0$. $\forall i$, $N_i = 125,000$, $\beta_{i,i+1} = 0.0002$, $\forall j \neq i+1, j \neq i$ $\beta_{i,j} = 0$.
B. Chained populations, heterogeneous population sizes. $H_{I1}(0) = 1$, $\forall i > 1$, $H_{Ii}(0) = 0$. $\forall i$, $\beta_{i,i+1} = 0.0002$, $\forall j \neq i+1, j \neq i$ $\beta_{i,j} = 0$. $\sum_i N_i = 9 \times 125,000$, $N_1 = 250,000$, $N_2 = 160000$, $N_3 = 180,000$, $N_4 = 70,000$, $N_5 = 170,000$, $N_6 = 8000$, $N_7 = 180000$, $N_8 = 72000$, $N_9 = 35000$.
C. Equally connected populations. $H_{I1}(0) = 1$, for $i > 1$, $H_{Ii}(0) = 0$. $\forall i$, $N_i = 125,000$, $\forall j \neq i$ $\beta_{i,j} = 0.0002$.
D. Equally connected populations, heterogeneous population sizes. $H_{I1}(0) = 1$, $\forall i > 1$, $H_{Ii}(0) = 0$. $\forall i$, $N_i = 125,000$, $\forall j \neq i$ $\beta_{i,j} = 0.0002$. $\sum_i N_i = 9 \times 125,000$, $N_1 = 250,000$, $N_2 = 160000$, $N_3 = 180,000$, $N_4 = 70,000$, $N_5 = 170,000$, $N_6 = 8000$, $N_7 = 180000$, $N_8 = 72000$, $N_9 = 35000$.

3.3.3 Parameter estimations from data

In order to confront the model to the data, we focus on the years 2000 to 2003, because we can make the simplifying assumption that the epidemics are dominated by a single serotype newly introduced in the city, and that the population is entirely susceptible at the beginning of the year 2000. Moreover, in order to limit the model size, we focus on a sub-area of the city (cf. Figure 3.2B), comprising the South-East and Downtown AP (administrative regions of Botafogo, Copacabana, Lagoa, Rocinha, Tijuca, Vila Isabel, Centro, Portuaria, Santa Teresa, San Cristovao, Rio Comprido; we subdivide the Tijuca AR by separating the Alto da Boa Vista neighborhood, which is covered by a forest). This sub-area of the city was chosen because it is close to the one considered in Stolerman et al. (2015), and it displayed interesting patterns in the data analysis phase (cf. Figure 3.4).

Seasonality

We first fitted a spatial model with independent subpopulations, homogeneous mosquito populations, and homogeneous disease introduction. This model can serve as a benchmark for comparison of the differing spatial models and it can also be used to explore the three implementations of seasonality. Using the model with a cosine forcing (cf. Figure 3.9A), the 2002 peak is well reproduced by the model, but the 2001 peak is predicted much earlier than observed in the data. We hypothesize that this effect could be triggered by climatic factors, the epidemic season being delayed this particular year. Indeed, the population of susceptible mosquitoes simulated by the model is in phase with the average temperature and precipitation in the region (cf. Figure 3.9D). The timing of the 2001 epidemic is better captured by the two models including a climatic component (cf. Figure 3.9B and C), and indeed, the growth in the susceptible mosquito population is slightly delayed in 2001, when climate is accounted for (cf. Figure 3.9E).

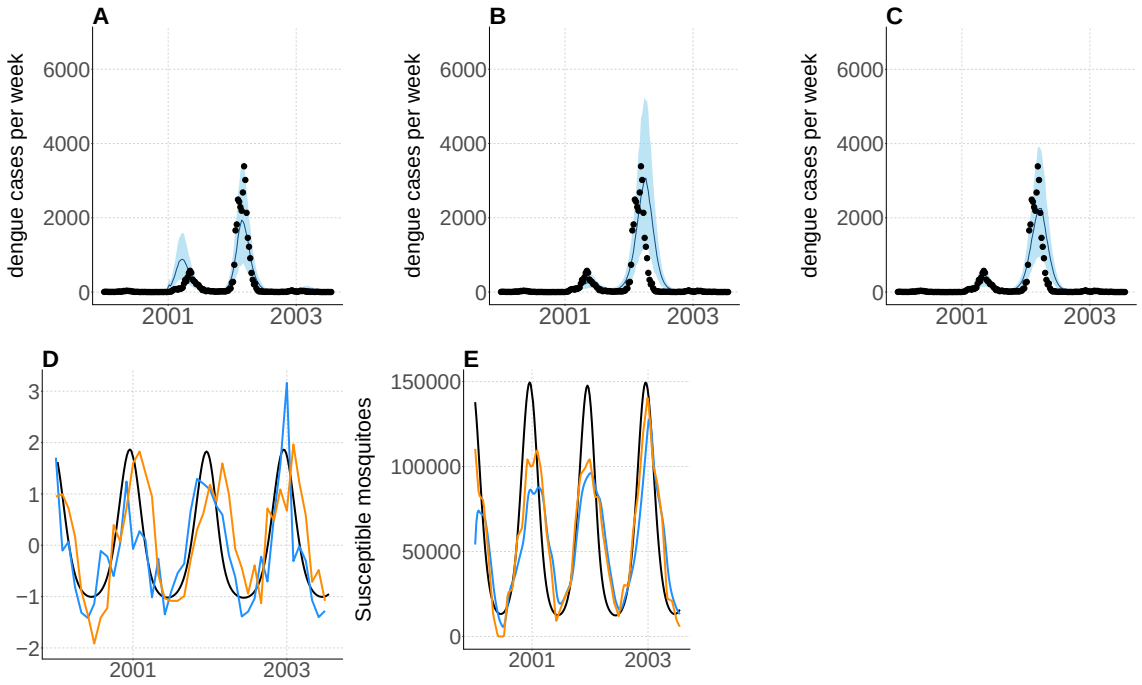


Figure 3.9 – **Comparing differing seasonality models. A to C. Total number of observed cases per week** Models with independent sub-populations, homogeneous mosquito populations, and homogeneous disease introduction. Median and 95% credible intervals based on simulations from the posterior distribution (plain line and shaded areas) and observations (black dots). **A.** Seasonality based on cosine function. **B.** Seasonality based on climate (model 1). **C.** Seasonality based on climate (model 2). **D. Mosquito dynamics in the cosine model and climate data.** Black line: standardized total number of susceptible mosquitoes in the cosine model, $\sum_i M_{Si}$ (cf. panel A), in a deterministic simulation based on the maximum a posteriori parameter set. Orange line: Standardized monthly average temperature. Blue line: Standardized monthly precipitation. **E. Susceptible mosquito populations in the three models.** Deterministic simulation based on the maximum a posteriori parameter set. Black line: total absolute number of susceptible mosquitoes in the cosine model, $\sum_i M_{Si}$ (cf. panel A). Blue line: total absolute number of susceptible mosquitoes in the climate 1 model, $\sum_i M_{Si}$ (cf. panel B). Orange line: total absolute number of susceptible mosquitoes multiplied by the seasonal term in the climate 2 model, $(1 + \alpha_t T + \alpha_p P)^+ \sum_i M_{Si}$ (cf. panel C).

Model	**Cosine**	**Climate 1**	**Climate 2**
nb observations	2220	2220	2220
nb parameters	6	7	6
log-likelihood	-5631	-5603	-5429
DIC	11278	11222	10872
obs. rate	5.7 (5.3-6.3)	6.2 (5.8-6.9)	5 (4.7-5.4)
m	0.12 (0.11-0.12)	0.03 (0.01-0.04)	0.05 (0.05-0.05)
$H_I(0)$	69 (64.4-79.5)	223.9 (170.1-348.7)	143.7 (122.7-161.8)
seroprevalence	49	57	60
ϕ	2.95 (2.32-3.55)	2.46 (1.9-2.78)	1.86 (1.58-2.28)

Table 3.2 – **Estimated parameters and information criteria in models with differing models for seasonality.** Maximum a posteriori and 95% credible intervals in parenthesis. Models include independent sub-populations, with homogeneous mosquito populations, and homogeneous disease introduction.

The climate model 2 performs better than the climate model 1 (it has the smallest DIC in table 3.2), therefore, we will focus on this model in the following sections. The fitted model (climate model 2) is compared to the spatial time series in Figure 3.10. The credible intervals due to observation noise comprise the observed values, except in two areas (Alto da Boa Vista and Portuaria). However, these intervals are very large, including zero in all areas, and overestimating the epidemic peak in many regions (especially Botafogo, Copacabana, Lagoa and Vila Isabel).

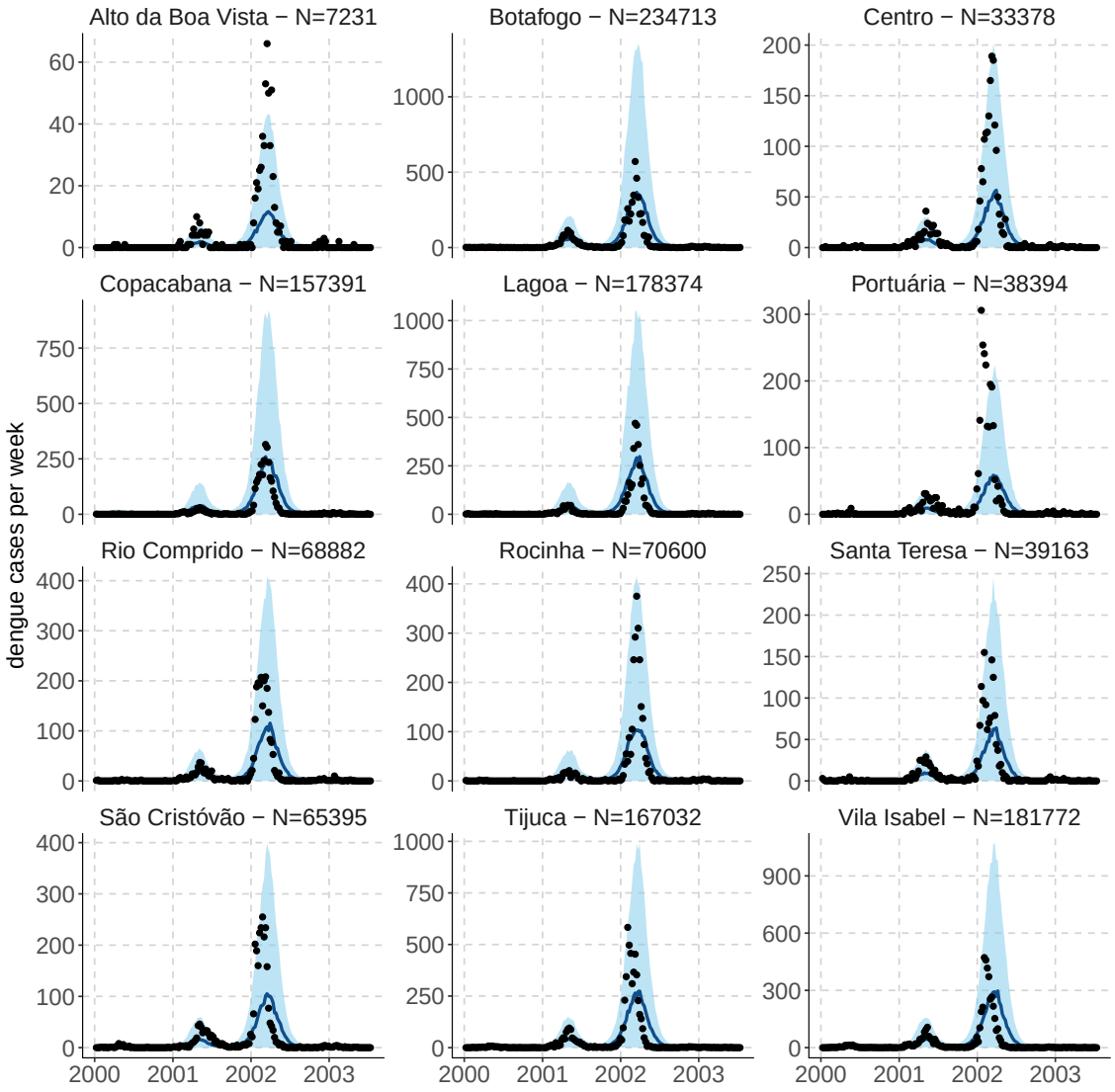


Figure 3.10 – **Number of observed cases per week per administrative region, in the model with seasonality based on climate (model 2)** The model includes independent sub-populations, homogeneous mosquito populations, and homogeneous disease introduction. Median and 95% credible intervals based on simulations from the posterior distribution (plain line and shaded areas) and data points (black dots).

Spatial models

We then introduce spatial structure by considering separately heterogeneity among sub-populations, connections among sub-populations and heterogeneity in disease introduction (cf. the first 6 columns in Table 3.3). The best model in terms of DIC is the one including employee-based connections. In this model, the sub-populations are strongly connected: the proportion of individuals moving outside their residing sub-population has a median η about 18%. The gravity model performed similarly to the independent population model: the parameter ρ indicating the decay of coupling due to distance is estimated very high, and indeed the median η is about 1%. Regarding heterogeneity in mosquito populations, the model including income and urbanization ("Income.Urb" model) performs slightly better than the basic model, but not as well as the more complex model in which one parameter per area is estimated ("Heterogeneity" model). The signs of the coefficients for income and urbanization indicate that a lower income and a higher urbanization rate are associated with higher local dengue transmission in this model. Heterogeneous disease introduction ("Introduction" model) improves the model fit, but also with a large increase in the number of parameters.

Model	**Basic**	**Income-Urb.**	**Heterogeneity**	**Introduction**	**Gravity**	**Employees**
nb observations	2220	2220	2220	2220	2220	2220
nb parameters	6	8	17	17	8	8
log-likelihood	-5429	-5399	-5377	-5333	-5429	-5301
DIC	10872	10812	10783	10695	10872	10618
α_t	0.39 (0.37-0.42)	0.4 (0.38-0.43)	0.4 (0.38-0.43)	0.37 (0.35-0.4)	0.4 (0.37-0.42)	0.36 (0.33-0.38)
α_p	0.34 (0.31-0.35)	0.35 (0.32-0.36)	0.35 (0.32-0.36)	0.36 (0.34-0.38)	0.33 (0.31-0.35)	0.42 (0.4-0.44)
m	0.05 (0.05-0.05)	0.05 (0.05-0.05)		0.05 (0.05-0.05)	0.05 (0.05-0.05)	0.03 (0.03-0.03)
$H_I(0)$	143.7	148.4	152.8		138.9	149.6
	(122.7-161.8)	(127.2-169.7)	(123.7-166.2)		(122.3-161.3)	(133.2-175.4)
r **(‰)**	5 (4.7-5.4)	5 (4.6-5.4)	5.1 (4.7-5.4)	5.3 (4.9-5.6)	5 (4.7-5.4)	4.5 (4.2-4.9)
ϕ	1.86 (1.58-2.28)	1.87 (1.55-2.31)	1.66 (1.42-2.08)	1.63 (1.38-2.03)	1.87 (1.58-2.28)	1.9 (1.56-2.31)
c_x **(‰)**		-0.02 (-0.03–0.01)				
c_u **(‰)**		0.16 (0.11-0.22)				
ρ					4.02 (1.83-4.93)	0.75 (0.69-0.82)
τ_2 **or** δ					0.14 (0.02-0.97)	0.25 (0.21-0.3)
median η **(min-max)**					0.01 (0-0.09)	0.19 (0.14-0.26)
seroprevalence (%)	60	59	59	59	60	53

Table 3.3 – **Estimated parameters in models with seasonality based on climate (model 2).** Maximum a posteriori and 95% credible intervals in parenthesis. **Basic:** Independent populations, homogeneous mosquito populations and homogeneous disease introduction. **Income-Urb.:** Heterogeneity in mosquito populations based on income and urbanization, independent populations and homogeneous disease introduction. **Heterogeneity:** Heterogeneous mosquito populations, independent populations and homogeneous disease introduction. **Introduction:** Heterogeneous disease introduction, independent populations and homogeneous mosquito populations. **Gravity:** Gravity model, homogeneous mosquito populations and homogeneous disease introduction. **Employees:** Model with employee-based connections, homogeneous mosquito populations and homogeneous disease introduction (data for employee number are based on the year 2000).

The median seroprevalence at the end of the epidemic is estimated about 60% in all models, except in the model with employee-based connections, in which it lower (53%) (the global abundance of mosquitoes is also lower and the observation rate higher, indicating that globally, less infections occur, cf. Table 3.3). In this model, the seroprevalence is also much more heterogeneous among sub-populations than in all other models (cf. Figure 3.11). In particular, the Centro region is highly affected, whereas the southern regions (Botafogo, Copacabana, Lagoa) have the lowest seroprevalence rate.

Figure 3.11 – **Proportion of recovered individuals at the end of the year 2003 (date 2003-12-28).** Deterministic simulation using the maximum a posteriori parameter set. Seasonality is based on climate (model 2). **Basic:** Independent populations, homogeneous mosquito populations and homogeneous disease introduction. **Income-Urb.:** Heterogeneity in mosquito populations based on income and urbanization, independent populations and homogeneous disease introduction. **Heterogeneity:** Heterogeneous mosquito populations, independent populations and homogeneous disease introduction. **Introduction:** Heterogeneous disease introduction, independent populations and homogeneous mosquito populations. **Gravity:** Gravity model, homogeneous mosquito populations and homogeneous disease introduction. **Employees:** Model with employee-based connections, homogeneous mosquito populations and homogeneous disease introduction (data for employee number are based on the year 2000).

Because the model with employee-based connections proved best, we then combined it with heterogeneity in mosquito populations and disease introduction. Adding heterogeneous mosquito populations resulted in the largest increase in model fit (cf. Table 3.4). It also resulted in a decrease in the seroprevalence at the end of the epidemic, which is also more heterogeneous among regions (cf. Figure 3.12). The regions are more connected, with about 30% of individuals moving out of their residential area (cf. Table 3.4). Accounting for heterogeneity in disease introduction lead to a smaller increase in model fit, and parameter estimates are similar to the ones in the model with employee-based connections only.

Model	**Employees**	**Empl.-Income.Urb.**	**Empl.-Heterogeneity**	**Empl.-Introduction**
nb observations	2220	2220	2220	2220
nb parameters	8	10	19	19
log-likelihood	-5301	-5221	-5182	-5278
DIC	10618	10460	10392	10589
α_t	0.36 (0.33-0.38)	0.36 (0.33-0.37)	0.37 (0.34-0.39)	0.36 (0.33-0.38)
α_p	0.42 (0.4-0.44)	0.46 (0.44-0.48)	0.44 (0.42-0.46)	0.41 (0.39-0.43)
m	0.03 (0.03-0.03)	0.03 (0.02-0.03)		0.03 (0.03-0.03)
$H_I(0)$	149.6 (133.2-175.4)	167.8 (153.7-202.2)	148.4 (122.5-166.6)	
r **(‰)**	4.5 (4.2-4.9)	5.1 (4.6-5.5)	5.4 (5-5.9)	4.5 (4.2-4.9)
ϕ	1.9 (1.56-2.31)	1.75 (1.48-2.26)	1.5 (1.28-1.89)	1.82 (1.51-2.31)
c_x **(‰)**		-0.3 (-0.35–0.25)		
c_u **(‰)**		1.04 (0.59-1.62)		
ρ	0.75 (0.69-0.82)	0.56 (0.5-0.64)	0.52 (0.47-0.63)	0.82 (0.71-0.86)
τ_2 **or** δ	0.25 (0.21-0.3)	0.13 (0.07-0.19)	0.08 (0.04-0.18)	0.28 (0.21-0.31)
median η **(min-max)**	0.19 (0.14-0.26)	0.31 (0.26-0.39)	0.36 (0.31-0.44)	0.16 (0.12-0.23)
seroprevalence (‰)	53	42	44	54

Table 3.4 – **Estimated parameters in models with employee-based connections, and seasonality based on climate (model 2).** Maximum a posteriori and 95% credible intervals in parenthesis. **Employees:** Model with employee-based connections, homogeneous mosquito populations and homogeneous disease introduction. Data for employee number are based on the year 2000. **Employees-Inc.Urb.:** Model with employee-based connections, heterogeneity in mosquito populations based on income and urbanization and homogeneous disease introduction. **Employees-Het.:** Model with employee-based connections, heterogeneous mosquito populations and homogeneous disease introduction. **Employees-Intr.:** Model with employee-based connection, heterogeneous disease introduction, and homogeneous mosquito populations.

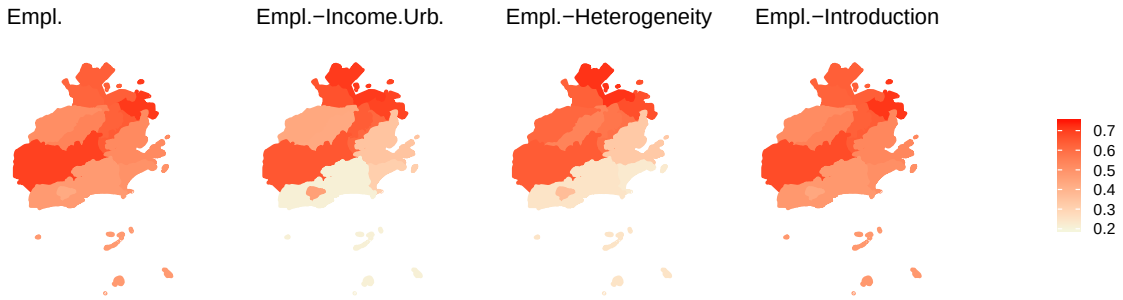


Figure 3.12 – **Proportion of recovered individuals at the end of the year 2003 (date 2003-12-28) in models with employee-based connections.** Simulation using the maximum a posteriori parameter set. Seasonality is based on climate (model 2). Data for employee number are based on the year 2000. **Employees:** Model with employee-based connections, homogeneous mosquito populations and homogeneous disease introduction. **Employees-Inc.Urb.:** Model with employee-based connections, heterogeneity in mosquito populations based on income and urbanization and homogeneous disease introduction. **Employees-Het.:** Model with employee-based connections, heterogeneous mosquito populations and homogeneous disease introduction. **Employees-Intr.:** Model with employee-based connection, heterogeneous disease introduction, and homogeneous mosquito populations.

The estimated abundance of mosquitoes relative to the human population in each patch (m_i) is displayed in Figure 3.13: in particular, we compared the values in models with and without employee-based connections. When connections are accounted for, the parameter estimates are lower and much more heterogeneous among areas.

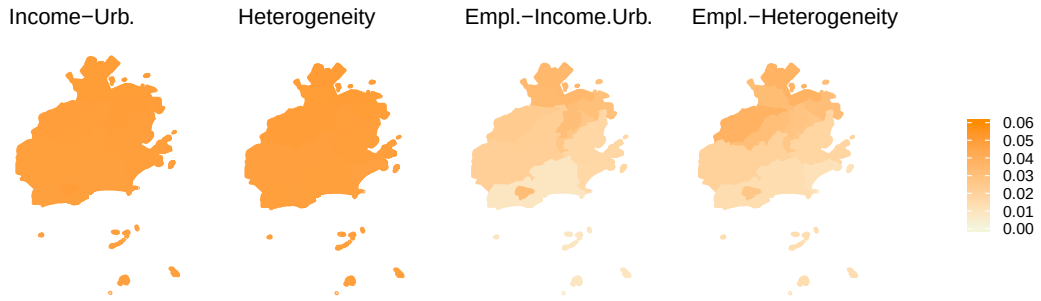


Figure 3.13 – **Abundance of mosquitoes relative to human population size (m_i).** Parameter values associated with the highest posterior for each model.

The fitted model with employee-based connections and heterogeneity in mosquito populations is compared to the data in Figure 3.14. Similarly to the independent model (Figure 3.10), credible intervals are very large (even though the overdispersion parameter is estimated to slightly lower values, cf. Table 3.3 and 3.4), but in some areas, for instance in Botafogo, Copacabana and Lagoa, the overestimation of the epidemic peak is smaller than in the independent case. However, in other areas, the credible intervals do not comprise all data points. Moreover, the epidemic is very synchronized among sub-populations, and the model does not reproduce the fact that the outbreak starts earlier in some regions (Tijuca and Vila Isabel for instance).

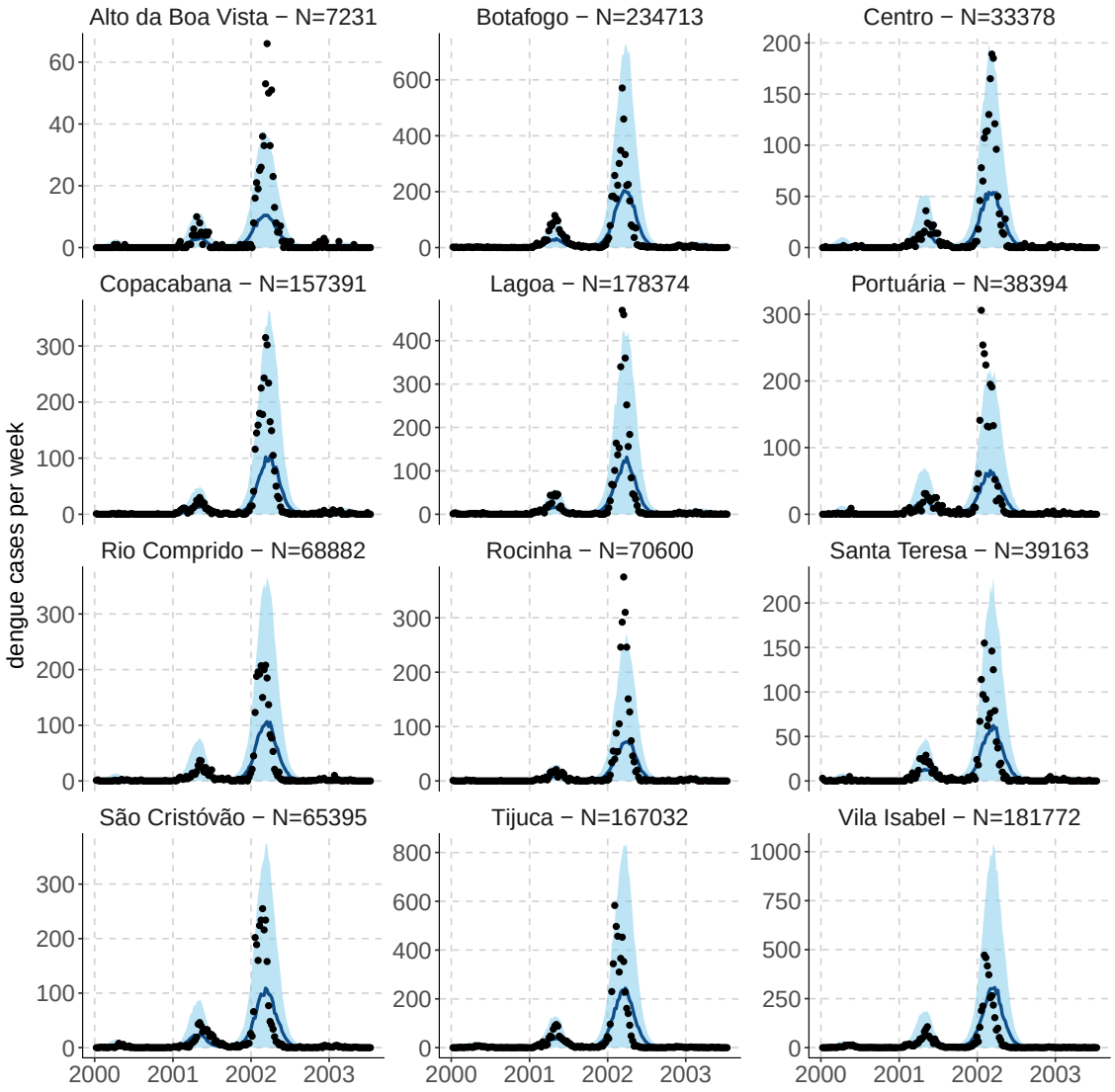


Figure 3.14 – **Number of observed cases per week per administrative region, in models with employee-based connections and heterogeneous mosquito populations.** Median and 95% credible intervals based on simulations from the posterior distribution (plain line and shaded areas) and data points (black dots). The model includes homogeneous disease introduction and seasonality is based on climate (model 2). Data for employee number are based on the year 2000.

3.4 Discussion

In this work, we explore the effects of seasonality and spatial structure on the Ross-Macdonald model, by confronting models to series of dengue cases in Rio de Janeiro.

Over the years 2000 to 2017, several dengue epidemics struck Rio de Janeiro, especially in 2002, 2008 and 2011-2013. During this period, the city experienced many changes, such as the successive re-introduction of DENV-3 and DENV-4, and later the introduction of chikungunya and Zika (Nunes et al., 2015; Villela et al., 2017; Passos et al., 2017), and they may have confounding effects on the interpretation of dengue curves. For these reasons, we decided to focus on the period

2000-2002 for estimations.

Including seasonality was decisive to fit the observed dynamics: the "dengue oscillator" effect of Amaku et al. (2016), namely the disease being stopped by the end of the suitable season for mosquito propagation and restarting the following year, enabled the model to reproduce global outbreaks of differing magnitudes (small peaks followed by a large one). The importance of seasonality highlights the need for a refined representation of seasonal effects, for example by introducing relationships with climate, as interannual climate variations explain a significant part of dengue variability (Lowe et al., 2015). In our estimations, a model including temperature and precipitation proved better to reproduce the observed dynamics. Nevertheless, more details in the model dependencies on climate could be introduced, for example at several stages of mosquito development (Lourenço and Recker, 2014; Lourenço et al., 2017) and semi-mechanistic models could also be used to further explore the type of relationship to include in the model (Cazelles et al., 2018).

For the studied epidemics, the global dynamics is well captured by the seasonal model without specific spatial effects, potentially because the local outbreaks are very synchronized in 2000-2002. However, at the regional scale, the outbreak magnitude is overestimated in some areas and underestimated in others, and introducing spatial structure has an important impact the model's fit. The best models included connections based on the number of employees: in these models, it is estimated that connections between areas are strong, with about 15 to 30% of the people in each area moving to other areas. The influence of human mobility on dengue propagation in cities has been highlighted in many modelling works, both theoretically (Adams and Kapan, 2009; Stoddard et al., 2009; Barmak et al., 2011) and empirically (Stoddard et al., 2013). In our estimations, outbreaks are very synchronized, and the models do not reproduce the small delays between areas. A model with sub-populations that are sequentially related to one another, which would help better fit sequential epidemic peaks, does not seem appropriate to model the disease at this spatial resolution, because of these strong estimated interconnections. However, patterns of sequentially infected areas could exist at more detailed spatial scales or refer more to heterogeneity among individuals or households in terms of connectivity than to administrative entities within the city.

Besides, the gravity model based on population sizes did not fit the data better than a model assuming identical independent populations. Population sizes may not be relevant at the city scale and employment availability seems to be a better indicator of people's mobility. However, the considered mobility model can be refined, first by using other measures of distance (effective distance, based on travel time or road structure, instead of direct kilometric distance) and second by comparing it with other frameworks, such as the radiation model (Simini et al., 2012), the impedance model (Sallah et al., 2017) or the population weighted opportunity model (Yan et al., 2014). If available, other indicators of mobility based on transportation, mobile phone usage (Wesolowski et al., 2015) or social networks (Kraemer et al., 2018) could also be used.

In addition to connectivity, heterogeneity among the sub-populations played an important role to improve data fitting, which was stronger when the sub-populations were connected. In particular, in our estimations, models with connections and heterogeneities had lower and more heterogeneous attack rates at the end of the epidemic, indicating that space can shape the herd immunity, even at a coarse spatial resolution. It would be interesting to study to what extend the map of heterogeneities in transmission potential (modelled here by the abundance of mosquitoes) differs from the one of

seroprevalence, to explore the difficulty to link dengue seroprevalence to socio-demographic local factors. The fact that regions with highest transmission potential are not always the ones with highest dengue incidence, or that areas with limited local transmission can nevertheless experience epidemics has been observed in several works (Luz et al., 2003; Stoddard et al., 2009; Stolerman et al., 2015; Kraemer et al., 2018). Interpretations of our results should however be limited firstly by the influence of the mobility model, as inference of local transmissibility may be sensitive on mobility assumptions (Kraemer et al., 2018). Secondly, in our model, heterogeneities were only considered in terms of transmission potential via the abundance of mosquitoes and other sources of heterogeneities, in disease propagation or reporting, may act as confounding effects. In particular, we assumed that the observation rate was the same in all sub-populations: even though the probability of showing symptoms may be homogeneous over spatial areas, the propensity to report them to public health infrastructures may vary strongly based on socio-demographic characteristics. Yet, it would be impossible to disentangle spatial differences in transmission from spatial differences in reporting without additional data. In order to model regional heterogeneities, we also tested a simple model based two covariates only, income and urbanization, as rough proxies of the heterogeneities in terms of socio-demographic and environmental factors. As expected, this simple model did not perform as well as a full model including population-specific transmission parameters. However, it could also be extended with additional covariates and would be easier to apply to more subpopulations, or to combine with population-specific disease introduction patterns. Other types of relationships between mosquito abundance and population density could also be included in the model (Romeo-Aznar et al., 2018).

This work has also several limitations. First of all, we simplified the model by considering only one dengue serotype because DENV-3 dominated the 2002 epidemic, even though other serotypes were present in the city (De Simone et al., 2004). Extending the model to account for several serotypes would require additional serotype specific data, and this might be decisive in order to study the following epidemics, especially in 2011-2013. In addition, for simplicity in model fitting and result interpretation, we focus on a sub-area of the city in isolation; however, the study of the disease introduction pattern and human mobility would require to extend the model to more sub-populations to represent the whole city.
Another limitation is the use of passively collected surveillance data. Firstly, given that most cases are only detected clinically, misdiagnosis may occur; however, this effect should be weaker during the years 2000-2002, because Zika and chikungunya were not yet introduced in the region. Secondly, surveillance data is characterized by a strong underreporting, because of the large proportion of infectious dengue cases with very mild clinical manifestations (ten Bosch et al., 2018). This effect was accounted for in our overdispersed observation model, and indeed, we estimated observation rates in the order of magnitude of 5%. Lastly, surveillance data is spatially structured in administrative areas, which may not be the most appropriate segmentation and resolution for studying dengue propagation. The spatial resolution is a key element when considering metapopulation models (Mills and Riley, 2014): the spatial scale for dengue transmission is very focal, most of inter-human transmission events occurring on distances less than 1 kilometer (Salje et al., 2012) and the probability of being from the same transmission chain decreasing sharply for cases separated by more than 200 meters in Bangkok (Salje et al., 2017). However, if the spatial scale is too detailed, the hypotheses of the deterministic meta-population model would not hold, so that individual based models and/or stochastic models would be more appropriate.

Fitting these complex spatial processes with deterministic models is a difficult task. First of all, it is difficult to infer mobility patterns from epidemic data, even though several works have

performed estimations of gravity models (see for example Xia et al. (2004); Sarzynska et al. (2013); Gog et al. (2014); Charu et al. (2017); O'Reilly et al. (2018)), often based on summary statistics on incidence curves such as onset dates (Gog et al., 2014; Charu et al., 2017), or peak timing and amplitude (O'Reilly et al., 2018). In our case, we simplified the gravity model by removing the dependence of the origin population, but we still observe correlations between parameters in the MCMC chain, indicating the difficulty to identify separately all parameters. Detailed study of the structural and practical identifiability properties (Miao et al., 2011) of these frequently used mobility models would be very helpful to guide parameter estimations. In addition to mobility patterns, considering location specific parameters induces a large increase in dimension when many sub-populations are considered, which requires appropriate sampling methods. Moreover, estimating more that one set of location specific parameters (heterogeneity both in mosquito population and disease introduction for instance) would raise additional identifiability issues.
Finally, observation noise is very large in these deterministic models that do not account for demographic and environmental stochasticity. In all the fitted models, the disease is already introduced in all considered regions in 2000, and, in the deterministic framework, once the disease is introduced in a given area, it remains present. In reality, local extinctions may be frequent, both in scarcely populated areas and during inter epidemic periods with limited transmission. These effects may influence strongly the behaviour of models, so that an important perspective is the extend of the analysis to the stochastic framework to include these sources of uncertainty.

Additional figures

3.4.1 Comparison of climate datasets

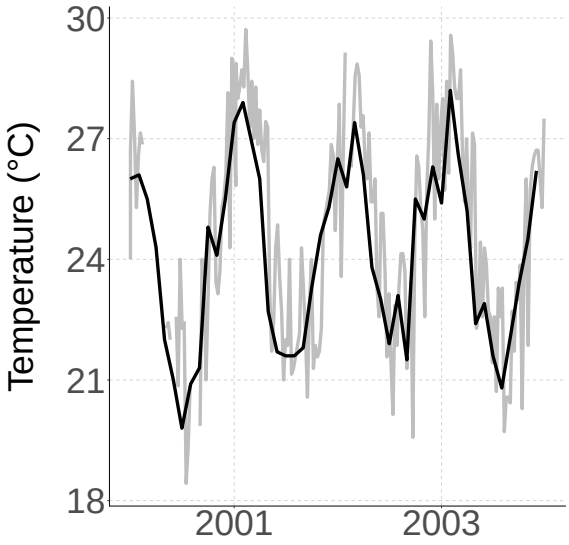


Figure 3.15 – **Comparison of temperature datasets.** Grey : Weather underground weekly data (WeatherUnderground, 2016). Black: CRU monthly data (Harris et al., 2014; Harris and Jones, 2017).

3.4.2 Other spatial indicators

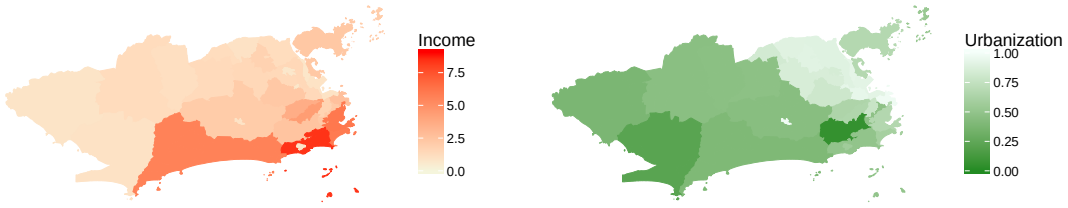


Figure 3.16 – **Income and urbanization indicators among administrative regions.** Household income in minimum salaries, based on 2010 census (data.rio, 2017a). Mean surface covered with urban areas per administrative region (PrefeituraRio, 2010).

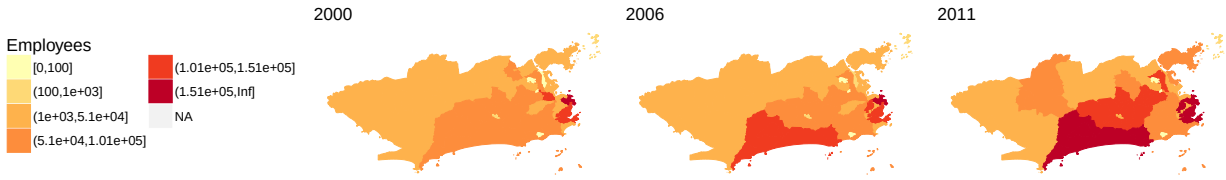


Figure 3.17 – **Number of employees per administrative region in years preceding large dengue outbreaks (2000, 2006 and 2011).** (data.rio, 2017b).

Chapter 4

Une autre arbovirose : le virus Zika

4.1 Introduction

In 2015, another arbovirus transmitted by *Aedes* species made the headlines in the media, when a microcephaly epidemic in newborns was reported in Brazil. The Zika virus (ZIKV) was actually known since the 1950's, but thought to trigger only mild symptoms that often do not require medical care. However, it was afterward associated with severe neurological desorders such as Guillain-Barré syndrome (Cao-Lormeau et al., 2016) and microcephaly in newborns (Schuler-Faccini, 2016), very rare complications that resulted in high absolute number of cases because of the large attack rates of the virus.

Before ZIKV reached the Americas, it circulated in Asia and the Pacific, and the well documented epidemics in the Pacific can be highly informative on the key parameters driving virus transmission, such as the basic reproduction number (R_0). At the time when this work was conducted, some estimates were already available for French Polynesia (Kucharski et al., 2016) and Micronesia (Funk et al., 2016), using deterministic models. In section 4.2, we extended these analyses by exploring the variability in estimations in two different ways: variability across geographical settings by considering four islands of differing sizes and variability associated with the formulation of epidemiological models, by comparing two representations of the mosquito vector population. Moreover, we introduced stochasticity in the transmission process, since demographic variations might be influential in islands with a small population size. Using PMCMC methods, R_0 for the Pacific ZIKV epidemics is estimated between 1.5 and 4.1, the smallest islands displaying higher and more variable values. In addition, we highlighted the importance of seroprevalence measures for appropriate quantification of pathogen propagation. Difficulty in predicting the final size of the epidemic, and therefore in reconciling seroprevalence and incidence data, enabled to shed light on potential model pitfalls and ways of improvements.

In consequence, this work has lead to some extensions. Since dengue virus co-circulated with Zika in French Polynesia and Brazil, in section 4.3, we modeled both pathogens together to study potential interaction effects at the population level. Identifiability issues preventing to detect these interactions were pointed out and we discussed the data needed to explore this question. In addition, the introduction of non-linear transmission rates in vector-host models to relax the homogeneous mixing assumption was studied by Billy Bauzile in his master thesis.

4.2 Structure in the variability of the basic reproductive number (R_0) for Zika epidemics in the Pacific islands [1]

Abstract

Before the outbreak that reached the Americas in 2015, Zika virus (ZIKV) circulated in Asia and the Pacific: these past epidemics can be highly informative on the key parameters driving virus transmission, such as the basic reproduction number (R_0). We compare two compartmental models with different mosquito representations, using surveillance and seroprevalence data for several ZIKV outbreaks in Pacific islands (Yap, Micronesia 2007, Tahiti and Moorea, French Polynesia 2013-2014, New Caledonia 2014). Models are estimated in a stochastic framework with recent Bayesian techniques. R_0 for the Pacific ZIKV epidemics is estimated between 1.5 and 4.1, the smallest islands displaying higher and more variable values. This relatively low range of R_0 suggests that intervention strategies developed for other flaviviruses should enable as, if not more effective control of ZIKV. Our study also highlights the importance of seroprevalence data for precise quantitative analysis of pathogen propagation, to design prevention and control strategies.

Introduction

In May 2015, the first local cases of Zika were recorded in Brazil and by December of the same year the number of cases had surpassed 1.5 million. On February 2016, the World Health Organization declared Zika as a public health emergency of international concern (WHO, 2016b) and in March 2016, local transmission of Zika was recognized in 34 countries. Previously the Zika virus had circulated in Africa and Asia but only sporadic human cases had been reported. In 2007 the outbreak on Yap (Micronesia) was the first Zika outbreak outside Africa and Asia (Duffy et al., 2009). Since, Zika outbreaks have been also reported in French Polynesia and in New Caledonia (Cao-Lormeau et al., 2014; Dupont-Rouzeyrol et al., 2015) between 2013 and 2014 and subsequently, there have been cases of Zika disease in the Cook Islands, the Solomon Islands, Samoa, Vanuatu, and Easter Island (Chile) (see Fig. 1 in (Petersen et al., 2016)).

Zika virus (ZIKV) is a flavivirus, mostly transmitted via the bites of infected *Aedes* mosquitoes, although non-vector-borne transmission has been documented (sexual and maternofoetal transmission, laboratory contamination, transmission through transfusion) (Musso and Gubler, 2016). The most common clinical manifestations include rash, fever, arthralgia, and conjunctivitis (Musso and Gubler, 2016) but a large proportion of infections are asymptomatic or trigger mild symptoms that can remain unnoticed. Nevertheless, the virus may be involved in many severe neurological complications, including Guillain-Barre syndrome (Cao-Lormeau et al., 2016) and microcephaly in newborns (Schuler-Faccini, 2016). These complications and the impressive speed of its geographically propagation make the Zika pandemic a public health threat (WHO, 2016b). This reinforces the urgent need to characterize the different facets of virus transmission and to evaluate its dispersal capacity. We address this here by estimating the key parameters of ZIKV transmission, including the basic reproduction number (R_0), based on previous epidemics in the Pacific islands.

[1] We reproduce here with small edits the paper: Champagne C., Salthouse D.G., Paul R., Cao-Lormeau V.-M., Roche B., Cazelles B., Structure in the variability of the basic reproductive number (R_0) for Zika epidemics in the Pacific islands. *eLife*, 2016

Defined as the average number of secondary cases caused by one typical infected individual in an entirely susceptible population, the basic reproduction number (R_0) is a central parameter in epidemiology used to quantify the magnitude of ongoing outbreaks and it provides insight when designing control interventions (Diekmann et al., 2010). It is nevertheless complex to estimate (Diekmann et al., 2010; van den Driessche and Watmough, 2002), and therefore, care must be taken when extrapolating the results obtained in a specific setting, using a specific mathematical model. In the present study, we explore the variability of R_0 using two models in several settings that had Zika epidemics in different years and that vary in population size (Yap, Micronesia 2007, Tahiti and Moorea, French Polynesia 2013-2014, and New Caledonia 2014). These three countries were successively affected by the virus, resulting in the first significant human outbreaks and they differ in several ways, including population size and location specific features. Hence, the comparison of their parameter estimates can be highly informative on the intrinsic variability of R_0. For each setting, we compare two compartmental models using different parameters defining the mosquito populations. Both models are considered in a stochastic framework, a necessary layer of complexity given the small population size and recent Bayesian inference techniques (Andrieu et al., 2010) are used for parameter estimation.

Results

We use mathematical transmission models and data from surveillance systems and seroprevalence surveys for several ZIKV outbreaks in Pacific islands (Yap, Micronesia 2007 (Duffy et al., 2009), Tahiti and Moorea, French Polynesia 2013-2014 (CHSP, 2014; Mallet et al., 2015; Aubry et al., 2015b), New Caledonia 2014 (DASS, 2014)) to quantify the ZIKV transmission variability.

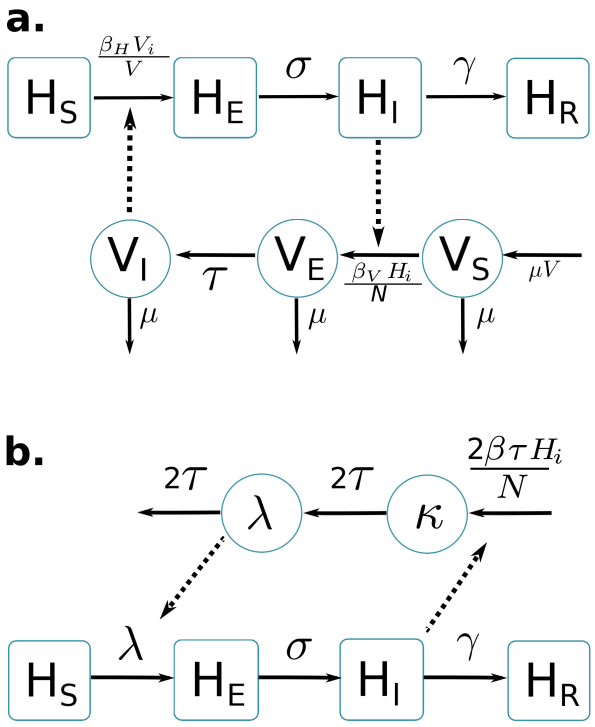


Figure 4.1 – **Graphical representation of compartmental models.** Squared boxes and circles correspond respectively to human and vector compartments. Plain arrows represent transitions from one state to the next. Dashed arrows indicate interactions between humans and vectors. a) Pandey model (Pandey et al., 2013). H_S susceptible individuals; H_E infected (not yet infectious) individuals; H_I infectious individuals; H_R recovered individuals; σ is the rate at which H_E-individuals move to infectious class H_I; infectious individuals (H_I) then recover at rate γ; V_S susceptible vectors; V_E infected (not yet infectious) vectors; V_I infectious vectors; V constant size of total mosquito population; τ is the rate at which V_E-vectors move to infectious class V_I; vectors die at rate μ. b) Laneri model (Laneri et al., 2010). H_S susceptible individuals; H_E infected (not yet infectious) individuals; H_I infectious individuals; H_R recovered individuals; σ is the rate at which H_E-individuals move to infectious class H_I; infectious individuals (H_I) then recover at rate γ; implicit vector-borne transmission is modeled with the compartments κ and λ; λ current force of infection; κ latent force of infection reflecting the exposed state for mosquitoes during the extrinsic incubation period; τ is the transition rate associated to the extrinsic incubation period.

Two compartmental models with vector-borne transmission are compared (cf. Figure 4.1). Both models use a Susceptible-Exposed-Infected-Resistant (SEIR) framework to describe the virus transmission in the human population, but differ in their representation of the mosquito population. Figure 1.a. is a schematic representation derived from Pandey et al. (Pandey et al., 2013) and formulates explicitly the mosquito population, with a Susceptible-Exposed-Infected (SEI) dynamic to account for the extrinsic incubation period (time taken for viral dissemination within the mosquito).

By contrast, in the second model (Figure 1.b.) based on Laneri et al. (Laneri et al., 2010) the vector is modeled implicitly: the two compartments κ and λ do not represent the mosquito population but the force of infection for vector to human transmission. This force of infection passes through two successive stages in order to include the delay associated with the extrinsic incubation period: κ stands for this latent phase of the force of infection whereas λ corresponds directly to the rate at which susceptible humans become infected.

The basic reproduction number of the models (R_0) is calculated using the next Generation Matrix method (Diekmann et al., 2010):

$$R_0^{Pandey} = \sqrt{\frac{\beta_H \beta_V \tau}{\gamma \mu (\mu + \tau)}}$$

$$R_0^{Laneri} = \sqrt{\frac{\beta}{\gamma}}$$

In addition, we consider that only a fraction ρ of the total population is involved in the epidemic, due to spatial heterogeneity, immuno-resistance, or cross-immunity. For both models we define $N = \rho \cdot H$ with H the total size of the population reported by census.

The dynamics of ZIKV transmission in these islands is highly influenced by several sources of uncertainties. In particular, the small population size (less than 7,000 inhabitants in Yap) leads to high variability in transmission rates. Therefore all these models are simulated in a discrete stochastic framework (Poisson with stochastic rates (Bretó et al., 2009)), to take this phenomenon into account. Stochasticity requires specific inference techniques: thus estimations are performed with PMCMC algorithm (particle Markov Chain Monte Carlo (Andrieu et al., 2010)).

Using declared Zika cases from different settings, the two stochastic models (Fig. 4.1) were fitted (Figs 2-3). These models allow us to describe the course of the observed number of cases and estimate the number of secondary cases generated, R_0. Our estimates of R_0 lie between 1.6 (1.5-1.7) and 3.2 (2.4-4.1) and vary notably with respect to settings and models (Figures 4.2-4.3 and Tables 4.1-4.2). Strikingly, Yap displays consistently higher values of R_0 in both models and in general, there is an inverse relationship between island size and both the value and variability of R_0. This phenomenon may be explained by the higher stochasticity and extinction probability associated with smaller populations and can also reflect the information contained in the available data. However, the two highly connected islands in French Polynesia, Tahiti and Moorea, display similar values despite their differing sizes.

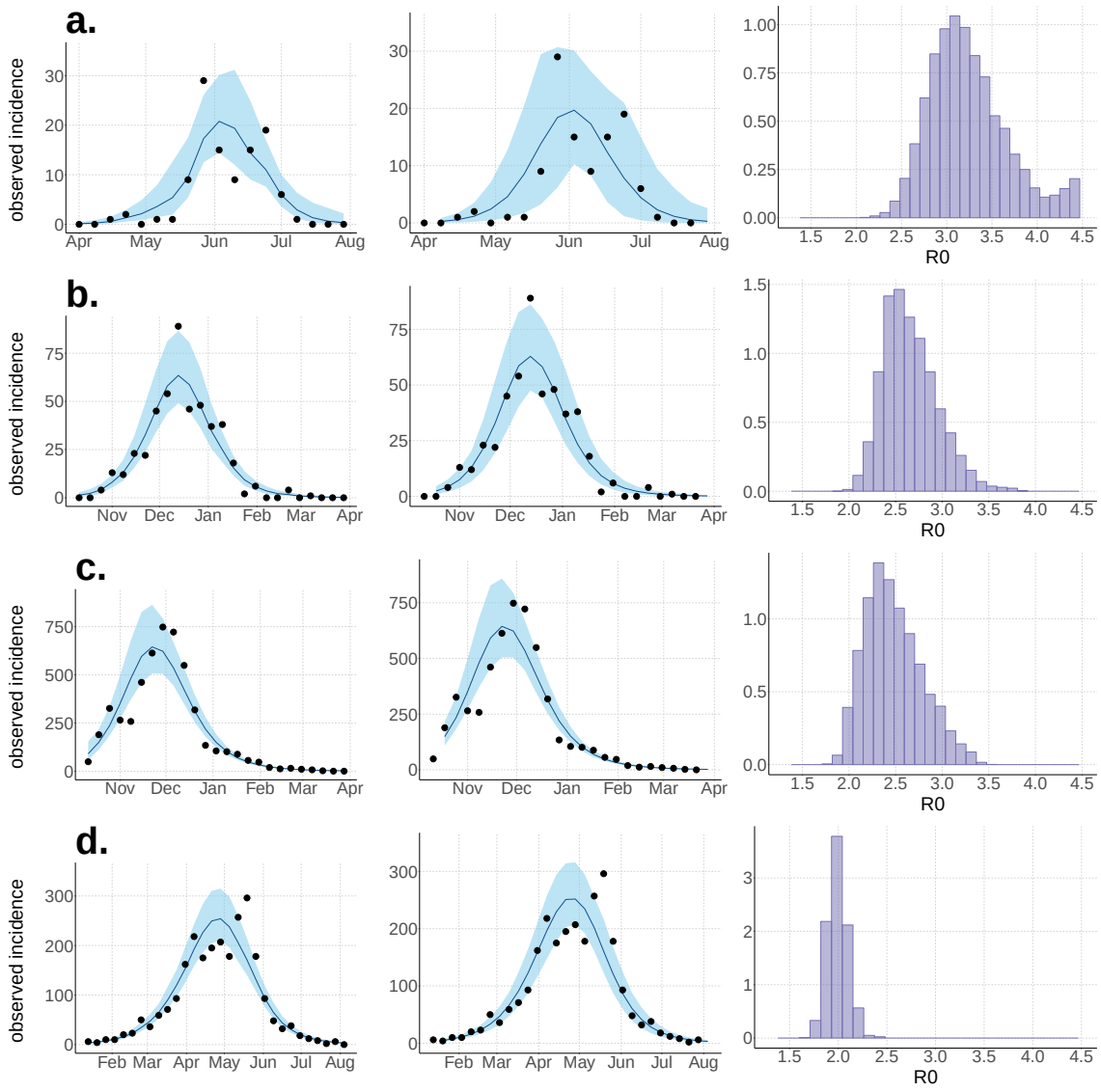


Figure 4.2 – **Results using the Pandey model.** Posterior median number of observed Zika cases (solid line), 95% credible intervals (shaded blue area) and data points (black dots). First column: particle filter fit. Second column: Simulations from the posterior density. Third column: R_0 posterior distribution. a) Yap. b) Moorea. c) Tahiti. d) New Caledonia. The estimated seroprevalences at the end of the epidemic (with 95% credibility intervals) are: a) 73% (CI95: 68-77, observed 73%); b) 49% (CI95: 45-53, observed 49%); c) 49% (CI95: 45-53, observed 49%); d) 39% (CI95: 8-92). See Figure 4.4.

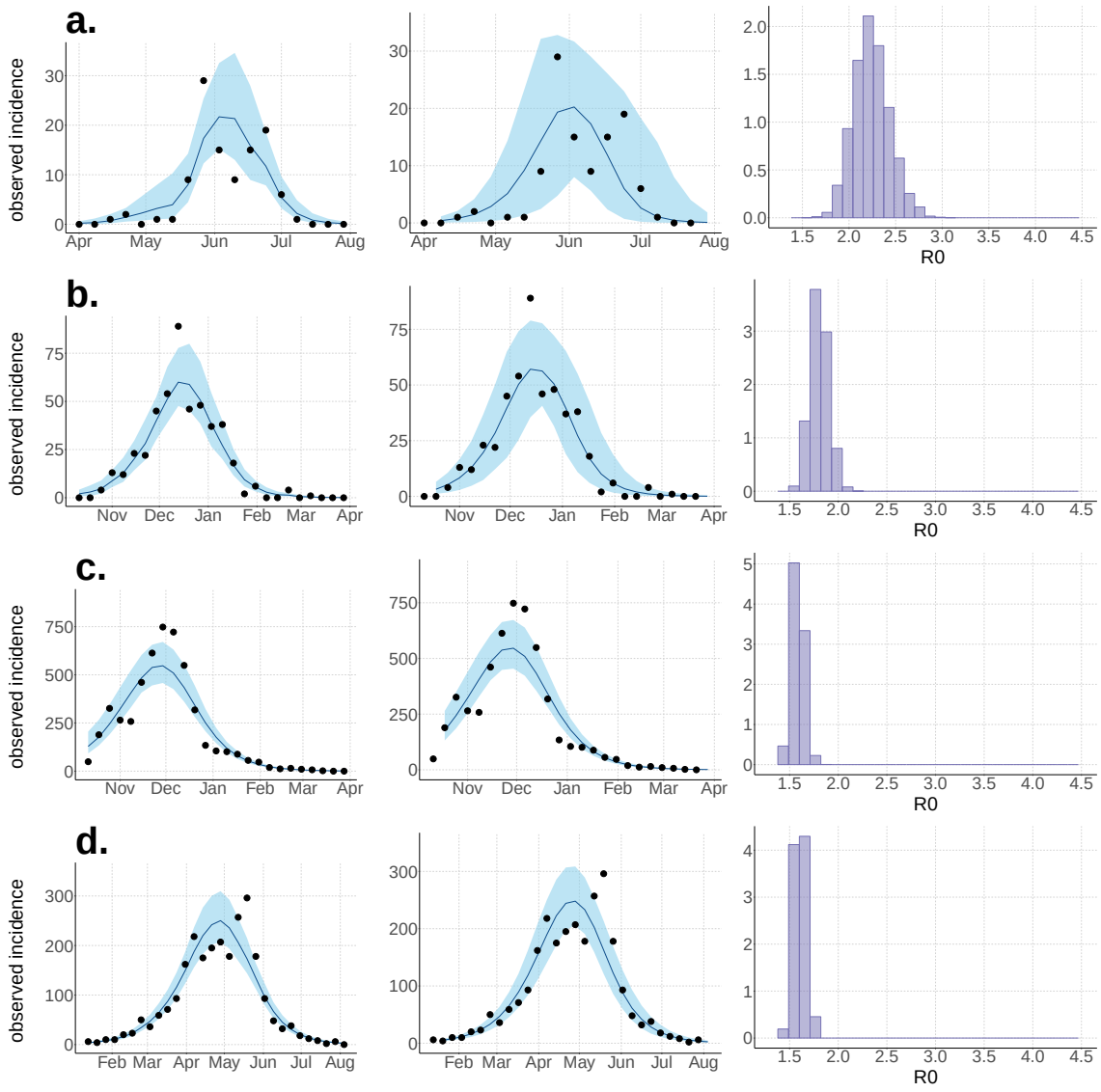


Figure 4.3 – **Results using the Laneri model.** Posterior median number of observed Zika cases (solid line), 95% credible intervals (shaded blue area) and data points (black dots). First column: particle filter fit. Second column: Simulations from the posterior density. Third column: R_0 posterior distribution. a) Yap. b) Moorea. c) Tahiti. d) New Caledonia. The estimated seroprevalences at the end of the epidemic (with 95% credibility intervals) are: a) 72% (CI95: 68-77, observed 73%); b) 49% (CI95: 45-53, observed 49%); c) 49% (CI95: 45-53, observed 49%); d) 65% (CI95: 24-91). See Figure 4.5.

Regarding model variability, R_0 estimates are always higher and coarser with the Pandey model than with the Laneri model (cf. Tables 4.1-4.2). The Pandey model has two additional estimated parameters (in particular, the mosquito lifespan), which can explain the higher variability of the output. It is worth noting that these parameters are very sensitive (see Materials and methods). The difference in R_0 may also be linked to the difference in the estimated initial number of infected individuals (H_{I0}), which is higher in the Laneri model than in the Pandey model. Because of the high proportion of asymptomatic cases (the ratio of asymptomatic:symptomatic is estimated to be 1:1.3, V.-M Cao-Lormeau personal communication), it is hard to determine which scenario is

more realistic, the time between introduction of the disease into the island and the first reported symptomatic case being unknown in most settings.

PANDEY MODEL		Yap	Moorea	Tahiti	New Caledonia
Population size	H	6,892	16,200	178,100	268,767
Basic reproduction number	R_0	3.2 (2.4 – 4.1)	2.6 (2.2 – 3.3)	2.4 (2.0 – 3.2)	2.0 (1.8 – 2.2)
Observation rate	r	0.024 (0.019 – 0.032)	0.058 (0.048 – 0.073)	0.060 (0.050 – 0.073)	0.024 (0.010 – 0.111)
Fraction of population involved	ρ	74% (69-81)	50% (48-54)	50% (46-54)	40% (9-96)
Initial number of infected individuals	Hi_0	2 (1-8)	5 (0-21)	329 (16-1047)	37 (1-386)
Infectious period in human (days)	γ^{-1}	5.2 (4.1 – 6.7)	5.2 (4.1 – 6.8)	5.2 (4.1 – 6.7)	5.5 (4.2 – 6.8)
Extrinsic incubation period in mosquito (days)	τ^{-1}	10.6 (8.7 – 12.5)	10.5 (8.6 – 12.4)	10.5 (8.6 – 12.6)	10.7 (8.9 – 12.5)
Mosquito lifespan (days)	μ^{-1}	15.6 (11.7 – 19.3)	15.3 (11.4 – 19.1)	15.1 (11.2 – 19.0)	15.4 (11.6 – 19.4)

Table 4.1 – **Parameter estimations for the Pandey model.** Posterior median (95% credible intervals). All the posterior parameter distributions are presented in Figures 4.6-4.9.

LANERI MODEL		Yap	Moorea	Tahiti	New Caledonia
Population size	H	6,892	16,200	178,100	268,767
Basic reproduction number	R_0	2.2 (1.9 – 2.6)	1.8 (1.6 – 2.0)	1.6 (1.5 – 1.7)	1.6 (1.5 – 1.7)
Observation rate	r	0.024 (0.019 – 0.033)	0.057 (0.047 – 0.071)	0.057 (0.049 – 0.069)	0.014 (0.010 – 0.037)
Fraction of population involved	ρ	73% (69 – 78)	51% (47 – 55)	54% (49 – 59)	71% (27 – 98)
Initial number of infected individuals	Hi_0	2 (1 – 10)	9 (1 – 28)	667 (22 – 1570)	82 (2 – 336)
Infectious period in human (days)	γ^{-1}	5.3 (4.1 – 6.6)	5.3 (4.1 – 6.7)	5.2 (4.1 – 6.7)	5.4 (4.1 – 6.8)
Extrinsic incubation period in mosquito (days)	τ^{-1}	10.7 (8.8 – 12.7)	10.6 (8.6 – 12.6)	10.5 (8.5 – 12.5)	10.8 (8.9 – 12.8)

Table 4.2 – **Parameter estimations for the Laneri model.** Posterior median (95% credible intervals). All the posterior parameter distributions are presented in Figures 4.10-4.13.

For the duration of infectious and intrinsic incubation (in human) and extrinsic incubation (in mosquito) periods, the posterior density ressembles the informative prior (cf. Figures 4.6-4.13), indicating the models' incapacity to identify properly these parameters without more informative data. Moreover, these parameters have a clear sensitivity (see Materials and methods) and precise field measures are therefore crucial for reliable model predictions.

The fraction ρ of the population involved in the epidemic is well estimated when the seroprevalence is known (in Yap and French Polynesia). In these cases, the proportion of the population involved is slightly greater than the seroprevalence rate, indicating a very high infection rate among involved individuals. In New Caledonia, as no information on seroprevalence was available, the fraction of population involved displays very large confidence intervals (cf. Tables 4.1 and 4.2), indicating that the model did not manage to identify properly this parameter with the available data. In this case, this parameter is highly correlated to the observation rate r (cf Figures 4.17 and 4.21): r and ρ seem unidentifiable without precise information on seroprevalence.

Discussion

The reproduction number R_0 is a key parameter in epidemiology that characterizes the epidemic dynamics and the initial spread of the pathogen at the start of an outbreak in a susceptible population. R_0 can be used to inform public health authorities on the level of risk posed by an infectious disease, vaccination strategy, and the potential effects of control interventions (Anderson and May, 1982). In the light of the potential public health crisis generated by the international propagation of ZIKV, characterization of the potential transmissibility of this pathogen is crucial for predicting epidemic size, rate of spread and efficacy of intervention.

Using data from both surveillance systems and seroprevalence surveys in four different geographical settings across the Pacific (Duffy et al., 2009; CHSP, 2014; Mallet et al., 2015; DASS, 2014; Aubry et al., 2015b), we have estimated the basic reproductive number R_0 (see Figs 4.2-4.3 and Tables 4.1-4.2). Our estimate of R_0 obtained by inference based on Particle MCMC (Andrieu et al., 2010) has values in the range 1.6 (1.5-1.7) - 3.2 (2.4-4.1). Our R_0 estimates vary notably across settings. Lower and finer R_0 values are found in larger islands. This phenomenon can at least in part be explained by large spatial heterogeneities and higher demographic stochasticity for islands with smaller populations, as well as the influence of stochasticity on biological and epidemiological processes linked to virus transmission. This phenomenon can also be specific to the selection of the studied islands or can reflect a highly clustered geographical pattern, the global incidence curve being the smoothed overview of a collection of more explosive small size outbreaks. However, it is notable that the two French Polynesian islands yield similar estimates of R_0 despite differing population sizes. Indeed, other important factors differ among French Polynesia, New Caledonia and Yap, such as the human genetic background and their immunological history linked to the circulation of others arboviruses. Moreover, whilst both New Caledonia and French Polynesia populations were infected by the same ZIKV lineage and transmitted by the same principle vector species, *Aedes aegypti*, the epidemic in Yap occurred much earlier with a different ZIKV lineage (Wang et al., 2016) and vectored by a different mosquito species *Aedes hensilli* (Ledermann et al., 2014). In French Polynesia, the vector *Aedes polynesiensis* is also present and dominates in Moorea with higher densities than in Tahiti. Finally, different vector control measures have been conducted in the three countries.

To date, studies investigating Zika outbreaks in the Pacific have always estimated R_0 using a deterministic framework. Using a similar version of the Pandey model in French Polynesia, Kucharski et al. (Kucharski et al., 2016) estimated R_0 between 1.6 and 2.3 (after scaling to square root for comparison) for Tahiti and between 1.8 and 2.9 in Moorea. These estimates are slightly lower and less variable than ours. This difference can be explained firstly by the chosen priors on mosquito parameters and secondly because our model includes demographic stochasticity. Moreover, they predicted a seroprevalence rate at the end of the epidemic of 95-97%, far from the 49% measured. In Yap island, a study (Funk et al., 2016) used a very detailed deterministic mosquito model, and estimated an R_0 for Zika between 2.9 and 8. In this case, our lower and less variable estimates may come from the fact that our model is more parsimonious in the number of uncertain parameters, especially concerning the mosquito population. Finally, a third study (Nishiura et al., 2016a) relied on another method for R_0 calculation (based on the early exponential growth rate of the epidemic) in French Polynesia as a whole and in Yap. Again, the obtained parameters are lower than ours in French Polynesia and higher in Yap. The first estimations for South America using a similar methodology (Nishiura et al., 2016b; Towers et al., 2016b; Gao et al., 2016) lead to similar R_0 values. In all these studies a deterministic framework is used

excluding the possibility of accounting for the high variability of biological and epidemiological processes exacerbated by the small size of the population. In these three studies, like in ours, it is worth noting that little insight is obtained regarding mosquito parameters. Posterior distribution mimics the chosen prior (cf. Figures 4.6-4.13). Both the simulation of the epidemics and the estimated R_0 are highly sensitive to the choice of priors on mosquito parameters, for which precise field measures are rare.

In the absence of sufficient data, the modeling of mosquito-borne pathogen transmission is a difficult task due to non-linearity and non-stationarity of the involved processes (Cazelles and Hales, 2006). This work has then several limitations. First, our study is limited by the completeness and quality of the data, with regard to both incidence and seroprevalence, but, above all, by the scarcity of information available on mosquitoes. Incidence data is aggregated at the island scale and cannot disentangle the effects of geography and observation noise to explain bimodal curves observed in Yap and New Caledonia. Moreover, although all data came from national surveillance systems, we had very little information about the potential degree of under-reporting, especially due to the high proportion of mildly symptomatic cases, at a time when the dangerous complications associated with the virus were unknown. Moreover, some cases might have been misdiagnosed as other flaviviruses, due to similarity in clinical manifestations or cross-reactivity in assays. Seroprevalence data were gathered from small sample sizes and were also sensitive to cross reactivity in populations non naive to dengue. In addition, they were missing in New Caledonia, which leads to strong correlation between our estimation of the observation rate and the fraction of the population involved in the epidemic. Because of the high proportion of asymptomatic or mildly symptomatic cases, the magnitude of the outbreaks is difficult to evaluate without precise seroprevalence data (Metcalf et al., 2016) or detection of mild, asymptomatic or pre-symptomatic infections (Thompson et al., 2016). Considering vectors, no demographic data were available and this partly explains the large variability of our R_0 estimations.
Secondly, incidence and seroprevalence data were difficult to reconcile; the use of incidence data led to higher infection rates than those observed in the seroprevalence data. This difficulty has been overcome by considering that only a fraction of the population (ρ) is involved in the epidemic and then our model manages to reproduce the observed seroprevalence rate. This exposed fraction could be the result of spatial heterogeneity and high clustering of cases and transmission, as observed for dengue. The small dispersal of the mosquito and the limited population inter-mingling likely leads to considerable spatial variation in the extent of exposure to the virus and pockets of refugia in Tahiti as elsewhere (Telle et al., 2016). For instance, previous dengue infection rates in French Polynesia display large spatial variations even within islands (Daudens et al., 2009). Finer scale incidence and seroprevalence data would be useful to explore this. Another explanation for higher predicted than observed infection rates could be due to interaction with other flaviviruses. The Zika outbreak was concomitant with dengue outbreaks in French Polynesia (CHSP, 2014; Mallet et al., 2015) and New Caledonia (DASS, 2014). Examples of coinfection have been reported (Dupont-Rouzeyrol et al., 2015) but competition between these close pathogens may also have occurred. Finally, mathematical models with vectorial transmission may tend to estimate high attack rates, sometimes leading to a contradiction between observed incidence and observed seroprevalence. Assumptions on the proportionality between infected mosquitoes and the force of infection, as well as the density-dependence assumption in these models could be questioned. Indeed even if these assumptions are at the heart of the mathematical models of mosquito-borne pathogen transmission (Reiner et al., 2013; Smith et al., 2014), a recent review (Halstead, 2008) and recent experimental results (Bowman et al., 2014; Harrington et al., 2014) question these important points.

On a final note, the estimates of R_0 for ZIKV are similar to but generally on the lower side of estimates made for two other flaviviruses of medical importance, dengue and Yellow Fever viruses (Favier et al., 2006; Imai et al., 2015; Massad et al., 2003), even though caution is needed in the comparison of studies with differing models, methods and data sources. Interventions strategies developed for dengue should thus enable as, if not more effective control of ZIKV, with the caveat that ZIKV remains principally a mosquito-borne pathogen with little epidemiological significance of the sexual transmission route. Even though further work and data are needed to support this hypothesis (Brauer et al., 2016), two recent studies indicated that sexual transmission alone is not sufficient to sustain an epidemic (Gao et al., 2016; Towers et al., 2016b).

In conclusion, using recent stochastic modeling methods, we have been able to determine estimates of R_0 for ZIKV with an unexpected relationship with population size. Further data from the current Zika epidemic in South America that is caused by the same lineage as French Polynesia lead to estimates in the same range of values (Nishiura et al., 2016b; Towers et al., 2016b; Gao et al., 2016). Our study highlights the importance of gathering seroprevalence data, especially for a virus that often leads to an asymptomatic outcome and it would provide a key component for precise quantitative analysis of pathogen propagation to enable improved planning and implementation of prevention and control strategies.

Materials and methods

Data

During the 2007 outbreak that struck Yap, 108 suspected or confirmed Zika cases (16 per 1,000 inhabitants) were reported by reviewing medical records and conducting prospective surveillance between April 1st and July 29th 2007 (Duffy et al., 2009). In French Polynesia, sentinel surveillance recorded more than 8,700 suspected cases (32 per 1,000 inhabitants) across the whole territory between October 2013 and April 2014 (CHSP, 2014; Mallet et al., 2015). In New Caledonia, the first Zika case was imported from French Polynesia on 2013 November 12th. Approximately 2,500 cases (9 per 1,000 inhabitants) were reported through surveillance between January (first autochtonous case) and August 2014 (DASS, 2014).

For Yap and French Polynesia, the post-epidemic seroprevalence was assessed. In Yap, a household survey was conducted after the epidemic, yielding an infection rate in the island of 73% (Duffy et al., 2009). In French Polynesia, three seroprevalence studies were conducted. The first one took place before the Zika outbreak, and concluded that most of the population was naive for Zika virus (Aubry et al., 2015a). The second seroprevalence survey was conducted between February and March 2014, at the end of the outbreak, and reported a seroprevalence rate around 49% (Aubry et al., 2015b). The third one concerned only schoolchildren in Tahiti and was therefore not included in the present study.

Demographic data on population size were based on censuses from Yap (Duffy et al., 2009), French Polynesia (Insee, 2012), and New Caledonia (Insee, 2014).

Models and inference

Model equations

Although the models are simulated in a stochastic framework, we present them with ordinary differential equations for clarity. The reactions involved in the stochastic models are the same as those governed by the deterministic equations, but the simulation process differs through the use of discrete compartments. It is described in the next section.

The equations describing Pandey model are:

$$\begin{aligned}
\frac{dH_S}{dt} &= -\beta_H v_I H_S \\
\frac{dH_E}{dt} &= \beta_H v_I H_S - \sigma H_E \\
\frac{dH_I}{dt} &= \sigma H_E - \gamma H_I \\
\frac{dH_R}{dt} &= \gamma H_I \\
\frac{dv_S}{dt} &= \mu - \frac{\beta_V H_I}{N} v_S - \mu v_S \\
\frac{dv_E}{dt} &= \frac{\beta_V H_I}{N} v_S - \tau v_E - \mu v_E \\
\frac{dv_I}{dt} &= \tau v_E - \mu v_I
\end{aligned}$$

where $v_s = \frac{V_S}{V}$ is the proportion of susceptible mosquitoes, $v_E = \frac{V_E}{V}$ the proportion of exposed mosquitoes, and $v_I = \frac{V_I}{V}$ the proportion of infected mosquitoes. Since we are using a discrete model, we cannot use directly the proportions v_S, v_E and v_I whose values are smaller than one. Therefore, we rescale using $V = H$, which leads to $V_S' = v_S \cdot H$, $V_E' = v_E \cdot H$, and $V_I' = v_I \cdot H$.
In this model, the force of infection for humans is $\lambda_H = \beta_H v_I$, and the force of infection for mosquitoes is $\lambda_V = \beta_V \frac{H_I}{N}$

The equations describing Laneri model are:

$$\begin{aligned}
\frac{dH_S}{dt} &= -\lambda H_S \\
\frac{dH_E}{dt} &= \lambda H_S - \sigma H_E \\
\frac{dH_I}{dt} &= \sigma H_E - \gamma H_I \\
\frac{dH_R}{dt} &= \gamma H_I \\
\frac{d\kappa}{dt} &= \frac{2\beta H_I \tau}{N} - 2\tau\kappa \\
\frac{d\lambda}{dt} &= 2\tau\kappa - 2\tau\lambda
\end{aligned}$$

In this model, the role of mosquitoes in transmission is represented only through the delay they introduce during the extrinsic incubation period (EIP, incubation period in the mosquito).

For modeling reasons, this delay is included by representing the force of infection from infected humans to susceptible humans with two compartments κ and λ: in this formalism, the duration between the moment when an exposed individual becomes infectious and the moment when another susceptible individual acquires the infection has a gamma distribution of mean τ^{-1} (Laneri et al., 2010; Roy et al., 2013; Lloyd, 2001). Therefore, λ represents the current force of infection for humans $\lambda_H = \lambda$. The compartment κ represents the same force of infection but at a previous stage, reflecting the exposed phase for mosquitoes during the extrinsic incubation period. As an analogy to Pandey model, the force of infection for mosquitoes is $\lambda_V = \frac{\beta H_I \tau}{v_s N}$, and therefore, the parameter β can be interpreted as the product of a transmission parameter β' by the proportion of susceptible mosquitoes: $\beta = v_s \beta'$. The force of infection for mosquitoes is then similar to Pandey's : $\lambda_V = \beta' \tau \frac{H_I}{N}$.
Again, since we are using a discrete model, we cannot use directly the proportions λ and κ whose values are smaller than one. Therefore, we rescale up to a factor N, which leads to $L = \lambda N$ and $K = \kappa N$.

In both models, following the dominant paradigm that diseases transmitted by *Aedes* mosquitoes are highly clustered, we restricted the total population H measured by census to a fraction $N = \rho.H$, in which the parameter ρ is estimated. This formulation makes the hypothesis that a fraction of the total population is not at risk from the epidemic, because of individual factors or because the individuals remain in areas where the virus is not present. Moreover as the vector's flying range is small, the clustering of ZIKV infection may be reinforced. This may result in escapees from infection within the population, even at a single island scale. The available data does not allow further exploration of the mechanisms underlying these phenomena, which seem fundamental to understand ZIKV propagation. At the very least, the restriction to a fraction ρ enables the model to reproduce the observed seroprevalence rates, and to provide coherent results with respect to both data sources (seroprevalence and surveillance data).

Stochastic framework

Both models are simulated in a stochastic and discrete framework, the Poisson with stochastic rates formulation (Bretó et al., 2009), to include the uncertainties related to small population size. In this framework, the number of reactions occurring in a time interval dt is approximated by a multinomial distribution. In a model with m possible reactions and c compartments, z_t being the state of the system at time t and θ the model parameters, the probability that each reaction r^k occurs n_k times in dt is calculated as follows (Dureau et al., 2013a):

$$p(n_1, ...n_m | z_t, \theta) = \prod_{i=1}^{c} \left\{ M_i (1 - \sum_{X(k)=i} p_k)^{\bar{n_i}} \prod_{X(k)=i} (p_k)^{n_k} \right\} + o(dt)$$

with, $z_t^{(i)}$ being the number of individual in compartment i at time t,

- $p_k = \left(1 - exp\left\{ - \sum_{X(l)=i} r^l(z_t, \theta) z_t^{X(l)} dt \right\}\right) \frac{r^k(z_t,\theta)}{\sum_{X(l)=i} r^l(z_t,\theta)}$
- $\bar{n_i} = z_t^{(i)} - \sum_{X(k)=i} n_k$ (the number of individuals staying in compartment i in dt)
- $M_i = \binom{z_t^{(i)}}{n_k \, X(k)=i \, \bar{n_i}}$ (multinomial coefficient)

Observation models

The only observed compartments are the infected humans (incidence measured every week) and the recovered humans (seroprevalence at the end of the outbreak when data is available). In order to link the model to the data, two observation models, for both incidence and seroprevalence data, are needed.

Observation model on incidence data

The observed weekly incidence is assumed to follow a negative binomial distribution (Bretó et al., 2009) whose mean equals the number of new cases predicted by the model times an estimated observation rate r.

The observation rate r accounts for non observed cases, due to non reporting from medical centers, mild symptoms unseen by health system, and asymptomatic infections. Without additional data, it is not possible to make a distinction between these three categories of cases. We also implicitely make the assumption that these cases transmit the disease as much as reported symptomatic cases.

The observation model for incidence data is therefore :

$$Inc_{obs} = NegBin(\phi^{-1}, \frac{1}{1 + \phi r Inc})$$

Inc_{obs} being the observed incidence, and Inc the incidence predicted by the model. The dispersion parameter (Bretó et al., 2009) ϕ is fixed at 0.1.

Observation model on seroprevalence data

Seroprevalence data is fitted for Tahiti, Moorea, and Yap settings. It is assumed that the observed seroprevalence at the end of the epidemic follows a normal distribution with fixed standard deviation, whose mean equals the number of individuals in the H_R compartment predicted by the model.

The observation model for seroprevalence data is therefore :

$$H_R^{obs} = Normal(H_R, \Lambda)$$

at the last time step, with notations detailed for each model in Table 4.3.

Island	Date	Standard deviation Λ	Observed seroprevalence H_R^{obs}
Yap	2007-07-29	150	5005 (Duffy et al., 2009)
Moorea	2014-03-28	325	$0.49 \times 16200 = 7938$ (Aubry et al., 2015b)
Tahiti	2014-03-28	3562	$0.49 \times 178100 = 87269$ (Aubry et al., 2015b)

Table 4.3 – **Details of the observation models for seroprevalence**

Prior distributions

Informative prior distributions are assumed for the mosquito lifespan, the duration of infectious period, and both intrinsic and extrinsic incubation periods. The initial numbers of infected mosquitoes and humans are estimated, and the initial number of exposed individuals is set to the initial number of infected to reduce parameter space. We assume that involved populations are naive to Zika virus prior to the epidemic and set the initial number of recovered humans to zero. The other priors and associated references are listed in Table 4.4.

Parameters		**Pandey model**	**Laneri model**	**References**
R_0^2	squared basic reproduction number	Uniform[0, 20]	Uniform[0, 20]	assumed
β_V	transmission from human to mosquito	Uniform[0, 10]	.	assumed
γ^{-1}	infectious period (days)	Normal(5, 1) in [4, 7]	Normal(5, 1) in [4, 7]	(Mallet et al., 2015)
σ^{-1}	intrinsic incubation period (days)	Normal(4, 1) in [2, 7]	Normal(4, 1) in [2, 7]	(Nishiura et al., 2016b; Bearcroft, 1956; Lessler et al., 2016)
τ^{-1}	extrinsic incubation period (days)	Normal(10.5, 1) in [4, 20]	Normal(10.5, 1) in [4, 20]	(Hayes, 2009; Chouin-Carneiro et al., 2016)
μ^{-1}	mosquito lifespan (days)	Normal(15, 2) in [4, 30]	.	(Brady et al., 2013; Liu-Helmersson et al., 2014)
ρ	fraction of population involved	Uniform[0, 1]	Uniform[0, 1]	
Initial conditions (t=0)		**Pandey model**	**Laneri model**	
$H_{I\,0}$	Infected humans	Uniform[10^{-6}, 1]·N	Uniform[10^{-6}, 1]·N	
$H_{E\,0}$	exposed humans	$H_{I\,0}$	$H_{I\,0}$	
$H_{R\,0}$	Recovered humans	0	0	
	Infected vectors	$V_{I\,0}$=Uniform[10^{-6}, 1]·H	L_0=Uniform[10^{-6}, 1]·N	
	exposed vectors	$V_{E\,0} = V_{I\,0}$	K_0=L_0	

Local conditions		**Yap**	**Moorea**	**Tahiti**	**New Caledonia**	**References**
r	Observation rate	Uniform[0, 1]	Uniform[0, 1]	Uniform[0, 0.3]	Uniform[0, 0.23]	(Mallet et al., 2015; DASS, 2014)
H	Population size	6,892	16,200	178,100	268,767	(Duffy et al., 2009; Insee, 2012, 2014)

Table 4.4 – **Prior distributions of parameters**. "Uniform[0,20]" indicates a uniform distribution between in the range [0,20]. "Normal(5,1) in [4,7]" indicates a normal distribution with mean 5 and standard deviation 1, restricted to the range [4,7].

The range for the prior on observation rate is reduced for Tahiti and New Caledonia, in order to reduce the parameter space and facilitate convergence. In both cases, we use the information provided with the data source. In French Polynesia, 8,750 cases we reported, but according to local health authorities, more than 32,000 people would have attended health facilities for Zika (Mallet et al., 2015) ($8750/32000 \leq 0.3$). In New Caledonia, approximately 2,500 cases were reported but more than 11,000 cases were estimated by heath authorities (DASS, 2014) ($2500/11000 \leq 0.23$). In both cases, these extrapolations are lower bounds on the real number of cases (in particular, they do not estimate the number of asymptomatic infections), and therefore can be used as upper bounds on the observation rate.

Estimations

Inference with PMCMC

The complete model is represented using the state space framework, with two equation systems: the transition equations refer to the transmission models, and the measurement equations are given

by the observation models.

In a deterministic framework, this model could be directly estimated using MCMC, with a Metropolis-Hastings algorithm targeting the posterior distribution of the parameters. This algorithm would require the calculation of the model likelihood at each iteration.

In our stochastic framework, the model output is given only through simulations and the likelihood is intractable. In consequence, estimations are performed with the PMCMC algorithm (particle Markov Chain Monte Carlo (Andrieu et al., 2010)), in the PMMH version (particle marginal Metropolis-Hastings). This algorithm uses the Metropolis-Hastings structure, but replaces the real likelihood by its estimation with Sequential Monte Carlo (SMC).

Algorithm 9 PMCMC (Andrieu et al., 2010) (PMMH version, as in SSM (Dureau et al., 2013a))

In a model with n observations and J particles.
$q(.|\theta^{(i)})$ is the transition kernel.

1: Initialize $\theta^{(0)}$.
2: Using SMC algorithm, compute $\hat{p}(y_{1:n}|\theta^{(0)})$ and sample $x^*_{0:n}$ from $\hat{p}(x_{0:n}|y_{1:n}, \theta^{(0)})$.
3: **for** $i = 1...N^\theta$ **do**
4: Sample θ^* from $q(.|\theta^{(i)})$
5: Using SMC algorithm, compute $L(\theta^*) = \hat{p}(y_{1:n}|\theta^*)$ and sample $x^*_{0:n}$ from $\hat{p}(x_{0:n}|y_{1:n}, \theta^*)$
6: Accept θ^* (et $x^*_{0:n}$) with probability $1 \wedge \frac{L(\theta^{(i)})q(|\theta^*)}{L(\theta^*)q(\theta^*|\theta^{(i)})}$
7: If accepted, $\theta^{(i+1)} = \theta^*$ and $x^{(i+1)}_{0:n} = x^*_{0:n}$. Otherwise $\theta^{(i+1)} = \theta^{(i)}$ and $x^{(i+1)}_{0:n} = x^{(i)}_{0:n}$.
8: **end for**

SMC (Doucet et al., 2001a) is a filtering method that enables to recover the latent variables and estimate the likelihood for a given set of parameters. The data is treated sequentially, by adding one more data point at each iteration. The initial distribution of the state variables is approximated by a sample a particles, and from one iteration to the next, all the particles are projected according to the dynamic given by the model. The particles receive a weight according to the quality of their prediction regarding the observations. Before the next iteration, all the particles are resampled using these weights, in order to eliminate low weight particles and concentrate the computational effort in high probability regions. Model likelihood is also computed sequentially at each iteration (Dureau et al., 2013a; Doucet and Johansen, 2011).

Algorithm 10 SMC (*Sequential Monte Carlo*, as implemented in SSM (Dureau et al., 2013a))
In a model with n observations and J particles.

L is the model likelihood $p(y_{1:T}|\theta)$. $W_k^{(j)}$ is the weight and $x_k^{(j)}$ is the state associated to particle j at iteration k.

Set $L = 1$, $W_0^{(j)} = 1/J$.
Sample $(x_0^{(j)})_{j=1:J}$ from $p(x|\theta_0)$.
for $k = 0 : n-1$ **do**
  for $j = 0 : J$ **do**
    Sample $(x_{k+1}^{(j)})_{j=1:J}$ from $p(x_{k+1}|x_k, \theta)$
    Set $\alpha^{(j)} = p(y_{k+1}|x_{k+1}^{(j)}, \theta)$
  end for
  Set $W_{k+1}^{(j)} = \frac{\alpha^{(j)}}{\sum_{l=1}^{J} \alpha^{(l)}}$ and $L = L\frac{1}{J}\sum_j \alpha^{(l)}$
  Resample $(x_{0:k+1}^{(j)})_{j=1:J}$ from $W_{k+1}^{(j)}$
end for

A gaussian kernel $q(.|\theta^{(i)})$ is used in the PMCMC algorithm, with mean $\theta^{(i)}$ and fixed variance Σ^q (random walk Metropolis Hastings).

Initialization

PMCMC algorithm is very sensitive to initialization of both the parameter values $\theta^{(0)}$ and the covariance matrix Σ^q. Several steps of initialization are therefore used.

Firstly, parameter values are initialized by maximum likelihood through simplex algorithm on a deterministic version of the model. We apply the simplex algorithm to a set of 1000 points sampled in the prior distributions and we select the parameter set with the highest likelihood.

Secondly, in order to initialise the covariance matrix, an adaptative MCMC (Metropolis Hastings) framework is used (Roberts and Rosenthal, 2009; Dureau et al., 2013a). It uses the empirical covariance of the chain $\Sigma^{(i)}$, and aims to calibrate the acceptance rate of the algorithm to an optimal value. The transition kernel is also mixed (with a probability $\alpha = 0.05$) with another gaussian using the identity matrix to improve mixing properties.

$$q^A(.|x^{(i)}) = \alpha N(x^{(i)}, \lambda\frac{2.38^2}{d}Id) + (1-\alpha)N(x^{(i)}, \lambda\frac{2.38^2}{d}\Sigma^{(i)})$$

The parameter λ is approximated by successive iterations using the empirical acceptance rate of the chain.

$$\lambda_{i+1} = \lambda_i a^i(AccRate_i - 0.234)$$

The adaptative PMCMC algorithm itself may have poor mixing properties without initialization. A first estimation of the covariance matrix is computed using KMCMC algorithm (Dureau et al., 2013a). In the KMCMC algorithm, the model is simulated with stochastic differential equations (intermediate between deterministic and Poisson with stochastic rates frameworks) and the SMC

part of the adaptative PMCMC is replaced by the extended Kalman filter. When convergence is reached with KMCMC, then, adaptative PMCMC is used.

The PMCMC algorithm is finally applied on the output of the adaptative PMCMC, using 50,000 iterations and 10,000 particles. Calculations are performed with SSM software (Dureau et al., 2013a) and R version 3.2.2.

R_0 calculation

R_0 is calculated using the Next Generation Matrix approach (NGM) (Diekmann et al., 2010).

R_0 calculation in Pandey model

$$F = \begin{pmatrix} 0 & 0 & 0 & \beta_H \\ 0 & 0 & 0 & 0 \\ 0 & \beta_V & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \qquad V = \begin{pmatrix} -\sigma & 0 & 0 & 0 \\ \sigma & -\gamma & 0 & 0 \\ 0 & 0 & -(\mu+\tau) & 0 \\ 0 & 0 & \tau & -\mu \end{pmatrix}$$

Then we have,

$$V^{-1} = \begin{pmatrix} -1/\sigma & 0 & 0 & 0 \\ -1/\gamma & -1/\gamma & 0 & 0 \\ 0 & 0 & -1/(\mu+\tau) & 0 \\ 0 & 0 & -\tau/[\mu(\tau+\mu)] & -1/\mu \end{pmatrix}$$

and

$$FV^{-1} = \begin{pmatrix} 0 & 0 & -\beta_H\tau/[\mu(\tau+\mu)] & -\beta_H/\mu \\ 0 & 0 & 0 & 0 \\ -\beta_V/\gamma & -\beta_V/\gamma & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

We calculate the eigen values α of $-FV^{-1}$:

$$\begin{vmatrix} -\alpha & 0 & \beta_H\tau/[\mu(\tau+\mu)] & \beta_H/\mu \\ 0 & -\alpha & 0 & 0 \\ \beta_V/\gamma & \beta_V/\gamma & -\alpha & 0 \\ 0 & 0 & 0 & -\alpha \end{vmatrix} = \alpha^2\left(\alpha^2 - \frac{\beta_H\beta_V\tau}{\gamma\mu(\tau+\mu)}\right) = 0$$

Then $\alpha = 0$ or $\alpha = \pm\sqrt{\frac{\beta_H\beta_V\tau}{\gamma\mu(\tau+\mu)}}$ and the highest eigenvalue is $R_0 = \sqrt{\frac{\beta_H\beta_V\tau}{\gamma\mu(\tau+\mu)}}$.

This formula defines R_0 as "the number of secondary cases per generation" (Dietz, 1993): i.e R_0 can be written as the geometric mean $R_0 = \sqrt{R_0^v R_0^h}$, where R_0^v is the number of infected mosquitoes after the introduction of one infected human in a naive population, and R_0^h is the number of infected humans after the introduction of one infected mosquito in a naive population. With this definition, herd immunity is reached when $(1 - R_0^{-2})$ of the population is vaccinated (Dietz, 1993).

R_0 calculation in Laneri model

Following the analogy with Pandey model, we compute the spectral radius of the NGM for the Laneri model.

$$F = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & \beta\tau & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \qquad V = \begin{pmatrix} -\sigma & 0 & 0 & 0 \\ \sigma & -\gamma & 0 & 0 \\ 0 & 0 & -\tau & 0 \\ 0 & 0 & \tau & -\tau \end{pmatrix}$$

Then we have,

$$V^{-1} = \begin{pmatrix} -1/\sigma & 0 & 0 & 0 \\ -1/\gamma & -1/\gamma & 0 & 0 \\ 0 & 0 & -1/\tau & 0 \\ 0 & 0 & -1/\tau & -1/\tau \end{pmatrix}$$

and

$$FV^{-1} = \begin{pmatrix} 0 & 0 & -1/\tau & -1/\tau \\ 0 & 0 & 0 & 0 \\ -\beta\tau/\gamma & -\beta\tau/\gamma & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

We calculate the eigen values α of $-FV^{-1}$:

$$\begin{vmatrix} -\alpha & 0 & 1/\tau & 1/\tau \\ 0 & -\alpha & 0 & 0 \\ \beta\tau/\gamma & \beta\tau/\gamma & -\alpha & 0 \\ 0 & 0 & 0 & -\alpha \end{vmatrix} = \alpha^2 \left(\alpha^2 - \frac{\beta\tau}{\gamma\tau} \right) = 0$$

Then $\alpha = 0$ or $\alpha = \pm\sqrt{\frac{\beta}{\gamma}}$ and the highest eigenvalue is $\alpha_R = \sqrt{\frac{\beta}{\gamma}}$.

Because λ and κ can be seen as parameters rather than state variables, the interpretation of the spectral radius as the R_0 of the model is not straightforward. Therefore, we computed the R_0 of the model through simulations, by counting the number of secondary infections after the introduction of a single infected individual in a naive population. Since Laneri model is considered here as a vector model, the number of infected humans after the introduction of a single infected is considered as R_0^2. We simulated 1000 deterministic trajectories, using parameter values sampled in the posterior distributions for all parameters except initial conditions. With this method, the confidence intervals for number of infected humans (R_0^2) are similar to the ones of α_R^2 estimated by the model. As a consequence, R_0 was approximated by the spectral radius of the NGM in our results with our stochastic framework (cf. Table 4.5).

As a robustness check, the same method was applied to Pandey model : the confidence intervals for the number of secondary cases in simulations are very similar to the ones of R_0^2 (cf. Table 4.5).

	Pandey model	Laneri model
Yap	3.1 (2.5-4.3)	2.2 (1.9-2.6)
Moorea	2.6 (2.2-3.3)	1.8 (1.6-2.0)
Tahiti	2.4 (2.0-3.2)	1.6 (1.5-1.7)
New Caledonia	2.0 (1.8-2.2)	1.6 (1.5-1.7)

Table 4.5 – **Square root of the number of secondary cases after the introduction of a single infected individual in a naive population.** Median and 95% credible intervals of 1000 deterministic simulations using parameters from the posterior distribution.

Sensitivity analysis

In order to analyse the influence of parameter values on the model's outputs, a sensitivity analysis was performed, using LHS/PRCC technique (Blower and Dowlatabadi, 1994), on Tahiti example. Similar results were obtained for the other settings. Three criteria were retained as outputs for the analysis: the seroprevalence at the last time point, the intensity of the peak of the outbreak and the date of the peak. We used uniform distributions for all parameters, which are listed in Tables 4.6 and 4.7. For model parameters, we used the same range as for the prior distribution. For initial conditions, the observation rate r and the fraction involved in the epidemic ρ, we used the 95% confidence interval obtained by PMCMC, in order to avoid unrealistic scenarios.

Parameters	Distribution	Seroprevalence	Peak intensity	Peak date
Model parameters				
R_0^2	Uniform[0,20]	0.87	0.90	-0.55
β_V	Uniform[0,10]	-0.66	-0.73	0.35
γ^{-1}	Uniform[4,7]	-0.25	0.10	0.20
σ^{-1}	Uniform[2,7]	-0.03	-0.10	0.15
τ^{-1}	Uniform[4,20]	-0.05	-0.07	0.06
μ^{-1}	Uniform[4,30]	-0.56	-0.70	0.49
Initial conditions				
Hi_0	Uniform[2.10^{-5},0.011]	0.05	-0.02	0.02
Vi_0	Uniform[10^{-4},0.034]	0.11	-0.00	-0.26
Fraction involved and observation model				
ρ	Uniform[0.46,0.54]	0.47	0.15	-0.03
r	Uniform[0.048,0.072]	-0.04	0.03	0.05

Table 4.6 – **Sensitivity analysis in Pandey model. Tahiti island.** 1000 parameter sets were sampled with latin hypercube sampling (LHS), using *lhs* R package (Carnell, 2016). On each parameter set, the model was simulated deterministically in order to explore variability in parameters without interference with variations due to stochasticity. PRCC were computed using the "sensitivity" R package (Pujol et al., 2016).

For all criteria, the key parameters in both models are transmission parameters (R_0 and β_V). High values for R_0 are positively correlated with a large seroprevalence and a high and early peak. On the contrary, high values for the parameters introducing a delay in the model, the incubation periods in human (σ^{-1}) and in mosquito (τ^{-1}), are associated with a lower and later peak, and have no significant effect on seroprevalence. Moreover, the simulations are clearly sensitive to the other model parameters, in particular the mosquito lifespan (μ^{-1}) in Pandey model.
Concerning other parameters, the initial conditions have a noticeable effect on the date of the peak only. As expected, the fraction involved in the epidemic (ρ) influences the magnitude of the

outbreak, by calibrating the proportion of people than can be infected, but it has no significant effect on the timing of the peak.

Parameters	Distribution	Seroprevalence	Peak intensity	Peak date
Model parameters				
R_0^2	Uniform[0,20]	0.62	0.93	-0.50
γ^{-1}	Uniform[4,7]	0.01	0.62	0.15
σ^{-1}	Uniform[2,7]	-0.03	-0.54	0.21
τ^{-1}	Uniform[4,20]	-0.03	-0.70	0.47
Initial conditions				
Hi_0	Uniform[10^{-5},0.015]	0.05	0.02	-0.32
L_0	Uniform[2.10^{-5},0.004]	0.05	0.00	-0.16
Fraction involved and observation model				
ρ	Uniform[0.49,0.59]	0.80	0.34	0.02
r	Uniform[0.048,0.068]	-0.01	0.01	-0.02

Table 4.7 – **Sensitivity analysis in Laneri model. Tahiti island.** 1000 parameter sets were sampled with latin hypercube sampling (LHS), using *lhs* R package (Carnell, 2016). On each parameter set, the model was simulated deterministically in order to explore variability in parameters without interference with variations due to stochasticity. PRCC were computed using the "sensitivity" R package (Pujol et al., 2016).

Complementary results

These complementary results include PMCMC results for both models in the four settings: the epidemic trajectories regarding the human compartments for infected and recovered individuals (Figures 4.4-4.5), the detailed posterior distributions for all parameters (Figures 4.6-4.13) and correlation plots for all models (Figures 4.14-4.21).

Infected and recovered

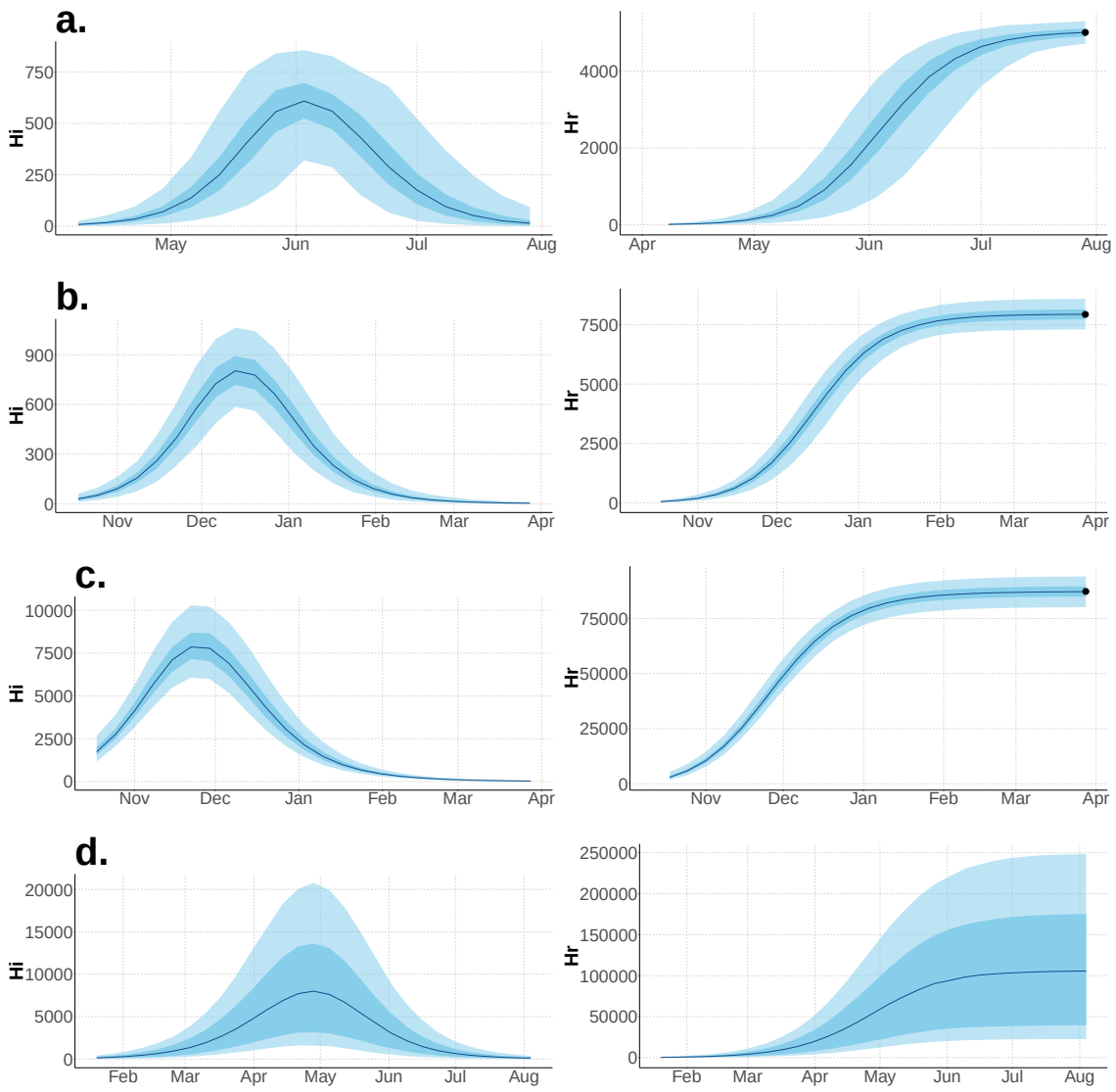


Figure 4.4 – **Infected and recovered humans evolution during the outbreak with Pandey model.** Simulations from the posterior density: posterior median (solid line), 95% and 50% credible intervals (shaded blue areas) and observed seroprevalence (black dots). First column: Infected humans (H_I). Second column: Recovered humans (H_R). a) Yap. b) Moorea. c) Tahiti. d) New Caledonia.

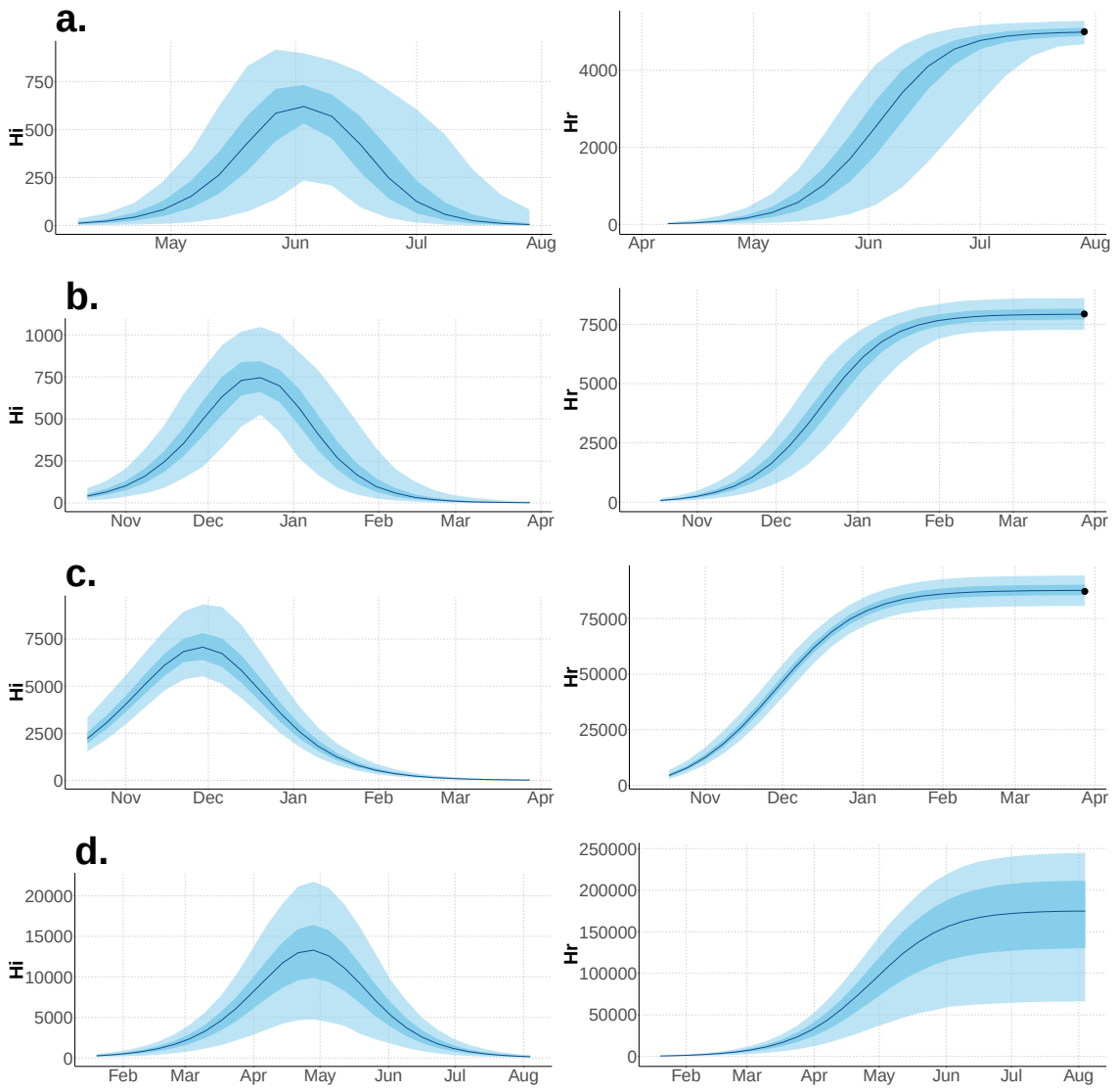


Figure 4.5 – **Infected and recovered humans evolution during the outbreak with Laneri model.** Simulations from the posterior density: posterior median (solid line), 95% and 50% credible intervals (shaded blue areas) and observed seroprevalence (black dots). First column: Infected humans (H_I). Second column: Recovered humans (H_R). a) Yap. b) Moorea. c) Tahiti. d) New Caledonia.

Posterior distributions

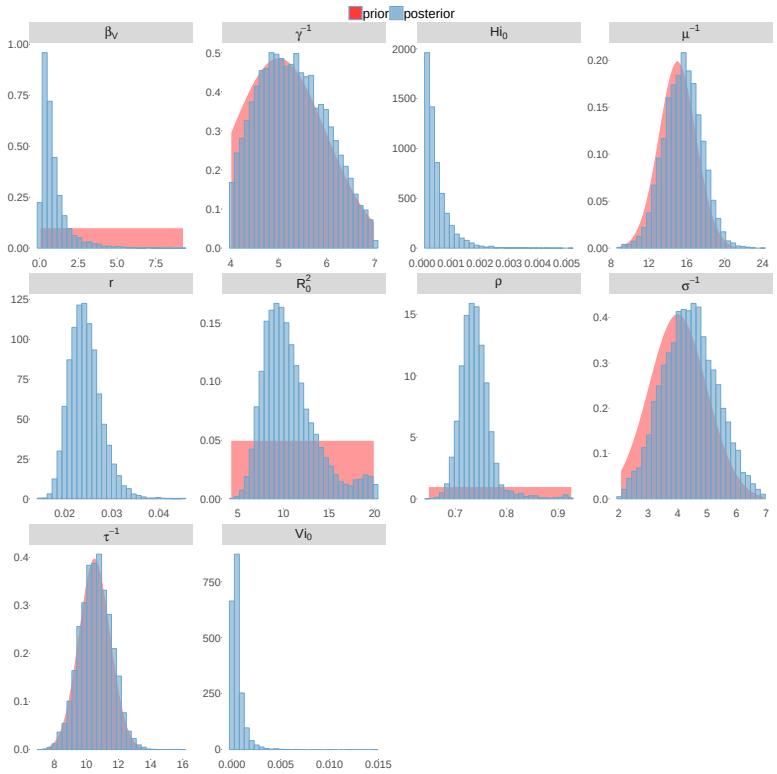


Figure 4.6 – **Posterior distributions. Pandey model, Yap island.**

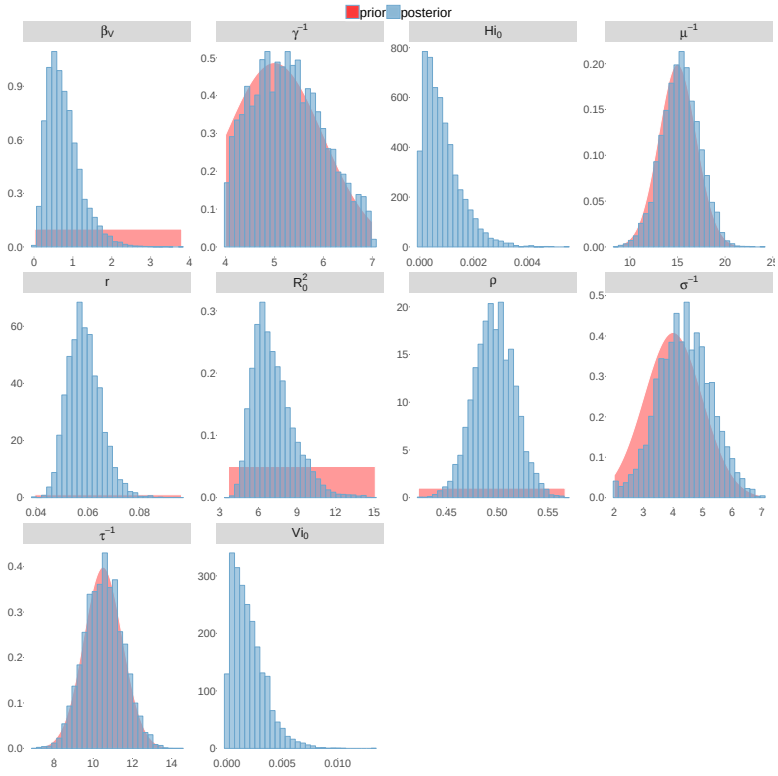


Figure 4.7 – **Posterior distributions. Pandey model, Moorea island.**

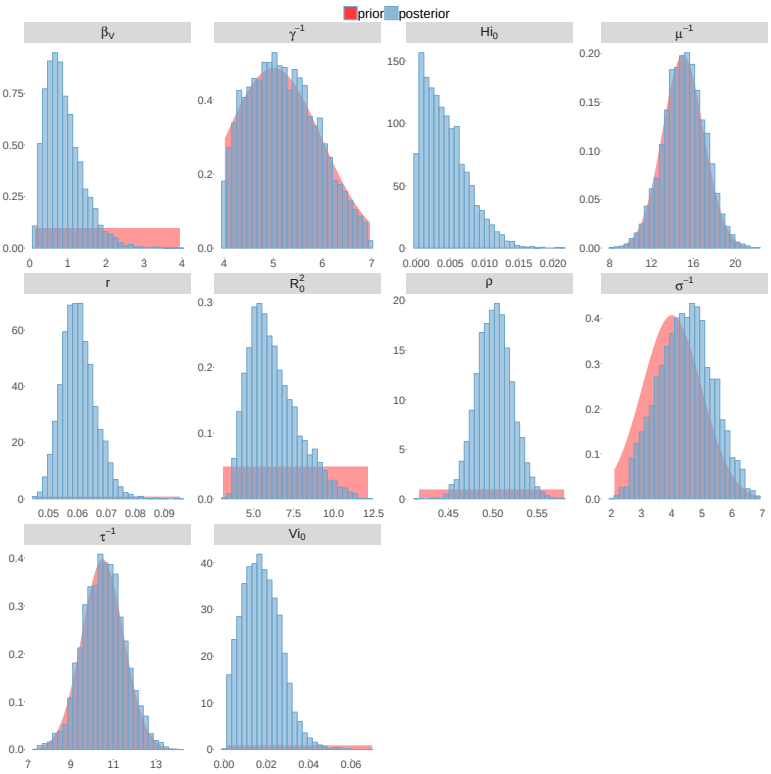


Figure 4.8 – **Posterior distributions. Pandey model, Tahiti island.**

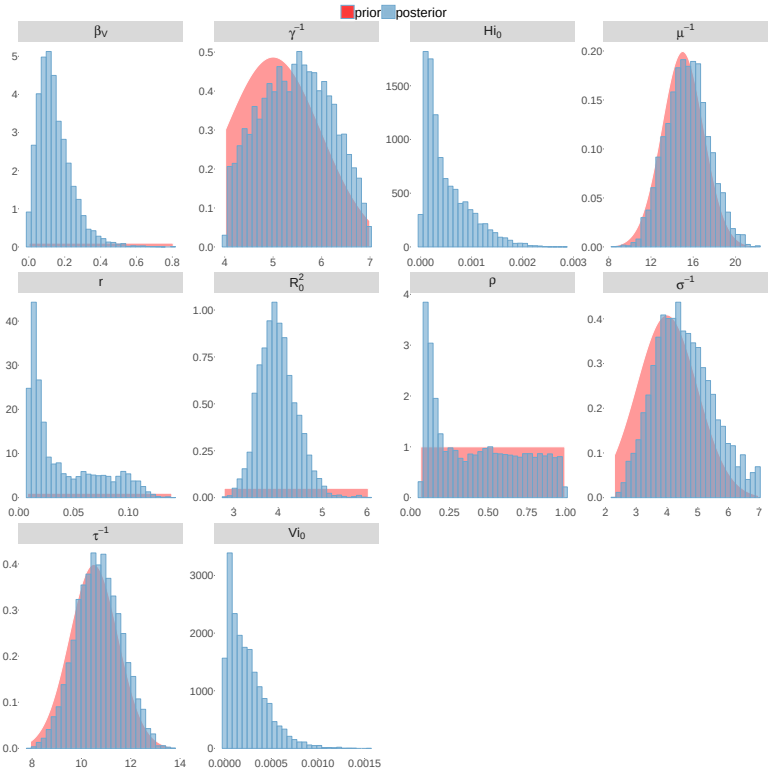


Figure 4.9 – **Posterior distributions. Pandey model, New Caledonia.**

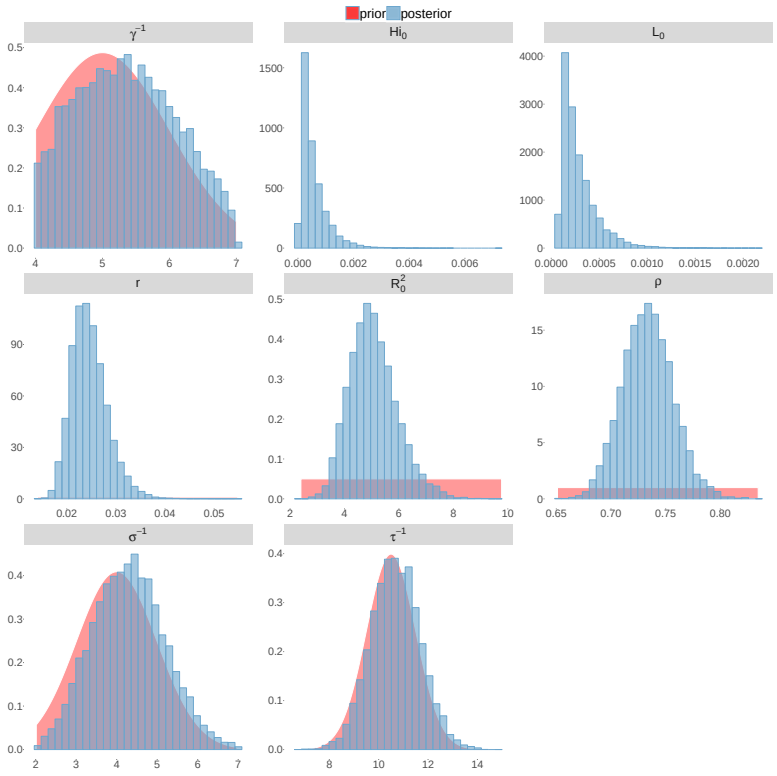


Figure 4.10 – **Posterior distributions. Laneri model, Yap island.**

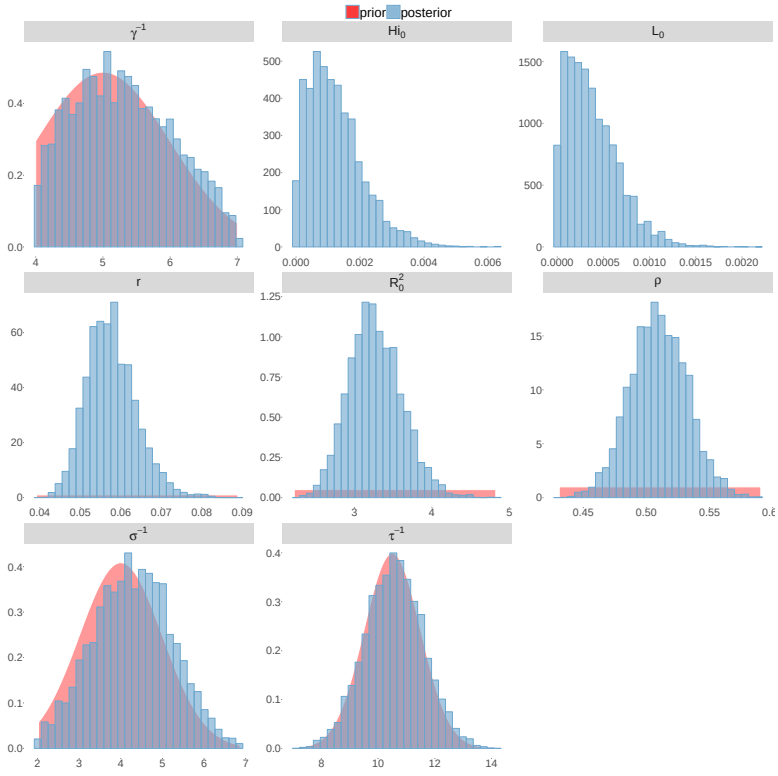


Figure 4.11 – **Posterior distributions. Laneri model, Moorea island.**

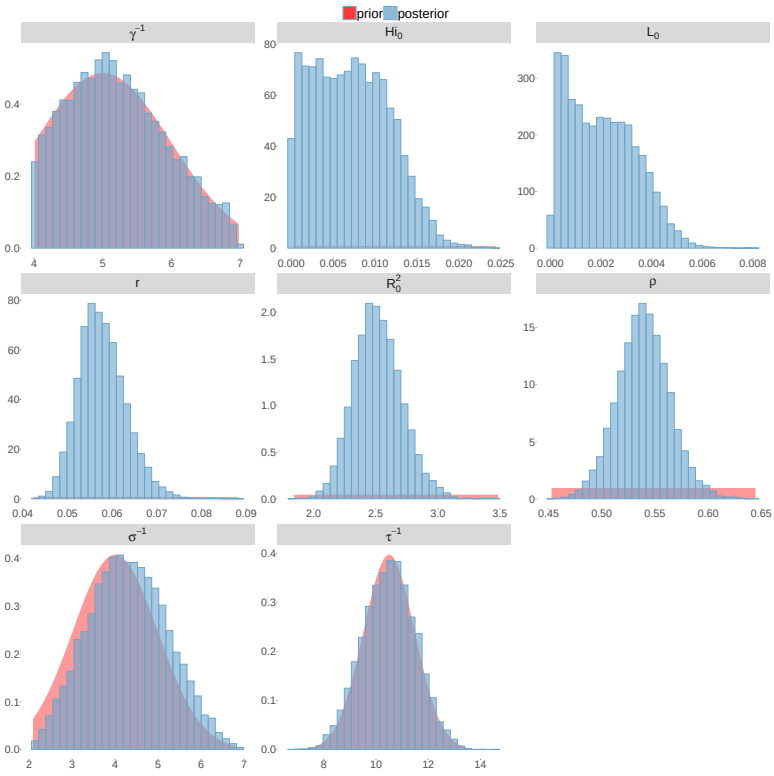


Figure 4.12 – **Posterior distributions. Laneri model, Tahiti island.**

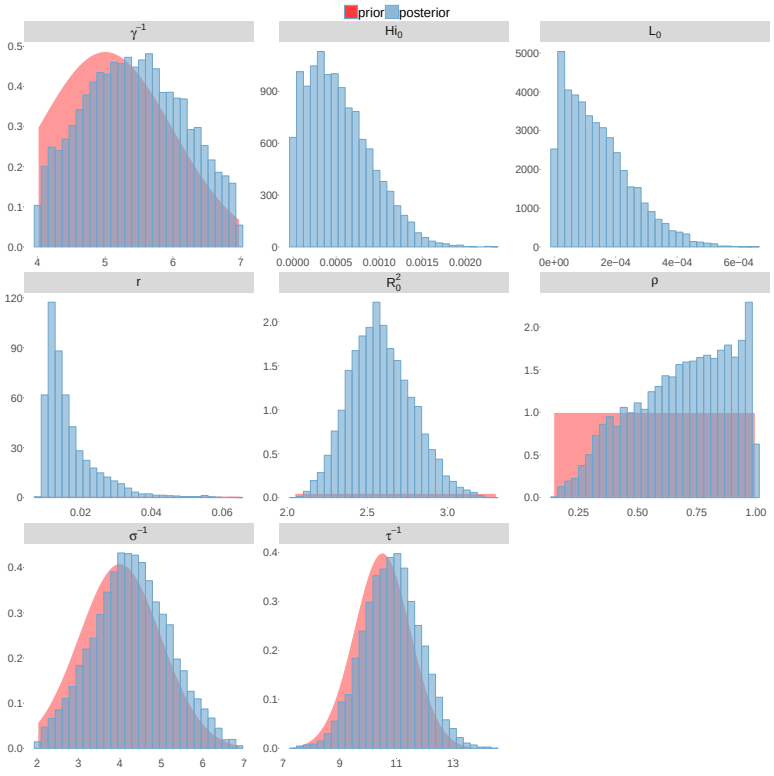


Figure 4.13 – **Posterior distributions. Laneri model, New Caledonia.**

Correlation between estimated parameters

The inference technique may fail to estimate some parameters due to identifiability issues. In particular, when two parameters are highly correlated to one another, the model manages to estimate the pair of parameters but not each one separately. The analysis of correlation between parameters' posterior distributions can reveal such cases. The following graphics display for each model the correlation coefficients between all pairs of parameters across the MCMC chain. For example, in models for New Caledonia, the observation rate and the fraction of the population involved in the epidemic are strongly negatively correlated (Figures 4.9 and 4.13): the inference technique does not manage to estimate properly these two parameters, due to the lack of information on seroprevalence.

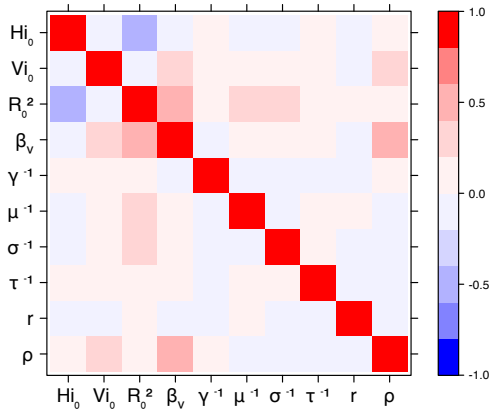


Figure 4.14 – **Correlation plot of MCMC output. Pandey model, Yap island.**

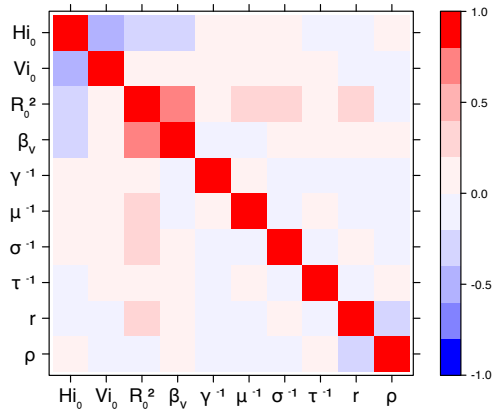


Figure 4.15 – **Correlation plot of MCMC output. Pandey model, Moorea island.**

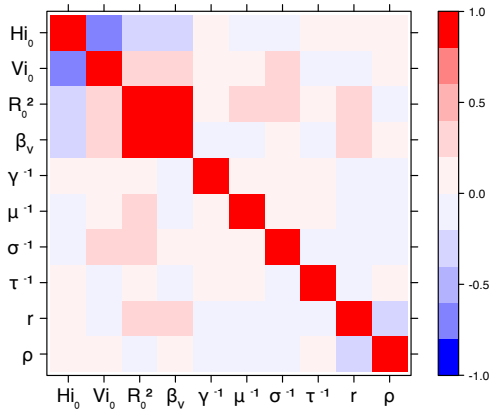


Figure 4.16 – **Correlation plot of MCMC output. Pandey model, Tahiti island.**

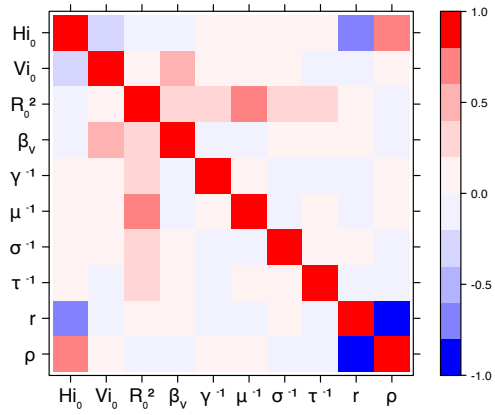


Figure 4.17 – **Correlation plot of MCMC output. Pandey model, New Caledonia.**

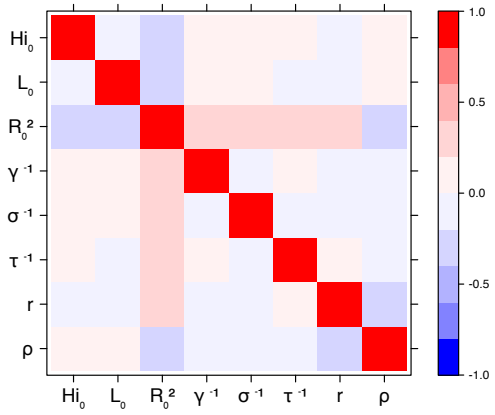


Figure 4.18 – **Correlation plot of MCMC output. Laneri model, Yap island.**

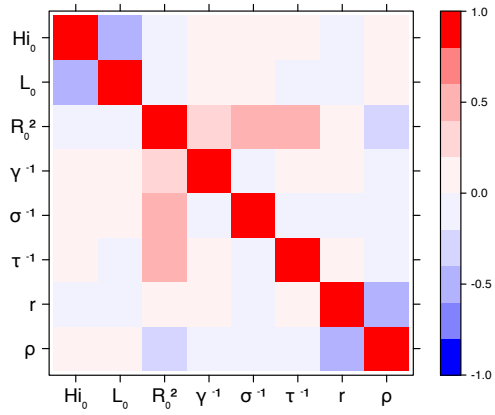


Figure 4.19 – **Correlation plot of MCMC output. Laneri model, Moorea island.**

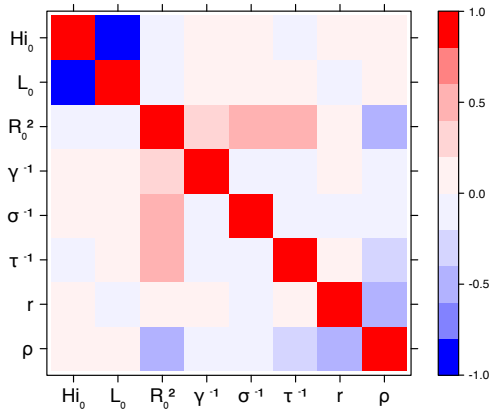


Figure 4.20 – **Correlation plot of MCMC output. Laneri model, Tahiti island.**

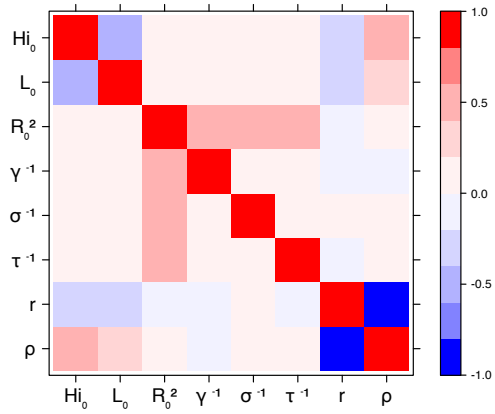


Figure 4.21 – **Correlation plot of MCMC output. Laneri model, New Caledonia.**

Acknowledgments

CC, DGS and BC are partially supported by the "Pépiniere interdisciplinaire Eco-Evo-Devo" from the Centre National de la Recherche Scientifique (CNRS). The research leading to these results has also received funding from the European Commission Seventh Framework Program [FP7/ 2007-2013] for the DENFREE project under Grant Agreement n° 282 378. The funders played no role in the study design, data collection, analysis, or preparation of the manuscript.

4.3 Interactions between Zika and dengue

Introduction

Many pathogens coexist and sometimes interact in an environment, and because these interactions can strongly influence disease dynamics, it is important to take them into account in mathematical models (Opatowski et al., 2018; Rohani et al., 1998; Vasco et al., 2007; Randuineau, 2015). This is particularly the case for dengue, due to the presence of four interacting virus serotypes: the infection by one serotype confers full immunity against this serotype but the effect against other serotypes is more complex, ranging from different forms of cross-protection to disease enhancement. Among the biological explanations for these phenomena, is a mechanism called antibody dependent enhancement (ADE) by which previous dengue antibody would have a counter-productive role and enhance a secondary infection by a heterologous serotype (Halstead, 2007). There is therefore an extensive literature on modeling these cross-reactions in dengue (Andraud et al., 2012; ten Bosch et al., 2016). On the one hand, cross immunity has been considered in many forms, such as partial immunity (Adams and Boots, 2006), temporary immunity (Wearing and Rohani, 2006), or clinical cross-protection (protection against the clinical manifestation of the disease but not against the infection itself) (Nagao and Koelle, 2008; Chikaki and Ishikawa, 2009). On the other hand, cross-enhancement has also been studied in several forms, such as the increase in infectiousness (Ferguson et al., 1999; Aguiar et al., 2011; Chikaki and Ishikawa, 2009), in susceptibility (Adams and Boots, 2006; Recker et al., 2009) or in severity (via disease induced-mortality (Adams and Boots, 2006) or hospitalization (Aguiar et al., 2011)).

Zika virus is a flavivirus closely related to dengue, that has expanded rapidly in the Pacific and Latin America since 2007. It was associated with severe complications such as Guillain Barré syndrome and microcephaly in newborns when it reached French Polynesia in 2013 and Brazil in 2015. In most of the settings affected by Zika, there was past dengue circulation and even sometimes co-circulation (Cao-Lormeau and Musso, 2014; Dupont-Rouzeyrol et al., 2015; Rodriguez-Morales et al., 2016), raising the question of possible interactions between the two pathogens. Serological analysis have already given support for such interactions (Martín-Acebes et al., 2018; Rey et al., 2018). In vitro, several studies found evidence that dengue antibodies could cause ADE on Zika virus (Dejnirattisai et al., 2016; Barba-Spaeth et al., 2016; Swanstrom et al., 2016; Durbin, 2016), or protection against it (Fernandez et al., 2017) and Kawiecki and Christofferson (2016) found that Zika may enhance dengue. In vivo, results are still more contrasted (Martín-Acebes et al., 2018; Rey et al., 2018), as enhancement of Zika infection by previous dengue immunity has be observed in mice at certain antibody concentrations (Bardina et al., 2017) but not in non-human primates (Pantoja et al., 2017), even though in this case, the reverse effect has been reported (George et al., 2017). In Zika-infected patients, however, Terzian et al. (2017) detected no sign of ADE induced by previous dengue antibodies.

These individual or clinical interaction effects may have specific consequences at the population level (Opatowski et al., 2018; Randuineau, 2015), that need to be taken into account when planning control strategies, such as vaccination campaigns (Aguiar et al., 2016). The use of epidemiological data and mathematical models has permitted the identification of pathogen interactions such as the ones between influenza and pneumococcal pneumonia (Shrestha et al., 2013; Opatowski et al., 2013). The objective of this chapter is to study the interaction between dengue and Zika at the population level through these methods, based on outbreaks that occurred in Rio de Janeiro (Brazil) and French Polynesia.

Methods

Data

Two outbreaks in two different settings are studied: the 2015-2016 outbreak in Rio de Janeiro, Brazil, and the 2013-2014 outbreak in French Polynesia.

Rio de Janeiro, Brazil

For dengue and Zika data in Rio de Janeiro (cf. Figure 4.22), we used the number of new cases per week between 2015-01-11 and 2016-11-13 from Rio's government website (updated on 2016-11-16) (SMS, 2016a,b). We also used temperature data from the Santos Dumont Airport, from the website Weather Underground (WeatherUnderground, 2016).

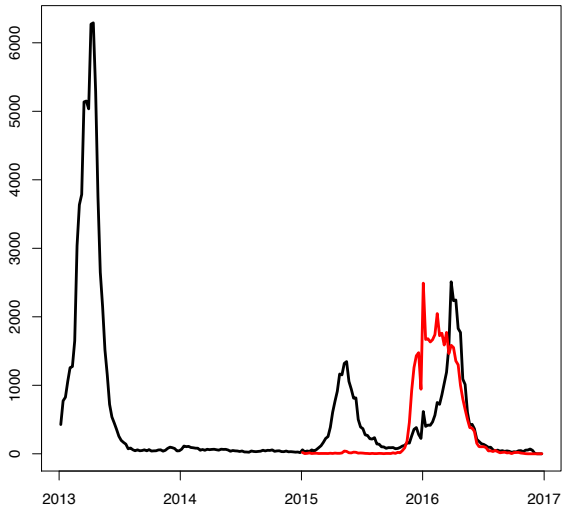
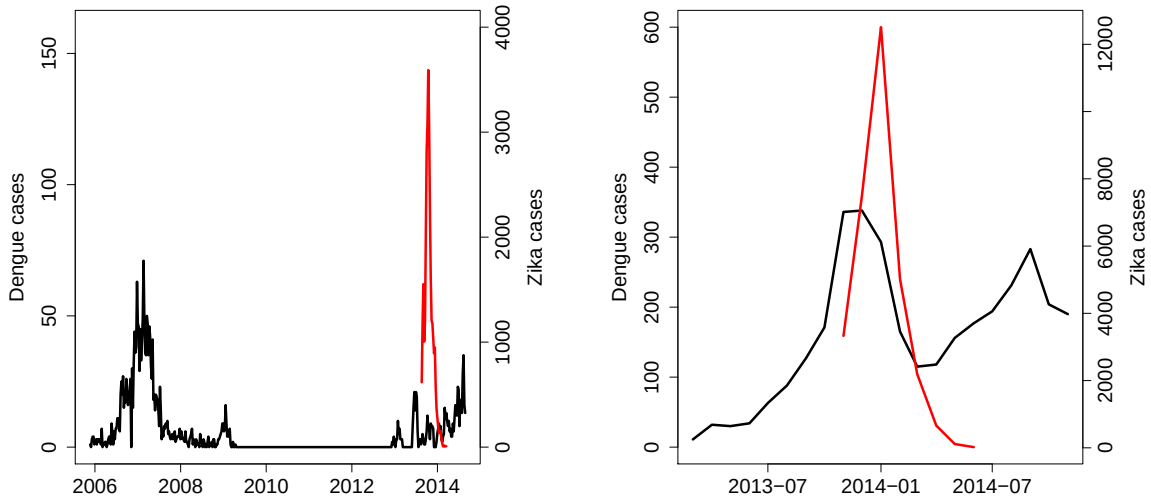


Figure 4.22 – **Surveillance data in Rio de Janeiro.** Dengue cases (black) and Zika cases (red) per week.

Polynesia

For Zika case data in French Polynesia, we used data from the surveillance system (Mallet et al., 2015). For dengue cases, two sources of data were compared. On the one hand, we used confirmed DENV1 cases from the Institut Louis Malardé, that were available since 2006: these confirmed cases may not be representative of the real dengue trend as the proportion of cases tested may have changed over the period. On the other hand, we used monthly positive cases through surveillance (updated on 2014-28-03) (CHSP, 2014). The total population of French Polynesia was $N = 271,800$ in 2014 (ISPF, 2014).

In addition, several sero-surveys were conducted in French Polynesia. Using the serosurvey by Aubry et al. (2017), we add an observation on the seroprevalence for Zika (133,180 on 2014-10-24, calculated as $0.49 * N$). Using the serosurvey among blood donor by Aubry et al. (2015a), we add an observation on the seroprevalence for DENV1 (176,668 recovered on 2013-02-01, calculated as $0.65 * N$). Finally, using the serosurvey among blood donor by Aubry et al. (2015a), we assume that the initial proportion of susceptibles to DENV1 in 2013 is 35%.



a. b.

Figure 4.23 – **Data in French Polynesia.** a) Dengue (DENV1) confirmed cases (black) and Zika surveillance cases (red), weekly data. b) Dengue (black) and Zika (red) surveillance cases, monthly data (Zika data is aggregated per month to have the same frequency of observation in both datasets, with linear interpolation in weeks overlapping two months).

Models

Transmission model

We consider a model with two pathogens, dengue and Zika (underscript D for dengue and Z for Zika), using an SEIR (Susceptible-Exposed-Infected-Recovered) dynamic for each disease, with possible co-infections (cf. Figure 4.24). For parsimony in the number of parameters, we did not model explicitly the vector population. The model was considered in the deterministic framework, with the following equations :

$$
\begin{aligned}
\frac{dH_{S_D S_Z}}{dt} &= \mu N - (\lambda_D + \lambda_Z) H_{S_D S_Z} - \mu H_{S_D S_Z} \\
\frac{dH_{E_D S_Z}}{dt} &= \lambda_D H_{S_D S_Z} - (\lambda_Z + \sigma_D + \mu) H_{E_D S_Z} \\
\frac{dH_{I_D S_Z}}{dt} &= \sigma_D H_{E_D S_Z} - (\tilde{\lambda_Z} + \gamma_D + \mu) H_{I_D S_Z} \\
\frac{dH_{R_D S_Z}}{dt} &= \gamma_D H_{I_D S_Z} - (\tilde{\lambda_Z} + \mu) H_{R_D S_Z} \\
\frac{dH_{S_D E_Z}}{dt} &= \lambda_Z H_{S_D S_Z} - (\lambda_D + \sigma_Z + \mu) H_{S_D E_Z} \\
\frac{dH_{E_D E_Z}}{dt} &= \lambda_D H_{S_D E_Z} + \lambda_Z H_{E_D S_Z} - (\sigma_Z + \sigma_D + \mu) H_{E_D E_Z} \\
\frac{dH_{I_D E_Z}}{dt} &= \sigma_D H_{E_D E_Z} + \tilde{\lambda_Z} H_{I_D S_Z} - (\sigma_Z + \gamma_D + \mu) H_{I_D E_Z} \\
\frac{dH_{R_D E_Z}}{dt} &= \gamma_D H_{I_D E_Z} + \tilde{\lambda_Z} H_{R_D S_Z} - (\sigma_Z + \mu) H_{R_D E_Z} \\
\frac{dH_{S_D I_Z}}{dt} &= \sigma_Z H_{S_D E_Z} - (\tilde{\lambda_D} + \gamma_Z + \mu) H_{S_D I_Z} \\
\frac{dH_{E_D I_Z}}{dt} &= \tilde{\lambda_D} H_{S_D I_Z} + \sigma_Z H_{E_D E_Z} - (\gamma_Z + \sigma_D + \mu) H_{E_D I_Z} \\
\frac{dH_{I_D I_Z}}{dt} &= \sigma_D H_{E_D I_Z} + \sigma_Z H_{I_D E_Z} - (\gamma_Z + \gamma_D + \mu) H_{I_D I_Z} \\
\frac{dH_{R_D I_Z}}{dt} &= \gamma_D H_{I_D I_Z} + \sigma_Z H_{R_D E_Z} - (\gamma_Z + \mu) H_{R_D I_Z} \\
\frac{dH_{S_D R_Z}}{dt} &= \gamma_Z H_{S_D I_Z} - (\tilde{\lambda_D} + \mu) H_{S_D R_Z} \\
\frac{dH_{E_D R_Z}}{dt} &= \tilde{\lambda_D} H_{S_D R_Z} + \gamma_Z H_{E_D I_Z} - (\sigma_D + \mu) H_{E_D R_Z} \\
\frac{dH_{I_D R_Z}}{dt} &= \sigma_D H_{E_D R_Z} + \gamma_Z H_{I_D I_Z} - (\gamma_D + \mu) H_{I_D R_Z} \\
\frac{dH_{R_D R_Z}}{dt} &= \gamma_D H_{I_D R_Z} + \gamma_Z H_{R_D I_Z} - \mu H_{R_D R_Z}
\end{aligned}
$$

with, in the model without interactions,

$$
\begin{aligned}
\lambda_D = \tilde{\lambda_D} &= \beta_D \frac{H_{I_D S_Z} + H_{I_D E_Z} + H_{I_D I_Z} + H_{I_D R_Z} + i_D}{N} \\
\lambda_Z = \tilde{\lambda_Z} &= \beta_Z \frac{H_{S_D I_Z} + H_{E_D I_Z} + H_{I_D I_Z} + H_{R_D I_Z} + i_Z}{N}
\end{aligned}
$$

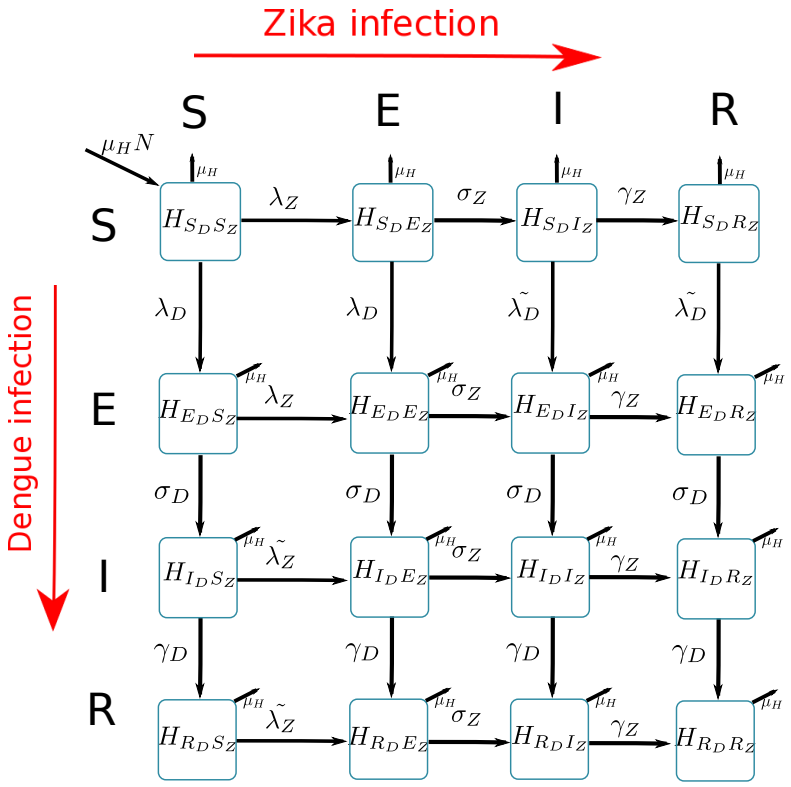


Figure 4.24 – **Zika-Dengue model, with co-infections.** Underscripts D and Z refer to dengue and Zika respectively. For each disease, there are H_S susceptible, H_E exposed (infected but not yet infectious), H_I infectious, and H_R recovered individuals. H_E individuals move to the infectious class H_I at rate σ and H_I then recover at rate γ. The force of infection is λ at primary infection, and $\tilde{\lambda}$ at secondary infection, and it depends on a transmission parameter β. The total population $\sum_{(i,j)\in\{S,E,I,R\}^2} H_{i_D j_Z} = N$ is constant, and the recruitment rate equals the mortality rate μ_H. i is an import parameter to prevent the extinction of the disease.

Observation model

Observation model in Rio de Janeiro In Rio de Janeiro, since Zika reporting became mandatory after the 43rd epidemic week (from 18 to 24th October 2015) (Villela et al., 2017), we separated the series in two periods. Before 2015-10-25, we estimate the total of dengue and Zika infections (considering possible misdiagnosis between both diseases, with observation rate ρ). After this date, we estimate the dengue series and the Zika series separately (with observations rate ρ_D and ρ_Z respectively). The observation model in every case is a negative binomial with dispersion parameter fixed to 0.1 and estimated observation rate.

Observation model in French Polynesia For both diseases, the observation model is a negative binomial dispersion parameter fixed to 0.1 and estimated observation rate. Because reporting for dengue changed after the 36th epidemic week (CHSP, 2014), we assume that, from the 2013-09-21 on, the observation rate for dengue in Malardé data is multiplied by an estimated factor $p \in [0, 1]$. The observation model for seroprevalence is a discretized normal distribution with fixed variance.

	Mean (m)	Variance (v)
Rio de Janeiro		
Zika + Dengue	$m_{DZ} = \rho(H_{S_DI_Z} + H_{E_DI_Z} + H_{I_DI_Z} + H_{R_DI_Z}$	$v_{DZ} = 1 + m_{DZ} + m_{DZ}^2\delta$
(15/01/11-15/10/18)	$+H_{I_DS_Z} + H_{I_DE_Z} + H_{I_DI_Z} + H_{I_DR_Z})$	
Zika (15/10/25-16/11/13)	$m_Z = \rho_Z(H_{S_DI_Z} + H_{E_DI_Z} + H_{I_DI_Z} + H_{R_DI_Z})$	$v_Z = 1 + m_Z + m_Z^2\delta$
Dengue (5/10/25-16/11/13)	$m_D = \rho_D(H_{I_DS_Z} + H_{I_DE_Z} + H_{I_DI_Z} + H_{I_DR_Z})$	$v_D = 1 + m_D + m_D^2\delta$
French Polynesia		
Zika	$m_Z = \rho_Z(H_{S_DI_Z} + H_{E_DI_Z} + H_{I_DI_Z} + H_{R_DI_Z})$	$v_Z = 1 + m_Z + m_Z^2\delta$
Dengue	$m_D = \rho_D(H_{I_DS_Z} + H_{I_DE_Z} + H_{I_DI_Z} + H_{I_DR_Z})$	$v_D = 1 + m_D + m_D^2\delta$
Seroprevalence	$m_R = H_{S_DR_Z} + H_{E_DR_Z} + H_{I_DR_Z} + H_{R_DR_Z}$	$v_R = 20000$

Table 4.8 – **Observation models.**

Dengue-Zika interaction

In order to model Dengue-Zika interaction, we consider that an infection with one of the disease can induce three types of change in the other disease: change in infectiousness, in susceptibility or in severity (Ferguson et al., 1999; Adams and Boots, 2006; Recker et al., 2009) . Two parameters ϕ^{ZD} and ϕ^{DZ} are introduced to model these changes: ϕ^{ZD} represents the change in dengue infections for individuals with a previous Zika infection and ϕ^{DZ} represents the change in Zika infections for individuals with a previous dengue infection. Parameters larger than 1 indicate enhancement, parameters inferior to 1 indicate protection, and parameters equal to one indicate an absence of interaction. We perform estimations in five cases :

A. $\phi^{ZD} = \phi^{DZ} = 1$: absence of interaction

B. $\phi^{ZD} = \phi^{DZ} \in [0, 2]$: symetric interaction

C. $\phi^{ZD} \in [0, 2]$ and $\phi^{DZ} \in [0, 2]$: interaction both ways , non necessarily symetric

D. $\phi^{ZD} = 1$ and $\phi^{DZ} \in [0, 2]$: influence of dengue on zika only

E. $\phi^{ZD} \in [0, 2]$ and $\phi^{DZ} = 1$: influence of zika on dengue only

Model with infectiousness enhancement After a primary infection with one pathogen, there is a change in infectiousness during infections with the second pathogen. In the baseline model (cf. Figure 4.24), the forces of infection are

$$\lambda_D = \tilde{\lambda_D} = \beta_D \frac{H_{I_DS_Z} + H_{I_DE_Z} + \phi^{ZD}(H_{I_DI_Z} + H_{I_DR_Z}) + i_D}{N}$$
$$\lambda_Z = \tilde{\lambda_Z} = \beta_Z \frac{H_{S_DI_Z} + H_{E_DI_Z} + \phi^{DZ}(H_{I_DI_Z} + H_{R_DI_Z}) + i_Z}{N}$$

Model with susceptibility enhancement After a primary infection with one pathogen, there is a change in susceptibility to infections with the second pathogen. In the baseline model (cf. Figure

4.24), the forces of infection are

$$\lambda_D = \beta_D \frac{H_{I_D S_Z} + H_{I_D E_Z} + H_{I_D I_Z} + H_{I_D R_Z} + i_D}{N}$$
$$\lambda_Z = \beta_Z \frac{H_{S_D I_Z} + H_{E_D I_Z} + H_{I_D I_Z} + H_{R_D I_Z} + i_Z}{N}$$
$$\tilde{\lambda_D} = \phi^{ZD} \beta_D \frac{H_{I_D S_Z} + H_{I_D E_Z} + H_{I_D I_Z} + H_{I_D R_Z} + i_D}{N}$$
$$\tilde{\lambda_Z} = \phi^{DZ} \beta_Z \frac{H_{S_D I_Z} + H_{E_D I_Z} + H_{I_D I_Z} + H_{R_D I_Z} + i_Z}{N}$$

Model with severity enhancement After a primary infection with one pathogen, there is a change in the reporting rate during infections with the second pathogen (severe infections are more commonly reported). In the baseline model (cf. Figure 4.24), the means of the observation models are :

$$m_D = \rho_D (H_{I_D S_Z} + H_{I_D E_Z} + \phi^{ZD} [H_{I_D I_Z} + H_{I_D R_Z}])$$
$$m_Z = \rho_Z (H_{S_D I_Z} + H_{E_D I_Z} + \phi^{DZ} [H_{I_D I_Z} + H_{R_D I_Z}])$$

In Rio de Janeiro, only change in infectiousness was considered (in configurations A, B, C, D and E). Susceptibility enhancement was not tested because seroprevalence data for dengue was not available, and severity enhancement was not tested because for a part of the series dengue and Zika were not observed separately.
In French Polynesia, change in infectiousness, susceptibility and severity were considered. In order to limit the number of models, only configurations A and C were explored because A corresponds to the absence of interaction and C is the most general case that comprises the other possibilities for interaction (B, D and E).

Estimation, parameters and priors

Estimations are performed in the Bayesian framework, with adaptive Metropolis-Hastings algorithm, using the SSM library (Dureau et al., 2013a). In order to initialize the MCMC chain, a simplex algorithm was run on 1000 parameter sets from the prior distribution. The parameter set with the highest posterior was used as initialization of the Metropolis Hastings algorithm.

The durations of infectiousness was fixed to 4.5 days for dengue (WHO, 2017) and 5 days for Zika (Mallet et al., 2015; Kucharski et al., 2016). The duration of incubation period was assumed to be the sum of the extrinsic (in mosquito) and intrinsic (in human) incubation periods, and was fixed to 12.5 days for dengue (Chan and Johansson, 2012) and 14.5 days for Zika (Hayes, 2009; Chouin-Carneiro et al., 2016; Bearcroft, 1956; Lessler et al., 2016; Champagne et al., 2016). The mortality rate was fixed to 1/75 year. The other parameters are specific to the geographical setting.

Rio de Janeiro Concerning the initial conditions, all individuals are assumed to be naive to Zika at the beginning of the studied period ($t = 0$), and therefore they are either in $H_{S_D S_Z}$, $H_{E_D S_Z}$,$H_{I_D S_Z}$ or $H_{R_D S_Z}$. $H_{I_D S_Z}(0)$ and $H_{S_D S_Z}(0)$ are estimated.
The model is then estimated in two steps. First, a preliminary model without interactions (case A) is estimated with priors as indicated in Table 4.9. Second, the observation rate for dengue ρ_D and Zika ρ_Z and the import parameters i_D and i_Z are fixed to the median a posteriori of the previous

estimation and the remaining parameters are estimated, along with interaction parameters (cases A, B, C, D and E).

Parameters		Rio de Janeiro		French Polynesia	
		Preliminary model	Final model	Malardé weekly data	Surveillance monthly data
Dengue and Zika parameters					
β_D	transmission parameter dengue	U[0.1,1]	U[0.1,1]	U[0,2]	U[0,2]
β_Z	transmission parameter Zika	U[0.1,1]	U[0.1,1]	U[0,2]	U[0,2]
ρ_D	observation rate dengue	U[0,0.25]	0.02	0.09729	U[0,0.25]
ρ_Z	observation rate Zika	U[0,0.25]	0.013	U[0,1]	U[0,0.25]
ρ	observation rate dengue+Zika	U[0,0.25]	U[0,0.25]	.	.
p	change in dengue observation rate	.	.	U[0,1]	.
i_D	import dengue	U[0,0.001]	0.00003	0	0
i_Z	import Zika	U[0,0.001]	0.00006	0	0
Initial conditions					
$H_{S_DS_Z}(0)$	initial condition	U[0,1]	U[0,1]	0.35*N	0.35*N
$H_{I_DS_Z}(0)$	initial condition	U[0,0.001]*N	U[0,0.1]*N	U[0,100]	U[0,100]
$H_{E_DS_Z}(0)$	initial condition	$H_{I_DS_Z}(0)$	$H_{I_DS_Z}(0)$	U[0,100]	U[0,100]
z	introduction Zika	.	.	U[0,10^{-3}]	$3.5.10^{-8}$
$H_{S_DI_Z}(0)$	initial condition	0	0	$z * H_{S_DS_Z}(0)$	$z * H_{S_DS_Z}(0)$
$H_{R_DI_Z}(0)$	initial condition	.	.	$z * H_{R_DS_Z}(0)$	$z * H_{R_DS_Z}(0)$
$H_{S_DE_Z}(0)$	initial condition	.	.	$z * H_{S_DS_Z}(0)$	$z * H_{S_DS_Z}(0)$
$H_{R_DE_Z}(0)$	initial condition	.	.	$z * H_{R_DS_Z}(0)$	$z * H_{R_DS_Z}(0)$

Table 4.9 – **Prior distributions and fixed parameters.**

Because seasonality is a key element in dengue dynamics in Rio de Janeiro, a seasonality parameter s is multiplied to the force of infection. The time spanned by the data is too short to estimate precisely the amplitude and phase of a sinusoidal forcing. Moreover, an atypic epidemic was observed in 2014, unusually small, and that was not well fitted by sinusoidal forcing. We therefore decided to rely on climatic indicators, such as temperature, following the example of Lourenço and Recker (2014).

The seasonal forcing is calculated as the moving average over 4 weeks of the following function (where T is the mean temperature):

$$s = \left(0.0729T^2 - 0.9037\right)\left(0.001044T(T - 12.286)\sqrt{32.461 - T}\right)$$

This formulation is derived from Lourenço and Recker (2014) using values from Lambrechts et al. (2011). However, in the work of Lourenço and Recker (2014), the mosquito vector is explicitly modeled, and therefore probability of infection from human to mosquito and from mosquito to human are separated from one another. In our case, since these two probabilities cannot be distinguished, and given that they are both multiplied in the resulting reproduction number, we took the product of both quantities.

Polynesia When Malardé weekly data is used, a preliminary model for dengue alone is fitted using data from 2006 to 2014, in order to estimate the observation rate for dengue ρ_D. Then, the model with dengue and Zika is fitted from 2013 to 2014, using prior distributions as in Table 4.9.

When surveillance monthly data is used, all parameters are estimated using prior distributions as in Table 4.9, except the introduction parameter for Zika (z), which is fixed to the value found in the Malardé weekly data model.

Results

Rio de Janeiro

In Rio de Janeiro, the five estimated models manage qualitatively to reproduce the observed data, except during the transition period due to the change in reporting (cf. Figure 4.25). According to the DIC criterion, the best models is the one including an influence of Zika on dengue infection, even though the DIC values are closed to one another (cf. Table 4.10).

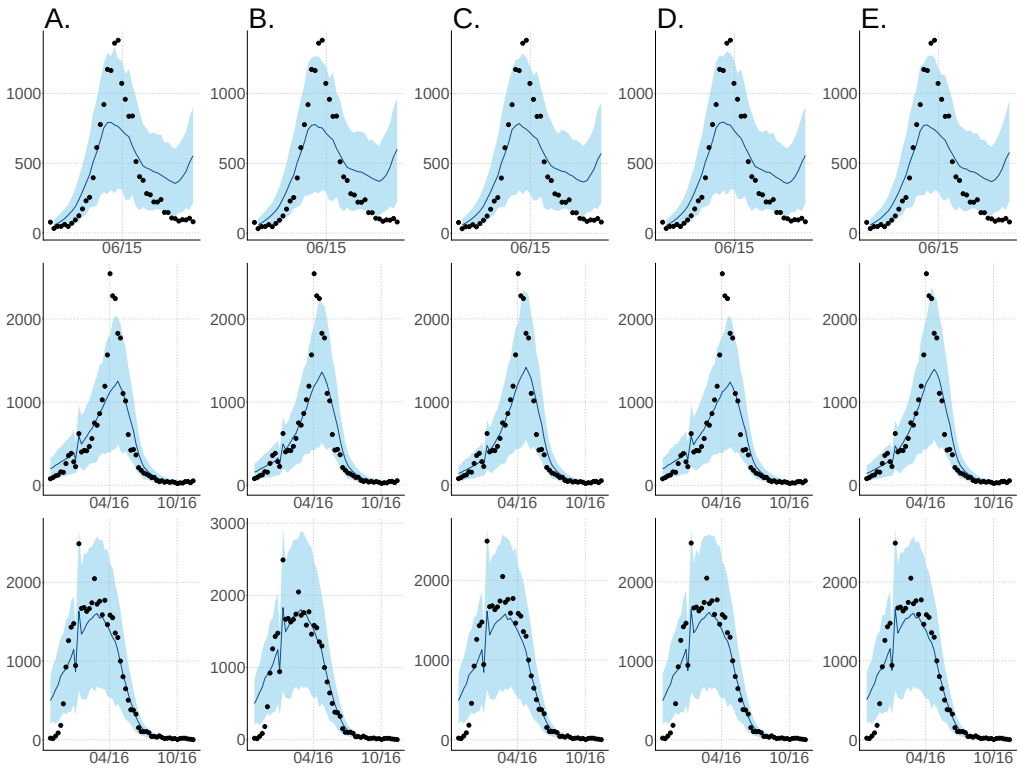


Figure 4.25 – **Simulations of the fitted model and observations in Rio de Janeiro (black dots).** Simulations with observation noise of the fitted models using parameter sets from the posterior distribution (median and 95% credible intervals). First row: total zika+dengue until October 2015. Second row: dengue since October 2015. Third row: zika since October 2015. The three data series (dengue+zika until October 2015, dengue since October 2015 and Zika since October 2015) were fitted simultaneously. Columns correspond respectively to cases A,B,C,D and E, with interaction parameters in the infectiousness enhancement case.

	A $\phi^{ZD} = \phi^{DZ} = 1$	**B** $\phi^{ZD} = \phi^{DZ} \in [0,2]$	**C** $\phi^{ZD} \neq \phi^{DZ} \in [0,2]$	**D** $\phi^{DZ} \in [0,2]$	**E** $\phi^{ZD} \in [0,2]$
log-likelihood	-975	-971	-967	-975	-967
DIC	1960	1953	1948	1962	1946
nb parameters	5	6	7	6	6
ϕ^{DZ}	1	1.3 (1.13-1.46)	0.95 (0.72-1.21)	1.0 (0.73-1.31)	1
ϕ^{ZD}	1	.	1.4 (1.24-1.55)	1	1.4 (1.23-1.55)
β_D	0.39 (0.36-0.43)	0.54 (0.45-0.63)	0.57 (0.50-0.64)	0.39 (0.36-0.43)	0.57 (0.50-0.64)
β_Z	0.36 (0.35-0.36)	0.32 (0.29-0.34)	0.36 (0.32-0.41)	0.36 (0.33-0.38)	0.36 (0.35-0.36)
$H_{S_D S_Z}/N$	0.87 (0.80-0.95)	0.61 (0.52-0.74)	0.57 (0.50-0.65)	0.87 (0.79-0.94)	0.57 (0.50-0.67)

Table 4.10 – **Estimated parameters on Rio de Janeiro data.** Posterior median and 95% credible interval. Interaction parameters are in the infectiousness enhancement case.

Regarding the influence of dengue on Zika (cases C and D, cf. Table 4.10), the interaction parameter ϕ^{DZ} is estimated around 1, with a large confidence interval and a correlation with the transmission parameter (cf. Figure 4.26). It is not clear whether there is no effect or whether we lack power to detect one. Regarding the influence of Zika on dengue (cases C and E, cf. Table 4.10), the interaction parameter ϕ^{ZD} is estimated significantly over 1, indicating an enhancement effect of Zika on dengue. In the data, there is indeed a large peak of dengue infections just after the Zika outbreak (cf. Figure 4.22), and the enhancement effect is a way for the model to reproduce this pattern without creating a large dengue outbreak the previous year. Caution is however needed when interpreting these results as some parameters are strongly correlated in the MCMC chain (cf. Figure 4.26): the proportion of susceptibles to dengue and the dengue transmission parameter are negatively correlated; the Zika transmission parameter and interaction term are negatively correlated ϕ^{DZ}, indicating that both couples of parameters are not identifiable.

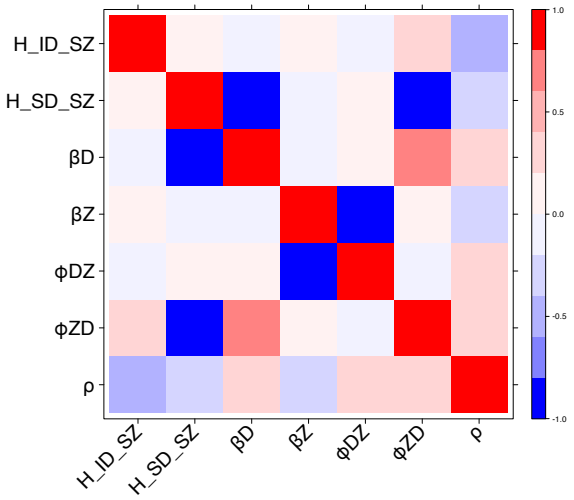


Figure 4.26 – **Correlation matrix of the MCMC chain.** Rio de Janeiro, infectiousness enhancement, case C.

French Polynesia

In French Polynesia, the models manage qualitatively to reproduce the observed incidence data, but they overestimate the seroprevalence at the end of the epidemic (cf. Figure 4.27 and 4.28), except in the susceptibility enhancement case, in which case the DIC criterion is the smallest (cf. Table 4.11). Regarding the influence of dengue on Zika, in the infectiousness and susceptibility enhancement cases, dengue seems strongly protective against Zika, which seems unrealistic as there were indeed people infected by Zika with previous exposition to dengue (Terzian et al., 2017). This may be a model artefact to better fit the Zika seroprevalence data at the end of the epidemic (cf. Figure 4.27 and 4.28). In addition, the estimation of the parameter ϕ^{DZ} also faces identifiability issues: it is very correlated to the transmission parameter for Zika β_Z in the infectiousness and susceptibility enhancement cases, and to the observation rate for Zika ρ_Z in the severity enhancement case (cf. Figure 4.29). Regarding the influence of Zika on dengue, models estimated on both dengue datasets display contradictory results for the interaction terms: there is cross-protection when considering surveillance data, and cross-enhancement when considering Malardé data (cf. Table 4.11). Indeed, due to different reporting strategies, both datasets do not show at all the same dynamic. Since there are few observed dengue cases compared to Zika, observation noise is very influent on the dynamic, and the data does not enable the study of this interaction.

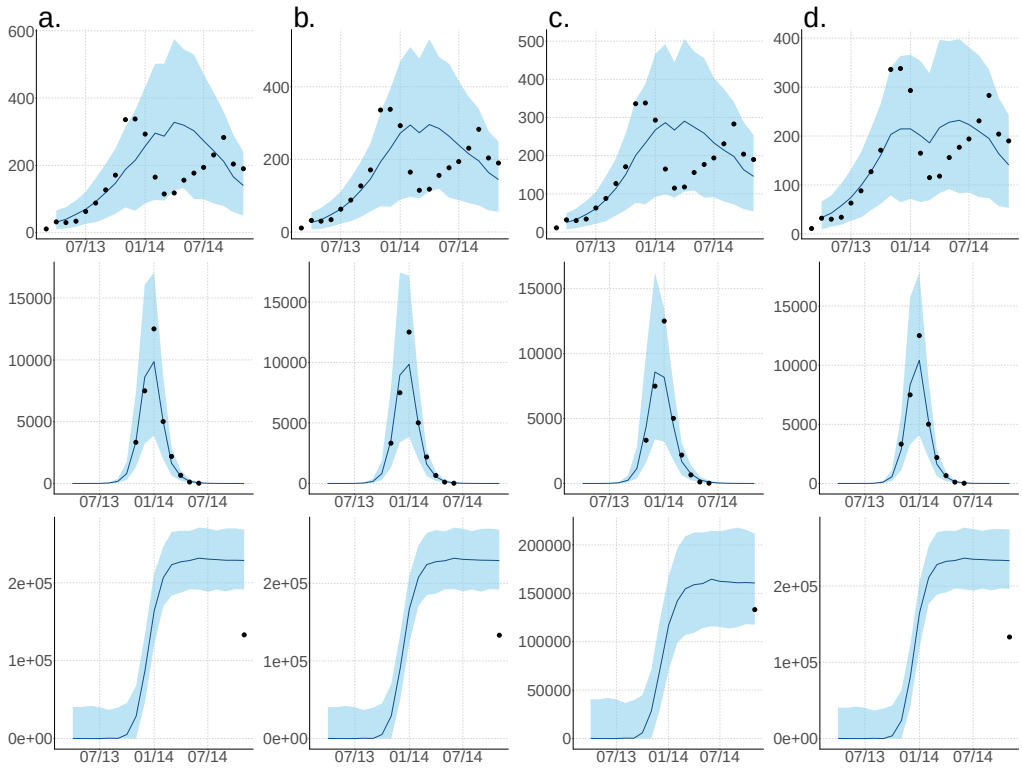


Figure 4.27 – **Simulations of the fitted model and observations (black dots).** Simulations with observation noise of the fitted models using parameter sets from the posterior distribution (median and 95% credible intervals). First row : dengue cases per month (surveillance data). Second row : Zika cases per month. Third row : Seroprevalence for Zika ($H_{S_D R_Z} + H_{E_D R_Z} + H_{I_D R_Z} + H_{R_D R_Z}$). a) Absence of interaction (case A). b) Infectiousness enhancement (case C). c) Susceptibility enhancement (case C). d) Severity enhancement (case C).

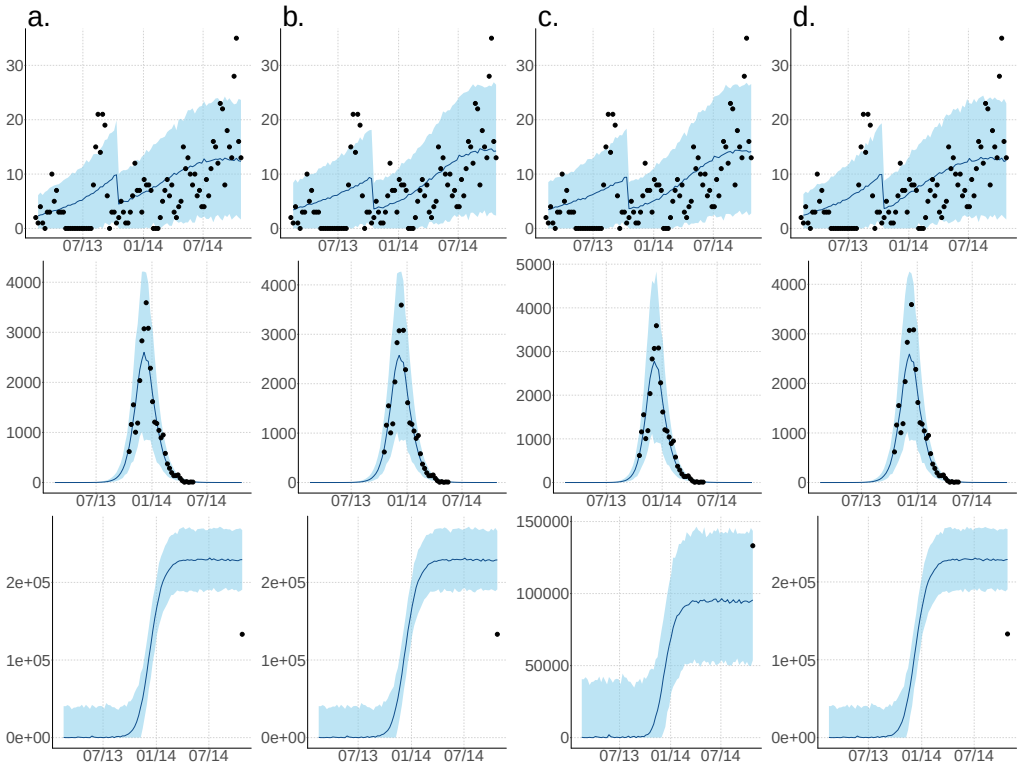


Figure 4.28 – **Simulations of the fitted model and observations (black dots).** Simulations with observation noise of the fitted models using parameter sets from the posterior distribution (median and 95% credible intervals). First row : dengue cases per month (Malardé data). Second row : Zika cases per month. Third row : Seroprevalence for Zika ($H_{S_D R_Z} + H_{E_D R_Z} + H_{I_D R_Z} + H_{R_D R_Z}$). a) Absence of interaction (case A). b) Infectiousness enhancement (case C). c) Susceptibility enhancement (case C). d) Severity enhancement (case C).

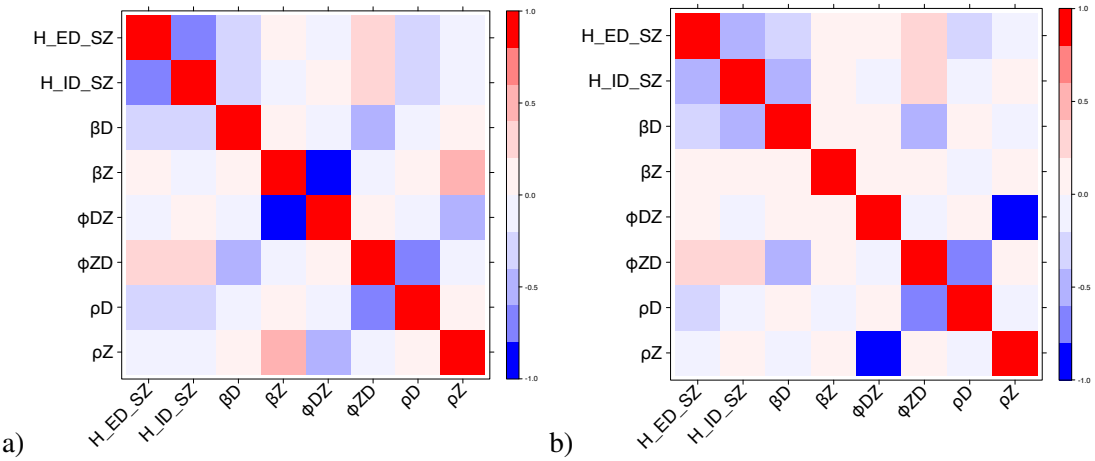


Figure 4.29 – **Correlation matrix of the MCMC chain.** French Polynesia, using surveillance dengue data. a) Susceptibility enhancement, (case C). b) Severity enhancement, (case C).

	Independence $\phi^{ZD} = \phi^{DZ} = 1$ **A**	**Infectiousness** $\phi^{ZD} \neq \phi^{DZ}$ **C**	**Susceptibility** $\phi^{ZD} \neq \phi^{DZ}$ **C**	**Severity** $\phi^{ZD} \neq \phi^{DZ}$ **C**
using Malardé weekly DENV-1 data				
log-likelihood	-499	-494	-484	-498
DIC	1011	1002	985	1011
nb parameters	7	9	9	9
ϕ^{DZ}	1	0.00 (0.00-0.04)	0.02 (0.00-0.07)	0.65 (0.00-1.85)
ϕ^{ZD}	1	1.07 (1.03-1.12)	1.07 (1.03-1.12)	1.49 (0.90-1.99)
β_D	0.70 (0.69-0.72)	0.68 (0.66-0.70)	0.68 (0.66-0.70)	0.70 (0.69-0.71)
β_Z	0.45 (0.42-0.48)	1.29 (1.15-1.38)	1.31(1.21-1.42)	0.45 (0.42-0.47)
ρ_Z	0.14 (0.12-0.16)	0.14 (0.12-0.16)	0.34 (0.24-0.42)	0.18 (0.08-0.37)
using surveillance monthly dengue data				
log-likelihood	-197	-192	-186	-188
DIC	404	398	387	394
nb parameters	6	8	8	8
ϕ^{DZ}	1	0.68 (0.21-1.40)	0.19 (0.13-0.31)	0.83 (0.20-1.82)
ϕ^{ZD}	1	0.90 (0.85-0.97)	0.90 (0.86-0.96)	0.36 (0.19-0.62)
β_D	0.74 (0.71-0.76)	0.74 (0.72-0.76)	0.74 (0.72-0.76)	0.73 (0.71-0.75)
β_Z	0.45 (0.44-0.45)	0.57 (0.35-0.90)	0.95 (0.82-1.00)	0.47 (0.46-0.47)
ρ_Z	0.13 (0.10-0.17)	0.14 (0.10-0.18)	0.19 (0.14-0.25)	0.15 (0.08-0.25)
ρ_D	0.14 (0.12-0.17)	0.21(0.15-0.25)	0.22 (0.16-0.25)	0.23 (0.19-0.25)

Table 4.11 – **Estimated parameters on French Polynesia data.** Posterior median and 95% credible interval.

What data is needed: exploration with simulations

In both settings, the available data does not enable the study of the interaction at the population level. Therefore, we try to identify which type of data would be needed to fit the models, using model simulations. We simulate a dataset in a setting comparable to French Polynesia without observation noise, and then reestimate the models parameters. We assume that seroprevalence studies are available so that we know the initial conditions, and that the observation rate is stable over time ($p = 1$). We also assume that dengue parameters (ρ_D, β_D, i_D) are known, as they could have been estimated on a previous dengue time series, but that the transmission parameter is unknown for Zika because it is an emerging pathogen. We perform estimations in three scenarios. First, we estimate β_Z, ϕ^{DZ}, ϕ^{ZD} and ρ_Z. Second, we assume that the observation rate for Zika is known, and we estimate β_Z, ϕ^{DZ} and ϕ^{ZD}. Third, we assume that, for each infection, we can determine whether it is a primary or a secondary infection.

	True value	Infectiousness		Susceptibility		Severity	
unknown observation rate ρ_Z							
ϕ^{DZ}	1.5	1.32	(0.75-2.00)	1.06	(0.40-1.82)	1.4	(0.82-2.00)
ϕ^{ZD}	1	1.02	(0.99-1.05)	1.02	(0.99-1.05)	0.90	(0.80-1.02)
β_Z	0.8	0.85	(0.69-1.09)	0.90	(0.73-1.11)	0.80	(0.77-0.82)
ρ_Z	0.13	0.13	(0.11-0.14)	0.12	(0.11-0.14)	0.13	(0.10-0.17)
known observation rate ρ_Z							
ϕ^{DZ}	1.5	1.13	(0.67-1.82)	0.97	(0.44-1.66)	1.29	(1.03-1.60)
ϕ^{ZD}	1	1.02	(0.99-1.05)	1.02	(0.98-1.05)	0.90	(0.79-1.01)
β_Z	0.8	0.9	(0.72-1.09)	0.92	(0.76-1.10)	0.80	(0.78-0.82)
identification of primary and secondary infections, known observation rate (=1)							
ϕ^{DZ}	1.5	1.11	(0.77-1.54)	1.4	(1.20-1.60)		
ϕ^{ZD}	1	1.02	(1.00-1.03)	1.02	(1.00-1.04)		
β_Z	0.8	0.92	(0.79-1.04)	0.83	(0.79-0.86)		

Table 4.12 – **Estimations of the simulated dengue-Zika model.** The model was simulated with the following parameters: $\beta_Z = \beta_D = 0.8$, $\rho_D = 0.86$, $\rho_Z = 0.13$, $i_D = 0.002$, $i_Z = 0$,$H_{S_D S_Z}(0) = 0.7 * N$, $H_{E_D S_Z}(0) = 28.8$, $H_{I_D S_Z}(0) = 10.9$, $z = 0.0002$, $\phi^{DZ} = 1.5$, $\phi^{ZD} = 1$.

In the infectiousness enhancement case, in the three considered estimation scenarios, the effect of Zika on dengue (ϕ^{ZD}) is well identified to 1 (the credible interval is small and contains the value 1). However, we do not manage to identify the effect of dengue on Zika ϕ^{DZ} from the transmission parameter for Zika β_Z, and to evaluate if it is superior or inferior to 1 (the credible interval is large and contains the value 1). In the susceptibility enhancement case, the effect of Zika on dengue (ϕ^{ZD}) is also well identified to 1. It is only when dengue naive and dengue non-naive Zika infections are distinguished that the effect of dengue on Zika ϕ^{DZ} is estimated superior to 1 (the credible interval does not contain 1 and contains the true value). In the severity enhancement case, the effect of Zika on dengue (ϕ^{ZD}) is also identified to 1, even though the credible intervals are greater than in the previous two cases, and less centered around 1. When the observation rate is known, β_Z is identified, and ϕ^{DZ} is estimated as superior to 1 (but its credible interval is large and only slightly above 1). When the observation rate is assumed unknown, β_Z is still well identified, but it is not possible to conclude on the presence of enhancement or inhibition of severity as 1 is comprised in the credible interval.

Discussion

We used data from two settings affected by both Zika and dengue, Rio de Janeiro (Brazil) and French Polynesia, to study possible interaction effects between the two pathogens at the population level. Regarding the influence of dengue on Zika, in both settings, we do not identify robustly any effect or absence of effect. Regarding the influence of Zika on dengue, results are not consistent across settings and datasets. In Rio de Janeiro, an enhancement effect is found, but it is sensitive

to the initial seroprevalence for dengue in the population, which is unknown. In French Polynesia, even though initial conditions are known via seroprevalence surveys, the estimations using two different datasets for dengue cases lead to contradictory results. It seems that the surveillance data better reflects the overall dengue dynamics, as we do not have the information on the proportion of cases sent to the Institut Louis Malardé for laboratory testing. With surveillance data, in contrary to estimations in Rio de Janeiro, Zika infection was found to have a mild protective effect on dengue. Besides, in another epidemiological study, Ribeiro et al. (2018) observed a decrease in dengue transmission in Salvador (Brazil) after the Zika epidemic and hypothesize a cross-protection effect. Further data and studies are therefore needed to shed light on these contrasted results. In addition, Chikungunya virus, another arbovirus sharing the same vector, was present in Latin America during the Zika outbreak (Rodriguez-Morales et al., 2016; Carrillo-Hernández et al., 2018), and could also have had an influence on both pathogens.

In our estimations, there is a clear identifiability issue between Zika parameters and interaction terms. In a proof-of-principle study, Shrestha et al. (2011) manage to recover interaction parameters from case data, using likelihood-based inference, even when the observation rate or initial conditions need to be estimated. Nevertheless, they highlight the importance of using long time series, as identifiability issues arise in some of their scenarios with less than 40 years of data. They also assume that the epidemiological parameters are known, in particular the transmission parameter, an hypothesis that is relaxed in other empirical studies (Shrestha et al., 2013; Opatowski et al., 2013). In our context, all these limitations are encountered together: time series are short due to the recent introduction of Zika in both settings, and epidemiological parameters and sometimes initial conditions are unknown. As highlighted in our estimations, in the infectiousness and susceptibility enhancement cases, the interaction term is strongly correlated to the transmission parameter for Zika, a parameter that is impossible to quantify beforehand for a pathogen emerging in a new area. In the severity enhancement case, it is correlated to the observation rate for Zika, which is also hard to measure as it depends on both the reporting system and the proportion of inapparent infections.

In order to identify what type of data would be useful to solve these identifiability issues, we mimicked ideal data scenarios through simulations. According to the re-estimations on simulated data, knowledge on the observation rate and the previous immune status of each case would help disentangle the severity and susceptibility (but not infectiousness) enhancement effects. These information would however be difficult to collect in a real context. Therefore, there are many uncertainties on Zika transmission so that, at the population level, it is hard to disentangle the interaction effect of dengue from the intrinsic Zika characteristics. Modeling together several geographical settings would probably not be sufficient either: there can be large variations in transmission between locations (Champagne et al., 2016) that are sometimes stronger than variations between pathogens (Funk et al., 2016), and therefore transmission parameters need to be setting specific, which does not permit to solve the identifiability issue. However, specific survey designs such as case-control studies (Halstead, 2017) could be implemented to explore interaction effects in the population.

Chapter 5

Discussion et perspectives

5.1 Synthèse des résultats (résumé substantiel en langue française - II)

Les arboviroses sont des maladies complexes qui font intervenir des facteurs humains et environnementaux, et les nombreux mécanismes en jeu dans leur transmission sont autant de pistes de modélisation. Parallèlement, de nombreux outils théoriques et computationnels existent aujourd'hui pour confronter des modèles épidémiologiques aux données observées. L'objectif de ce travail de thèse a donc été d'apporter l'éclairage des données sur les modèles de propagation des arboviroses, afin de comparer différents modèles (chapitres 2 et 3) et d'adapter les modèles à la situation épidémiologique d'une maladie émergente (chapitre 4).

Dans le chapitre 2, il s'est agi de comparer différents modèles, en complexifiant d'un côté le détail de l'histoire de la maladie et de l'autre côté la prise en compte des incertitudes. Afin de modéliser les dynamiques de long terme, on a pu souligner l'importance des interactions entre sérotypes, et à l'inverse, le faible intérêt pour la représentation explicite des moustiques vecteurs et des individus asymptomatiques. Sur des échelles de temps moins longues, la prise en compte de la stochasticité souligne le fait qu'une large part de la dynamique n'est pas expliquée par le squelette déterministe du modèle et est seulement capturée par la stochasticité. Ce phénomène indique la présence d'autres éléments non modélisés ayant une forte influence sur les dynamiques observées.

Aussi le chapitre 3 explore-t-il d'autres aspects de la transmission de la dengue, à travers l'étude d'un autre contexte où des données spatiales sont disponibles. Saisonnalité et structure spatiale sont introduites dans le modèle classique de Ross-Macdonald pour la transmission vectorielle, selon le cadre proposé par Amaku et al. (2016). La saisonnalité a un impact décisif sur la capacité du modèle à reproduire les données observées, ce qui invite à la modéliser plus finement, notamment en incorporant des variables climatiques. La structure spatiale a également une influence notable sur les résultats obtenus, même s'il est difficile de représenter à la fois les contacts entre sous-populations, l'hétérogénéité entre les régions et les schémas d'introduction de la maladie.

Enfin, dans le chapitre 4, on a pu appliquer les méthodes d'estimation à l'étude d'une maladie émergente, le virus Zika, afin d'apporter un éclairage supplémentaire à la compréhension de ce nouveau problème de santé publique. On a ainsi pu explorer la variabilité des estimations selon le contexte géographique, le choix de structure du modèle ou l'interaction avec une autre arbovirose.

Ces travaux ont toutefois soulevé de nouvelles interrogations, discutées ici sous deux angles : d'une part celui des méthodes d'inférence, d'autre part celui des difficultés liées à la représentation des moustiques vecteurs dans la modélisation des arboviroses.

5.2 Discussion et difficultés rencontrées

5.2.1 Inférence des modèles épidémiologiques

Implémentation des méthodes d'inférence

Les méthodes d'estimation utilisées, MCMC et PMCMC ont pour grand avantage leur généralité et leur capacité à traiter des modèles complexes et variés. Toutefois, elles présentent souvent des difficultés d'implémentation pratique. Il s'agit en effet d'algorithmes coûteux (surtout dans le cas du PMCMC), pour lesquels les phases de calibration et d'initialisation sont très importantes: elles ont une forte influence sur l'applicabilité de la méthode et parfois ses résultats. A ce sujet, on peut évoquer la difficulté à échantillonner des distributions a posteriori multimodales, souvent rencontrées dans les modèles multi-souches pour lesquelles les conditions initiales sont inconnues. L'exploration de l'espace des paramètres par des méthodes comme l'échantillonnage par hypercube latin a permis de contourner ce problème, mais il serait également utile de se tourner vers des algorithmes plus robustes à la multimodalité. Par exemple, Yang and Rosenthal (2017) proposent un schéma de MCMC adaptatif spécifique pour les distributions multimodales. Buchholz and Chopin (2017) évoquent également ce type de problèmes dans le contexte de l'algorithme ABC.

Toutefois, de nombreuses difficultés rencontrées dans l'implémentation de ces méthodes d'estimation complexes révèlent également un problème plus profond et structurel : les données disponibles ne sont pas toujours suffisantes pour estimer convenablement les modèles utilisés, en raison de problèmes d'identifiabilité. Dans les analyses menées dans ce travail de thèse, les problèmes d'identifiabilité ont été rencontrés de façon récurrente. Les résultats des modèles multi-souches sont très sensibles aux conditions initiales, bien souvent inconnues dans les contextes endémiques et difficiles à reconstituer à partir des données d'incidence. Les durées d'infection ou d'incubation, chez les hôtes comme chez les vecteurs, présentent des distributions a posteriori marginales très proches des distributions a priori, indiquant que les données utilisées n'ont pas permis d'apporter de l'information supplémentaire sur ces paramètres. L'analyse des matrices de corrélation de la chaîne de MCMC a permis d'identifier des couples de paramètres fortement corrélés entre eux : bien souvent, seul le couple de paramètres est identifiable à partir des données observés, et non la valeur de chacun d'entre eux.

Comparaison des modèles épidémiologiques

Parallèlement, tout au long des chapitres précédents, s'est posée la question de la comparaison de plusieurs modèles épidémiologiques, afin de sélectionner les éléments les plus importants pour reproduire les données observées (chapitre 2 et 3), ou bien afin d'évaluer la variabilité des estimations selon les choix de structure du modèle (chapitre 4).

Si de nombreux travaux se sont intéressés à la comparaison de modèles épidémiologiques (Gibson et al., 2018), cette question demeure complexe. Les méthodes de comparaison entre

différents modèles doivent s'adapter à l'objectif visé (test d'hypothèses scientifiques ou incorporation d'incertitude dans les estimations de paramètres ou de quantités prédites), mais elles sont aussi souvent liées au contexte méthodologique considéré. Il faut en effet noter qu'il existe plusieurs degrés de différences entre les modèles à comparer (même si les frontières de l'un à l'autre sont poreuses).
Dans certains cas, la comparaison de modèles peut se réduire à un test sur la valeur d'un paramètre, ou sur l'égalité de deux paramètres, les modèles étant alors emboîtés. C'est le cas, par exemple, lorsqu'on s'interroge sur la présence d'effets d'interaction entre deux pathogènes (cf. chapitre 4) : en l'absence d'interaction, le paramètre d'interaction a une valeur égale à 1.
Dans d'autres cas, il s'agit de comparer deux spécifications fonctionnelles au sein d'un même modèle. Par exemple, on peut comparer les choix de différentes distributions de la durée passée dans les compartiments du modèle (par exemple entre les distributions exponentielle et Gamma). On peut aussi comparer plusieurs formulations de la propagation spatiale, par exemple, entre différents modèles de gravité (cf. chapitre 3), ou en comparaison avec un modèle de radiation.
Enfin, les modèles à comparer peuvent différer fortement dans l'ensemble de leur structure, par exemple, avec ou sans transmission vectorielle, avec ou sans représentation spatiale, modèles mécanistes ou semi-mécanistes (cf. chapitre 2). Il peut alors s'avérer plus difficile de rapporter les modèles à un cadre commun permettant la comparaison.
De manière générale, les méthodes adaptées dans un contexte sont parfois difficilement applicables dans un autre, en raison de limitations théoriques ou pratiques.

5.2.2 Modéliser la transmission vectorielle pour les arboviroses

L'étude des arboviroses est complexe car elle fait intervenir et interagir deux populations : les hôtes et les vecteurs. Au cours de ce travail, plusieurs modèles qui incorporent cette spécificité ont été considérés. Notamment, le vecteur a été représenté soit de façon explicite selon un formalisme de type Ross-Macdonald, soit de façon implicite à travers une force d'infection externe. Toutefois, ces modèles ont à plusieurs reprises montré des limites. Dans le chapitre 2, des modèles sans vecteur incorporant un forçage saisonnier se sont montrés aussi performants que les modèles à transmission vectorielle, tout en étant beaucoup plus parcimonieux en termes de paramètres. On peut d'ailleurs s'interroger sur la façon d'incorporer la stochasticité démographique dans la population de moustiques, alors que la taille même de cette population est inconnue. En outre, les données disponibles rendent très difficile l'estimation des nombreux paramètres de ces modèles, qui reposent donc fortement sur les choix de distributions a priori opérées à partir de données de la littérature. En particulier, le taux de mortalité du vecteur joue un rôle important dans la dynamique du modèle et beaucoup de résultats y sont sensibles (comme on peut le voir à travers l'analyse de sensibilité du chapitre 4 et dans les travaux de Pandey et al. (2013); Funk et al. (2016)). Or, ce paramètre est difficile à mesurer sur le terrain, et les mesures effectuées en laboratoire peuvent mal refléter la réalité (lorsque des mesures de contrôle, par exemple, sont mises en place).

De plus, dans l'étude du Zika en Polynésie et en Micronésie au chapitre 4, les modèles à transmission vectorielle estimés sur les données d'incidence ont conduit à des taux d'attaque bien supérieurs à ce qui a été mesuré dans les données de séroprévalence. Cet effet a également été observé dans d'autres études (Kucharski et al., 2016; Funk et al., 2016; Shutt et al., 2017). En revanche, ce problème n'a pas été rencontré dans une analyse utilisant des modèles sans transmission vectorielle explicite, dans lequel les durées d'incubation et d'infections sont plus courtes (Andronico et al., 2017). On peut tout d'abord jeter un doute sur les mesures de

séroprévalence, en raison de la difficulté des diagnostics sérologiques lorsque plusieurs arboviroses co-circulent. Toutefois, les taux d'attaques estimés à l'aide des seules données d'incidence sont supérieurs à 90% de la population, induisant une forte déplétion de la population susceptible, ce qui semble anormalement élevé. Ce phénomène est d'autant plus problématique que le taux d'attaque, et par suite l'immunité dans la population et les possibilités de transmission future du pathogène, font justement partie des questions auxquelles la modélisation est censée apporter des réponses.

Netto et al. (2017) introduisent une prise en compte de la saisonnalité dans le modèle, permettant ainsi une diminution exogène de la force d'infection par les facteurs climatiques et une réconciliation des deux sources de données. Toutefois, le climat ne présente pas des variations saisonnières marquées dans toutes les régions affectées par la dengue. D'autres phénomènes peuvent introduire des variations temporelles dans la transmission, comme les changements de comportements humains ou les mesures de contrôle. Ces phénomènes sont plus difficiles à capturer et introduire des paramètres variables dans le temps est une façon de les étudier.

Une autre piste pour expliquer des taux d'attaque plus faibles que ceux prédits par le modèle est la remise en cause de l'hypothèse de mixage homogène entre hôtes et vecteurs. Cette hypothèse simplificatrice est au coeur de nombreux modèles à transmission vectorielle (Reiner et al., 2013) mais elle apparaît réductrice à plusieurs échelles spatiales et temporelles, au vu de la complexité de l'écologie des vecteurs, qui piquent les individus de façon hétérogène (Smith et al., 2014). Dans le modèle du chapitre 4, un paramètre a été ajouté afin de refléter le fait que tous les individus ne participent pas à la transmission de la maladie, pour des raisons d'immunité, ou d'hétérogénéité spatiale. Une approche similaire a été adoptée par Shutt et al. (2017) en calculant la population à risque à l'aide de données d'épidémies antérieures. Toutefois, ce paramètre ne propose pas une explication du mécanisme par lequel certains individus échappent à l'épidémie. L'hypothèse des interactions entre la dengue et le Zika a été explorée (chapitre 4), mais cette piste est limitée par l'absence de données sur le temps long et les problèmes d'identifiabilité. La piste des taux de transmission non linéaires a été explorée par Billy Bauzile dans son travail de master. Enfin, l'hypothèse de l'hétérogénéité spatiale a été abordée à travers l'étude des dynamiques de la dengue à Rio de Janeiro.
Malgré tout, cette question nécessite d'autres travaux pour être mieux comprise.

5.3 Perspectives

Afin de répondre aux difficultés évoquées dans la section précédente, plusieurs pistes de réflexion peuvent être envisagées, que l'on articulera ici autour de trois axes. Dans un premier temps, on s'intéressera à la question de l'identifiabilité des modèles épidémiologiques. Dans un second temps, on discutera des méthodes permettant d'incorporer plusieurs sources de données. Enfin, évoquera la question de la comparaison de modèles.

5.3.1 Identifiabilité

Bien que cette question soit peu souvent abordée dans les applications pratiques en épidémiologie, il existe une abondante littérature sur l'identifiabilité des modèles dynamiques, notamment ceux qui représentent la propagation des maladies infectieuses. On peut diviser les problèmes d'identifiabilité en deux catégories : la non-identifiabilité structurelle et la non-identifiabilité pratique (Raue et al., 2009). La non-identifiabilité structurelle correspond à une "paramétrisation redondante dans la

structure formelle" du modèle (Raue et al., 2009). Miao et al. (2011) indiquent plusieurs nuances dans cette définition de la non-identifiabilité structurelle dans la littérature, selon si celle-ci est globale (valable sur tout l'espace des paramètres), locale (valable seulement sur un voisinage d'un certain jeu de paramètres), ou dépend de la connaissance des conditions initiales du système.
La non-identifiabilité pratique, elle, survient lorsque les données sont observées avec bruit et que leur quantité ou leur qualité est insuffisante, ce qui conduit à des intervalles de confiance de taille infinie (Raue et al., 2009). L'identifiabililté structurelle est donc un pré-requis nécessaire mais non suffisant à l'identifiabilité pratique.

L'étude de l'identifiabilité des modèles dynamiques non linéaires requiert des méthodes non triviales (Miao et al., 2011). L'analyse de l'identifiabilité structurelle peut être menée avec différentes techniques, parmi lesquelles l'algèbre différentielle qui a été utilisé à plusieurs reprises en épidémiologie (Eisenberg et al., 2013; Moulay et al., 2012). L'identifiabilité pratique, elle, nécessite la confrontation à des données expérimentales et peut être étudiée à l'aide de méthodes de Monte Carlo (toutefois coûteuses) ou à travers l'étude des profils de vraisemblance. Miao et al. (2011) détaillent également des méthodes en lien avec la sensibilité, qui s'intéressent à l'identifiabilité liée à un jeu de paramètres préalablement fixé ("at-a-point identifiability"). Egalement, dans cette perspective, Pant (2018) utilise les fonctions d'information de sensibilité (information sensitivity functions, ISF). Elles permettent d'identifier les régions de la série de données qui comportent le plus d'information sur la valeur des paramètres, de révéler les corrélations potentielles entre eux, ou d'étudier comment l'identifiabilité de chaque paramètre dépend de la qualité des données observées (en termes de bruit et de fréquence). Ces diagnostics servent à indiquer quelles sont les données les plus importantes à collecter afin de mieux estimer les modèles.

Dans le contexte bayésien, la notion d'identifiabilité est plus élusive (cf. discussion sur le blog d'A. Gelman (Gelman, 2014)). On pourrait considérer qu'un modèle non-identifiable est un modèle dont la distribution a posteriori est impropre. Toutefois cette caractérisation, contradictoire avec les prérequis de l'inférence bayésienne (B. Goodrich in Gelman (2014)), ne semble pas assez opérationnelle pour analyser l'éventail des difficultés rencontrées en pratique. De ce fait, la documentation du logiciel d'inférence bayésienne STAN ({Stan Development Team}, 2017), plutôt que de parler d'identifiabilité, traite la question des "distributions a posteriori problématiques". Celles-ci se décomposent en plusieurs catégories parmi lesquelles la colinéarité, les distributions a posteriori plates et la multimodalité.
Dans le cas de la colinéarité, plusieurs paramètres sont très corrélés entre eux et impossible à distinguer les uns des autres. Dans le cas des distributions marginales a posteriori plates, les données n'ont pas apporté d'information sur le paramètre considéré: pour ce paramètre, toute l'inférence repose sur la distribution a priori, dont la définition est alors très influente sur les résultats. Le cas de la multimodalité recouvre plusieurs situations. Certaines peuvent être reliées au problème de non-identifiabilité structurelle, comme par exemple en présence d'échangeabilité de labels. Dans d'autres situations, la multimodalité ne correspond pas à proprement parler à une non-identifiabilité, car la distribution multimodale elle-même est le résultat de l'inférence. Toutefois, ceci est à la source de deux types de problèmes. D'une part, les méthodes d'inférence ne sont souvent pas adaptées pour explorer convenablement une distribution fortement multimodale (souvent, elles restent piégées dans un mode). D'autre part, l'estimation ponctuelle des paramètres est moins aisée car les statistiques habituelles (moyenne, variance) reflètent mal les distributions multimodales. La présence de multiples solutions rend plus difficile l'interprétation des résultats du modèle et de ses potentielles prédictions.

Plusieurs travaux ont étudié spécifiquement l'identifiabilité des modèles épidémiologiques. Dans le cadre d'un simple modèle SIR avec forçage saisonnier dans lequel seuls les individus infectés sont observés, Evans et al. (2002) indiquent qu'il n'est pas possible d'identifier à la fois la taille de la population, le taux d'observation, le taux de transmission et les conditions initiales du système. Toutefois, le modèle est identifiable dès que l'une de ces quantités est connue. Plusieurs analyses ont été menées dans des contextes plus appliqués, comme l'étude de modèles intra-hôtes pour le VIH (virus de l'immuno-déficience humaine) (Wu et al., 2008) ou sur différents modèles de choléra (Eisenberg et al., 2013). Ces études permettent notamment d'identifier les paramètres sur lesquels concentrer les efforts de collecte de données.
Les modèles à transmission vectorielle sont également sujets à ces problèmes. Leur identifiabilité a été étudiée par plusieurs travaux, dont beaucoup sont cités par Kao and Eisenberg (2018). Moulay et al. (2012); Shousheng et al. (2014) étudient à l'aide de méthodes algébriques les propriétés d'identifiabilité structurelle de modèles à transmission vectorielle pour le chikungunya, qui incorporent les stades aquatiques du moustique, en supposant l'observation du nombre d'oeufs ou de larves. Kao and Eisenberg (2018) en étudient de plus les propriétés d'identifiabilité pratique sous différentes hypothèses de données observées à l'aide de profils de vraisemblance : elles indiquent que, si certains paramètres ne sont pas identifiables individuellement, le nombre de reproduction de base R_0 est une combinaison de paramètres qui l'est.

Face au problème de l'identifiabilité, notamment de l'identifiabilité structurelle, la première solution consiste à reparamétrer les modèles. C'est ce qui a été fait à de nombreuses reprises. Par exemple, Pandey et al. (2013) proposent une reparamétrisation du modèle de Ross-Macdonald, en regroupant des combinaisons de paramètres qui ne peuvent pas être identifiées (c'est cette version du modèle de Ross-Macdonald qui a été utilisé à plusieurs reprises au cours des chapitres précédents). Les nouveaux modèles ainsi obtenus sont alors estimables mais ils peuvent ne pas suffire à répondre à toutes les questions épidémiologiques pour lesquelles ils ont été formulés. Par exemple, Kao and Eisenberg (2018) signalent que certaines paramétrisations présentant la même adéquation aux données (et donc non identifiables) peuvent conduire à des conclusions différentes concernant les scénarios de contrôle.
Une autre solution à ces problèmes est donc d'incorporer d'autres sources d'information pour lever les incertitudes présentes dans les données disponibles.

5.3.2 Incorporer d'autres sources d'information

Distributions a priori informatives

Les distributions a priori informatives sont particulièrement adaptées pour incorporer une information préalable et incertaine sur les paramètres du modèle. Elles peuvent être construites à partir de données de la littérature sur des grandeurs mesurables expérimentalement (durées d'infection, durées de vie, etc.). Dans les chapitres précédents, les distributions a priori informatives ont été utilisées à plusieurs reprises, notamment pour la proportion initiale d'individus susceptibles ou les paramètres liés au cycle de vie du vecteur. Certains paramètres, comme la durée de vie du moustique, étant particulièrement influents sur les résultats du modèle, il est donc décisif que les mesures expérimentales soient précises et transposables au contexte étudié. En effet, si les éléments comme la moyenne sont bien connus, parfois, la forme de la distribution ou sa dispersion le sont moins. De plus, les données disponibles ont parfois été récoltées sur une autre zone géographique, un autre pathogène, ou dans des contextes expérimentaux de laboratoire qui peuvent ne pas refléter

la réalité du terrain.

Les distributions a priori peuvent également être construites comme résultat d'estimations préliminaires sur un autre jeu de données, par exemple pour un autre pathogène proche sur la région étudiée dans le cadre de modèles bayésiens hiérarchiques (Riou et al., 2018).

Combinaison de jeux de données

Les autres données peuvent également être incorporées en tant qu'observations, en modifiant l'expression de la vraisemblance du modèle. Chaque jeu de données comporte à les fois des défauts et des informations spécifiques, et c'est la prise en compte de leur ensemble qui permet de restreindre les zones d'incertitude, comme cela a pu être illustré par de nombreux travaux, comme par exemple Birrell et al. (2011); Rasmussen et al. (2011). La revue de Birrell et al. (2018) présente différentes approches permettant d'incorporer différentes sources de données dans les modèles épidémiologiques, qu'ils soient statistiques et mécanistes. De plus, cette perspective bénéficie du développement de plateformes informatiques permettant de combiner et d'analyser différentes formes de données simultanément (Jombart et al., 2014a,b). Dans notre cas, nous avons incorporé plusieurs sources de données d'incidence, pour la comparaison de modèles de dengue au Cambodge (chapitre 2), et des données d'incidence et de séroprévalence dans l'étude du virus Zika dans le Pacifique (chapitre 4).

De Angelis et al. (2015) questionnent les enjeux méthodologiques liées à la combinaison de plusieurs jeux de données. Tout d'abord, il n'est pas toujours évident de quantifier la contribution relative des différentes sources de données dans la vraisemblance. En effet, le poids de chaque jeu de données dépend à la fois du nombre d'observations disponibles et de l'incertitude qui lui est associée dans le modèle d'observation. Le choix du modèle d'observation le plus approprié en devient d'autant plus important. De Angelis et al. (2015) évoquent également les biais systématiques présents dans certains jeux de données (par exemple, les données internet) ainsi que l'importance d'une prise en compte explicite de ces biais dans l'élaboration du modèle.
De plus, si l'hypothèse d'indépendance conditionnelle des différents jeux de données permet une écriture factorielle simple de la vraisemblance et a été très souvent utilisée, cette hypothèse n'est pas forcément appropriée dans toutes les situations (par exemple, quand les données de séquences génétiques sont issues des mêmes individus que les données d'incidence). De Angelis et al. (2015) évoquent certaines pistes pour prendre en compte la dépendance entre les jeux de données.
Un autre problème surgit lorsque le modèle relatif à l'un des jeux de données est mal spécifié et risque de contaminer l'ensemble de l'analyse, ce qui a notamment été étudié par Jacob et al. (2017). Si l'on représente l'ensemble des dépendances entre jeux de données et paramètres sous forme d'un graphe, plusieurs distributions-cibles peuvent être formulées : il est possible de prendre en compte toutes les dépendances simultanément (a posteriori complet), ou bien de considérer des distributions "modulaires" qui restreignent la propagation de l'information au sein du graphe (par exemple, pour estimer certains paramètres, seuls certains des jeux de données, et non tous, sont pris en compte). Les auteurs indiquent des situations dans lesquelles ces distributions "modulaires" s'avèrent préférables à l'a posteriori complet et proposent une méthode pour choisir parmi ces différentes formulations à partir des données.
De manière générale, ces travaux soulignent l'importance de disposer d'un cadre formel pour l'intégration de données diverses et hétérogènes.

Malgré ces nuances, qui sont autant de pistes de prolongements, la confrontation des modèles à plusieurs jeux de données s'est révélée décisive dans ce travail de thèse. D'une part, elle a permis de réduire les problèmes d'identifiabilité et de mieux estimer des modèles complexes dans l'étude des modèles de dengue à Kampong Cham. D'autre part, dans le cas des modèles de Zika, elle a permis de révéler des situations de conflit entre les jeux de données (incidence et séroprévalence), une notion également évoquée par De Angelis et al. (2015). Dans ce dernier exemple, la confrontation aux différentes données observées permet de poser un regard critique sur les modèles et leurs problèmes de spécification. Remettre en cause les modèles et en construire de nouveaux à la complexité croissante nécessite alors un cadre formel pour les comparer.

5.3.3 Comparer et questionner la structure du modèle épidémiologique

Outre les méthodes fondées sur les critères d'information ou les vérifications "posterior predictive", une autre approche au problème de comparaison de modèles dans le cadre bayésien est l'utilisation du facteur de Bayes (Kass and Raftery, 1995). Afin de comparer deux hypothèses ayant pu générer un jeu de données, on s'intéresse à la transformation permettant de passer du rapport de leurs distributions a priori à celui de leurs distributions a posteriori, afin de quantifier la certitude en faveur d'une hypothèse par rapport à l'autre. Ces facteurs ont néanmoins le défaut d'être sensibles au choix des distributions a priori dans chacun des modèles.

En pratique, ces facteurs, qui reposent sur l'expression de la vraisemblance marginale, sont souvent difficiles à calculer ; toutefois, de nombreuses méthodes existent. D'une part, certaines méthodes permettent d'approximer la vraisemblance marginale (Kass and Raftery, 1995). Touloupou et al. (2018) comparent plusieurs de ces méthodes sur des données épidémiologiques et en proposent une nouvelle, qui permet de traiter les situations à variables latentes, fondée sur l'*importance sampling*. En particulier, Alzahrani et al. (2018) l'utilisent dans le cadre de modèles de séries temporelles estimés par PMCMC.
D'autre part, certains algorithmes permettent de simuler le modèle dans l'espace agrandi formé de l'union des espaces de paramètres de différents modèles (Gibson et al., 2018). A chaque itération, un modèle, puis ses paramètres, sont échantillonnés et potentiellement acceptés ; ainsi, la probabilité a posteriori d'un modèle dépend de la proportion d'itérations pour lesquelles il a été retenu. Pour cela, certains utilisent des méthodes de MCMC à sauts réversibles (RJMCMC), la difficulté étant de trouver une densité de proposition permettant d'échantillonner efficacement l'espace agrandi. Malgré les difficultés computationnelles, Gibson et al. (2018) citent deux exemples où cette méthode a été employée. Toni et al. (2009) utilisent, eux, la méthode ABC-SMC pour approximer la distribution a posteriori de chaque modèle, et l'appliquent à des systèmes déterministes et stochastiques en épidémiologie et en écologie (ils comparent notamment plusieurs modèles épidémiologiques de grippe à la structure différente). Néanmoins, ils mentionnent eux aussi la sensibilité de cette méthodologie aux distributions a priori, à la fois par la sensibilité inhérente des facteurs de Bayes (Kass and Raftery, 1995), et par le risque d'un faible taux d'acceptation pour les modèles dans lesquels les paramètres sont échantillonnés dans une étendue trop grande et donc très souvent rejetés. Il signalent de plus la sensibilité du facteur de Bayes à certains paramètres de calibration de l'ABC-SMC (seuil de tolérance et fonction de perturbation). Toutefois, de nombreuses extensions de l'utilisation de l'ABC pour la sélection de modèles existent (Pudlo et al., 2016). Récemment, O'Reilly et al. (2018) utilisent également des méthodes d'ABC pour comparer des hypothèses sur les contacts entre les populations de leur modèle spatial de Zika en Amérique Latine.

Là encore, en plus des limitations méthodologiques dans la comparaison de modèles, une forte limite pratique est souvent liée au fait que les données ne contiennent parfois pas l'information suffisante pour discriminer les modèles ("model distinguishability", Lee et al. (2017)). Ce problème est différent selon l'objectif sous-jacent au contexte de modélisation, qui peut être une meilleure compréhension des mécanismes en jeu ou la mise en place de mesures de contrôle (Koopman, 2004). Dans le premier cas, la question est de savoir si l'hypothèse en jeu est nécessaire pour expliquer les phénomènes observés, et l'absence de distinction entre deux modèles peut conduire à privilégier le plus simple. Dans le second cas, la question est de savoir si l'hypothèse en jeu a une influence non seulement sur les données observées mais aussi pour prédire le futur ou des situations contrefactuelles (scénarios de contrôle). Si les données observées ne suffisent pas à discriminer des modèles aux prédictions divergentes, il est important de mettre en place des stratégies de prédiction qui incorporent l'incertitude liée à la structure du modèle.

L'incorporation des différentes formes de stochasticité dans le processus épidémiologique est une étape importante dans cette prise en compte de l'incertitude. Toutefois, en plus de l'incertitude inhérente à chaque modèle, on peut également vouloir incorporer l'incertitude liée aux différentes structures de modèles possibles. Dans le cas d'une comparaison entre plusieurs modèles, Kass and Raftery (1995) détaillent dans leur revue des méthodes pour tenir compte de l'"incertitude du modèle". A chaque modèle est associée une distribution a priori et, par le biais du facteur de Bayes, une distribution a posteriori. Afin de prédire une quantité en tenant compte de l'incertitude de modélisation, il est alors possible de construire une moyenne pondérée des prédictions des différents modèles, le poids associé à chaque modèle reflétant sa qualité de prédictions sur les données déjà observées. Raftery et al. (2005) adaptent cette méthodologie aux systèmes dynamiques et ce type d'approche a été utilisé par exemple par Viboud et al. (2018) pour combiner les prédictions d'épidémies d'Ebola formulées indépendamment par différents groupes de recherche, dans le cadre du challenge Ebola. Ce cadre d'analyse semble une perspective intéressante pour mieux tenir compte de l'incertitude de modèle dans les prédictions et l'étude de scénarios de contrôle.

Ainsi, l'intégration de nombreuses sources de données et de diverses formes de modélisation soulève de nombreux défis méthodologiques et pratiques. Toutefois, elle constitue des pistes prometteuses pour enrichir l'analyse et la compréhension de la propagation des arboviroses.

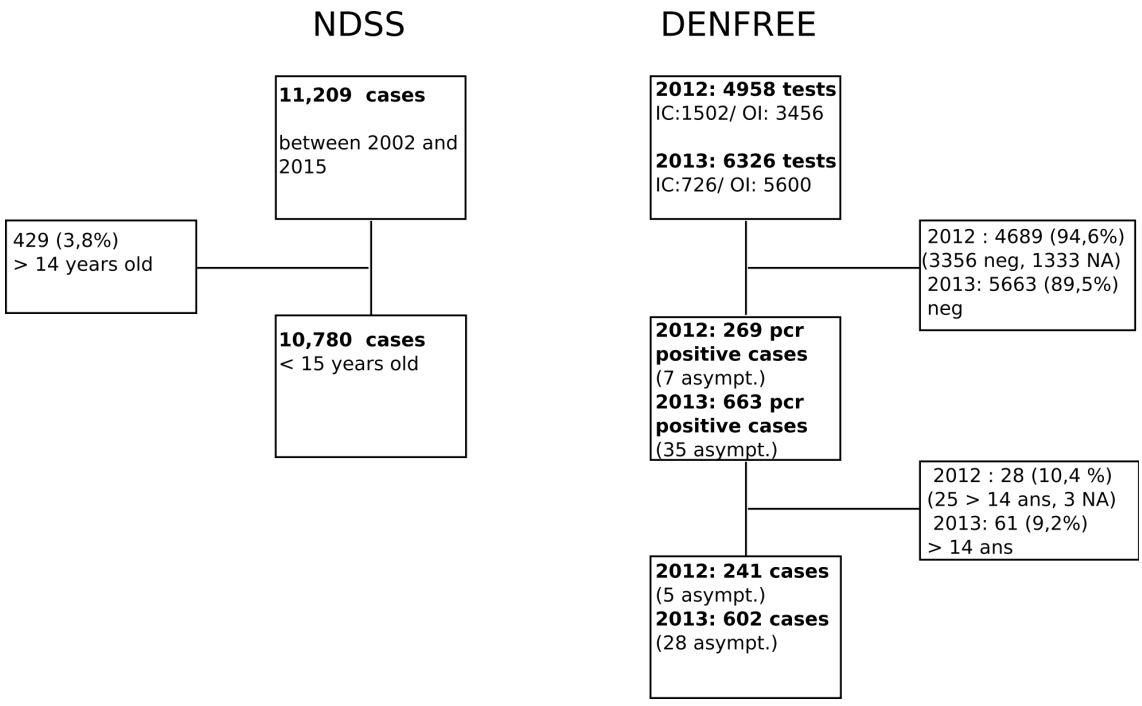
Appendix A

Annexe: Dengue modeling in rural Cambodia: statistical performance versus epidemiological relevance

A.1 Data and codes

The datasets and codes supporting this article are available at https://github.com/clchampag/KC-dengue.

A.2 Flowchart



A.3 Convergence diagnosis

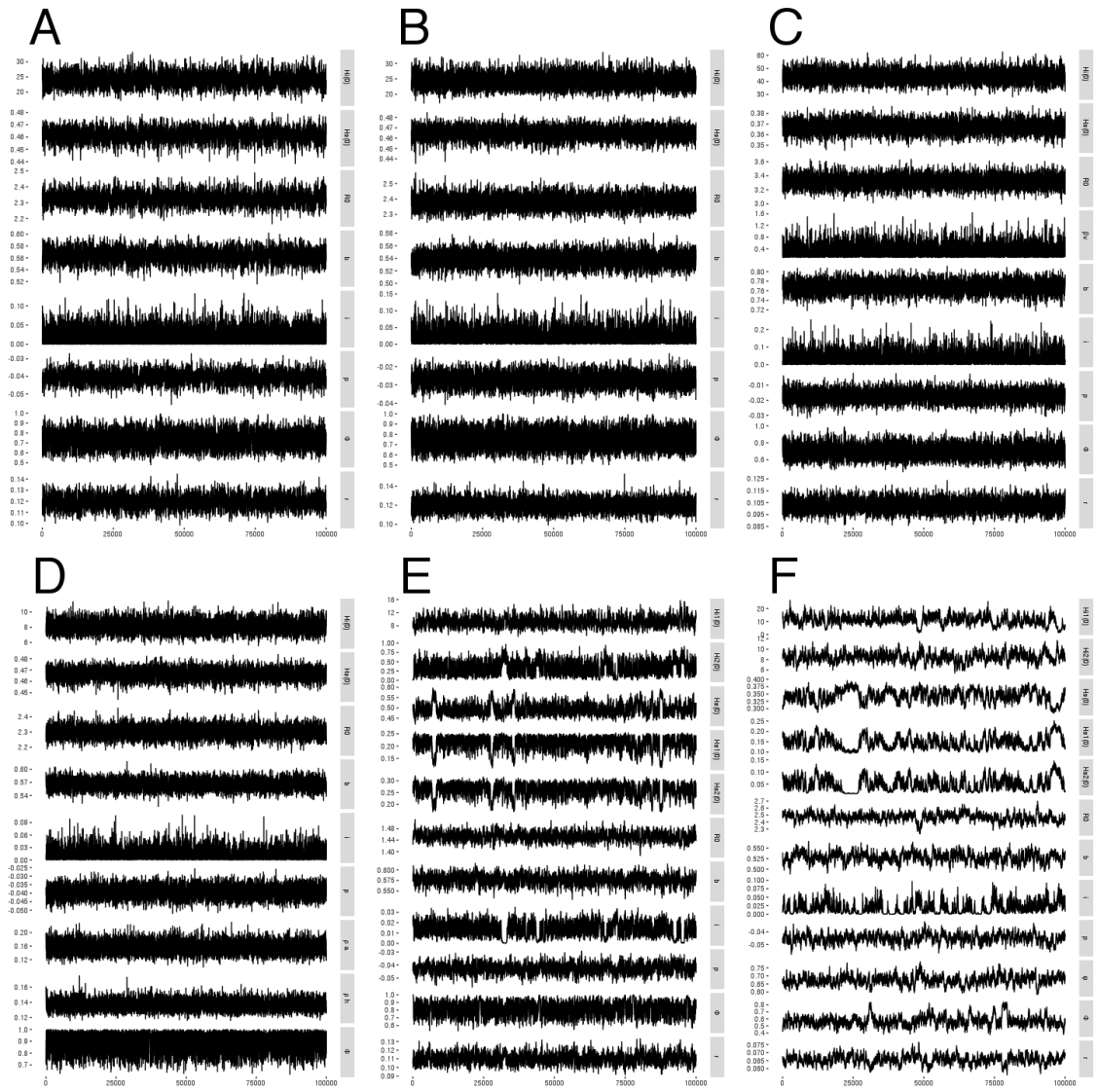


Figure A.1 – **Trace plots for the MCMC algorithm, with 100,000 iterations.** MCMC chain associated with the highest posterior. a) SEIR. b) Laneri. c) Pandey. d) SEIAR. e) SEIR2. f) SEIR2psi

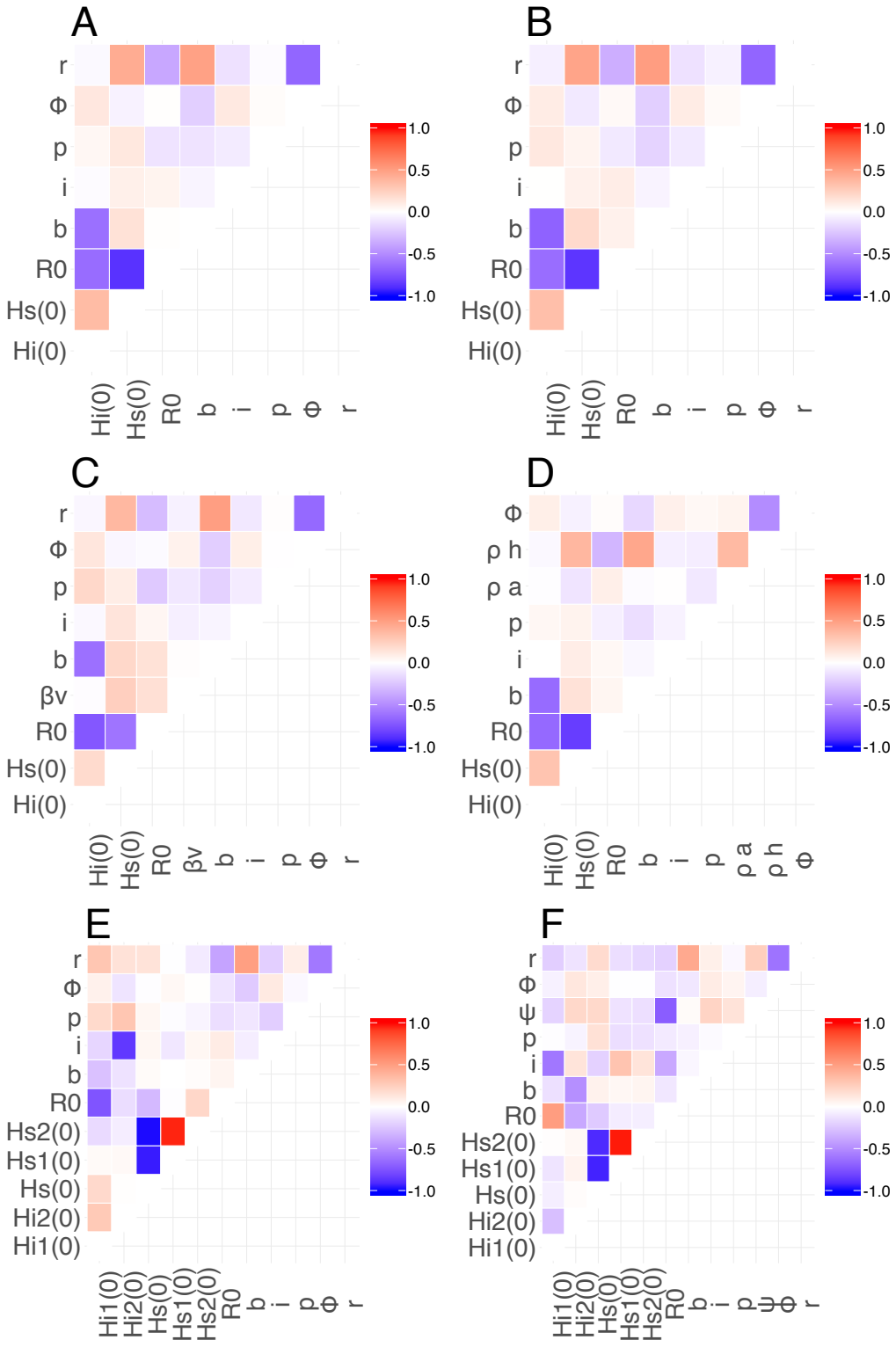


Figure A.2 – **Correlation plots of the MCMC algorithm, with 100,000 iterations.** MCMC chain associated with the highest posterior. a) SEIR. b) Laneri. c) Pandey. d) SEIAR. e) SEIR2. f) SEIR2psi

A.4 Effective reproduction number over time

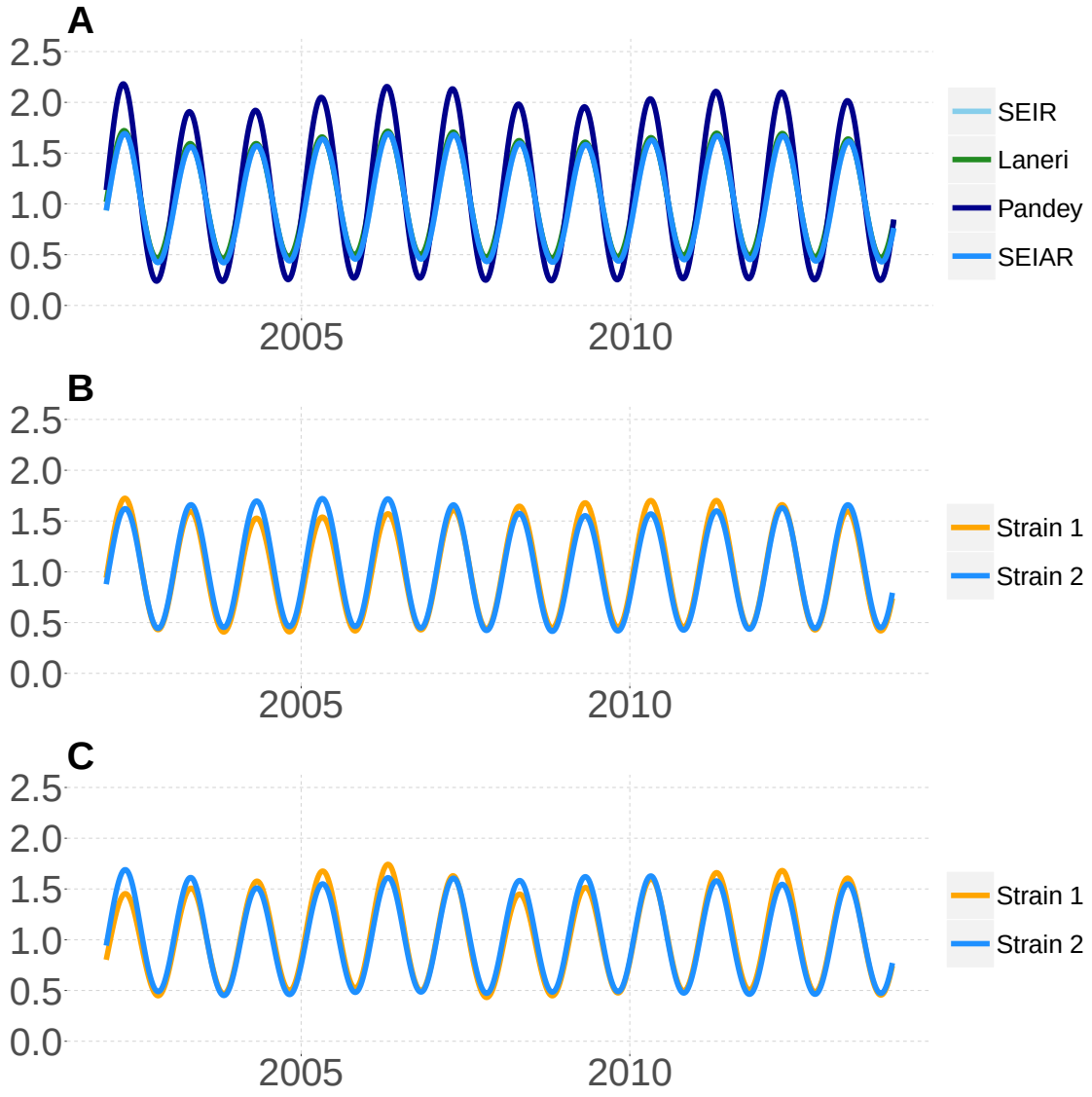


Figure A.3 – **Effective reproduction number over time.** Median based on simulations without negative binomial noise using parameters from the MCMC chain associated with the highest posterior, calculated using both NDSS and DENFREE datasets. A. Single strain models (SEIR, Laneri and SEIAR are almost overlapping). B. SEIR2 model. C. SEIR2psi model.

A.5 Multimodality in multistrain models

Multistrain models are difficult to parameterize, in particular due to the high number of unknown initial conditions. The adaptive MCMC estimation remains trapped in local maxima, and we therefore explored the parameter space using the simplex algorithm over a latin hypercube sample of 100,000 parameter sets. However, in some cases some local maxima are close to each other in terms of likelihood and difficult to identify.

To illustrate this point, in Figure A.4 we display two parameter sets corresponding to different initializations of the MCMC chain. The chains have similar marginal likelihoods (cf. overlapping histograms in panels A) but differing estimated values, for example in terms of basic reproduction number or observation rate for NDSS data (cf. non overlapping histograms in panels B, C and D), highlighting the multimodal nature of the posterior distribution.

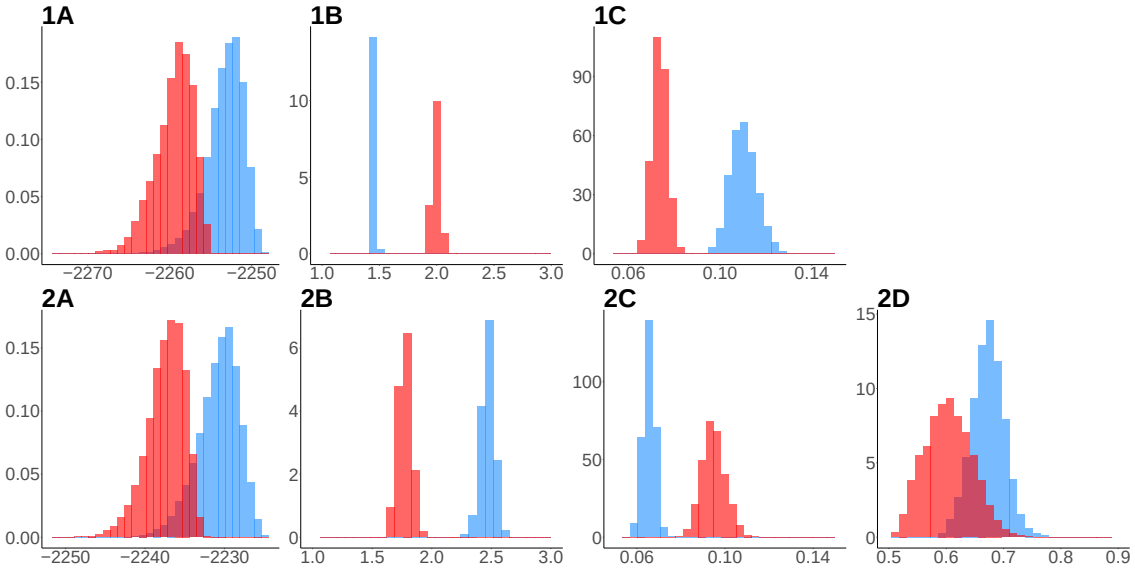


Figure A.4 – **Posterior distributions of some parameters for two parameterization of the SEIR2 model.** Colors red and blue correspond to two differing parameterizations. The blue one is the one commented in the main text. Panel 1 refers to the SEIR2 model, and panel 2 to the SEIR2psi model. A. Posterior values (marginal likelihood multiplied by prior). B. Basic reproduction number R_0. C. Observation rate for NDSS data. D. Interaction parameter ϕ in the SEIR2psi model.

Moreover, these two parameterizations lead to different trajectories that are both epidemiologically coherent in terms of strain asynchrony and replacement (cf. Figures A.5 and A.6). Long term serotype-specific case data and seroprevalence surveys would be crucial to estimate properly these models.

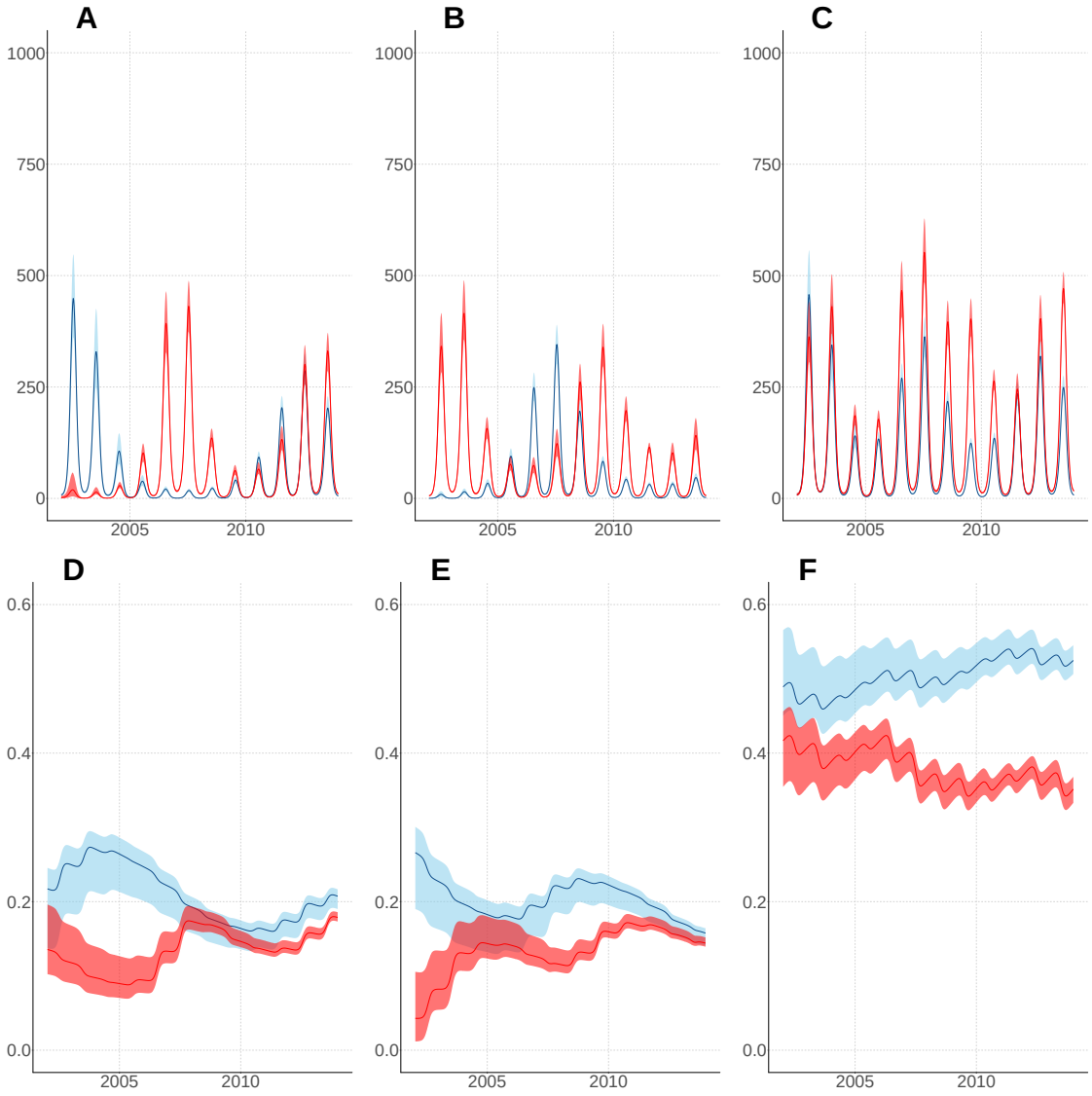


Figure A.5 – **Simulated trajectories for two parameterizations of the SEIR2 model.** Colors red and blue correspond to two differing parameterizations. The blue one is the one commented in the main text. A. Individuals infected with strain 1 ($H_{I1} + H_{I21}$). B. Individuals infected with strain 2 ($H_{I2} + H_{I12}$). C. Individuals infected with any strain ($H_{I1} + H_{I21} + H_{I2} + H_{I12}$). D. Proportion of susceptible individuals with immunity to strain 1 (H_{S1}/N). E. Proportion of susceptible individuals with immunity to strain 2 (H_{S2}/N). F. Proportion of individuals susceptible to both strains (H_S/N).

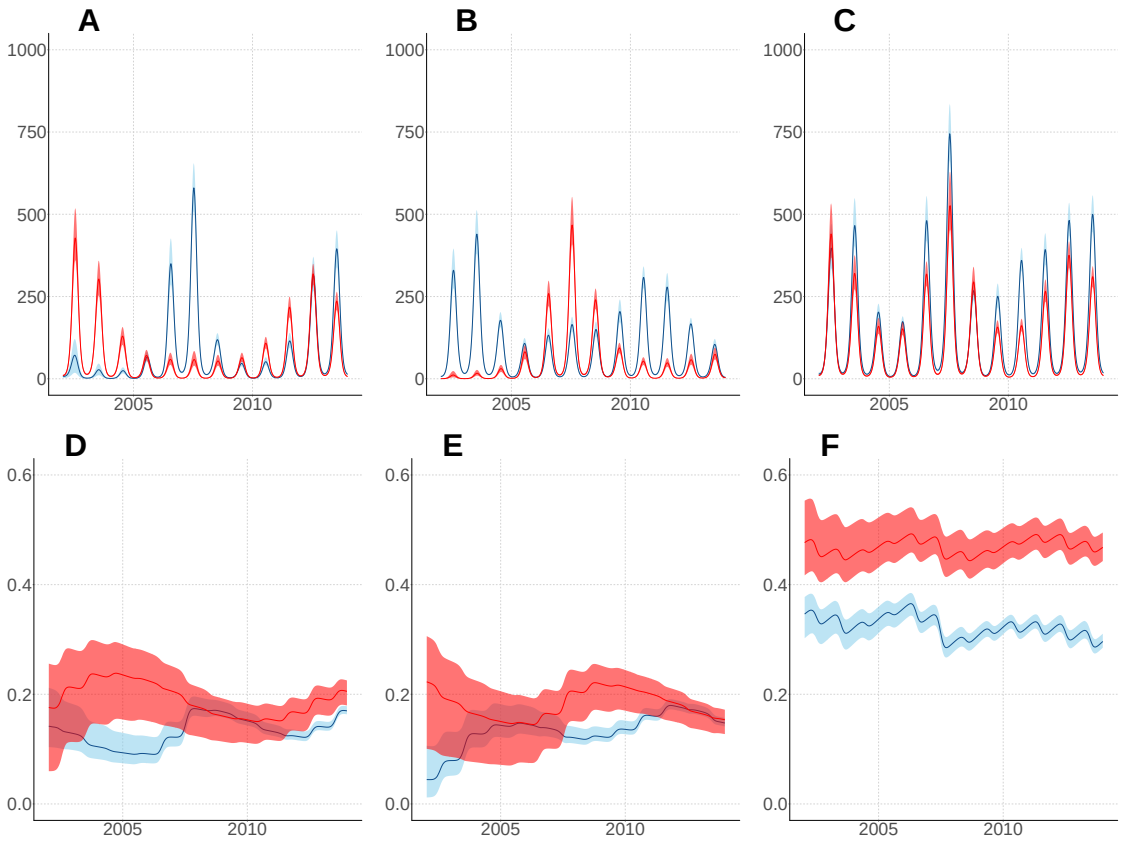


Figure A.6 – **Simulated trajectories for two parameterizations of the SEIR2psi model.** Colors red and blue correspond to two differing parameterizations. The blue one is the one commented in the main text. A. Individuals infected with strain 1 ($H_{I1} + H_{I21}$). B. Individuals infected with strain 2 ($H_{I2} + H_{I12}$). C. Individuals infected with any strain ($H_{I1} + H_{I21} + H_{I2} + H_{I12}$). D. Proportion of susceptible individuals with immunity to strain 1 (H_{S1}/N). E. Proportion of susceptible individuals with immunity to strain 2 (H_{S2}/N). F. Proportion of individuals susceptible to both strains (H_S/N).

Despite these differences, the estimated values, in terms of reproductive number, incidence proportion and observation rate remain in the same order of magnitude as the other models (cf. Table A.1).

Model		**SEIR2 A**	**SEIR2 B**	**SEIR2psi A**	**SEIR2psi B**
nb parameters		11	11	12	12
nb observations		704	704	704	704
ESTIMATION SET					
DIC		4492	4506	4446	4460
RMSE NDSS		25 (22-29)	23 (21-25)	21 (19-24)	24 (21-26)
RMSE DENFREE		23 (19-27)	24 (18-30)	25 (19-31)	22 (18-26)
TEST SET					
RMSE 2014-2015		17 (13-22)	22 (17-28)	13 (10-16)	17 (13-22)
mean R_0	median (95%CI)	1.45 (1.42-1.48)	1.99 (1.93-2.06)	2.47 (2.35-2.57)	1.77 (1.67-1.88)
max R_0	median (95%CI)	2.29 (2.23-2.35)	3.09 (2.98-3.21)	3.78 (3.59-3.94)	2.77 (2.61-2.94)
ψ	median (95%CI)			0.67 (0.62-0.73)	0.6 (0.53-0.69)
$H_S(0)/N$ (%)	median (95%CI)	49 (45-56)	41 (35-45)	35 (30-38)	48 (42-55)
$H_{S1}(0)/N$ (%)	median (95%CI)	22 (14-25)	14 (10-20)	14 (10-21)	18 (6-26)
$H_{S2}(0)/N$ (%)	median (95%CI)	27 (19-30)	4 (1-11)	4 (1-11)	22 (10-31)
Observation rate (%)	median (95%CI)	11 (10-12)	7 (7-8)	7 (6-7)	9 (8-11)
Median annual incidence proportion					
primary infection (%)	median 2002-2015 (min-max)	6 (3-10)	13 (6-18)	14 (7-27)	7 (4-13)
secondary infection (%)	median 2002-2015 (min-max)	3 (1-5)	7 (3-9)	7 (3-14)	4 (2-7)
mean Re strain 1	median (95%CI)	1.03 (1.02-1.03)	1.03 (1.02-1.03)	1.03 (1.02-1.03)	1.03 (1.02-1.03)
mean Re strain 2	median (95%CI)	1.04 (1.03-1.04)	1.02 (1.02-1.03)	1.02 (1.02-1.03)	1.04 (1.03-1.04)
max Re strain 1	median (95%CI)	1.73 (1.69-1.77)	1.73 (1.7-1.77)	1.74 (1.71-1.77)	1.69 (1.66-1.73)
max Re strain 2	median (95%CI)	1.72 (1.69-1.75)	1.72 (1.67-1.76)	1.69 (1.65-1.73)	1.73 (1.7-1.76)

Table A.1 – **Information criteria and epidemiological indicators, comparing two parameterizations of multi-strain models.** Models A refers to the ones with the best likelihood, described in the main text. Models B refer to the alternative parameterizations coloured in red in Figures A.4, A.5 and A.6.

Appendix B

Annexe: Comparison of stochastic and deterministic frameworks in dengue modeling

B.1 Model reactions

Here are the reactions describing the stochastic counterparts of all epidemiological models. The description of parameters and notations can be found in the main text.

B.1.1 SEIR model

Reaction	Effect	Rate
Infection	$(H_S, H_E, H_I, H_R) \to (H_S - 1, H_E + 1, H_I, H_R)$	$\beta(t)\frac{H_I+\delta}{N} \cdot \xi(t)$
Incubation	$(H_S, H_E, H_I, H_R) \to (H_S, H_E - 1, H_I + 1, H_R)$	σ
Recovery	$(H_S, H_E, H_I, H_R) \to (H_S, H_E, H_I - 1, H_R + 1)$	γ
Birth	$(H_S, H_E, H_I, H_R) \to (H_S + 1, H_E, H_I, H_R)$	$\mu_H N$
Age	$(H_S, H_E, H_I, H_R) \to (H_S - 1, H_E, H_I, H_R)$	μ_H
Age	$(H_S, H_E, H_I, H_R) \to (H_S, H_E - 1, H_I, H_R)$	μ_H
Age	$(H_S, H_E, H_I, H_R) \to (H_S, H_E, H_I - 1, H_R)$	μ_H
Age	$(H_S, H_E, H_I, H_R) \to (H_S, H_E, H_I, H_R - 1)$	μ_H

Table B.1 – **Reactions of the SEIR model.** In the framework without environmental stochasticity, $\xi(t) = 1$. When environmental stochasticity is added, $\xi(t)$ is a Gamma white noise with infinitesimal variance s^2.

B.1.2 SEIR-vector model

Reaction	Effect	Rate
Infection (host)	$(H_S, H_E, H_I, H_R, V_S, V_E, V_I) \to (H_S - 1, H_E + 1, H_I, H_R, V_S, V_E, V_I)$	$\beta_H \frac{V_I}{V} \cdot \xi(t)$
Incubation (host)	$(H_S, H_E, H_I, H_R, V_S, V_E, V_I) \to (H_S, H_E - 1, H_I + 1, H_R, V_S, V_E, V_I)$	σ
Recovery (host)	$(H_S, H_E, H_I, H_R, V_S, V_E, V_I) \to (H_S, H_E, H_I - 1, H_R + 1, V_S, V_E, V_I)$	γ
Birth (host)	$(H_S, H_E, H_I, H_R, V_S, V_E, V_I) \to (H_S + 1, H_E, H_I, H_R, V_S, V_E, V_I)$	$\mu_H N$
Age (host)	$(H_S, H_E, H_I, H_R, V_S, V_E, V_I) \to (H_S - 1, H_E, H_I, H_R, V_S, V_E, V_I)$	μ_H
Age (host)	$(H_S, H_E, H_I, H_R, V_S, V_E, V_I) \to (H_S, H_E - 1, H_I, H_R, V_S, V_E, V_I)$	μ_H
Age (host)	$(H_S, H_E, H_I, H_R, V_S, V_E, V_I) \to (H_S, H_E, H_I - 1, H_R, V_S, V_E, V_I)$	μ_H
Age (host)	$(H_S, H_E, H_I, H_R, V_S, V_E, V_I) \to (H_S, H_E, H_I, H_R - 1, V_S, V_E, V_I)$	μ_H
Infection (vector)	$(H_S, H_E, H_I, H_R, V_S, V_E, V_I) \to (H_S, H_E, H_I, H_R, V_S - 1, V_E + 1, V_I)$	$\beta_V(t) \frac{(H_I + \delta)}{N}$
EIP (vector)	$(H_S, H_E, H_I, H_R, V_S, V_E, V_I) \to (H_S, H_E, H_I, H_R, V_S, V_E - 1, V_I + 1)$	τ
Birth (vector)	$(H_S, H_E, H_I, H_R, V_S, V_E, V_I) \to (H_S, H_E, H_I, H_R, V_S + 1, V_E, V_I)$	$\mu_V V$
Death (vector)	$(H_S, H_E, H_I, H_R, V_S, V_E, V_I) \to (H_S, H_E, H_I, H_R, V_S - 1, V_E, V_I)$	μ_V
Death (vector)	$(H_S, H_E, H_I, H_R, V_S, V_E, V_I) \to (H_S, H_E, H_I, H_R, V_S, V_E - 1, V_I)$	μ_V
Death (vector)	$(H_S, H_E, H_I, H_R, V_S, V_E, V_I) \to (H_S, H_E, H_I, H_R, V_S, V_E, V_I - 1)$	μ_V

Table B.2 – **Reactions of the SEIR-vector model.** In the framework without environmental stochasticity, $\xi(t) = 1$. When environmental stochasticity is added, $\xi(t)$ is a Gamma white noise with infinitesimal variance s^2. $V = V_S + V_E + V_I$.

B.1.3 Laneri model

Reaction	Effect	Rate
Infection (host)	$(H_S, H_E, H_I, H_R, \kappa', \lambda') \to (H_S - 1, H_E + 1, H_I, H_R, \kappa', \lambda')$	$\lambda \cdot \xi(t)$
Incubation (host)	$(H_S, H_E, H_I, H_R, \kappa', \lambda') \to (H_S, H_E - 1, H_I + 1, H_R, \kappa', \lambda')$	σ
Recovery (host)	$(H_S, H_E, H_I, H_R, \kappa', \lambda') \to (H_S, H_E, H_I - 1, H_R + 1, \kappa', \lambda')$	γ
Birth (host)	$(H_S, H_E, H_I, H_R, \kappa', \lambda') \to (H_S + 1, H_E, H_I, H_R, \kappa', \lambda')$	$\mu_H N$
Age (host)	$(H_S, H_E, H_I, H_R, \kappa', \lambda') \to (H_S - 1, H_E, H_I, H_R, \kappa', \lambda')$	μ_H
Age (host)	$(H_S, H_E, H_I, H_R, \kappa', \lambda') \to (H_S, H_E - 1, H_I, H_R, \kappa', \lambda')$	μ_H
Age (host)	$(H_S, H_E, H_I, H_R, \kappa', \lambda') \to (H_S, H_E, H_I - 1, H_R, \kappa', \lambda')$	μ_H
Age (host)	$(H_S, H_E, H_I, H_R, \kappa', \lambda') \to (H_S, H_E, H_I, H_R - 1, \kappa', \lambda')$	μ_H
"Infection" (vector)	$(H_S, H_E, H_I, H_R, \kappa', \lambda') \to (H_S, H_E, H_I, H_R, \kappa' + 1, \lambda')$	$\beta(t) 2(H_I + \delta)\tau$
EIP (vector)	$(H_S, H_E, H_I, H_R, \kappa', \lambda') \to (H_S, H_E, H_I, H_R, \kappa' - 1, \lambda' + 1)$	2τ
"Death" (vector)	$(H_S, H_E, H_I, H_R, \kappa', \lambda') \to (H_S, H_E, H_I, H_R, \kappa', \lambda' - 1)$	2τ

Table B.3 – **Reactions of the Laneri model.** In the framework without environmental stochasticity, $\xi(t) = 1$. The quantities $\kappa' = \alpha\kappa$ and $\lambda' = \alpha\lambda$ are used. When environmental stochasticity is added, $\xi(t)$ is a Gamma white noise with infinitesimal variance s^2.

B.1.4 SEIAR model

Reaction	Effect	Rate
Infection	$(H_S, H_E, H_I, H_A, H_H, H_R) \to (H_S - 1, H_E + 1, H_I, H_A, H_H, H_R)$	$\beta(t)\frac{H_I+H_A+H_H+\delta}{N} \cdot \xi(t)$
Incubation, asymptomatic	$(H_S, H_E, H_I, H_A, H_H, H_R) \to (H_S, H_E - 1, H_I, H_A + 1, H_H, H_R)$	$\rho_A\sigma$
Incubation, hospital	$(H_S, H_E, H_I, H_A, H_H, H_R) \to (H_S, H_E - 1, H_I, H_A, H_H + 1, H_R)$	$\rho_H(1-\rho_A)\sigma$
Incubation, symptomatic	$(H_S, H_E, H_I, H_A, H_H, H_R) \to (H_S, H_E - 1, H_I + 1, H_A, H_H, H_R)$	$(1-\rho_H)(1-\rho_A)\sigma$
End of infectiousness (asymptomatic)	$(H_S, H_E, H_I, H_A, H_H, H_R) \to (H_S, H_E, H_I, H_A - 1, H_H, H_R + 1)$	γ
Recovery, hospital	$(H_S, H_E, H_I, H_A, H_H, H_R) \to (H_S, H_E, H_I, H_A, H_H - 1, H_R + 1)$	γ
Recovery, symptomatic	$(H_S, H_E, H_I, H_A, H_H, H_R) \to (H_S, H_E, H_I - 1, H_A, H_H, H_R + 1)$	γ
Birth	$(H_S, H_E, H_I, H_A, H_H, H_R) \to (H_S + 1, H_E, H_I, H_A, H_H, H_R)$	$\mu_H N$
Age	$(H_S, H_E, H_I, H_A, H_H, H_R) \to (H_S - 1, H_E, H_I, H_A, H_H, H_R)$	μ_H
Age	$(H_S, H_E, H_I, H_A, H_H, H_R) \to (H_S, H_E - 1, H_I, H_A, H_H, H_R)$	μ_H
Age	$(H_S, H_E, H_I, H_A, H_H, H_R) \to (H_S, H_E, H_I, H_A - 1, H_H, H_R)$	μ_H
Age	$(H_S, H_E, H_I, H_A, H_H, H_R) \to (H_S, H_E, H_I, H_A, H_H - 1, H_R)$	μ_H
Age	$(H_S, H_E, H_I, H_A, H_H, H_R) \to (H_S, H_E, H_I - 1, H_A, H_H, H_R)$	μ_H
Age	$(H_S, H_E, H_I, H_A, H_H, H_R) \to (H_S, H_E, H_I, H_A, H_H, H_R - 1)$	μ_H

Table B.4 – **Reactions of the SEIAR model.** In the framework without environmental stochasticity, $\xi(t) = 1$. When environmental stochasticity is added, $\xi(t)$ is a Gamma white noise with infinitesimal variance s^2.

B.1.5 SEIR2 model

Reaction	Effect	Rate
1st infection with strain 1	$(H_S, H_{E1}, H_{I1}, H_{S1}, H_{E2}, H_{I2}, H_{S2}, H_{E12}, H_{I12}, H_{E21}, H_{I21}, H_R)$ $\to (H_S - 1, H_{E1} + 1, H_{I1}, H_{S1}, H_{E2}, H_{I2}, H_{S2}, H_{E12}, H_{I12}, H_{E21}, H_{I21}, H_R)$	$\beta(t)\frac{H_{I1}+\psi H_{I21}+\delta_1}{N} \cdot \xi_1(t)$
Incubation strain 1	$(H_S, H_{E1}, H_{I1}, H_{S1}, H_{E2}, H_{I2}, H_{S2}, H_{E12}, H_{I12}, H_{E21}, H_{I21}, H_R)$ $\to (H_S, H_{E1} - 1, H_{I1} + 1, H_{S1}, H_{E2}, H_{I2}, H_{S2}, H_{E12}, H_{I12}, H_{E21}, H_{I21}, H_R)$	σ
Recovery strain 1	$(H_S, H_{E1}, H_{I1}, H_{S1}, H_{E2}, H_{I2}, H_{S2}, H_{E12}, H_{I12}, H_{E21}, H_{I21}, H_R)$ $\to (H_S, H_{E1}, H_{I1} - 1, H_{S1} + 1, H_{E2}, H_{I2}, H_{S2}, H_{E12}, H_{I12}, H_{E21}, H_{I21}, H_R)$	γ
2nd infection with strain 2	$(H_S, H_{E1}, H_{I1}, H_{S1}, H_{E2}, H_{I2}, H_{S2}, H_{E12}, H_{I12}, H_{E21}, H_{I21}, H_R)$ $\to (H_S, H_{E1}, H_{I1}, H_{S1} - 1, H_{E2}, H_{I2}, H_{S2}, H_{E12} + 1, H_{I12}, H_{E21}, H_{I21}, H_R)$	$\beta(t)\frac{H_{I2}+\psi H_{I12}+\delta_2}{N} \cdot \xi_2(t)$
Incubation strain 2	$(H_S, H_{E1}, H_{I1}, H_{S1}, H_{E2}, H_{I2}, H_{S2}, H_{E12}, H_{I12}, H_{E21}, H_{I21}, H_R)$ $\to (H_S, H_{E1}, H_{I1}, H_{S1}, H_{E2}, H_{I2}, H_{S2}, H_{E12} - 1, H_{I12} + 1, H_{E21}, H_{I21}, H_R)$	σ
Recovery strain 2	$(H_S, H_{E1}, H_{I1}, H_{S1}, H_{E2}, H_{I2}, H_{S2}, H_{E12}, H_{I12}, H_{E21}, H_{I21}, H_R)$ $\to (H_S, H_{E1}, H_{I1}, H_{S1}, H_{E2}, H_{I2}, H_{S2}, H_{E12}, H_{I12} - 1, H_{E21}, H_{I21}, H_R + 1)$	γ
1st infection with strain 2	$(H_S, H_{E1}, H_{I1}, H_{S1}, H_{E2}, H_{I2}, H_{S2}, H_{E12}, H_{I12}, H_{E21}, H_{I21}, H_R)$ $\to (H_S - 1, H_{E1}, H_{I1}, H_{S1}, H_{E2} + 1, H_{I2}, H_{S2}, H_{E12}, H_{I12}, H_{E21}, H_{I21}, H_R)$	$\beta(t)\frac{H_{I2}+\psi H_{I12}+\delta_2}{N} \cdot \xi_2(t)$
Incubation strain 2	$(H_S, H_{E1}, H_{I1}, H_{S1}, H_{E2}, H_{I2}, H_{S2}, H_{E12}, H_{I12}, H_{E21}, H_{I21}, H_R)$ $\to (H_S, H_{E1}, H_{I1}, H_{S1}, H_{E2} - 1, H_{I2} + 1, H_{S2}, H_{E12}, H_{I12}, H_{E21}, H_{I21}, H_R)$	σ
Recovery strain 2	$(H_S, H_{E1}, H_{I1}, H_{S1}, H_{E2}, H_{I2}, H_{S2}, H_{E12}, H_{I12}, H_{E21}, H_{I21}, H_R)$ $\to (H_S, H_{E1}, H_{I1}, H_{S1}, H_{E2}, H_{I2} - 1, H_{S2} + 1, H_{E12}, H_{I12}, H_{E21}, H_{I21}, H_R)$	γ
2nd infection with strain 1	$(H_S, H_{E1}, H_{I1}, H_{S1}, H_{E2}, H_{I2}, H_{S2}, H_{E12}, H_{I12}, H_{E21}, H_{I21}, H_R)$ $\to (H_S, H_{E1}, H_{I1}, H_{S1}, H_{E2}, H_{I2}, H_{S2} - 1, H_{E12}, H_{I12}, H_{E21} + 1, H_{I21}, H_R)$	$\beta(t)\frac{H_{I1}+\psi H_{I21}+\delta_1}{N} \cdot \xi_1(t)$
Incubation strain 1	$(H_S, H_{E1}, H_{I1}, H_{S1}, H_{E2}, H_{I2}, H_{S2}, H_{E12}, H_{I12}, H_{E21}, H_{I21}, H_R)$ $\to (H_S, H_{E1}, H_{I1}, H_{S1}, H_{E2}, H_{I2}, H_{S2}, H_{E12}, H_{I12}, H_{E21} - 1, H_{I21} + 1, H_R)$	σ
Recovery strain 1	$(H_S, H_{E1}, H_{I1}, H_{S1}, H_{E2}, H_{I2}, H_{S2}, H_{E12}, H_{I12}, H_{E21}, H_{I21}, H_R)$ $\to (H_S, H_{E1}, H_{I1}, H_{S1}, H_{E2}, H_{I2}, H_{S2}, H_{E12}, H_{I12}, H_{E21}, H_{I21} - 1, H_R + 1)$	γ
Birth	$(H_S, H_{E1}, H_{I1}, H_{S1}, H_{E2}, H_{I2}, H_{S2}, H_{E12}, H_{I12}, H_{E21}, H_{I21}, H_R)$ $\to (H_S + 1, H_{E1}, H_{I1}, H_{S1}, H_{E2}, H_{I2}, H_{S2}, H_{E12}, H_{I12}, H_{E21}, H_{I21}, H_R)$	$\mu_H N$
Age	$(H_S, H_{E1}, H_{I1}, H_{S1}, H_{E2}, H_{I2}, H_{S2}, H_{E12}, H_{I12}, H_{E21}, H_{I21}, H_R)$	μ_H

	$\to (H_S - 1, H_{E1}, H_{I1}, H_{S1}, H_{E2}, H_{I2}, H_{S2}, H_{E12}, H_{I12}, H_{E21}, H_{I21}, H_R)$	
Age	$(H_S, H_{E1}, H_{I1}, H_{S1}, H_{E2}, H_{I2}, H_{S2}, H_{E12}, H_{I12}, H_{E21}, H_{I21}, H_R)$	μ_H
	$\to (H_S, H_{E1} - 1, H_{I1}, H_{S1}, H_{E2}, H_{I2}, H_{S2}, H_{E12}, H_{I12}, H_{E21}, H_{I21}, H_R)$	
Age	$(H_S, H_{E1}, H_{I1}, H_{S1}, H_{E2}, H_{I2}, H_{S2}, H_{E12}, H_{I12}, H_{E21}, H_{I21}, H_R)$	μ_H
	$\to (H_S, H_{E1}, H_{I1} - 1, H_{S1}, H_{E2}, H_{I2}, H_{S2}, H_{E12}, H_{I12}, H_{E21}, H_{I21}, H_R)$	
Age	$(H_S, H_{E1}, H_{I1}, H_{S1}, H_{E2}, H_{I2}, H_{S2}, H_{E12}, H_{I12}, H_{E21}, H_{I21}, H_R)$	μ_H
	$\to (H_S, H_{E1}, H_{I1}, H_{S1} - 1, H_{E2}, H_{I2}, H_{S2}, H_{E12}, H_{I12}, H_{E21}, H_{I21}, H_R)$	
Age	$(H_S, H_{E1}, H_{I1}, H_{S1}, H_{E2}, H_{I2}, H_{S2}, H_{E12}, H_{I12}, H_{E21}, H_{I21}, H_R)$	μ_H
	$\to (H_S, H_{E1}, H_{I1}, H_{S1}, H_{E2} - 1, H_{I2}, H_{S2}, H_{E12}, H_{I12}, H_{E21}, H_{I21}, H_R)$	
Age	$(H_S, H_{E1}, H_{I1}, H_{S1}, H_{E2}, H_{I2}, H_{S2}, H_{E12}, H_{I12}, H_{E21}, H_{I21}, H_R)$	μ_H
	$\to (H_S, H_{E1}, H_{I1}, H_{S1}, H_{E2}, H_{I2} - 1, H_{S2}, H_{E12}, H_{I12}, H_{E21}, H_{I21}, H_R)$	
Age	$(H_S, H_{E1}, H_{I1}, H_{S1}, H_{E2}, H_{I2}, H_{S2}, H_{E12}, H_{I12}, H_{E21}, H_{I21}, H_R)$	μ_H
	$\to (H_S, H_{E1}, H_{I1}, H_{S1}, H_{E2}, H_{I2}, H_{S2} - 1, H_{E12}, H_{I12}, H_{E21}, H_{I21}, H_R)$	
Age	$(H_S, H_{E1}, H_{I1}, H_{S1}, H_{E2}, H_{I2}, H_{S2}, H_{E12}, H_{I12}, H_{E21}, H_{I21}, H_R)$	μ_H
	$\to (H_S, H_{E1}, H_{I1}, H_{S1}, H_{E2}, H_{I2}, H_{S2}, H_{E12} - 1, H_{I12}, H_{E21}, H_{I21}, H_R)$	
Age	$(H_S, H_{E1}, H_{I1}, H_{S1}, H_{E2}, H_{I2}, H_{S2}, H_{E12}, H_{I12}, H_{E21}, H_{I21}, H_R)$	μ_H
	$\to (H_S, H_{E1}, H_{I1}, H_{S1}, H_{E2}, H_{I2}, H_{S2}, H_{E12}, H_{I12} - 1, H_{E21}, H_{I21}, H_R)$	
Age	$(H_S, H_{E1}, H_{I1}, H_{S1}, H_{E2}, H_{I2}, H_{S2}, H_{E12}, H_{I12}, H_{E21}, H_{I21}, H_R)$	μ_H
	$\to (H_S, H_{E1}, H_{I1}, H_{S1}, H_{E2}, H_{I2}, H_{S2}, H_{E12}, H_{I12}, H_{E21} - 1, H_{I21}, H_R)$	
Age	$(H_S, H_{E1}, H_{I1}, H_{S1}, H_{E2}, H_{I2}, H_{S2}, H_{E12}, H_{I12}, H_{E21}, H_{I21}, H_R)$	μ_H
	$\to (H_S, H_{E1}, H_{I1}, H_{S1}, H_{E2}, H_{I2}, H_{S2}, H_{E12}, H_{I12}, H_{E21}, H_{I21} - 1, H_R)$	
Age	$(H_S, H_{E1}, H_{I1}, H_{S1}, H_{E2}, H_{I2}, H_{S2}, H_{E12}, H_{I12}, H_{E21}, H_{I21}, H_R)$	μ_H
	$\to (H_S, H_{E1}, H_{I1}, H_{S1}, H_{E2}, H_{I2}, H_{S2}, H_{E12}, H_{I12}, H_{E21}, H_{I21}, H_R - 1)$	

Table B.5 – **Reactions of the SEIR2 model.** In the framework without environmental stochasticity, $\xi_1(t) = \xi_2(t) = 1$. When environmental stochasticity is added, $\xi_1(t)$ and $\xi_2(t)$ are Gamma white noises with infinitesimal variance s^2.

B.2 Values for fixed initial conditions

Model	$H_S(0)$	$H_E(0)$	$H_I(0)$	$H_H(0)$	$H_A(0)$
SEIR	72,853	30	9	.	.
Laneri	73,432	11	9	.	.
SEIR-vector	56,517	13	10	.	.
SEIAR	73,875	29	6	1	1

Model	$H_S(0)$	$H_{S1}(0)$	$H_{S2}(0)$	$H_{I1}(0)$	$H_{I2}(0)$	$H_{I21}(0)$	$H_{I12}(0)$	$H_{E1}(0)$	$H_{E2}(0)$	$H_{E21}(0)$	$H_{E12}(0)$
SEIR2	51,480	19,919	28,821	5	8	3	3	16	26	9	10

Table B.6 – **Initial conditions for the date 2011-12-26.** These values are obtained using a deterministic model estimated by MCMC on 14 years of NDSS data and the two-year DENFREE data (Champagne et al., 2018).

B.3 Convergence diagnosis

B.3.1 Trace plots

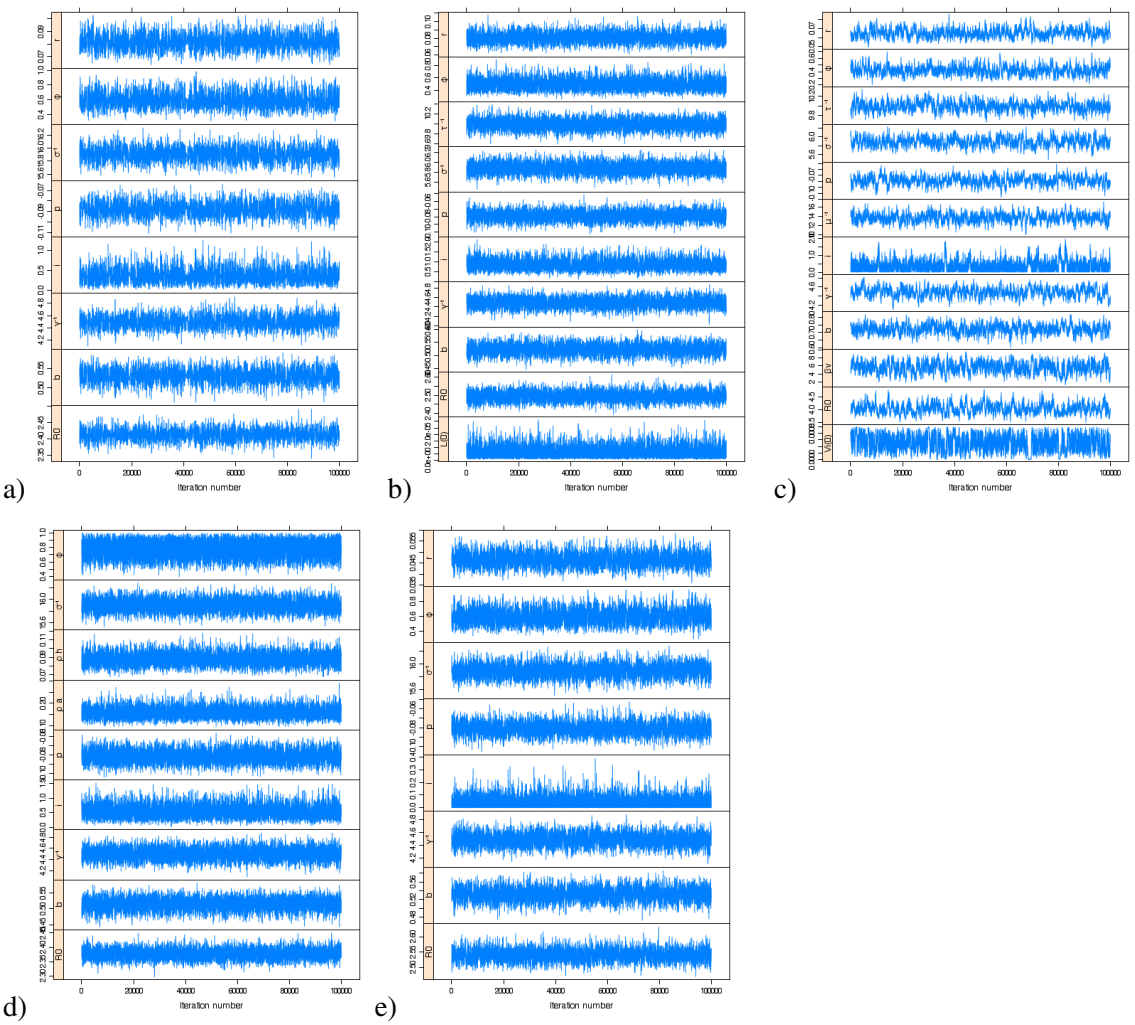


Figure B.1 – **Trace plots in deterministic framework.** Trace plots for the MCMC algorithm, with 100,000 iterations. a) SEIR. b) Laneri. c) SEIR-vector. d) SEIAR. e) SEIR2

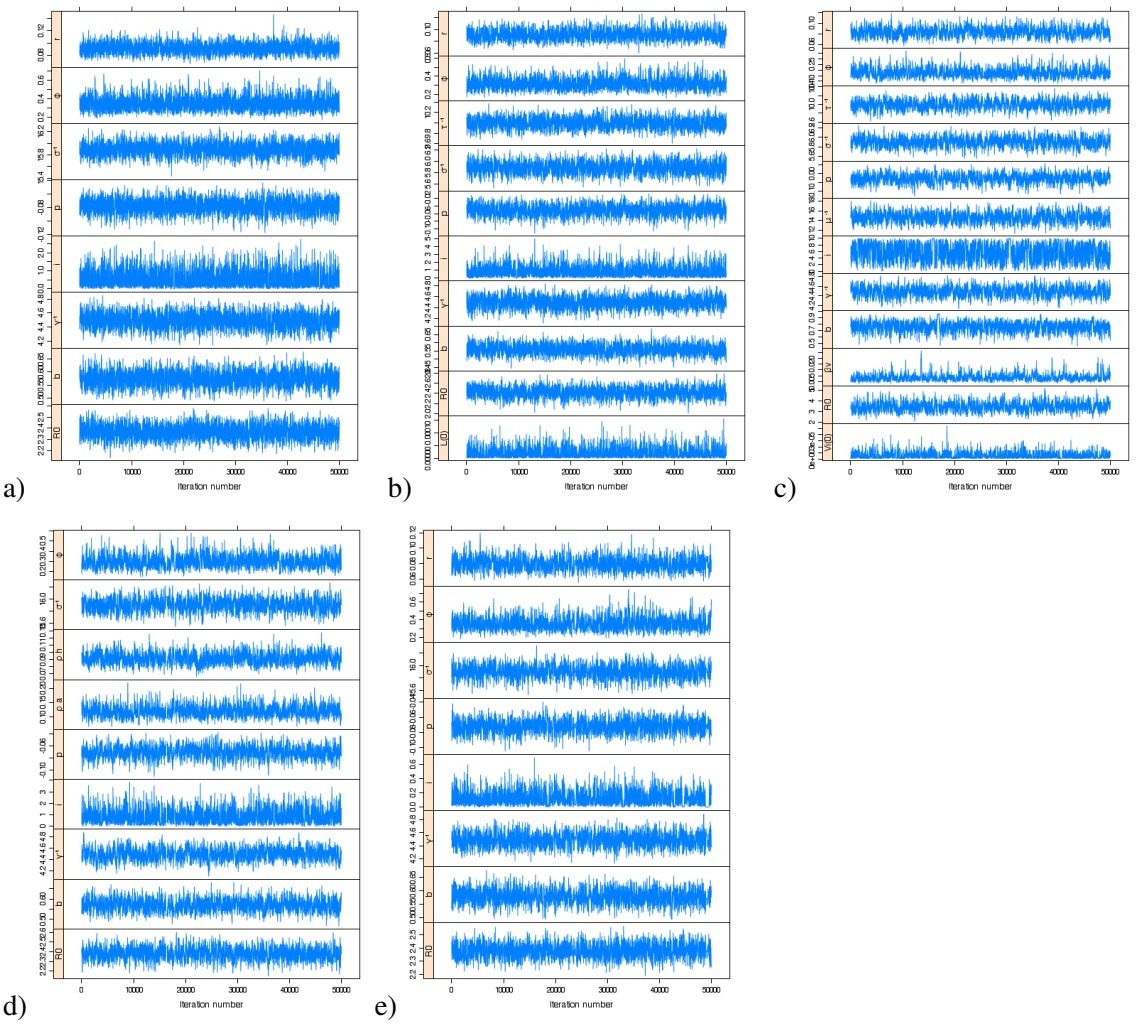


Figure B.2 – **Trace plots in stochastic framework.** Trace plots for the adaptative PMCMC algorithm, with 50,000 iterations and 10,000 particles. a) SEIR. b) Laneri. c) SEIR-vector. d) SEIAR. e) SEIR2.

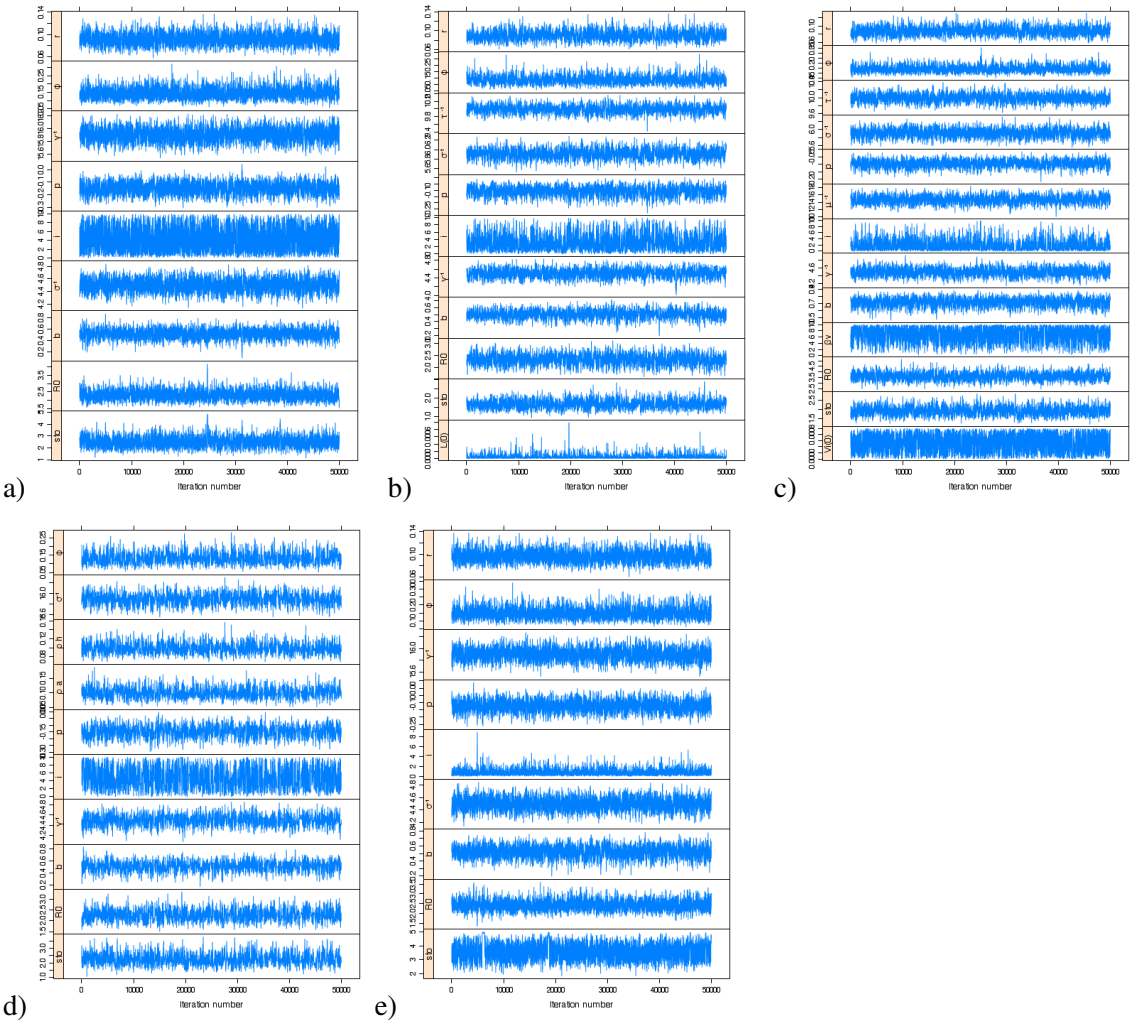


Figure B.3 – **Trace plots in stochastic framework with white noise on the infection reaction.** Trace plots for the adaptative PMCMC algorithm, with 50,000 iterations and 10,000 particles. a) SEIR. b) Laneri. c) SEIR-vector. d) SEIAR. e) SEIR2.

B.4 Posterior distributions (other parameters)

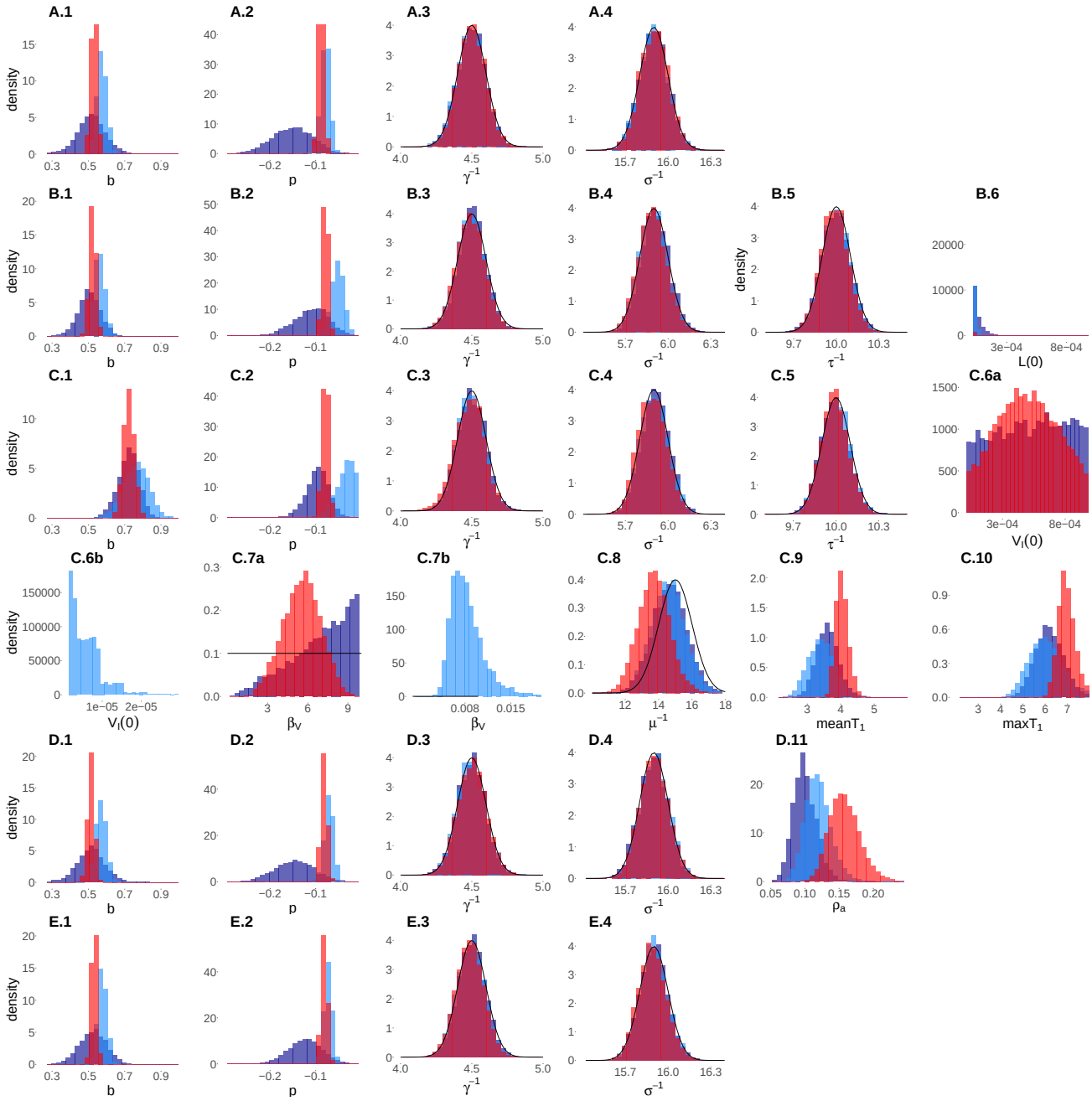


Figure B.4 – **Posterior distributions.** Posterior distributions in the five models for the remaining parameters. Red: deterministic case. Light blue: stochastic case with demographic stochasticity only. Dark blue: stochastic case with demographic and environmental stochasticity. Informative prior distributions are represented with a plain black line. A. SEIR model. B. Laneri model. C. SEIR-vector model. D. SEIAR model. E. SEIR2 model. Parameters : 1. amplitude of seasonality b , 2. phase of seasonality p , 3. duration of infectiousness for human host γ^{-1} , 4. incubation period in human host σ^{-1} , 5. incubation period in the vector τ^{-1}, 6. initial number of vectors $L(0)$ in Laneri model, and $V_I(0)$ SEIR-vector models (histogram for the stochastic case with demographic stochasticity only is presented separately in C.6b for appropriate scaling), 7. transmission parameter from host to vector β_V (histogram for the stochastic case with demographic stochasticity only is presented separately in C.7b for appropriate scaling), 8. mosquito lifespan μ^{-1}, 9. Mean T_1, 10. Max T_1, 11. proportion of asymptomatic infections ρ_A.

B.5 Information criteria, using absolute error

$$\text{mean MAE NDSS} = \frac{1}{M}\sum_{s=1}^{S}\text{MAE NDSS}_s = \frac{1}{M}\sum_{s=1}^{S}\frac{1}{K}\sum_{k=1}^{K}|C_N^{\text{sim}_s}(k) - C_N^{\text{obs}}(k)|$$

$$\text{mean MAE DENFREE} = \frac{1}{M}\sum_{s=1}^{S}\text{MAE DENFREE}_s = \frac{1}{M}\sum_{s=1}^{S}\frac{1}{K}\sum_{k=1}^{K}|C_{DS}^{\text{sim}_s}(k) - C_{DS}^{\text{obs}}(k)|$$

Model	**SEIR**	**Laneri**	**SEIR-vector**	**SEIAR**	**SEIR2**
nb observations	247	247	247	286	286
DETERMINISTIC					
nb parameters	8	10	12	9	8
log-likelihood	-758	-756	-751	-817	-848
DIC	1529	1528	1529	1647	1712
mean MAE NDSS	9.6	9.8	9.3	9.9	9.4
mean MAE DENFREE	17.8	18.7	19.9	17.4	21.9
STOCHASTIC (demographic)					
nb parameters	8	10	12	9	8
log-likelihood	-733	-731	-719	-789	-800
DIC	1481	1480	1460	1598	1620
mean MAE NDSS	10.7	12.7	16.5	10.5	12.2
mean MAE DENFREE	18.2	20.2	22.1	17.8	17.9
STOCHASTIC (demographic + environmental)					
nb parameters	9	11	13	10	9
log-likelihood	-706	-708	-709	-760	-771
DIC	1427	1435	1440	1536	1558
mean MAE NDSS	17.4	15.4	13.4	16.8	21.6
mean MAE DENFREE	24.4	22.9	18.6	22.8	23.3
White noise SD s (median + %95 CI)	2.5 (1.8-3.4)	1.7 (1.2-2.1)	1.9 (1.5-2.4)	2.2 (1.5-3.1)	3.6 (2.6-4.8)

Table B.7 – **Information criteria.** Information criteria for the five models in the three frameworks. When environmental stochasticity is included, the median and credible intervals for the infinitesimal standard deviation are indicated.

Appendix C

Accounting for non-stationarity in epidemiology by embedding time-varying parameters in stochastic models

PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Accounting for non-stationarity in epidemiology by embedding time-varying parameters in stochastic models

Bernard Cazelles[1,2,3]*, **Clara Champagne**[1,4], **Joseph Dureau**[5]

1 Institut de Biologie de l'Ecole Normale Supérieure (IBENS), Ecole Normale Supérieure, CNRS UMR 8197, Paris, France, **2** International Center for Mathematical and Computational Modeling of Complex Systems (UMMISCO), UMI 209, UPMC/IRD, France, **3** Hosts, Vectors and Infectious Agents, CNRS URA 3012, Institut Pasteur, Paris, France, **4** CREST, ENSAE, Université Paris Saclay, Palaiseau, France, **5** SNIPS, Paris, France

* cazelles@biologie.ens.fr

Citation: Cazelles B, Champagne C, Dureau J (2018) Accounting for non-stationarity in epidemiology by embedding time-varying parameters in stochastic models. PLoS Comput Biol 14(8): e1006211. https://doi.org/10.1371/journal.pcbi.1006211

Editor: Katia Koelle, Emory University, UNITED STATES

Received: December 16, 2017

Accepted: May 18, 2018

Published: August 15, 2018

Data Availability Statement: Data are in the Supporting Information in a zip file.

Funding: BC and CC have been supported by a grant from Agence Nationale de la Recherche for the PANIC project (ANR-14-CE02-0015-01). This research has also received funding from the European Commission Seventh Framework Program FP7 for the DENFREE project under Grant Agreement 282378. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

The spread of disease through human populations is complex. The characteristics of disease propagation evolve with time, as a result of a multitude of environmental and anthropic factors, this non-stationarity is a key factor in this huge complexity. In the absence of appropriate external data sources, to correctly describe the disease propagation, we explore a flexible approach, based on stochastic models for the disease dynamics, and on diffusion processes for the parameter dynamics. Using such a diffusion process has the advantage of not requiring a specific mathematical function for the parameter dynamics. Coupled with particle MCMC, this approach allows us to reconstruct the time evolution of some key parameters (average transmission rate for instance). Thus, by capturing the time-varying nature of the different mechanisms involved in disease propagation, the epidemic can be described. Firstly we demonstrate the efficiency of this methodology on a toy model, where the parameters and the observation process are known. Applied then to real datasets, our methodology is able, based solely on simple stochastic models, to reconstruct complex epidemics, such as flu or dengue, over long time periods. Hence we demonstrate that time-varying parameters can improve the accuracy of model performances, and we suggest that our methodology can be used as a first step towards a better understanding of a complex epidemic, in situation where data is limited and/or uncertain.

Author summary

As our world becomes more and more globalized, infectious disease poses an ever-increasing threat to human health. The multitude of environmental and behavioral factors, which account for the spread of infectious diseases, are ever-evolving and thus infectious diseases propagation is complex.

In the face of this complexity, mathematical models offer valuable tools to study the dynamics of epidemic diseases. Developing adequate statistical and mathematical tools, that take account of the time-varying nature of the different mechanisms responsible for

Competing interests: The authors have declared that no competing interests exist.

disease propagation, remains a major challenge. To take this increasingly important aspect into consideration, we propose a flexible methodology that encompasses time-varying aspects of the epidemic. It does this via diffusion process equations for time-varying parameters. Considering the relative paucity of available data, our principal assertion is that it is preferable to use this flexible framework with time-varying parameters, that tracks epidemiological patterns, and updates the key parameters according to data, than to use a more complex model.

Introduction

Our world constantly faces the threat of emerging and re-emerging diseases and it has been shown that this has intensified over the past 50 years. This intensification is due, in part, to climate change, urbanization and globalization [1] meaning that infectious diseases remain a constant and unpredictable threat to human health.

Numerous factors contribute to the propagation of an infectious disease. These include increased human connectivity, limited availability of economic resources for adequate intervention, increasing antimicrobial resistance, evolution of the dominant strains and increasing parasite and vector resistance to the most widely used drugs and insecticides, etc. A key factor in this huge complexity is non-stationarity [2], meaning that the characteristics of the dynamical epidemiological processes evolve with time. Thus, the mechanisms of transmission are uncertain, making it difficult to obtain quantitative predictions. One of the classic aspects of non-stationarity, is the seasonality of epidemiological dynamics, linked to environment and climate [3–4] but the environmental variability can shape the disease propagation in unforeseeable ways on small and large spatial scales [5–8]. Intervention and control may also modify the course of an epidemic. A less well-described but equally important cause of non-stationarity is linked to social cycles, *e.g.* school terms, religious holidays and agricultural cycles [9–13]. Research increasingly focuses on the effect of behavioral change in the presence of epidemiological risk as a source of non-stationarity [14–16]. Societal responses and changing human behavior play an important role in our connected society. Thus, during an epidemic, depending on the availability of information on the disease, people exhibit a variety of behaviors including anxiety and social distancing that might greatly influence the course of an epidemic.

For all of the above reasons, the spread of pathogens through human populations can be complex and hard to predict. In the face of this complexity, mathematical models offer valuable tools to study the dynamics of epidemic diseases, in order to synthesize information to understand observed epidemiological patterns and to test different hypothesis on the underlying key mechanisms [17]. Moreover, mathematical models play a crucial role in infectious disease prevention by assessing the impact of different control measures, *e.g.* vaccination strategies [18–19].

Nonetheless, there are very few, if indeed any, cases where modelers can access all the necessary information to reliably predict the course of an epidemic. This is particularly the case when we consider the non-stationarity features of epidemics and their transient nature poses a challenging problem for modeling. Further to this, different hypothesis must be formulated. In the case of influenza, for example, some researchers have suggested using a quantitative relationship between climatic variables and the effective transmission rate [20]. Another recent example illustrates non-stationarity in epidemiology. Between November 2010 and February 2011, despite a low level of population susceptibility, an unexpected third wave of infection by the H1N1pdm09 pandemic virus was observed in the United Kingdom. Using a

compartmental mathematical model of influenza transmission, this third wave was explained, by a substantial increase in the transmissibility of the H1N1pdm09 virus [21]. It has been proposed that this modification of the transmissibility was caused by the virus evolution with a better adaptation to the human host, or by climatic factors, namely the very cold weather experienced in the United Kingdom at that time, or by a combination of these factors [21].

To tackle the problem of non-stationarity in epidemiology, some approaches use a linear function to reconstruct the effective reproduction number (average number of secondary cases per primary case, R_{eff}). Wallinga and Teunis [22] proposed a generic method that requires only case incidence data and the distribution of the serial interval (the time between the onset of symptoms in a primary case and the onset of symptoms of secondary cases) to estimate R_{eff} over the course of an epidemic. This approach has been improved by numerous authors and applied to real time estimation of R_{eff} [23–25]. Other authors estimated using mathematical models R_{eff} for each season [26,27]. To calculate the time-varying infection rate the reconstructed time series of R_{eff} derived from the notification data can be used [28].

However, the time evolution of R_{eff} by definition depends not only on the time evolution of the epidemiological parameters but also on the number of susceptibles. More complex approaches have therefore been proposed. These approaches use semi-mechanistic models that incorporate the known compartmental structure of disease transmission but do not specify the form of the transmission rate equation that is estimated based on the data. In an early paper, the force of infection is estimated by neuronal network or kernel regression [29]. Now it is more common to use B-spline [30–33]. An alternative approach is to use diffusion models driven by fractional Brownian motion to model time-varying parameter of major epidemiological significance [34–36]. The models developed assign diffusion processes to the time-varying parameters embedded in a state-space framework. With the Kalman filter, the time-evolution of some key parameters (average transmission rate, mean incubation rate, and basic reproduction rate) were estimated during the course of the HIV/AIDS epidemics in the *Paris region* [34–35]. Dureau et al. [36] generalized this approach using a Bayesian framework with an adjusted adaptive particle Markov chain Monte Carlo algorithm (PMCMC), but only applied to the transmission rate, for short epidemics, with application to the 2009 pandemic flu. Very recently, an algorithm relying on robustly estimating the time-varying infection rate, based on the method of the unknown input observers from control theory, has been proposed [37]. Similarly, an approach for the reconstruction of time-dependent transmission rates, by projecting onto a finite subspace, spanned by Legendre polynomials, has been introduced [38].

In our previous works [34–36], we have introduced an approach for reconstructing the time evolution of some key parameters with just the weak hypothesis according to which they follow a basic stochastic process. The parameter time evolution is estimated solely based on observations of the incidence or the prevalence. Here, we propose to expand this approach to recurrent epidemics over time periods longer than just one season. The underlying idea of this approach is to capture unknown influences by considering time-varying parameters. As with other semi-mechanistic approaches, the key advantage of this approach, for the parameter dynamics, is that it is data-driven, and thus the shape of change does not need to be specified beforehand. We applied our framework both to a toy model, where parameters and observation process are known, and to two real data sets. This allows us to demonstrate that this data-driven approach is very effective for tackling the non-stationarity of recurrent epidemics, even with long time series. It has other benefits too. For instance, with limited access to information, it can capture unknown influences. By so doing, and by analyzing the parameter time evolution, this framework allows a more thorough analysis of the different influences, facilitating their introduction in more complex models with pertinent hypotheses based on observations.

Models with time-varying parameters

Our approach is based on three main components: an epidemiological model embedded in a state-space framework, a diffusion process for each time-varying parameter and an up-to-date Bayesian inference technique based on adaptive PMCMC.

The main advantage of the state-space framework is the use of two sets of equations, the first set describes the propagation of the disease in the population and the second is for the observation process. This allows for consideration of unknowns and uncertainty both in the epidemiological mechanisms and in the partial observation of the disease:

$$\begin{cases} \dot{x}(t) = g(t, x(t), \theta'(t), u(t)) \\ y(t)|x(t) \sim f(h(x(t)), \theta'(t)) \end{cases} \quad (1)$$

The first equation is for the epidemiological model, with $x(t)$ representing the state variables (for instance, $S(t)$ the susceptibles, $I(t)$ the infectious and $R(t)$ the removed for the classical SIR model) and $\theta'(t)$ the epidemiological parameters. The second is the observational process defined by probabilistic law f and a reporting rate on transformation of some state variable $h(x(t))$ because we may not be able to directly measure all state variables but just some or a function of them. In these equations, $y(t)$ are partial observations of $x(t)$, $u(t)$ is the process noise describing different form of stochasticity and the observational noise is included in f. In our applications, $h(x(t))$ will be the cumulative sum of new cases over the observation time step, that is generally the quantity observed by Public Health systems.

Considering the time-varying parameters $\theta(t)$ as a subset of $\theta'(t)$, we make the assumption that they evolve more or less randomly and do not follow a defined mathematical function. In the absence of prior information the use of diffusion motion allows us to impose few restrictions on the evolution of $\theta(t)$. We consider that they follow a continuous diffusion process (a discrete diffusion process was used in [35]):

$$\begin{aligned} & d\theta(t) = \sigma dB(t) \\ & or \\ & d\log(\theta(t)) = \sigma dB(t) \end{aligned} \quad (2)$$

where σ is the volatility of the Brownian process ($dB(t)$) and will be estimated during the fitting process. The use of a Brownian process can be viewed as a weak hypothesis for the imposed motion of $\theta(t)$ and the volatility σ being a regularized factor. Intuitively, the higher the values of σ the larger the changes in $\theta(t)$. The logarithm transformation avoids negative values which have no biological meaning. When prior knowledge on $\theta(t)$ is available this Brownian process can be modified to account for a drift in (2) (see [36]).

For the time-varying parameter, we focus on the parameter of the force of infection classically defined as:

$$\lambda(t) = \beta(t).\frac{S(t).I(t)}{N} \quad (3)$$

with $\beta(t)$ the transmission rate usually defined by a sinusoidal function. The control or the behavior modification can also be taken into account:

$$\lambda(t) = \beta(t).\frac{(S(t)^{\varepsilon_S(t)}).(I(t)^{\varepsilon_I(t)})}{N} \quad (4)$$

$\varepsilon^i(t)$ describe the clustering of the population [39,40] but can also describe a reduction in the population due to voluntary avoidance behavior or social distancing. However due to the

absence of structural identifiability properties [41, 42] it should be very difficult to estimate simultaneously both $\beta(t)$ and $\varepsilon^i(t)$.

For model estimation we use Bayesian methods, coupling particle filter and MCMC for partially observed stochastic non-linear systems [36,43] (see Methods). The implementation provided in SSM software [36] is used.

Results

Reconstructing dynamics and time-varying parameter for a toy SIRS model

We start our demonstration by showing that it is possible to reconstruct both the trajectory of a SIRS model (SIRS stands for Susceptibles, Infectious, Removed and Susceptibles again) and that of the sinusoidal transmission rate. In this example, the trajectory of each variable has been simulated with a model for which all the parameters were known. Moreover we also knew the observation process that has generated the data, a Poisson law for the incidence with an observation rate equal to 1. Fig 1 displays the reconstructed trajectories of both the

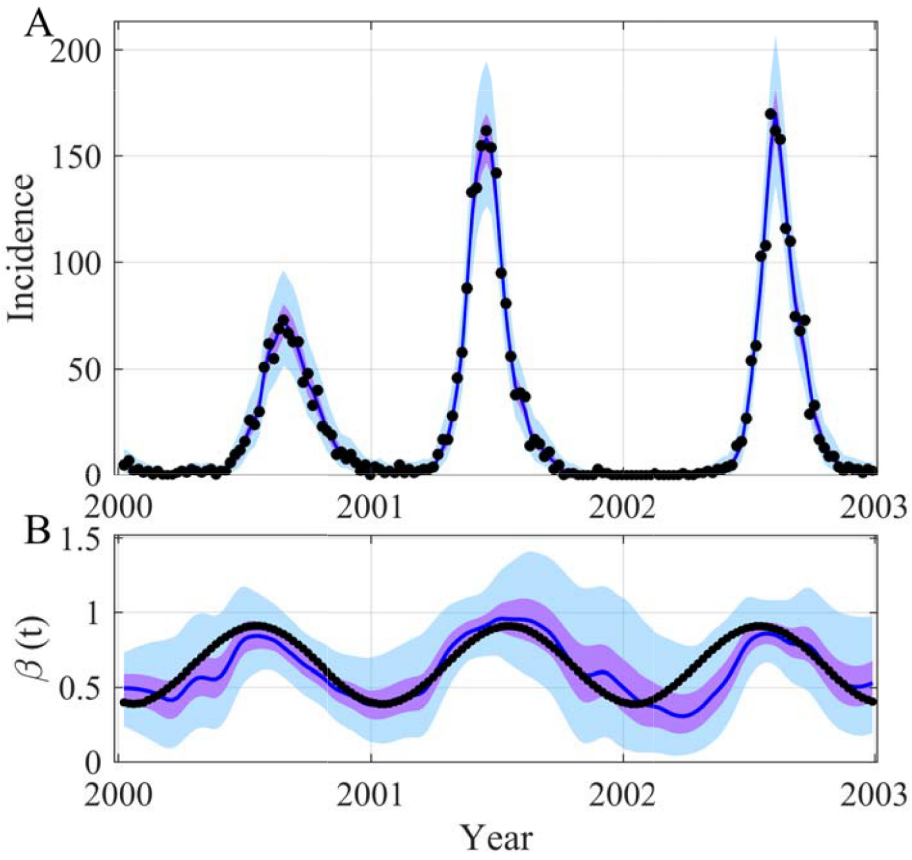


Fig 1. Reconstruction of both the incidence (A) and the time evolution $\beta(t)$ (B) for the SIRS model. In (A) the black points are observations generated with a Poisson process with a mean equal to the incidence simulated by the model. In (B) the black points are the true values of $\beta(t) = \beta_0.(1 + \beta_1 \sin(2\pi\, t/365+2\pi\phi))$. The blue lines are the median of the posterior, the mauve areas are the 50% Credible Intervals (CI) and the light blue areas the 95% CI. For all the figures, the observation process is also applied to the inferred incidence trajectory. The time unit of the model is *day*, the initial date is arbitrary (2000-01-09) and parameters used for the SIRS model are as follows: $\mu = 1/(50*365)$, $\alpha = 1/(7*365)$, $\gamma = 1/14$, $\beta_0 = 0.65$, $\beta_1 = 0.4$, $\phi = -0.2$, $\rho = 1$, $N = 10000$, $S(0) = 600$, $I(0) = 30$. The prior and posterior distributions of the inferred parameters are in S1 Fig.

https://doi.org/10.1371/journal.pcbi.1006211.g001

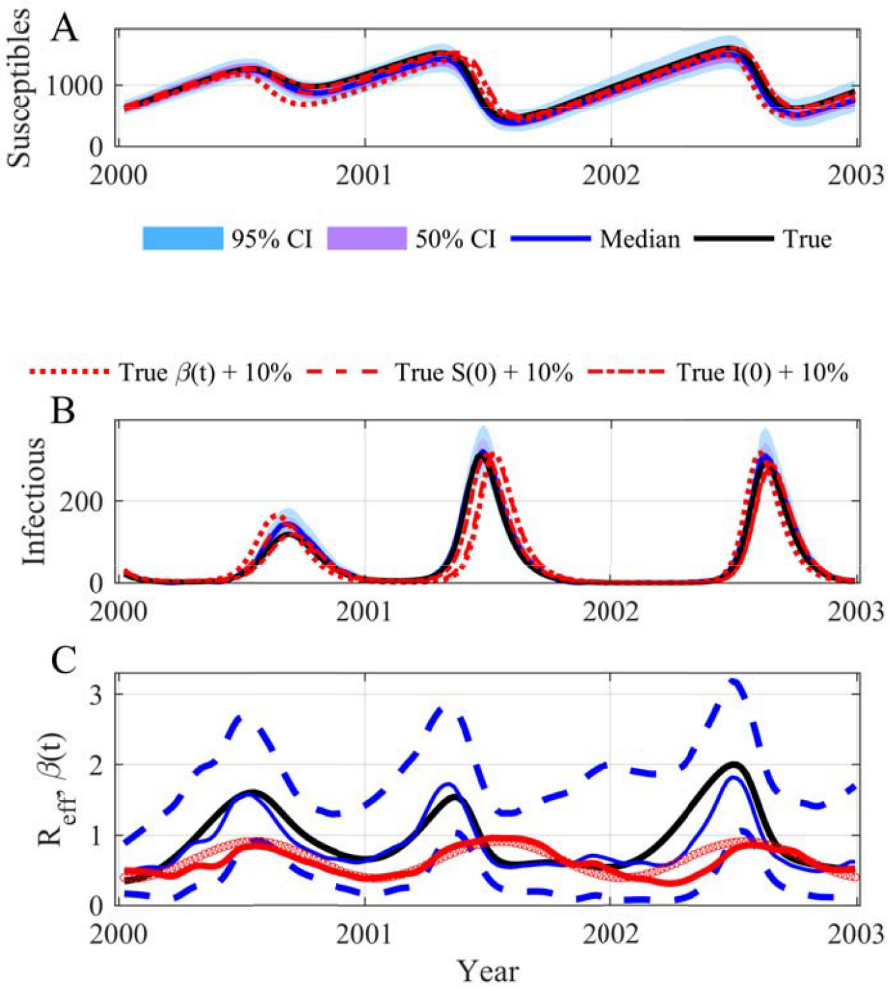


Fig 2. Simulation of the SIRS model: (A) Susceptibles; (B) Infectious; (C) Time evolution of both R_{eff} and $\beta(t)$. In (A) and (B) the black lines are the true values, the blue lines are the median of the posterior, the mauve areas are the 50% CI and the light blue areas the 95% CI. In (A) and (B) the susceptibles and infectious trajectories with a modification of 10% of the value of $S(0)$, $I(0)$ and $\beta(t)$ have been added to show the weak sensitivity of the SIRS model to these values. In (C) the black line is the true values of R_{eff}, the blue line is the median of the posterior, and the dashed lines the 95% CI of R_{eff}, the red dot line is the true time evolution of $\beta(t)$ and the red line the median of its posterior. Model parameters as in Fig 1: The time unit of the model is *day*, the initial date is arbitrary (2000-01-09) and parameters used for the SIRS model are the following: $\mu = 1/(50*365), \alpha = 1/(7*365), \gamma = 1/14, \beta_0 = 0.65, \beta_1 = 0.4, \phi = -0.2, \rho = 1, N = 10000, S(0) = 600, I(0) = 30$.

https://doi.org/10.1371/journal.pcbi.1006211.g002

incidence and the transmission rate highlighting the potential of the method. The parameter estimations are in perfect agreement with the values used to generate the observations and the estimation process has correctly converged (S1 and S2 Figs). This clearly demonstrates the feasibility of accurately ascertaining the time evolution of the transmission rate and correctly estimating the R_{eff} (see Fig 2). It is worth emphasizing that the SIRS model is a complicated example for different reasons. First, even with a constant transmission rate the SIRS model can generate oscillations (damped oscillations, see [17,44]). Secondly, the model trajectories are not very sensitive, a modification of ± 10% can induce minor modifications of the trajectories

that are inside or near the 95% CI of our inferences (Fig 2). Moreover in this example we have used initial conditions outside the attractor of the dynamics to generate transients that appear more realistic for real applications, but are more complex to reconstruct. The robustness of our approach has also been tested: (i) using long time series and initial conditions near the attractor (Fig 3A and S3A Fig); (ii) modifying the number of inferring parameters (S4–S6 Figs), for instance estimating just the volatility parameter (S7 and S8 Figs); (iii) considering the possibility of not using the transformation log in the diffusion process (S9 and S10 Figs) and (iv) using a true $\beta(t)$ with 2 or 3 periodic components (Fig 3B and 3C and S3B and S3C Fig).

We have also explored the performance of our approach by comparing their inferences to those of the true model. The re-estimation of the true model on its own data is displayed in S11–S13 Figs. Table 1 presents indices of the goodness-of-fit of the true model and models with time-varying β(t) with different number of parameters inferred. As expected, the error on $\beta(t)$ is smaller when the true equation is used (Table 1). However, regarding the estimated incidence, the true model and our approach give similar results both in terms of mean and variance (Table 1). It could be argued that the price of the flexibility of our approach is a greater variability in some of the trajectory estimations (Table 1). Nevertheless the average dynamics are always estimated correctly.

As misspecification is an important problem (*e.g.* [45]) we have also compared the performance of our approach to those of a misspecified seasonal SIRS model. We have thus used the example of a sinusoidal $\beta(t)$ with two periodic components (see Fig 3B) and computed the indices of the goodness-of-fit of the true model with the SIRS model with 1 year sinusoidal $\beta(t)$ and with our time-varying periodic $\beta(t)$. The results clearly show that our approach performed better than the misspecified model for the three trajectories analyzed, Incidence, β *and* R_{eff} (Table 2). Once again the price of the flexibility of our approach is a greater variability in some of the trajectory estimations. However this is preferable to a large error in the median trajectories as occurred in those observed with the misspecified model (Table 2).

Our methodology is also applicable to other more complex or simpler tasks. For instance, it can follow the time evolution of a parameter describing the availability of susceptibles, $\varepsilon_S(t)$ (Fig 4 and S14 and S15 Figs). Fig 4 shows the accurate reconstruction of the trajectory of the incidence and also of the trajectory of $\varepsilon_S(t)$ that shifted at a given time point and decreased slightly thereafter. This highlights once again the potential of our approach as it is never easy to estimate a discontinuous dynamic with a continuous process (2).

Application to real datasets: Flu

In previous works, the dynamics of influenza in Israel have been analyzed using a discrete deterministic SIRS model and weekly data from Israel's Maccabi health maintenance organization [20,46]. To describe the seasonality of this recurrent epidemic, the authors used a linear model between the transmission rate and local climatic variables, daily temperature and relative humidity [20,46]. We have re-analyzed their dataset (but limited to 1998–2003 due to a modification in the reporting) to reconstruct the time evolution of $\beta(t)$. Our results (Fig 5 and S16 Fig) clearly show the potential of our method, highlighting that the $\beta(t)$ fluctuations are more irregular and complex than a simple sinusoidal function.

Application to real datasets: Dengue

Our last example is on dengue in Cambodia. Again the idea is to relax the assumption of a sinusoidal $\beta(t)$ in a SEIR model. Monthly data from the capital Phnom Penh [47], for which the meteorological data is available from the international airport, was used. We can accurately describe the 12 year time series and reconstruct the time evolution of $\beta(t)$ (Fig 6 and S17 Fig).

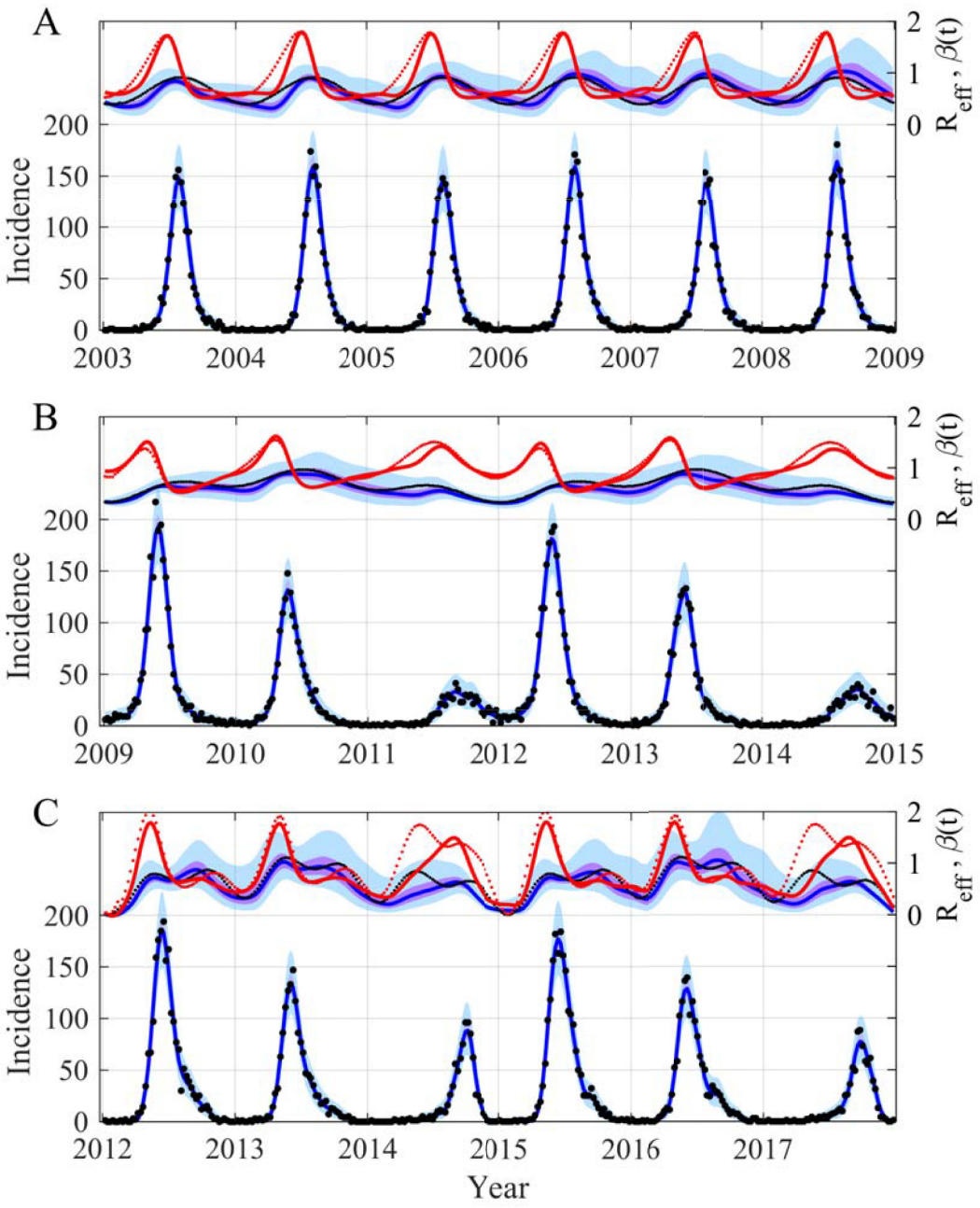


Fig 3. Reconstruction of both the incidence (bottom panel) and the time evolution of $\beta(t)$ and R_{eff} (top panel) for a SIRS model with the initial conditions near the attractor of the dynamics. The true β is generated by $\beta(t) = \beta_0.(1 + \beta_1 \sin(2\pi t/365 + 2\pi\phi) + \beta_2 \sin(2\pi t/(3\ 365) + 2\pi\phi)) + \beta_3 \sin(2\pi t/(0.5\ 365) + 2\pi\phi))$ with in (A) $\beta_1 = 0.4, \beta_2 = 0, \beta_3 = 0, S(0) = 911.5, I(0) = 3.5$; in (B) $\beta_1 = 0.4, \beta_2 = 0.3, \beta_3 = 0, S(0) = 1735, I(0) = 20$; and in (C) $\beta_1 = 0.1, \beta_2 = 0.1, \beta_3 = 0.1, S(0) = 3365, I(0) = 3$. The other parameters used are as follows: $\mu = 1/(50*365), \alpha = 1/(7*365), \gamma = 1/14, \beta_0 = 0.65, \phi = -0.2, \rho = 1, N = 10000$. In both panels blue lines are the median of the posterior, the mauve areas are the 50% CI and the light blue areas the 95% CI. In the top panel, the red line is the reconstructed R_{eff}, the points are the true values of R_{eff} (red) and of $\beta(t)$ (black). In the bottom panel the black points are observations generated with a Poisson process. The prior and posterior distributions of the inferred parameters are in S3 Fig.

https://doi.org/10.1371/journal.pcbi.1006211.g003

Table 1. Comparison of goodness-of fit indices for different models and different numbers of parameters inferred. The indices are computed on Incidence, β and R_{eff} trajectories: RMSE: root mean square error using the median; MAPE: maximum absolute percentage error using the median; MIQR: mean inter-quartile range. The parameter values used are in the captions of the figures. For comparison purposes we used a stochastic version of the SIRS model with sinusoidal β and for all figures the observation process is applied to the inferred incidence trajectory.

Model		Sinusoidal β	Time-varying $\beta(t)$	Time-varying $\beta(t)$	Time-varying $\beta(t)$
Figure		S11 Fig	Fig 1	S4 Fig	S7 Fig
# of parameters inferred		8	7	5	1
Incidence	RMSE	4.18	4.14	4.13	4.12
	MAPE	2.24	2.34	2.39	2.24
	MIQR	1.03	0.95	1.04	1.03
$\beta(t)$	RMSE	0.07	0.11	0.14	0.13
	MAPE	0.29	0.48	0.52	0.50
	MIQR	0.07	0.43	0.36	0.33
R_{eff}	RMSE	0.06	0.19	0.24	0.24
	MAPE	0.18	0.47	0.50	0.50
	MIQR	0.23	0.64	0.53	0.41

https://doi.org/10.1371/journal.pcbi.1006211.t001

Our results stress that the $\beta(t)$ oscillations are more complex than a simple sinusoidal function. Sometimes bi-modality occurs over one season. In general one observes a fast growth of $\beta(t)$ and a slow decrease. Moreover the amplitude of the $\beta(t)$ varies from year to year, perhaps depending on the fluctuations in the mosquito population and in the environment. Interestingly the peak in $\beta(t)$ appears 1 to 2 months before the incidence peak. This delay can be explained by the extrinsic incubation period and might be used in a warning system.

To explain the $\beta(t)$ oscillations we have explored the potential effects of local and global climatic variables using wavelet decomposition [48] as one of our main underlying hypotheses is non-stationarity. We observed very significant coherency between $\beta(t)$ and climate for the local climate for the seasonal mode (Fig 7 and S18–S20 Figs) and also for the 2–3 year components with global climatic variable (S21 Fig). Thus, the rhythm of $\beta(t)$ can be explained perfectly by climatic factors. Nevertheless, again mainly due to large non-stationarity, by using solely one or two climatic variables we are able to correctly describe dengue evolution in the short-term (Fig 7C, red area) but not over a large time period (Fig 7C, blue area). This reflects

Table 2. Comparison of goodness-of fit indices between a misspecified model and a model with a time-varying parameter. The reference is a sinusoidal model with 2 periodic components. The parameter values used are in the captions of Fig 3. The indices are computed on Incidence, β and R_{eff} trajectories: RMSE: root mean square error using the median; MAPE: maximum absolute percentage error using the median; MIQR: mean inter-quartile range. For comparison purposes we used a stochastic version of the SIRS models with sinusoidal β and for all figures the observation process is applied to the inferred incidence trajectory.

Model		Reference: Sinusoidal β with 2 periods (see Fig 3B caption)	Sinusoidal β with 1 year period (see Fig 3A caption)	Time-varying $\beta(t)$ (5)
# of parameters inferred		8 including 4 for β	7 including 3 for β	5 including 1 (σ) for β
Incidence	RMSE	4.80	5.21	4.83
	MAPE	2.21	1.98	2.40
	MIQR	0.87	1.03	0.98
$\beta(t)$	RMSE	0.05	0.26	0.08
	MAPE	0.22	1.43	0.27
	MIQR	0.06	0.06	0.28
R_{eff}	RMSE	0.06	0.22	0.08
	MAPE	0.12	0.44	0.20
	MIQR	0.20	0.20	0.46

https://doi.org/10.1371/journal.pcbi.1006211.t002

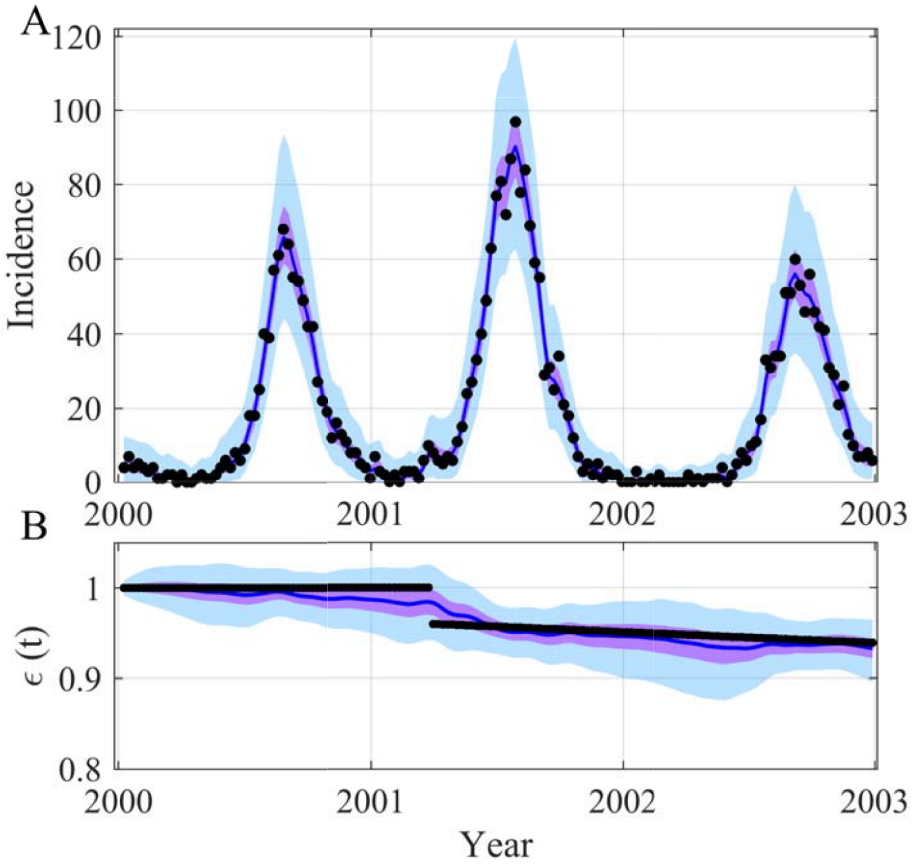


Fig 4. Reconstruction of both the incidence (A) and the time evolution of $\varepsilon_S(t)$ (B) for the SIRS model. In (A) the black points are observations generated with a Poisson process with a mean equal to the incidence simulated by the model. In (B) the black points are the true value of $\varepsilon_S(t)$: $\varepsilon_S(t) = 1$ and shift to $\varepsilon_S(t) = 0.96 - (0.012/365)*(t—t_{shift})$ at $t_{shift} = 450$ days after t_0 in days. The blue lines are the median of the posterior, the mauve areas are the 50% CI and the light blue areas the 95% CI. The time unit of the model is *day*, the initial date is arbitrary (2000-01-09), $\beta(t) = \beta_0.(1 + \beta_1 \sin(2\pi t/365+2\pi\phi))$ and parameters used for the SIRS model are as follows: $\mu = 1/(50*365)$, $\alpha = 1/(7*365)$, $\gamma = 1/14$, $\beta_0 = 0.65$, $\beta_1 = 0.4$, $\phi = -0.2$, $\rho = 1$, $N = 10000$, $S(0) = 600$, $I(0) = 30$. The prior and posterior distributions of the inferred parameters are in S15 Fig.

https://doi.org/10.1371/journal.pcbi.1006211.g004

the complexity of such a disease where the ecology of the vectors, the environment, the climate, the immune status of the human population and its behavior are all involved. This large non-stationarity association between dengue and climatic factors has recently been demonstrated using statistical models (dynamic generalized linear models) and data from a medium-sized city in Colombia [49]. The authors showed that dengue cases correlate with climatic variables (temperature, rainfall, solar radiation and relative humidity) but these correlations change over time, some intervals showing a positive association, while in others the association is negative [49]. The non-stationarity association between dengue and climate may be explained by the fact that a climatic variable has different effects depending on the biological cycle of the pathogen or of the vector. Moreover the effects of one climatic variable can also depend on other climatic variables potentially enhancing the non-stationarity association.

Discussion

As there remain numerous uncertainties during the course of each epidemic, we are increasingly aware of the importance of developing adequate statistical and mathematical tools. Such

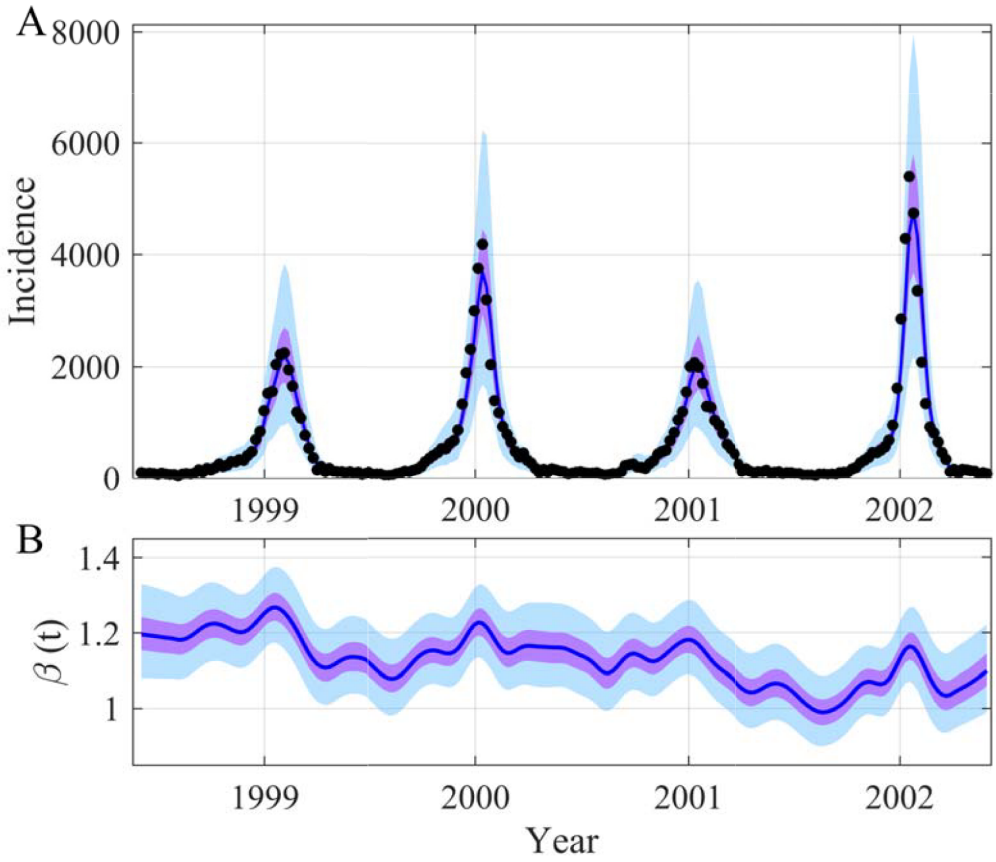


Fig 5. Reconstruction of both the incidence (A) and the time evolution $\beta(t)$ (B) in the case of the 1998–2002 seasonal flu epidemics in Israel. In (A) the black points are influenza-like illness incidence collected by Israel's Maccabi health maintenance organization [46]. The blue lines are the median of the posterior, the mauve areas are the 50% CI and the light blue areas the 95% CI. The prior and posterior distributions of the inferred parameters are in S16 Fig.

https://doi.org/10.1371/journal.pcbi.1006211.g005

tools need to take account of the time-varying nature of the underlying ecological and biological mechanisms as well as social and behavioral influences involved in an epidemic. Because of this, time-varying parameters modeled with a diffusion process, that track epidemiological patterns and update the key parameters according to data appear to be a worthwhile approach. Indeed developing a more complex model would be difficult considering the relative paucity of available data.

We propose a flexible modeling framework that encompasses time-varying aspects of the epidemic. It does this via diffusion process equations for time-varying parameters and also considers uncertainty associated with key parameters and data. This data-driven framework for time-varying parameters has been coupled with simple stochastic models and a robust Bayesian procedure for inference. To test its efficiency, our proposed methodology was first applied to a toy model and then to real epidemiological examples.

Our results clearly demonstrate the potential of our framework. Firstly, our methodology was able to accurately reconstruct both the incidence and the sinusoidal transmission rate of a simple SIRS model just based on noisy observations (Figs 1–4 and S4,S5,S7,S9 and S14 Figs). Based on these reconstructions one can also closely estimate R_{eff} which is one of the key relevant epidemiological parameters. Our results also highlight the flexibility of our developed methodology. It can reconstruct the time evolution of a shifting parameter ($\varepsilon_S(t)$, see Fig 4 and

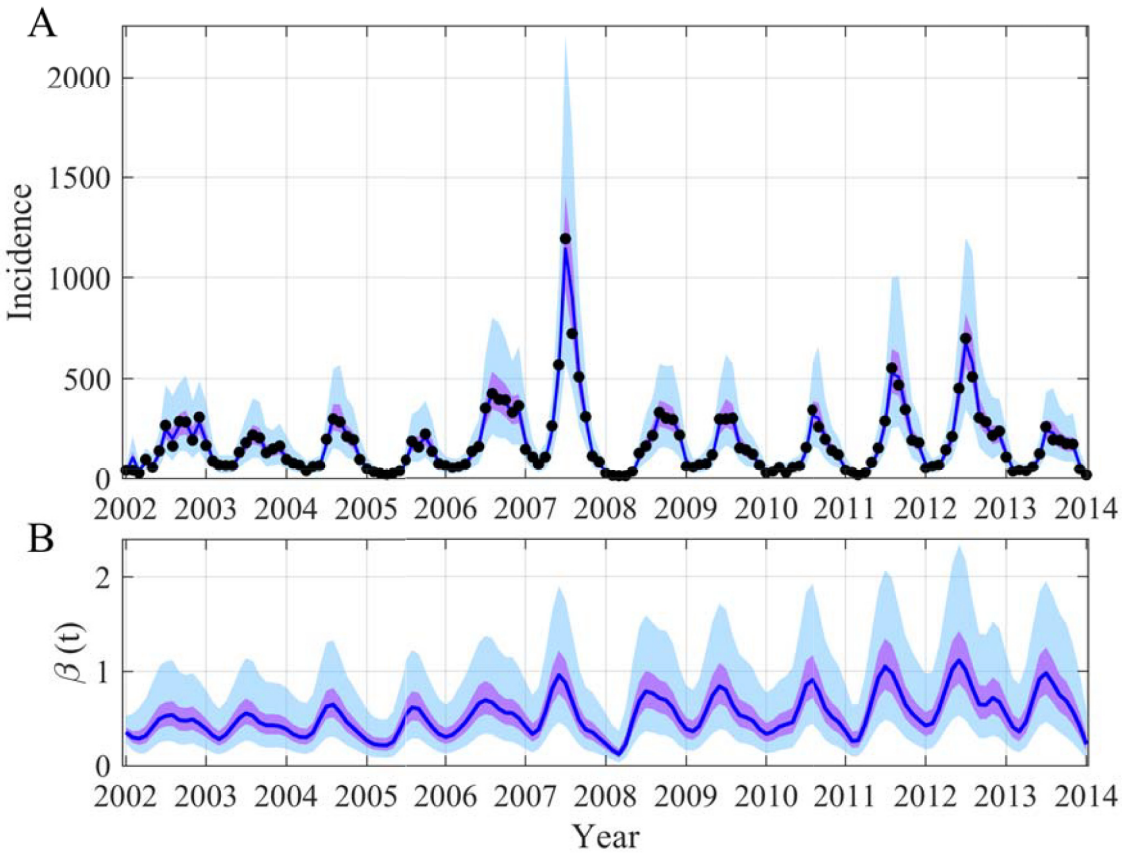


Fig 6. Reconstruction of both the incidence (A) and the time evolution of $\beta(t)$ (B) in the case of the 2002–2013 dengue epidemics in the province of Phnom Penh (Cambodia). In (A) the black points are dengue incidence recorded by the Cambodian National surveillance (National Dengue Control Program from the Ministry of Health, see [47]). The blue lines are the median of the posterior, the mauve areas are the 50% CI and the light blue areas the 95% CI. The prior and posterior distributions of the inferred parameters are in S17 Fig.

https://doi.org/10.1371/journal.pcbi.1006211.g006

S14 Fig) as well as an oscillating parameter that influences the nonlinear part of the model ($\beta(t)$, see Figs 1–3 and S4,S5,S7 and S9 Figs). The comparison using goodness-of-fit indices with the inferred true model allows us to highlight the fact that our methodology performs as well for the observed incidence. Its flexibility results in greater variability in some other trajectories mainly $\beta(t)$ and R_{eff} (Table 1). Moreover, in the absence of knowledge of the true evolution of the transmission rate, our approach appears to capture the dynamic observed more accurately than a misspecified model (Table 2). Secondly, applied to real datasets, our framework is able, based solely on simple stochastic models, to reconstruct complex epidemics such as flu or dengue over long time periods (Figs 5 and 6). In such cases, the reconstruction of the time evolution of the transmission rate clearly stresses that, on real datasets, it is difficult to assimilate the dynamic of this parameter as a simple sinusoidal function. It is more irregular in amplitude and sometimes multi-modal over one season.

Considering the paucity of information available regarding the complexity of the mechanisms involved during an epidemic, describing and fitting a full model for a given transmissible disease is always challenging. Our data-driven methodology can be used as a first step towards a better understanding of a complex epidemic, where data is limited or lacks certainty. Indeed most of the unknowns and uncertainties can be put into time-varying parameters. The

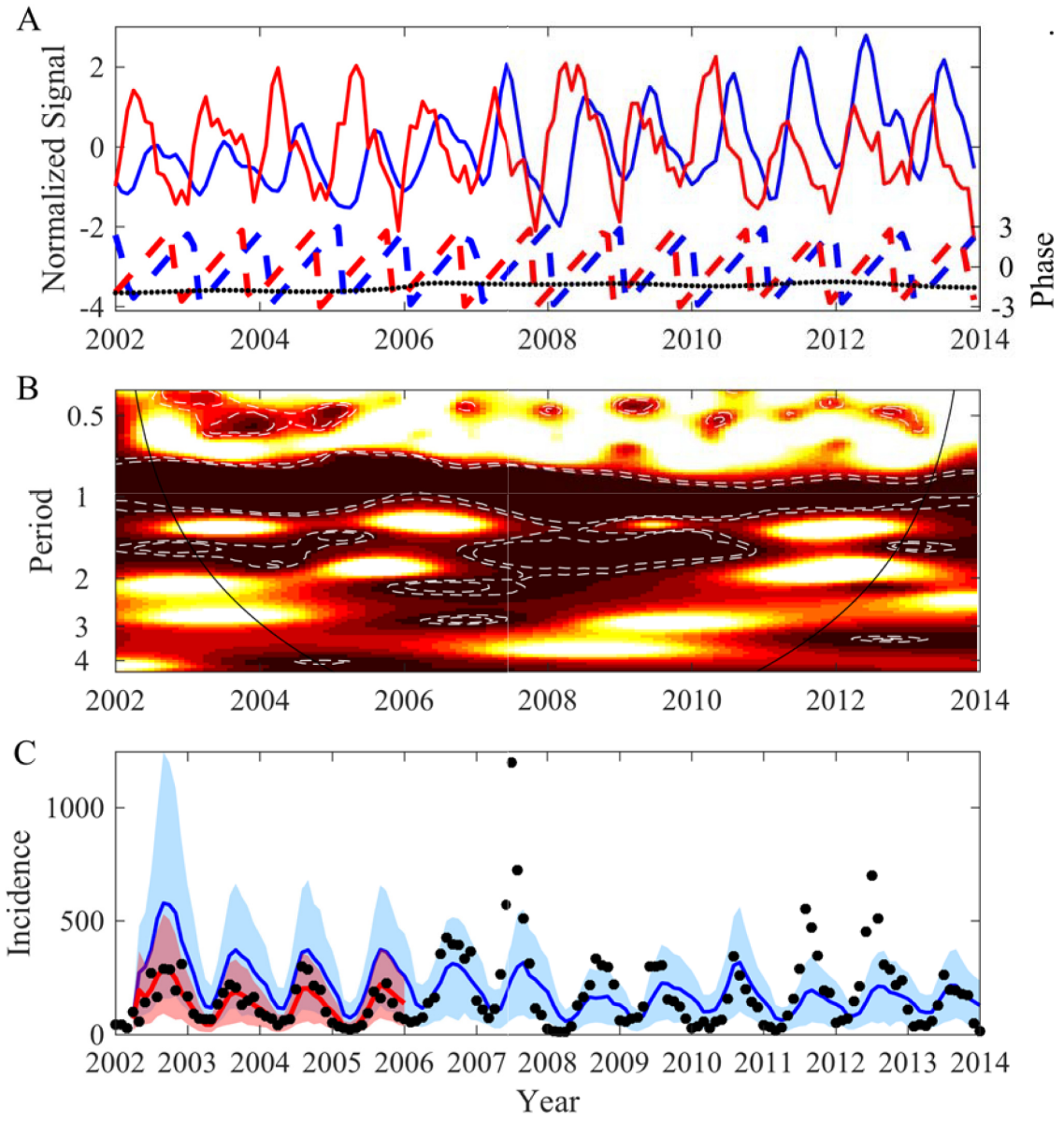


Fig 7. Association between dengue transmission rate and monthly average maximum temperature in the province of Phnom Penh (Cambodia). (A) Time evolution of the normalized median of $\beta(t)$ (blue line) and average temperature (red line) as well as the evolution of their phase computed based on wavelet decomposition (see Method and [48,70]), blue dashed line for the normalized $\beta(t)$, red dashed line for the normalized averaged temperature and black dotted line for their phase difference. (B) Wavelet coherence (see Method and [48,70]) between the reconstructed $\beta(t)$ and average temperature. The colors code for low values in white to high values in dark red. The white dashed lines show the 90% and the 95% CI computed with adapted bootstrappes [71]. (C) Model simulations using a linear model describing $\beta(t)$ with monthly average maximum temperature (S18 Fig) and monthly average minimum temperature (S19 Fig) ($\beta(t) = a_0 + a_1.MaxTemp(t)+a_2.MinTemp(t)$). The red line is the median of the posterior and the red area is the 95% CI when parameters are estimated for the period 2002–2005. The blue line is the median of the posterior and the light blue area is the 95% CI when parameters are estimated for the full time period. The black points are dengue incidence recorded by the Cambodian National surveillance [47].

https://doi.org/10.1371/journal.pcbi.1006211.g007

potential effects of all these uncertainties can then be explored by analyzing the reconstructed time evolution of the time-varying parameters. See Fig 7 for such preliminary analysis of dengue in Phnom Penh. This allows a more thorough analysis of the influences and the

interactions between both the human behavior and complex environmental drivers. In a recent paper [50], the authors reviewed evidence of interactions between seasonal influenza virus and other pathogens (bacteria or virus). They concluded that it is important to incorporate these different coinfecting pathogens in models of seasonal flu in order to get a better estimate of the burden of influenza. Our framework could be an alternative to the development of complex models with all the potential interactions between pathogens and to estimate the strength of the interactions. After reconstructing the time evolution of the transmission rate the statistical association between the coinfecting pathogens and the transmission rate could be tested. This screening may facilitate the construction of more complex models that could incorporate only the most significant coinfecting pathogens in the seasonal flu model.

Our methodology also has other advantages. Taking account of the simplicity of the model used, and the fact that weak hypotheses on the dynamics of the time-varying parameters have been included, our proposed methodology can retrospectively test the impact of interventions. This has previously been done in the case of HIV epidemics [34–35], where it was hypothesized that the reduction in the transmissibility was due to a modification of the sexual behavior in the population and the increase in the seropositive period duration due to the introduction of the first antiviral treatments. Evaluation of interventions has also been done recently in the case of the Ebola epidemic in West Africa [51]. The relative simplicity of our methodology is also suitable for short-term predictions and it can then easily be used to predict an epidemic in real time. Starting with a given estimated state defining the system, the fitting process can be run again each time new data is available and the new posteriors are used for new predictions [36]. This can inform public health decisions and indeed has been done recently to great effect in the case of the Ebola epidemics in West Africa [52].

A major challenge in model fitting is the reliability of data collected and also the non-identifiability of the mechanistic models that always have very rich dynamical behavior. The question of identifiability is too often avoided in epidemiological models applied to a topical Public Health issues. There is, however, considerable literature on this subject (*e.g.* [41,42,53–55]). Identifiability is not evident even for a simple seasonal SIR model [56]. To solve this problem one can fit a combination of parameters or fix some of them (the population size for instance) [57]. In our applications there is a clear limitation due to practical non-identifiability of reporting rate and initial conditions. To fix these problems we have used informative priors (see Method). Using informative priors or fixing some parameters gives very similar results (compare Fig 1 and S4–S7 Figs). Related to this is the misspecification of models [45]. In our cases, as with other semi-mechanistic models the time-varying parameter methodology captures some of the information in the data but not in the mechanistic part of the model. If the model is misspecified due to lack of precision, it compensates for it and the dynamics of $\beta(t)$ will drive improvements in the model to make it more complex and realistic (Table 2). If the model is misspecified to the extent that it creates mechanisms that do not exist, the reconstructed $\beta(t)$ would compensate for these effects but it will be harder to interpret.

In this work we have used simple mechanistic models. The proposed methodology is not limited to simple models. For instance, a two-strain dengue model has also been tested. In this case the main problem was linked to the unavailability of both seroprevalence and incidence for each strain. Indeed, one of the major difficulties with these multi-strain models is the identification of the initial conditions (*e.g.* [58]). Nevertheless it is worth emphasizing that the Bayesian inference method used in our framework, PMCMC, the approximation of the likelihood is limited for a large number of parameters and/or equations [59]. In such cases testing other methodologies like ABC [60,61] is advisable.

It is always difficult to fit complex models with rich behaviors based on very limited information. In this regard we agree with Metcalf et al. [62] who stressed that nowadays we need

seroprevalence studies to quantify the immunological status of the population, because in most cases the magnitude of the outbreak is difficult to evaluate without precise seroprevalence data.

To tackle the uncertainty and the non-stationarity of epidemics, our methodology, although it appears non-standard, makes important progress towards a better understanding of the mechanisms responsible for disease propagation. We believe that, should it form part of the development of the next predictive tools for Public Health, it will make a significant contribution to improving the understanding and control of infectious diseases in our increasingly uncertain world.

Methods

Models

SIRS model. Our first model is a classical SIRS model with an observation rate $\rho = 1$ and Poisson law as the observational process:

$$\begin{aligned} \dot{S} &= \mu.(N - S) - \beta(t).S.I/N + \alpha.R \\ \dot{I} &= \beta(t).S.I/N - (\gamma + \mu).I \\ \dot{C} &= \beta(t).S.I/N \\ \dot{R} &= \gamma.I - (\alpha + \mu).R \end{aligned} \tag{5}$$

where S, I and R are the susceptibles, the infectious and the removed respectively, the transmission rate $\beta(t) = \beta_0.(1 + \beta_1 \sin(2\pi t/365 + 2\pi\phi))$, $1/\alpha$ is the average duration of immunity, γ is the recovery rate and μ is the recruitment or mortality rate. In (5), C is the number of new cases, then $h(x(t))$ is the cumulative sum of $C(t)$ over the observation time step, 7 days. With this model $R_{eff}(t) = \beta(t).S(t)/(N.\gamma)$. The parameter values are in the caption of Fig 1. For the fit of our simulated data, Gaussian priors are used for epidemiological parameters (α and γ). Initially non-informative priors were used for the volatility σ, the reporting rate ρ and the initial conditions $\gamma(0)$ by $\beta(0)$ but to reduce problems linked to practical non-identifiability materialized by correlation between some estimates, informative priors were used for ρ (see S1 Fig). Some other simulations have been done fixing $\beta(0)$ and ρ or just inferring σ (see S4–S8 Figs).

SIRS flu model. For analyzing Israel flu data we have used a continuous SIRS model identical to (5), we simply added imported infectious people i in the force of infection:

$$\lambda(t) = \beta(t).\frac{S(t).(I(t) + i)}{N} \tag{6}$$

The initial guess values for the parameters are from [46]. In this example the observational process is a Negative-Binomial law with an over-dispersion parameter equal to 0.05 and the reporting rate $\rho = 0.15$ [46]. For the fit, Gaussian priors are used for epidemiological parameters (i, α, γ) and non-informative priors for the volatility σ and the initial conditions ($S(0)$, $I(0)$). S16 Fig displays the priors and the posteriors.

SEIR one strain dengue model. To describe the dengue epidemics, taking account of the available data for Phnom Penh, we have fitted a one strain model using a SEIR model:

$$\begin{aligned} \dot{S} &= \mu.(N - S) - \beta(t).S.(I + i)/N \\ \dot{E} &= \beta(t).S.(I + i)/N - (\delta + \mu).E \\ \dot{I} &= \delta.E - (\gamma + \mu).I \\ \dot{R} &= \gamma.I - \mu.R \end{aligned} \tag{7}$$

where E is for infected but not yet infectious, $\beta(t)$ is the transmission rate, $1/\delta$ is the average duration of the latent period, γ is the recovery rate and μ is the recruitment or mortality rate. The initial guess for parameter values comes from the literature [63]. In this example the observational process is a Negative-Binomial law with an over-dispersion parameter equal to 0.05 and the reporting rate ρ has been estimated using a narrow Gaussian prior. Non-informative priors are used for the volatility σ, initial condition for infected $E(0)$ and imported infectious people i. Gaussian priors are used for other parameters and initial conditions. When $E(0)$ is fitted, $I(0)$ is estimated as a steady-state value $I(0) = \delta.E(0)/(\gamma+\mu)$. S17 Fig. displays the priors and the posteriors.

Inference

Stochastic framework. Due to the use of a diffusion Eq (2) for the dynamic of the time-varying parameters, the stochastic versions of the previous models have been fitted. Thus the models are considered in a stochastic framework in which the compartments are discrete and the number of reactions occurring in a time interval dt is approximated by a multinomial distribution. It is fully described in [64,65].

SMC algorithm as implemented in SSM [65]. In a model with n observations and J particles. L is the model likelihood $p(y_{1:n}|\theta)$. $W_k^{(j)}$ is the weight and $x_k^{(j)}$ is the state associated to particle j at iteration k.

1. Set $L = 1$, $W_0^{(j)} = 1/J$.
2. Sample $(x_0^{(j)})_{j=1:J}$ from $p(x_0 | \theta)$
3. **for** (k = 0:n−1) **do**
4. **for** (j = 1:J) **do**
5. Sample $x_{k+1}^{(j)}$ from $p(x_{k+1}|x_k^{(j)}, \theta)$
6. Set $\alpha^{(j)} = p(y_{k+1}|x_{k+1}^{(j)}, \theta)$
7. **end for**
8. Set $W_{k+1}^{(j)} = \alpha^{(j)} / \sum_{l=1}^{J} \alpha^{(l)}$ and $L = L \frac{1}{J} \sum_{l=1}^{J} \alpha^{(l)}$
9. Resample $(x_{0:k+1}^{(j)})_{j=1:J}$ from $W_{k+1}^{(j)}$.
10. **end for**

Estimation with particle Markov Chain Monte Carlo (PMCMC). Since the epidemiological propagation models are considered in a stochastic framework, their likelihood is intractable and it is estimated with particle filtering methods (Sequential Monte Carlo, SMC). With a given set of parameters, the SMC algorithm reconstructs sequentially the trajectory of the state variables and the time-varying parameters, and computes the associated likelihood. Firstly, the distribution of the initial conditions of the system is approximated with a sample of particles. Then, at each iteration, the particles are projected according to the propagation model up to the next observation point, they receive a weight reflecting the quality of their prediction compared to the observation and the total likelihood is updated. A resampling step using the weights is performed before the next iteration, in order to discard the trajectories associated with low weight particles.

In order to estimate the parameters of the system, the particle filter is embedded in a Markov Chain Monte Carlo framework, leading to the PMCMC algorithm [43]. More precisely, the likelihood estimated by SMC is used in a Metropolis Hasting scheme (particle marginal Metropolis Hastings) [43]. The proposal distribution is a Gaussian whose co-variance matrix is adapted following the framework described in [65].

The starting point of the MCMC chain is initialized using optimal values obtained from the KSimplex algorithm on a large number of parameter sets. Then, a pre-adaptation of the proposal co-variance matrix is performed with Kalman MCMC (KMCMC). Each time the idea

relies on less computationally intense algorithms in order to facilitate the exploration of parameter space. But as we use stochastic models we approximate the likelihood using the extended Kalman filter both in the simplex algorithm (KSimplex) [65] and in the MCMC (KMCMC) [36]. Then the adaptive PMCMC is executed on the output of the KMCMC with 100000 iterations and 10000 particles for the final figures. For instance, the results such as those of Fig 1 take less than 24 hours, on a blade server from PowerEdge M-Series with 40 processor cores.

PMCMC algorithm as implemented in SSM [65]. In a model with n observations and J particles.

$q(.|\theta^{(i)})$ is the transition kernel

1. Initialize $\theta^{(0)}$
2. Using SMC algorithm, compute $\hat{p}(y_{1:n}|\theta^{(0)})$ and sample $x_{0:n}^{(0)}$ from $\hat{p}(x_{0:n}|y_{1:n},\theta^{(0)})$
3. **for** (i = 0:N−1) **do**
4. Sample θ^{*} from $q(.|\theta^{(i)})$
5. Using SMC algorithm, compute $L(\theta^{*}) = \hat{p}(y_{1:n}|\theta^{*})$ and sample $x_{0:n}^{*}$ from $\hat{p}(x_{0:n}|y_{1:n},\theta^{*})$
6. Accept θ^{*} (and $x_{0:n}^{*}$) with probability $1 \wedge \frac{L(\theta^{*})p(\theta^{*})q(\theta^{(i)}|\theta^{*})}{L(\theta^{(i)})p(\theta^{(i)})q(\theta^{*}|\theta^{(i)})}$
7. If accepted, $\theta^{(i+1)} = \theta^{*}$ and $x_{0:n}^{(i+1)} = x_{0:n}^{*}$. Otherwise, $\theta^{(i+1)} = \theta^{(i)}$ and $x_{0:n}^{(i+1)} = x_{0:n}^{(i)}$.
8. **end for**

In order to assess convergence of the chain, the visual inspection of the chains (*e.g.* S2 Fig or S8 Fig) was complemented by diagnosis provided in the Coda package in R [66]. Due to the large computational cost of the algorithm, we did not run multiple independent chains, rather we relied on diagnosis using one MCMC chain and testing its stationarity: Geweke diagnosis [67] and Heidelberger and Welch's diagnosis [68]. The results are presented in S1 and S2 Tables.

Wavelet analysis

Among the various approaches developed to study nonstationary data, wavelet analysis is probably the most efficient. In particular, this method gives us the possibility of investigating and quantifying the evolution in time of the periodic components of a time series (see [69]). Wavelets constitute a family of functions derived from a single function, the "mother wavelet", $\Psi_{a,\tau}(t)$, that can be expressed as the function of two parameters, one for the time position τ, and the other for the scale of the wavelets a, related to the frequency. More explicitly, wavelets are defined as:

$$\Psi_{a,\tau}(t) = \frac{1}{\sqrt{a}}\psi\left(\frac{t-\tau}{a}\right)$$

The wavelet transform of a time series $x(t)$ with respect to a chosen mother wavelet is performed as follows:

$$W_x(a,\tau) = \frac{1}{\sqrt{a}}.\int_{-\infty}^{+\infty} x(t).\Psi^{*}\left(\frac{t-\tau}{a}\right).dt = \int_{-\infty}^{+\infty} x(t).\Psi_{a,\tau}^{*}.dt$$

where * denotes the complex conjugate form. The wavelet transform $W_x(a,\tau)$ represents the contribution of the scale a to the signal at different time positions τ. The computation of the wavelet transform is done along the signal $x(t)$ simply by increasing the parameter τ over a range of scales a until all coherent structures within the signal can be identified. Here, as mother wavelet, we have used the Morlet wavelet [69].

With the wavelet approach, we can estimate the repartition of variance at different scale a and different time location τ. This is known as the wavelet power spectrum: $S_x(a,\tau) = |W_x(a,\tau)|^2$.

An important point with the continuous wavelet is that the relationship between the wavelet frequency f_0 and the wavelet scale a can be derived analytically. For the Morlet wavelet this relationship is given by:

$$\frac{1}{f} = \frac{4\pi a}{f_0 + \sqrt{2 + f_0^2}}$$

Then when $f_0 = 2\pi$, the wavelet scale a is inversely related to the frequency, $f \approx 1/a$. This greatly simplifies the interpretation of the wavelet analysis and one can replace, on all equations, the scale a by the frequency f or the period $1/f$.

To determine the statistical relationship between two time series, wavelet coherence can be computed (*e.g.* [48,70]):

$$R_{x,y}(f,\tau) = \left(\frac{|\langle W_{x,y}(f,\tau)\rangle|^2}{|\langle W_x(f,\tau)\rangle|^2 . |\langle W_y(f,\tau)\rangle|^2} \right)^{1/2}$$

where the angle brackets around terms indicate smoothing in both time and frequency, $W_x(f,\tau)$ is the wavelet transform of series $x(t)$, $W_y(f,\tau)$ is the wavelet transform of series $y(t)$, and $W_{x,y}(f,\tau)$ is the cross-wavelet spectrum. The values of wavelet coherence are between $0 < R_{x,y}(f,\tau) < 1$. The wavelet coherency is equal to 1 when there is a perfect linear relation at particular time and scale between the two signals, and equal to 0 if $x(t)$ and $y(t)$ are independent.

To complement this, phase analysis can be used to characterise the association between signals (*e.g.* [48,70]). The phase difference provides information on the sign of the relationship (*i.e.*, in phase or out of phase) and can be computed, for complex mother wavelet, with the wavelet transform $W_x(f,\tau)$ as:

$$\phi_x(f,\tau) = \tan^{-1} \frac{\mathrm{Im}(W_x(f,\tau))}{\mathrm{Re}(W_x(f,\tau))}$$

Similarly with the cross-wavelet transform $W_{x,y}(f,\tau)$ the phase difference between the two time series can be computed:

$$\phi_{x,y}(f,\tau) = \tan^{-1} \frac{\mathrm{Im}(W_{x,y}(f,\tau))}{\mathrm{Re}(W_{x,y}(f,\tau))}$$

Supporting information

S1 Code. Examples of model files and code for using SSM.
(ZIP)

S1 Table. Test of the MCMC chains: Geweke diagnosis [67] that tests for the non-stationarity of the chains, the parameter means computed using the first 10% and the last 50% of the chain are compared through a Z-score (stationarity is not rejected if the Z-scores are below the critical values at 5%, NS (non-significant) in the Table). The test was implemented with the Coda package in R [66].
(PDF)

S2 Table. Test of the MCMC chains: Heidelberger and Welch's diagnosis [68] that tests for the non-stationarity of the chains (NS (non-significant) in the Table meaning stationarity is not rejected at the 5% level). The test was implemented with the Coda package in R [66].
(PDF)

S1 Fig. Prior and posterior distributions for the SIRS model inferences of Fig 1. $I(0)$, $S(0)$ initial values, $\beta(0)$ initial value of $\beta(t)$, $1/\alpha$ is the average duration of immunity, γ is the recovery rate, ρ is the reporting rate and σ is the volatility of the Brownian process of $\beta(t)$. The blue distributions are the priors and the discrete histograms are the posteriors. The medians of the prior distributions for $I(0)$, $S(0)$, $\beta(0)$, $1/\alpha$, $1/\gamma$ and ρ are the "true values" used for the simulations of the observed incidences.
(PDF)

S2 Fig. The traces of the MCMC chain for the SIRS model inferences of Fig 1. $I(0)$, $S(0)$ initial values, $\beta(0)$ initial value of $\beta(t)$, $1/\alpha$ is the average duration of immunity, γ is the recovery rate, ρ is the reporting rate and σ is the volatility of the Brownian process of $\beta(t)$.
(PDF)

S3 Fig. Prior and posterior distributions for the SIRS model inferences displayed on Fig 3 when the initial conditions are near the attractor of the dynamics. A/ Observed data generated with a SIRS model and a sinusoidal β with 1 periodic component (5). B/ Observed data generated with a SIRS model and a sinusoidal β with 2 periodic components. In A/ and B/, $I(0)$, $S(0)$ are initial values, $1/\alpha$ is the average duration of immunity, γ is the recovery rate and σ is the volatility of the Brownian process of $\beta(t)$. C/ Observed data generated with a SIRS model and a sinusoidal β with 3 periodic components, σ the volatility of the Brownian process of $\beta(t)$ is the only parameter inferred. The blue distributions are the priors and the discrete histograms are the posteriors. The medians of the prior distributions are the "true values" used for the simulations of the observed incidences.
(PDF)

S4 Fig. Reconstruction of both the incidence (A) and the time evolution $\beta(t)$ (B) for the SIRS model as in Fig 1 but only 5 parameters have been inferred, $\beta(0)$ and ρ were fixed. Model parameters as in Fig 1 and S6 Fig.
(PDF)

S5 Fig. Simulation of the SIRS model when the initial conditions are near the attractor of the dynamics and just 5 parameters inferred: (A) Susceptibles; (B) Infectious; (C) Time evolution of both R_{eff} and $\beta(t)$. In (A) and (B) the black lines are the true values, the blue lines are the median of the posterior, the mauve areas are the 50% CI and the light blue areas the 95% CI. In (C) the black line is the true values of R_{eff}, the blue line is the median of the posterior, and the dashed lines the 95% CI of R_{eff}; the red dot line is the true time evolution of $\beta(t)$ and the red line the median of its posterior. Model parameters as in Fig 1 and S6 Fig.
(PDF)

S6 Fig. Prior and posterior distributions for the SIRS model inferences of S4 Fig. $I(0)$, $S(0)$ initial values, $1/\alpha$ is the average duration of immunity, γ is the recovery rate and σ is the volatility of the Brownian process of $\beta(t)$. The blue distributions are the priors and the discrete histograms are the posteriors. The medians of the prior distributions for $I(0)$, $S(0)$, $1/\alpha$ and $1/\gamma$, are the "true values" used for the simulations of the observed incidences.
(PDF)

S7 Fig. Reconstruction of both the incidence (A) and the time evolution $\beta(t)$ (B) for the SIRS model as in Fig 1 but σ the volatility of the Brownian process of $\beta(t)$ is the only parameter inferred. Model parameters as in Fig 1 and S8 Fig.
(PDF)

S8 Fig. The trace of the MCMC chain and the prior and posterior distributions for the SIRS model inferences of S7 Fig when σ is the volatility of the Brownian process of $\beta(t)$ is the only parameter inferred.
(PDF)

S9 Fig. Reconstruction of both the incidence (A) and the time evolution $\beta(t)$ (B) for the SIRS model as in S4 Fig but the logarithm transformation of the Brownian process of $\beta(t)$ is not used: $d\theta(t) = \sigma.dB(t)$. Model parameters as in Fig 1 and S4 Fig.
(PDF)

S10 Fig. As in S6 Fig but the logarithm transformation of the Brownian process of $\beta(t)$ is not used: $d\theta(t) = \sigma.dB(t)$.
(PDF)

S11 Fig. Reconstruction of both the incidence (A) and the time evolution $\beta(t)$ (B) with the true SIRS model. In (A) the black points are observations generated with a Poisson process with a mean equal to the incidence simulated by the model. In (B) the black points are the true values of $\beta(t) = \beta_0.(1 + \beta_1 \sin(2\pi t/365+2\pi\phi))$. The blue lines are the median of the posterior, the mauve areas are the 50% Credible Intervals (CI) and the light blue areas the 95% CI. For all the figures, the observation process is also applied to the inferred incidence trajectory. The time unit of the model is *day*, the initial date is arbitrary (2000-01-09) and parameters used for the SIRS model are as follows: $\mu = 1/(50*365)$, $\alpha = 1/(7*365)$, $\gamma = 1/14$, $\beta_0 = 0.65$, $\beta_1 = 0.4$, $\phi = -0.2$, $\rho = 1$, $N = 10000$, $S(0) = 600$, $I(0) = 30$. The prior and posterior distributions of the inferred parameters are in S13 Fig
(PDF)

S12 Fig. Simulation of the true SIRS model: (A) Susceptibles; (B) Infectious; (C) Time evolution of both R_{eff} and $\beta(t)$. In (A) and (B) the black lines are the true values, the blue lines are the median of the posterior, the mauve areas are the 50% CI and the light blue areas the 95% CI. In (C) the black line is the true values of R_{eff}, the blue line is the median of the posterior, and the dashed lines the 95% CI of R_{eff}; the red dot line is the true time evolution of $\beta(t)$ and the red line the median of its posterior. Model parameters as in S11 Fig and S13 Fig.
(PDF)

S13 Fig. Prior and posterior distributions for the true SIRS model inferences of S11 Fig. $I(0)$, $S(0)$ initial values, β_0, β_1, ϕ, the parameters of the sinusoidal β, $1/\alpha$ is the average duration of immunity, γ is the recovery rate and ρ is the reporting rate. The blue distributions are the priors and the discrete histograms are the posteriors. The medians of the prior distributions are the "true values" used for the simulations of the observed incidences.
(PDF)

S14 Fig. Simulation of the SIRS model: (A) Susceptibles; (B) Infectious; (C) Time evolution of both R_{eff} and $\varepsilon_S(t)$. In (A) and (B) the black lines are the true values, the blue lines are the median of the posterior, the mauve areas are the 50% CI and the light blue areas the 95% CI. In (C) the black line is the true values of R_{eff}, the blue line is the median of the posterior and the dashed lines the 95% CI of R_{eff}; the red dot line is the true time evolution of $\varepsilon_S(t)$ and the red line the median of its posterior. Model parameters as in Fig 4 and S15 Fig.
(PDF)

S15 Fig. Prior and posterior distributions for the SIRS model inferences of Fig 4. $I(0)$, $S(0)$ initial values, $1/\alpha$ is the average duration of immunity, γ is the recovery rate and σ is the volatility of the Brownian process of $\varepsilon_S(t)$. The blue distributions are the priors and the discrete

histograms are the posteriors. The medians of the prior distributions for $I(0)$, $S(0)$, $1/\alpha$ and $1/\gamma$, are the "true values" used for the simulations of the observed incidences.
(PDF)

S16 Fig. Prior and posterior distributions for the parameters of the SIRS flu model. $I(0)$, $S(0)$ initial values expressed in percentage of the population N, $\beta(0)$ initial value of $\beta(t)$, i imported infectious, $1/\alpha$ is the average duration of immunity, γ is the recovery rate and σ is the volatility of the Brownian process of $\beta(t)$. The blue distributions are the priors and the discrete histograms are the posteriors. Prior values are adapted from [46].
(PDF)

S17 Fig. Prior and posterior distributions for the parameters of the SEIR dengue model. $E(0)$, $S(0)$ expressed in percentage of the population N, $\beta(0)$ the initial value of $\beta(t)$, γ is the recovery rate, i imported infectious, $1/\alpha$ is the average duration of immunity, σ is the volatility of the Brownian process of $\beta(t)$ and ρ the reporting rate. The blue distributions are the priors and the discrete histograms are the posteriors. Prior values are adapted from [63].
(PDF)

S18 Fig. Association between dengue transmission rate and monthly average maximum temperature recorded at the Phnom Penh International Airport (Cambodia). (A) Time evolution of the normalized $\beta(t)$ (blue line) and normalized average temperature (red line). (B) and (C) Wavelet Power Spectrum (WPS) [48,70] of the two time series. The graph on the right shows the average WPS. (D) Wavelet coherence [48,70] between the reconstructed $\beta(t)$ and average temperature. In (B), (C) and (D) the colors code for low values in white to high values in dark red. The dashed lines show the 95% CI computed with adapted bootstrappes [71], in (C) the 90% and the 95% CI have been plotted. (E) The evolution of the phase of the two time series computed based on wavelet decomposition for the seasonal mode, blue dashed line for the normalized $\beta(t)$ red dashed line for the normalized averaged temperature and black dotted line for their phase difference. The graph on the right shows the distribution of the phase differences.
(PDF)

S19 Fig. Association between dengue transmission rate and monthly average minimum temperature recorded at the Phnom Penh International Airport (Cambodia). (A) Time evolution of the normalized $\beta(t)$ (blue line) and normalized average temperature (red line). (B) and (C) Wavelet Power Spectrum (WPS) [48,70] of the two time series. The graph on the right shows the average WPS. (D) Wavelet coherence [48,70] between the reconstructed $\beta(t)$ and average temperature. In (B), (C) and (D) the colors code for low values in white to high values in dark red. The dashed lines show the 95% CI computed with adapted bootstrappes [71], in (C) the 90% and the 95% CI have been plotted. (E) The evolution of the phase of the two time series computed based on wavelet decomposition for the seasonal mode, blue dashed line for the normalized $\beta(t)$ red dashed line for the normalized averaged temperature and black dotted line for their phase difference. The graph on the right shows the distribution of the phase differences.
(PDF)

S20 Fig. Association between dengue transmission rate and monthly rainfall recorded at the Phnom Penh International Airport (Cambodia). (A) Time evolution of the normalized $\beta(t)$ (blue line) and normalized monthly rainfall (red line). (B) and (C) Wavelet Power Spectrum (WPS) [48,70] of the two time series. The graph on the right shows the average WPS. (D) Wavelet coherence [48,70] between the reconstructed $\beta(t)$ and monthly rainfall. In (B), (C)

and (D) the colors code for low values in white to high values in dark red. The dashed lines show the 95% CI computed with adapted bootstrappes [71], in (C) the 90% and the 95% CI have been plotted. (E) The evolution of the phase of the two time series computed based on wavelet decomposition for the seasonal mode, blue dashed line for the normalized $\beta(t)$ red dashed line for the normalized monthly rainfall and black dotted line for their phase difference. The graph on the right shows the distribution of the phase differences.
(PDF)

S21 Fig. Association between dengue transmission rate and the Dipole Mode Index (DMI), a proxy of Ocean Indian Dipole (see [72] and http://www.jamstec.go.jp/frcgc/research/d1/iod/HTML/Dipole%20Mode%20Index.html). (A) Time evolution of the normalized $\beta(t)$ (blue line) and normalized DMI (red line). (B) and (C) Wavelet Power Spectrum (WPS) [48,70] of the two time series. The graph on the right shows the average WPS. (D) Wavelet coherence [48,70] between the reconstructed $\beta(t)$ and DMI. In (B), (C) and (D) the colors code for low values in white to high values in dark red. The dashed lines show the 95% CI computed with adapted bootstrappes [71], in (C) the 90% and the 95% CI have been plotted. (E) The evolution of the phase of the two time series computed based on wavelet decomposition for the seasonal mode, blue dashed line for the normalized $\beta(t)$ red dashed line for the normalized DMI and black dotted line for their phase difference. The graph on the right shows the distribution of the phase differences.
(PDF)

Acknowledgments

BC would like to express his gratitude to the "Institut Ecologie et Environnement" of CNRS for facilitating his research activity thanks to a "Délégation CNRS" for 2 years in the URA 3012 "Hosts, Vectors and Infectious Agents". "Lovely thanks" to Una Ni Mhaoldhomhnaigh who has smoothed our "French English".

Author Contributions

Conceptualization: Bernard Cazelles.

Data curation: Bernard Cazelles.

Formal analysis: Bernard Cazelles, Clara Champagne.

Funding acquisition: Bernard Cazelles.

Investigation: Bernard Cazelles, Clara Champagne.

Methodology: Bernard Cazelles, Clara Champagne, Joseph Dureau.

Project administration: Bernard Cazelles.

Resources: Bernard Cazelles.

Software: Joseph Dureau.

Supervision: Bernard Cazelles.

Validation: Bernard Cazelles.

Visualization: Bernard Cazelles.

Writing – original draft: Bernard Cazelles, Clara Champagne, Joseph Dureau.

Writing – review & editing: Bernard Cazelles, Clara Champagne, Joseph Dureau.

References

1. Jones KE, Patel NG, Levy MA, Storeygard A, Balk D, Gittleman JL, Daszak P. Global trends in emerging infectious diseases. Nature. 2008; 451: 990–994. https://doi.org/10.1038/nature06536 PMID: 18288193
2. Cazelles B, Hales S. Infectious diseases, climate influences, and nonstationarity. PLoS Med. 2006; 3 (8): e328. https://doi.org/10.1371/journal.pmed.0030328 PMID: 16903777
3. Altizer S, Dobson A, Hosseini P, Hudson P, Pascual M, Rohani P. Seasonality and the dynamics of infectious diseases. Ecology letters. 2006; 9(4): 467–484. https://doi.org/10.1111/j.1461-0248.2005.00879.x PMID: 16623732
4. Fisman DN. Seasonality of infectious diseases. Annu. Rev. Public Health. 2007; 28: 127–143. https://doi.org/10.1146/annurev.publhealth.28.021406.144128 PMID: 17222079
5. Cazelles B, Chavez M, McMichael AJ, Hales S. Nonstationary influence of El Nino on the synchronous dengue epidemics in Thailand. PLoS Med. 2005; 2(4): e106. https://doi.org/10.1371/journal.pmed.0020106 PMID: 15839751
6. Constantin de Magny GC, Cazelles B, Guégan JF. Cholera threat to humans in Ghana is influenced by both global and regional climatic variability. EcoHealth. 2006; 3(4): 223–231.
7. Laneri K, Bhadra A, Ionides EL, Bouma M, Dhiman RC, Yadav RS, Pascual M. 2010 Forcing versus feedback: epidemic malaria and monsoon rains in northwest India. PLoS Comput. Biol. 2010; 6: e1000898. https://doi.org/10.1371/journal.pcbi.1000898 PMID: 20824122
8. Metcalf CJE, Walter KS, Wesolowski A, Buckee CO, Shevliakova E, Tatem AJ, Boos WR, Weinberger DM, Pitzer VE. Identifying climate drivers of infectious disease dynamics: Recent advances and challenges ahead. Proc. R. Soc. B. 2017; 284: 20170901. https://doi.org/10.1098/rspb.2017.0901 PMID: 28814655
9. Cauchemez S, Ferguson NM, Wachtel C, Tegnell A, Saour G, Duncan B, Nicoll A. Closure of schools during an influenza pandemic. The Lancet Infectious Diseases. 2009; 9(8): 473–481. https://doi.org/10.1016/S1473-3099(09)70176-8 PMID: 19628172
10. Wu JT, Cowling BJ, Lau EH, Ip DK, Ho LM, Tsang T, Chuang SK, Leung PY, Lo SV, Liu SH, Riley S. School closure and mitigation of pandemic (H1N1) 2009, Hong Kong. Emerging Infectious Diseases. 2010; 16(3): 538–541. https://doi.org/10.3201/eid1603.091216 PMID: 20202441
11. Ewing A, Lee EC, Viboud C, Bansal S. Contact, travel, and transmission: The impact of winter holidays on influenza dynamics in the United States. The Journal of Infectious Diseases. 2016; 215(5): 732–739. https://doi.org/10.1093/infdis/jiw642
12. Khan K, Memish ZA, Chabbra A, Liauw J, Hu W, Janes DA, Sears J., Arino J, Macdonald M., Calderon F., Raposo P, Heidebrecht C, Wang J, Chan A, Brownstein J, Gardam M. Global public health implications of a mass gathering in Mecca, Saudi Arabia during the midst of an influenza pandemic. Journal of Travel Medicine. 2010; 17(2): 75–81. https://doi.org/10.1111/j.1708-8305.2010.00397.x PMID: 20412172
13. Ferrari MJ, Djibo A, Grais RF, Bharti N, Grenfell BT, Bjørnstad ON. Rural–urban gradient in seasonal forcing of measles transmission in Niger. Proc. R. Soc. B. 2010; 277: 2775–2782. https://doi.org/10.1098/rspb.2010.0536 PMID: 20427338
14. Funk S, Salathé M, Jansen VA. Modelling the influence of human behaviour on the spread of infectious diseases: A review. J. R. Soc. Interface. 2010; 7: 1247–1256. https://doi.org/10.1098/rsif.2010.0142 PMID: 20504800
15. Verelst F, Willem L, Beutels P. Behavioural change models for infectious disease transmission: A systematic review (2010–2015). J. R. Soc. Interface. 2016; 13: 20160820. https://doi.org/10.1098/rsif.2016.0820 PMID: 28003528
16. Kucharski AJ, Kwok KO, Wei VW, Cowling BJ, Read JM, Lessler J, Cummings DA, Riley S. The contribution of social behavior to the transmission of influenza A in a human population. PLoS Pathogen. 2014; 10: e1004206. https://doi.org/10.1371/journal.ppat.1004206
17. Keeling MJ, Rohani P. Modeling infectious diseases in humans and animals. Princeton University Press; 2008.
18. Heesterbeek H, Anderson RM, Andreasen V, Bansal S, De Angelis D, Dye C et al. Modeling infectious disease dynamics in the complex landscape of global health. Science. 2015; 347(6227), aaa4339. https://doi.org/10.1126/science.aaa4339 PMID: 25766240
19. Metcalf CJE, Lessler J. Opportunities and challenges in modeling emerging infectious diseases. Science. 2017; 357(6347), 149–152. https://doi.org/10.1126/science.aam8335 PMID: 28706037
20. Axelsen JB, Yaari R, Grenfell BT, Stone L. Multiannual forecasting of seasonal influenza dynamics reveals climatic and evolutionary drivers. Proc Natl Acad Sci USA. 2014; 111(26): 9538–9542. https://doi.org/10.1073/pnas.1321656111 PMID: 24979763

21. Dorigatti I, Cauchemez S, Ferguson NM. Increased transmissibility explains the third wave of infection by the 2009 H1N1 pandemic virus in England. Proc Natl Acad Sci USA. 2013; 110(33): 13422–13427. https://doi.org/10.1073/pnas.1303117110 PMID: 23882078

22. Wallinga J, Teunis P. Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. American Journal of Epidemiology. 2014; 160(6): 509–516. https://doi.org/10.1093/aje/kwh255

23. Cauchemez S, Boëlle PY, Donnelly CA, Ferguson NM, Thomas G, Leung GM, Hedley AJ, Anderson R., Valleron AJ. Real-time estimates in early detection of SARS. Emerging Infectious Diseases. 2006; 12(1): 110–113. https://doi.org/10.3201/eid1201.050593 PMID: 16494726

24. Nishiura H, Chowell G, Heesterbeek H, Wallinga J. The ideal reporting interval for an epidemic to objectively interpret the epidemiological time course. J. R. Soc. Interface. 2009; 7: 297–307. https://doi.org/10.1098/rsif.2009.0153 PMID: 19570792

25. Cori A, Ferguson NM, Fraser C, Cauchemez S. A new framework and software to estimate time-varying reproduction numbers during epidemics. American Journal of Epidemiology. 2013; 178(9), 1505–1512. https://doi.org/10.1093/aje/kwt133 PMID: 24043437

26. Ionides EL, Bretó C, King AA. Inference for nonlinear dynamical systems. Proc Natl Acad Sci USA. 2006; 103(49): 18438–18443. https://doi.org/10.1073/pnas.0603181103 PMID: 17121996

27. Churcher TS, Cohen JM, Novotny J, Ntshalintshali N, Kunene S, Cauchemez S. Measuring the path toward malaria elimination. Science. 2014; 344(6189): 1230–1232. https://doi.org/10.1126/science.1251449 PMID: 24926005

28. Coelho FC, De Carvalho LM. Estimating the attack ratio of dengue epidemics under time-varying force of infection using aggregated notification data. Scientific report. 2015; 5: 18455. http://dx.doi.org/10.1038/srep18455

29. Ellner SP, Bailey BA, Bobashev GV, Gallant AR, Grenfell B T, Nychka DW. Noise and nonlinearity in measles epidemics: combining mechanistic and statistical approaches to population modeling. The American Naturalist 1998; 151: 425–440. https://doi.org/10.1086/286130 PMID: 18811317

30. King AA, Ionides EL, Pascual M, Bouma MJ. Inapparent infections and cholera dynamics. Nature. 2008; 454: 877–880. https://doi.org/10.1038/nature07084 PMID: 18704085

31. Bhadra A, Ionides EL, Laneri K, Pascual M, Bouma M, Dhiman RC (2011). Malaria in Northwest India: Data analysis via partially observed stochastic differential equation models driven by Lévy noise. Journal of the American Statistical Association. 2011; 106: 440–451. https://doi.org/10.1198/jasa.2011.ap10323

32. He D, Dushoff J, Day T, Ma J, Earn DJ. Mechanistic modelling of the three waves of the 1918 influenza pandemic. Theoretical Ecology. 2011; 4: 283–288. https://doi.org/10.1007/s12080-011-0123-3

33. Martinez-Bakker M, King AA, Rohani P. Unraveling the transmission ecology of polio. PLoS Biol. 2015; 13: e1002172. https://doi.org/10.1371/journal.pbio.1002172 PMID: 26090784

34. Cazelles B, Chau NP. Adaptive dynamic modelling of HIV/AIDS epidemics using extended Kalman filter. Journal of Biological Systems. 1995; 3: 759–768. https://doi.org/10.1142/S0218339095000691

35. Cazelles B, Chau, NP. Using the Kalman filter and dynamic models to assess the changing HIV/AIDS epidemic. Mathematical Biosciences. 1997; 140(2): 131–154. https://doi.org/10.1016/S0025-5564(96)00155-1 PMID: 9046772

36. Dureau J, Kalogeropoulos K, Baguelin M. Capturing the time-varying drivers of an epidemic using stochastic dynamical systems. Biostatistics. 2013; 14(3): 541–555. https://doi.org/10.1093/biostatistics/kxs052 PMID: 23292757

37. Angulo MT, Velasco-Hernandez JX. Robust qualitative estimation of time-varying contact rates in uncertain epidemics. Epidemics. 2018; https://doi.org/10.1016/j.epidem.2018.03.001 PMID: 29567063

38. Smirnova A, Chowell G, de Camp L. Forecasting epidemics through nonparametric estimation of time-dependent transmission rates using the SEIR model. Bulletin of Mathematical Biology. 2017; on-line https://doi.org/10.1007/s11538-017-0284-3 PMID: 28466232

39. Finkenstädt BF, Grenfell BT. Time series modelling of childhood diseases: a dynamical systems approach. Journal of the Royal Statistical Society: Series C. 2000; 49(2): 187–205.

40. Bjørnstad ON, Finkenstädt BF, Grenfell BT. Dynamics of measles epidemics: estimating scaling of transmission rates using a time series SIR model. Ecological Monographs. 2002; 72(2): 169–184.

41. Cobelli C, DiStefano JJ. Parameter and structural identifiability concepts and ambiguities: a critical review and analysis. Am. J. Physiol.—Regul. Integr. Comp. Physiol. 1980; 239: R7–R24.

42. Miao H, Xia X, Perelson AS, Wu H. On identifiability of nonlinear ODE models and applications in viral dynamics. SIAM review 2011; 53: 3–39. https://doi.org/10.1137/090757009 PMID: 21785515

43. Andrieu C, Doucet A, Holenstein R. Particle Markov chain Monte Carlo methods. Journal of the Royal Statistical Society: Series B. 2010; 72: 269–342.

44. Hethcote HW. The mathematics of infectious diseases. SIAM review. 2000; 42(4): 599–653.
45. Lee EC, Kelly MR, Ochocki BM, Akinwumi SM, Hamre KES, Tien JH, Eisenberg MC. Model distinguishability and inference robustness in mechanisms of cholera transmission and loss of immunity. Journal of Theoretical Biology. 2017; 420: 68–81. https://doi.org/10.1016/j.jtbi.2017.01.032 PMID: 28130096
46. Yaari R, Katriel G, Huppert A, Axelsen JB, Stone L. Modelling seasonal influenza: the role of weather and punctuated antigenic drift. J R Soc Interface. 2013; 10: 20130298. https://doi.org/10.1098/rsif.2013.0298 PMID: 23676899
47. Teurlai M, Huy R, Cazelles B, Duboz R, Baehr C, Vong S. Can Human Movements Explain Heterogeneous Propagation of Dengue Fever in Cambodia? PLoS Negl Trop Dis. 2012; 6(12): e1957. https://doi.org/10.1371/journal.pntd.0001957 PMID: 23236536
48. Cazelles B, Chavez M, Constantin de Magny G, Guégan JF, Hales S. Time-dependent spectral analysis of epidemiological time-series with wavelets. J. R. Soc. Interface. 2007; 4: 625–636. https://doi.org/10.1098/rsif.2007.0212 PMID: 17301013
49. Martínez-Bello DA, López-Quílez A, Torres-Prieto A. Bayesian dynamic modeling of time series of dengue disease case counts. PLoS Negl Trop Dis. 2017; 11(7): e0005696. https://doi.org/10.1371/journal.pntd.0005696 PMID: 28671941
50. Opatowski L, Baguelin M, Eggo RM. Influenza interaction with cocirculating pathogens and its impact on surveillance, pathogenesis, and epidemic profile: A key role for mathematical modelling. PLoS Pathogen. 2018; 14: e1006770. https://doi.org/10.1371/journal.ppat.1006770
51. Funk S, Camacho A, Kucharski AJ, Eggo RM, Edmunds WJ. Real-time forecasting of infectious disease dynamics with a stochastic semi-mechanistic model. Epidemics. 2018; 22: 56–61. https://doi.org/10.1016/j.epidem.2016.11.003 PMID: 28038870
52. Camacho A, Kucharski A, Aki-Sawyerr Y, White MA, Flasche S, Baguelin M, Pollington T, Carney JR, Glover R, Smout E, Tiffany A, Edmunds WJ, Funk. Temporal changes in Ebola transmission in Sierra Leone and implications for control requirements: a real-time modelling study. PLoS Curr. 2015; 7 (http://dx.doi.org/10.1371/currents.outbreaks.406ae55e83ec0b5193e30856b9235ed2).
53. Jacquez JA, Greif P. Numerical parameter identifiability and estimability: Integrating identifiability, estimability, and optimal sampling design. Mathematical Biosciences. 1985; 77: 201–227.
54. Tunali E, Tarn TJ. New results for identifiability of nonlinear systems. IEEE Transactions on Automatic Control. 1987; 32: 146–154.
55. Audoly S, Bellu G, D'Angio L, Saccomani MP, Cobelli C. Global identifiability of nonlinear models of biological systems. IEEE Transactions on Biomedical Engineering. 2001; 48: 55–65. https://doi.org/10.1109/10.900248 PMID: 11235592
56. Evans ND, White LJ, Chapman MJ, Godfrey KR, Chappell MJ. The structural identifiability of the susceptible infected recovered model with seasonal forcing. Mathematical Biosciences. 2005; 194: 175–197. https://doi.org/10.1016/j.mbs.2004.10.011 PMID: 15854675
57. Chapman JD, Evans ND. The structural identifiability of susceptible–infective–recovered type epidemic models with incomplete immunity and birth targeted vaccination. Biomedical Signal Processing and Control. 2009; 4: 278–284.
58. Champagne C, Salthouse DG, Paul R, Cao-Lormeau VM, Roche B, Cazelles B. Structure in the variability of the basic reproductive number (R_0) for Zika epidemics in the Pacific islands. eLife. 2016; 5: e19874. https://doi.org/10.7554/eLife.19874 PMID: 27897973
59. Bengtsson T, Bickel P, and Bo Li B. Curse-of-dimensionality revisited: Collapse of the particle filter in very large scale systems. In Probability and Statistics: Essays in Honor of David A. Freedman, Vol. 2; 2008. pp. 316–334.
60. Toni T, Welch D, Strelkowa N, Ipsen A, Stumpf M. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. J R Soc Interface. 2007; 6: 187–202. https://doi.org/10.1098/rsif.2008.0172
61. Sunnåker M, Busetto AG, Numminen E, Corander J, Foll M, Dessimoz C. Approximate Bayesian Computation. PLoS Comput Biol. 2013; 9: e1002803. https://doi.org/10.1371/journal.pcbi.1002803 PMID: 23341757
62. Metcalf CJE, Farrar J, Cutts FT, Basta NE, Graham AL, Lessler J, Ferguson NM, Burke DS, Grenfell BT. Use of serological surveys to generate key insights into the changing global landscape of infectious disease. The Lancet. 2016; 388(10045): 728–730. https://doi.org/10.1016/S0140-6736(16)30164-7
63. Champagne C, Paul R, Ly S, Duong V, Leang R, Cazelles B. Dengue modeling in rural Cambodia: statistical performance versus epidemiological relevance. bioRxiv. 2017; 208876.
64. Bretó C, He D, Ionides EL, King AA. Time Series Analysis via Mechanistic Models. The Annals of Applied Statistics. 2009; 3: 319–348.

65. Dureau J, Ballesteros S, Bogich T. SSM: Inference for time series analysis with State Space Models. 2013; https://github.com/JDureau/ssm/blob/master/doc/doc.pdf.

66. Plummer M, Best N, Cowles K, Vines K. Coda: convergence diagnosis and output analysis for MCMC. R news. 2006; 6: 7–11.

67. Geweke J. Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In Bernado JM, Berger JO, Dawid AP, Smith AFM, editors. Bayesian Statistics 4. Clarendon Press, Oxford, UK; 1992.

68. Heidelberger P, Welch PD. Simulation run length control in the presence of an initial transient. Opns Res. 1983; 31: 1109–11044.

69. Torrence C, Compo GP. A practical guide to wavelet analysis. Bull Am Meteorol Soc. 1998; 79: 61–78.

70. Cazelles B, Chavez M, Berteaux D, Ménard F, Vik JO, Jenouvrier S, Stenseth NC. Wavelet analysis of ecological time series. Oecologia. 2008; 156: 287–304. https://doi.org/10.1007/s00442-008-0993-2 PMID: 18322705

71. Cazelles B, Cazelles K, Chavez M. Wavelet analysis in ecology and epidemiology: impact of statistical tests. J. R. Soc. Interface. 2014; 11: 20130585. https://doi.org/10.1098/rsif.2013.0585 PMID: 24284892

72. Saji NH, Goswami BN, Vinayachandran PN, Yamagata T. A dipole mode in the tropical Indian Ocean. Nature. 1999; 401(6751): 360–363. https://doi.org/10.1038/43854 PMID: 16862108

Bibliographie

Bibliography

B. Adams and M. Boots. Modelling the relationship between antibody-dependent enhancement and immunological distance with application to dengue. *Journal of Theoretical Biology*, 242 (2):337–346, Sept. 2006. ISSN 0022-5193. doi: 10.1016/j.jtbi.2006.03.002. URL http://www.sciencedirect.com/science/article/pii/S002251930600107X.

B. Adams and M. Boots. How important is vertical transmission in mosquitoes for the persistence of dengue? Insights from a mathematical model. *Epidemics*, 2(1):1–10, Mar. 2010. ISSN 1878-0067. doi: 10.1016/j.epidem.2010.01.001.

B. Adams and D. D. Kapan. Man Bites Mosquito: Understanding the Contribution of Human Movement to Vector-Borne Disease Dynamics. *PLoS ONE*, 4(8):e6763, Aug. 2009. ISSN 1932-6203. doi: 10.1371/journal.pone.0006763. URL http://dx.plos.org/10.1371/journal.pone.0006763.

B. Adams, E. C. Holmes, C. Zhang, M. P. Mammen, S. Nimmannitya, S. Kalayanarooj, and M. Boots. Cross-protective immunity can account for the alternating epidemic pattern of dengue virus serotypes circulating in Bangkok. *Proceedings of the National Academy of Sciences*, 103 (38):14234–14239, Sept. 2006. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.0602768103. URL http://www.pnas.org/content/103/38/14234.

M. Aguiar, B. Kooi, and N. Stollenwerk. Epidemiology of Dengue Fever: A Model with Temporary Cross-Immunity and Possible Secondary Infection Shows Bifurcations and Chaotic Behaviour in Wide Parameter Regions. *Mathematical Modelling of Natural Phenomena*, 3(4):48–70, 2008. ISSN 0973-5348, 1760-6101. doi: 10.1051/mmnp:2008070. URL http://www.mmnp-journal.org/10.1051/mmnp:2008070.

M. Aguiar, S. Ballesteros, B. W. Kooi, and N. Stollenwerk. The role of seasonality and import in a minimalistic multi-strain dengue model capturing differences between primary and secondary infections: complex dynamics and its implications for data analysis. *Journal of Theoretical Biology*, 289:181–196, Nov. 2011. ISSN 1095-8541. doi: 10.1016/j.jtbi.2011.08.043.

M. Aguiar, B. W. Kooi, F. Rocha, P. Ghaffari, and N. Stollenwerk. How much complexity is needed to describe the fluctuations observed in dengue hemorrhagic fever incidence data? *Ecological Complexity*, 16:31–40, Dec. 2013. ISSN 1476-945X. doi: 10.1016/j.ecocom.2012.09.001. URL http://www.sciencedirect.com/science/article/pii/S1476945X12000670.

M. Aguiar, R. Paul, A. Sakuntabhai, and N. Stollenwerk. Are we modelling the correct dataset? Minimizing false predictions for dengue fever in Thailand. *Epidemiology & Infection*, 142

(11):2447–2459, Nov. 2014. ISSN 1469-4409. doi: 10.1017/S0950268813003348. URL http://journals.cambridge.org/article_S0950268813003348.

M. Aguiar, N. Stollenwerk, and S. B. Halstead. The risks behind Dengvaxia recommendation. *The Lancet. Infectious Diseases*, 16(8):882–883, 2016. ISSN 1474-4457. doi: 10.1016/S1473-3099(16)30168-2.

H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, Dec. 1974. ISSN 0018-9286. doi: 10.1109/TAC.1974.1100705.

J. Aldstadt, I.-K. Yoon, D. Tannitisupawong, R. G. Jarman, S. J. Thomas, R. V. Gibbons, A. Uppapong, S. Iamsirithaworn, A. L. Rothman, T. W. Scott, and T. Endy. Space-time analysis of hospitalised dengue patients in rural Thailand reveals important temporal intervals in the pattern of dengue virus transmission. *Tropical medicine & international health: TM & IH*, 17 (9):1076–1085, Sept. 2012. ISSN 1365-3156. doi: 10.1111/j.1365-3156.2012.03040.x.

L. J. S. Allen. An Introduction to Stochastic Epidemic Models. In *Mathematical Epidemiology*, Lecture Notes in Mathematics, pages 81–130. Springer, Berlin, Heidelberg, 2008. ISBN 978-3-540-78910-9 978-3-540-78911-6. doi: 10.1007/978-3-540-78911-6_3. URL https://link.springer.com/chapter/10.1007/978-3-540-78911-6_3.

D. Alonso, A. J. McKane, and M. Pascual. Stochastic amplification in epidemics. *Journal of the Royal Society Interface*, 4(14):575–582, June 2007. ISSN 1742-5689. doi: 10.1098/rsif.2006.0192. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2373404/.

N. Alzahrani, P. Neal, S. E. Spencer, T. J. McKinley, and P. Touloupou. Model selection for time series of count data. *Computational Statistics & Data Analysis*, 122:33–44, June 2018. ISSN 01679473. doi: 10.1016/j.csda.2018.01.002. URL http://linkinghub.elsevier.com/retrieve/pii/S0167947318300033.

M. Amaku, F. Azevedo, M. N. Burattini, G. E. Coelho, F. a. B. Coutinho, D. Greenhalgh, L. F. Lopez, R. S. Motitsuki, A. Wilder-Smith, and E. Massad. Magnitude and frequency variations of vector-borne infection outbreaks using the Ross-Macdonald model: explaining and predicting outbreaks of dengue fever. *Epidemiology and Infection*, pages 1–16, Aug. 2016. ISSN 1469-4409. doi: 10.1017/S0950268816001448.

R. M. Anderson and R. M. May. Directly transmitted infections diseases: control by vaccination. *Science*, 215(4536):1053–1060, Feb. 1982. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.7063839. URL http://science.sciencemag.org/content/215/4536/1053.

R. M. Anderson and R. M. May. *Infectious Diseases of Humans: Dynamics and Control*. Oxford University Press, Oxford, New York, Aug. 1992. ISBN 978-0-19-854040-3.

H. Andersson and T. Britton. *Stochastic Epidemic Models and Their Statistical Analysis*. Lecture Notes in Statistics. Springer-Verlag, New York, 2000. ISBN 978-0-387-95050-1. URL //www.springer.com/us/book/9780387950501.

M. Andraud, N. Hens, C. Marais, and P. Beutels. Dynamic Epidemiological Models for Dengue Transmission: A Systematic Review of Structural Approaches. *PLoS ONE*, 7(11):e49085, Nov. 2012. ISSN 1932-6203. doi: 10.1371/journal.pone.0049085. URL http://dx.plos.org/10.1371/journal.pone.0049085.

C. Andrieu, A. Doucet, and R. Holenstein. Particle Markov chain Monte Carlo methods: Particle Markov Chain Monte Carlo Methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(3):269–342, June 2010. ISSN 13697412, 14679868. doi: 10.1111/j.1467-9868.2009.00736.x. URL http://doi.wiley.com/10.1111/j.1467-9868.2009.00736.x.

A. Andronico, F. Dorléans, J.-L. Fergé, H. Salje, F. Ghawché, A. Signate, E. Daudens-Vaysse, L. Baudouin, T. Dub, M. Aubry, V.-M. Cao-Lormeau, M. Ledrans, H. Noel, H.-P. Mallet, A. Fontanet, A. Cabié, and S. Cauchemez. Real-Time Assessment of Health-Care Requirements During the Zika Virus Epidemic in Martinique. *American Journal of Epidemiology*, 186(10): 1194–1203, Nov. 2017. ISSN 1476-6256. doi: 10.1093/aje/kwx008.

M. Aubry, J. Finke, A. Teissier, C. Roche, J. Broult, S. Paulous, P. Desprès, V.-M. Cao-Lormeau, and D. Musso. Seroprevalence of arboviruses among blood donors in French Polynesia, 2011-2013. *International journal of infectious diseases: IJID: official publication of the International Society for Infectious Diseases*, 41:11–12, Dec. 2015a. ISSN 1878-3511. doi: 10.1016/j.ijid.2015.10.005.

M. Aubry, A. Teissier, C. Roche, S. Teururai, P. Desprès, H.-P. Mallet, S. Merceron, M. Huart, S. Sicard, X. Deparis, D. Musso, and V. Cao-Lormeau. Serosurvey of dengue, Zika and other mosquito-borne viruses in French Polynesia, abstr 765, Oct. 2015b. URL http://www.abstractsonline.com/Plan/ViewAbstract.aspx?sKey=2e5199c4-aceb-4d61-8769-342586917c5a&cKey=74ffe328-3a83-4f28-84c6-5656971501c8&mKey=%7bAB652FDF-0111-45C7-A5E5-0BA9D4AF5E12%7d.

M. Aubry, A. Teissier, M. Huart, S. Merceron, J. Vanhomwegen, C. Roche, A.-L. Vial, S. Teururai, S. Sicard, S. Paulous, P. Desprès, J.-C. Manuguerra, H.-P. Mallet, D. Musso, X. Deparis, and V.-M. Cao-Lormeau. Zika Virus Seroprevalence, French Polynesia, 2014-2015. *Emerging Infectious Diseases*, 23(4):669–672, Apr. 2017. doi: 10.3201/eid2304.161549. URL https://hal.archives-ouvertes.fr/hal-01521256.

K. Auranen, E. Arjas, T. Leino, and A. K. Takala. Transmission of Pneumococcal Carriage in Families: A Latent Markov Process Model for Binary Longitudinal Data. *Journal of the American Statistical Association*, 95(452):1044, Dec. 2000. ISSN 01621459. doi: 10.2307/2669741. URL https://www.jstor.org/stable/2669741?origin=crossref.

N. T. J. Bailey. The mathematical theory of infectious diseases and its applications. 2nd edition. *The mathematical theory of infectious diseases and its applications. 2nd edition.*, 1975. URL https://www.cabdirect.org/cabdirect/abstract/19762902036.

G. Barba-Spaeth, W. Dejnirattisai, A. Rouvinski, M.-C. Vaney, I. Medits, A. Sharma, E. Simon-Lorière, A. Sakuntabhai, V.-M. Cao-Lormeau, A. Haouz, P. England, K. Stiasny, J. Mongkolsapaya, F. X. Heinz, G. R. Screaton, and F. A. Rey. Structural basis of potent Zika-dengue virus antibody cross-neutralization. *Nature*, 536(7614):48–53, 2016. ISSN 1476-4687. doi: 10.1038/nature18938.

S. V. Bardina, P. Bunduc, S. Tripathi, J. Duehr, J. J. Frere, J. A. Brown, R. Nachbagauer, G. A. Foster, D. Krysztof, D. Tortorella, S. L. Stramer, A. García-Sastre, F. Krammer, and J. K. Lim. Enhancement of Zika virus pathogenesis by preexisting antiflavivirus immunity. *Science (New York, N.Y.)*, 356(6334):175–180, 2017. ISSN 1095-9203. doi: 10.1126/science.aal4365.

D. H. Barmak, C. O. Dorso, M. Otero, and H. G. Solari. Dengue epidemics and human mobility. *Physical Review E*, 84(1), July 2011. ISSN 1539-3755, 1550-2376. doi: 10.1103/PhysRevE.84.011901. URL https://link.aps.org/doi/10.1103/PhysRevE.84.011901.

J. M. Barrios, W. W. Verstraeten, P. Maes, J.-M. Aerts, J. Farifteh, and P. Coppin. Using the Gravity Model to Estimate the Spatial Spread of Vector-Borne Diseases. *International Journal of Environmental Research and Public Health*, 9(12):4346–4364, Dec. 2012. ISSN 1661-7827. doi: 10.3390/ijerph9124346. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3546766/.

M. S. Bartlett. Measles Periodicity and Community Size. *Journal of the Royal Statistical Society. Series A (General)*, 120(1):48–70, 1957. ISSN 0035-9238. doi: 10.2307/2342553. URL http://www.jstor.org/stable/2342553.

M. S. Bartlett. The Critical Community Size for Measles in the United States. *Journal of the Royal Statistical Society. Series A (General)*, 123(1):37–44, 1960. ISSN 0035-9238. doi: 10.2307/2343186. URL http://www.jstor.org/stable/2343186.

M. S. Bartlett. The Relevance of Stochastic Models for Large-Scale Epidemiological Phenomena. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 13(1):2–8, 1964. ISSN 0035-9254. doi: 10.2307/2985217. URL http://www.jstor.org/stable/2985217.

W. G. C. Bearcroft. Zika virus infection experimentally induced in a human volunteer. *Transactions of The Royal Society of Tropical Medicine and Hygiene*, 50(5):442–448, Sept. 1956. ISSN 0035-9203. doi: 10.1016/0035-9203(56)90091-8. URL https://academic.oup.com/trstmh/article/50/5/442/1908586.

D. Bernoulli. Essai d'une nouvelle analyse de la mortalité causée par la petite vérole et des avantages de l'inoculation pour la prévenir. In *Mémoires de mathématiques et de physique, Académie royale des sciences*, pages 1–45. Imprimerie de Du Pont, Paris, 1760. URL http://gallica.bnf.fr/ark:/12148/bpt6k3558n. ark:/12148/bpt6k3558n.

A. Bhadra. Discussion of "Particle Markov Chain Monte Carlo Methods" by C. Andrieu, A. Doucet and R. Holenstein. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, pages 314–315, 2010. doi: doi:10.1111/j.1467-9868.2009.00736.x.

S. Bhatt, P. W. Gething, O. J. Brady, J. P. Messina, A. W. Farlow, C. L. Moyes, J. M. Drake, J. S. Brownstein, A. G. Hoen, O. Sankoh, M. F. Myers, D. B. George, T. Jaenisch, G. R. W. Wint, C. P. Simmons, T. W. Scott, J. J. Farrar, and S. I. Hay. The global distribution and burden of dengue. *Nature*, 496(7446):504–507, Apr. 2013. ISSN 0028-0836. doi: 10.1038/nature12060. URL http://www.nature.com/nature/journal/v496/n7446/full/nature12060.html.

P. J. Birrell, G. Ketsetzis, N. J. Gay, B. S. Cooper, A. M. Presanis, R. J. Harris, A. Charlett, X.-S. Zhang, P. J. White, R. G. Pebody, and D. D. Angelis. Bayesian modeling to unmask and predict influenza A/H1n1pdm dynamics in London. *Proceedings of the National Academy of Sciences*, 108(45):18238–18243, Nov. 2011. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1103002108. URL http://www.pnas.org/content/108/45/18238.

P. J. Birrell, D. D. Angelis, and A. M. Presanis. Evidence Synthesis for Stochastic Epidemic Models. *Statistical Science*, 33(1):34–43, Feb. 2018. ISSN 0883-4237, 2168-8745. doi: 10.1214/17-STS631. URL https://projecteuclid.org/euclid.ss/1517562023.

S. M. Blower and H. Dowlatabadi. Sensitivity and Uncertainty Analysis of Complex Models of Disease Transmission: An HIV Model, as an Example. *International Statistical Review / Revue Internationale de Statistique*, 62(2):229–243, 1994. ISSN 0306-7734. doi: 10.2307/1403510. URL http://www.jstor.org/stable/1403510.

L. R. Bowman, S. Runge-Ranzinger, and P. J. McCall. Assessing the Relationship between Vector Indices and Dengue Transmission: A Systematic Review of the Evidence. *PLoS Neglected Tropical Diseases*, 8(5):e2848, May 2014. ISSN 1935-2735. doi: 10.1371/journal.pntd.0002848. URL http://dx.plos.org/10.1371/journal.pntd.0002848.

L. R. Bowman, S. Donegan, and P. J. McCall. Is Dengue Vector Control Deficient in Effectiveness or Evidence?: Systematic Review and Meta-analysis. *PLOS Neglected Tropical Diseases*, 10(3):e0004551, Mar. 2016. ISSN 1935-2735. doi: 10.1371/journal.pntd.0004551. URL http://journals.plos.org/plosntds/article?id=10.1371/journal.pntd.0004551.

O. J. Brady, M. A. Johansson, C. A. Guerra, S. Bhatt, N. Golding, D. M. Pigott, H. Delatte, M. G. Grech, P. T. Leisnham, R. Maciel-de Freitas, L. M. Styer, D. L. Smith, T. W. Scott, P. W. Gething, and S. I. Hay. Modelling adult Aedes aegypti and Aedes albopictus survival at different temperatures in laboratory and field settings. *Parasites & Vectors*, 6:351, 2013. ISSN 1756-3305. doi: 10.1186/1756-3305-6-351.

C. Braga, C. F. Luna, C. M. Martelli, W. V. de Souza, M. T. Cordeiro, N. Alexander, M. d. F. P. M. de Albuquerque, J. C. S. Júnior, and E. T. Marques. Seroprevalence and risk factors for dengue infection in socio-economically distinct areas of Recife, Brazil. *Acta Tropica*, 113(3):234–240, Mar. 2010. ISSN 1873-6254. doi: 10.1016/j.actatropica.2009.10.021.

F. Brauer, C. Castillo-Chavez, A. Mubayi, and S. Towers. Some models for epidemics of vector-transmitted diseases. *Infectious Disease Modelling*, 1(1):79–87, Oct. 2016. ISSN 2468-0427. doi: 10.1016/j.idm.2016.08.001. URL http://www.sciencedirect.com/science/article/pii/S2468042716300070.

C. Bretó. Modeling and Inference for Infectious Disease Dynamics: A Likelihood-Based Approach. *Statistical Science*, 33(1):57–69, Feb. 2018. ISSN 0883-4237, 2168-8745. doi: 10.1214/17-STS636. URL https://projecteuclid.org/euclid.ss/1517562025.

C. Bretó, D. He, E. L. Ionides, and A. A. King. Time series analysis via mechanistic models. *The Annals of Applied Statistics*, 3(1):319–348, Mar. 2009. ISSN 1932-6157, 1941-7330. doi: 10.1214/08-AOAS201. URL https://projecteuclid.org/euclid.aoas/1239888373.

T. Britton. Stochastic epidemic models: A survey. *Mathematical Biosciences*, 225(1):24–35, May 2010. ISSN 0025-5564. doi: 10.1016/j.mbs.2010.01.006. URL http://www.sciencedirect.com/science/article/pii/S0025556410000143.

T. Britton and F. Giardina. Introduction to statistical inference for infectious diseases. *Journal de la Société Française de Statistique*, 157(1):53–70, 2016. URL http://journal-sfds.fr/article/view/521.

T. Britton and D. Lindenstrand. Epidemic modelling: Aspects where stochasticity matters. *Mathematical Biosciences*, 222(2):109–116, Dec. 2009. ISSN 00255564. doi: 10.1016/

j.mbs.2009.10.001. URL http://linkinghub.elsevier.com/retrieve/pii/S0025556409001643.

C. S. Broussard, C. K. Shapiro-Mendoza, G. Peacock, S. A. Rasmussen, C. T. Mai, E. E. Petersen, R. R. Galang, K. Newsome, M. R. Reynolds, S. M. Gilboa, C. A. Boyle, and C. A. Moore. Public Health Approach to Addressing the Needs of Children Affected by Congenital Zika Syndrome. *Pediatrics*, 141(Supplement 2):S146–S153, Feb. 2018. ISSN 0031-4005, 1098-4275. doi: 10.1542/peds.2017-2038C. URL http://pediatrics.aappublications.org/content/141/Supplement_2/S146.

J. E. Brown, B. R. Evans, W. Zheng, V. Obas, L. Barrera-Martinez, A. Egizi, H. Zhao, A. Caccone, and J. R. Powell. Human impacts have shaped historical and recent evolution in Aedes aegypti, the dengue and yellow fever mosquito. *Evolution*, 68(2):514–525, Feb. 2014. ISSN 1558-5646. doi: 10.1111/evo.12281. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/evo.12281.

A. Buchholz and N. Chopin. Improving approximate Bayesian computation via quasi-Monte Carlo. *arXiv:1710.01057 [stat]*, Oct. 2017. URL http://arxiv.org/abs/1710.01057. arXiv: 1710.01057.

M. Burattini, M. Chen, A. Chow, F. Coutinho, K. Goh, L. Lopez, S. Ma, and E. Massad. Modelling the control strategies against dengue in Singapore. *Epidemiology and Infection*, 136(3):309–319, Mar. 2008. ISSN 0950-2688. doi: 10.1017/S0950268807008667. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2870819/.

K. P. Burnham and D. R. Anderson. Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociological Methods & Research*, 33(2):261–304, Nov. 2004. ISSN 0049-1241, 1552-8294. doi: 10.1177/0049124104268644. URL http://journals.sagepub.com/doi/10.1177/0049124104268644.

A. Camacho. *Approches stochastiques pour la modélisation en épidémiologie : application à la grippe humaine*. PhD thesis, Université Pierre et Marie Curie-Paris VI, Paris, 2011.

A. Camacho, S. Ballesteros, A. L. Graham, F. Carrat, O. Ratmann, and B. Cazelles. Explaining rapid reinfections in multiple-wave influenza outbreaks: Tristan da Cunha 1971 epidemic as a case study. *Proceedings of the Royal Society of London B: Biological Sciences*, page rspb20110300, Apr. 2011. ISSN 0962-8452, 1471-2954. doi: 10.1098/rspb.2011.0300. URL http://rspb.royalsocietypublishing.org/content/early/2011/04/22/rspb.2011.0300.

A. Camacho, A. Kucharski, Y. Aki-Sawyerr, M. A. White, S. Flasche, M. Baguelin, T. Pollington, J. R. Carney, R. Glover, E. Smout, A. Tiffany, W. J. Edmunds, and S. Funk. Temporal Changes in Ebola Transmission in Sierra Leone and Implications for Control Requirements: a Real-time Modelling Study. *PLoS Currents*, 7, Feb. 2015. ISSN 2157-3999. doi: 10.1371/currents.outbreaks.406ae55e83ec0b5193e30856b9235ed2. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4339317/.

V.-M. Cao-Lormeau and D. Musso. Emerging arboviruses in the Pacific. *The Lancet*, 384(9954):1571–1572, Nov. 2014. ISSN 0140-6736, 1474-547X. doi: 10.1016/S0140-6736(14)61977-2. URL http://www.thelancet.com/journals/lancet/article/PIIS0140-6736(14)61977-2/abstract.

V.-M. Cao-Lormeau, C. Roche, A. Teissier, E. Robin, A.-L. Berry, H.-P. Mallet, A. A. Sall, and D. Musso. Zika Virus, French Polynesia, South Pacific, 2013. *Emerging Infectious Diseases*, 20 (6):1084–1086, June 2014. ISSN 1080-6040, 1080-6059. doi: 10.3201/eid2006.140138. URL http://wwwnc.cdc.gov/eid/article/20/6/14-0138_article.htm.

V. M. Cao-Lormeau, A. Blake, S. Mons, S. Lastere, C. Roche, J. Vanhomwegen, T. Dub, L. Baudouin, A. Teissier, P. Larre, A. L. Vial, C. Decam, V. Choumet, S. K. Halstead, H. J. Willison, L. Musset, J. C. Manuguerra, P. Despres, E. Fournier, H. P. Mallet, D. Musso, A. Fontanet, J. Neil, and F. Ghawché. Guillain-Barré Syndrome outbreak associated with Zika virus infection in French Polynesia: a case-control study. *Lancet (London, England)*, 387 (10027):1531–1539, Apr. 2016. ISSN 1474-547X. doi: 10.1016/S0140-6736(16)00562-6.

R. Carnell. lhs: Latin Hypercube Samples, 2016. URL https://CRAN.R-project.org/package=lhs.

M. Y. Carrillo-Hernández, J. Ruiz-Saenz, L. J. Villamizar, S. Y. Gómez-Rangel, and M. Martínez-Gutierrez. Co-circulation and simultaneous co-infection of dengue, chikungunya, and zika viruses in patients with febrile syndrome at the Colombian-Venezuelan border. *BMC infectious diseases*, 18(1):61, 2018. ISSN 1471-2334. doi: 10.1186/s12879-018-2976-1.

S. Cauchemez, F. Carrat, C. Viboud, A. J. Valleron, and P. Y. Boëlle. A Bayesian MCMC approach to study transmission of influenza: application to household longitudinal data. *Statistics in Medicine*, 23(22):3469–3487, Nov. 2004. ISSN 0277-6715. doi: 10.1002/sim.1912.

B. Cazelles and K. Cazelles. Major urban centers have weak influence on the timing of dengue epidemics in Southeast Asia. *International Journal of Infectious Diseases*, 21, Supplement 1:217, Apr. 2014. ISSN 1201-9712. doi: 10.1016/j.ijid.2014.03.873. URL http://www.sciencedirect.com/science/article/pii/S1201971214009321.

B. Cazelles and N. P. Chau. Using the Kalman filter and dynamic models to assess the changing HIV/AIDS epidemic. *Mathematical Biosciences*, 140(2):131–154, Mar. 1997. ISSN 0025-5564. doi: 10.1016/S0025-5564(96)00155-1. URL http://www.sciencedirect.com/science/article/pii/S0025556496001551.

B. Cazelles and S. Hales. Infectious Diseases, Climate Influences, and Nonstationarity. *PLOS Med*, 3(8):e328, Aug. 2006. ISSN 1549-1676. doi: 10.1371/journal.pmed.0030328. URL http://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.0030328.

B. Cazelles, M. Chavez, A. J. McMichael, and S. Hales. Nonstationary Influence of El Niño on the Synchronous Dengue Epidemics in Thailand. *PLoS Medicine*, 2(4), Apr. 2005. ISSN 1549-1277. doi: 10.1371/journal.pmed.0020106. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1087219/.

B. Cazelles, C. Champagne, and J. Dureau. Accounting for non-stationarity in epidemiology by embedding time-varying parameters in stochastic models. *PLOS Computational Biology*, 14(8):e1006211, Aug. 2018. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1006211. URL http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1006211.

G. Celeux, F. Forbes, C. P. Robert, and D. M. Titterington. Deviance information criteria for missing data models. *Bayesian Analysis*, 1(4):651–673, Dec. 2006. ISSN 1936-0975,

1931-6690. doi: 10.1214/06-BA122. URL https://projecteuclid.org/euclid.ba/1340370933.

C. Champagne, D. G. Salthouse, R. Paul, V.-M. Cao-Lormeau, B. Roche, and B. Cazelles. Structure in the variability of the basic reproductive number (R0) for Zika epidemics in the Pacific islands. *eLife*, 5:e19874, Nov. 2016. ISSN 2050-084X. doi: 10.7554/eLife.19874. URL https://elifesciences.org/articles/19874.

C. Champagne, R. Paul, S. Ly, V. Duong, R. Leang, and B. Cazelles. Dengue modeling in rural Cambodia: Statistical performance versus epidemiological relevance. *Epidemics*, Aug. 2018. ISSN 1755-4365. doi: 10.1016/j.epidem.2018.08.004. URL http://www.sciencedirect.com/science/article/pii/S1755436517301706.

M. Chan and M. A. Johansson. The incubation periods of Dengue viruses. *PloS One*, 7(11): e50972, 2012. ISSN 1932-6203. doi: 10.1371/journal.pone.0050972.

O. Chareonsook, H. M. Foy, A. Teeraratkul, and N. Silarug. Changing epidemiology of dengue hemorrhagic fever in Thailand. *Epidemiology and Infection*, 122(1):161–166, Feb. 1999. ISSN 0950-2688.

V. Charu, S. Zeger, J. Gog, O. N. Bjørnstad, S. Kissler, L. Simonsen, B. T. Grenfell, and C. Viboud. Human mobility and the spatial transmission of influenza in the United States. *PLOS Computational Biology*, 13(2):e1005382, Feb. 2017. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1005382. URL http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005382.

C. Chen, N. S. Chong, and R. Smith. A Filippov model describing the effects of media coverage and quarantine on the spread of human influenza. *Mathematical Biosciences*, 296:98–112, Feb. 2018. ISSN 00255564. doi: 10.1016/j.mbs.2017.12.002. URL http://linkinghub.elsevier.com/retrieve/pii/S0025556417305126.

Q. Cheng, Q. Jing, R. C. Spear, J. M. Marshall, Z. Yang, and P. Gong. The interplay of climate, intervention and imported cases as determinants of the 2014 dengue outbreak in Guangzhou. *PLOS Neglected Tropical Diseases*, 11(6):e0005701, June 2017. ISSN 1935-2735. doi: 10.1371/journal.pntd.0005701. URL http://journals.plos.org/plosntds/article?id=10.1371/journal.pntd.0005701.

E. Chikaki and H. Ishikawa. A dengue transmission model in Thailand considering sequential infections with all four serotypes. *Journal of Infection in Developing Countries*, 3(9):711–722, Oct. 2009. ISSN 1972-2680.

N. Chopin, P. E. Jacob, and O. Papaspiliopoulos. SMC2: an efficient algorithm for sequential analysis of state space models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(3):397–426, June 2013. ISSN 1467-9868. doi: 10.1111/j.1467-9868.2012.01046.x. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2012.01046.x.

T. Chouin-Carneiro, A. Vega-Rua, M. Vazeille, A. Yebakima, R. Girod, D. Goindin, M. Dupont-Rouzeyrol, R. Lourenço-de Oliveira, and A.-B. Failloux. Differential Susceptibilities of Aedes aegypti and Aedes albopictus from the Americas to Zika Virus. *PLoS neglected tropical diseases*, 10(3):e0004543, Mar. 2016. ISSN 1935-2735. doi: 10.1371/journal.pntd.0004543.

G. Chowell, C. A. Torre, C. Munayco-Escate, L. Suárez-Ognio, R. López-Cruz, J. M. Hyman, and C. Castillo-Chavez. Spatial and temporal dynamics of dengue fever in Peru: 1994-2006. *Epidemiology and Infection*, 136(12):1667–1677, Dec. 2008. ISSN 0950-2688. doi: 10.1017/S0950268808000290.

CHSP. Centre d'hygiène et de salubrité publique. Surveillance et Veille Sanitaire en Polynesie Française. Technical report, 2014. URL http://www.hygiene-publique.gov.pf/IMG/pdf/bulletin_surv_pf_sem_42-_2014.pdf.

H. Clapham, D. A. T. Cummings, A. Nisalak, S. Kalayanarooj, B. Thaisomboonsuk, C. Klungthong, S. Fernandez, A. Srikiatkhachorn, L. R. Macareo, J. Lessler, J. Reiser, and I.-K. Yoon. Epidemiology of Infant Dengue Cases Illuminates Serotype-Specificity in the Interaction between Immunity and Disease, and Changes in Transmission Dynamics. *PLOS Neglected Tropical Diseases*, 9(12):e0004262, Dec. 2015. ISSN 1935-2735. doi: 10.1371/journal.pntd.0004262. URL http://dx.plos.org/10.1371/journal.pntd.0004262.

L. Coudeville and G. P. Garnett. Transmission Dynamics of the Four Dengue Serotypes in Southern Vietnam and the Potential Impact of Vaccination. *PLOS ONE*, 7(12):e51244, Dec. 2012. ISSN 1932-6203. doi: 10.1371/journal.pone.0051244. URL http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0051244.

K. Cowles and B. P. Carlin. Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review. *Journal of the American Statistical Association*, 91, July 1995. doi: 10.1080/01621459.1996.10476956.

H. Q. Cuong, N. T. Vu, B. Cazelles, M. F. Boni, K. T. Thai, M. A. Rabaa, L. C. Quang, C. P. Simmons, T. N. Huu, and K. L. Anders. Spatiotemporal Dynamics of Dengue Epidemics, Southern Vietnam. *Emerging Infectious Diseases*, 19(6):945–953, June 2013. ISSN 1080-6040. doi: 10.3201/eid1906.121323. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3713821/.

DASS. Direction des Affaires Sanitaires et Sociales. Situation sanitaire en Nouvelle Calédonie. Les arboviroses : dengue, chikungunya, zika. Technical report, 2014. URL http://www.dass.gouv.nc/portal/page/portal/dass/librairie/fichiers/32142267.PDF.

data.rio. Índice de Desenvolvimento Social (IDS) por Áreas de Planejamento (AP), Regiões de Planejamento (RP), Regiões Administrativas (RA), Bairros e Favelas do Município do Rio de Janeiro - Tabela 2248 - Fontes: dados - IBGE.Censo 2010; calculos - IPP/DIG., 2017a. URL https://www.arcgis.com/sharing/rest/content/items/f7ec7718c6ab403e9171d898538ab798/info/metadata/metadata.xml?format=default&output=html.

data.rio. Número de empregados por atividade econômica segundo as Áreas de Planejamento (AP), Regiões Administrativas (RA) e Bairros no Município do Rio de Janeiro em 2000/2002-2013 - Tabela 2631 - Fonte: Ministério do Trabalho e Emprego - MTE, Relação Anual de Informações Sociais - RAIS., 2017b. URL https://www.arcgis.com/sharing/rest/content/items/08731fa6de334bd79b07572636f0299e/info/metadata/metadata.xml?format=default&output=html.

E. Daudens, S. Lastere, C. Hirschauer, V. M. Cao-Lormeau, R. Louette, C. Roche, N. Chee Ayee, N. Goffard, E. Vrousos, L. Renou, and A. Wiegandt. *Épidémiologie de la dengue et stratégies de lutte en Polynésie française, 2006-2008. Numéro thématique. Polynésie française : une situation particulière*. 2009. URL http://opac.invs.sante.fr/index.php?lvl=notice_display&id=916.

D. De Angelis, A. M. Presanis, P. J. Birrell, G. S. Tomba, and T. House. Four key challenges in infectious disease modelling using data from multiple sources. *Epidemics*, 10:83–87, Mar. 2015. ISSN 1755-4365. doi: 10.1016/j.epidem.2014.09.004. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4383805/.

T. S. De Simone, R. M. R. Nogueira, E. S. M. Araújo, F. R. Guimarães, F. B. Santos, H. G. Schatzmayr, R. V. Souza, G. Teixeira Filho, and M. P. Miagostovich. Dengue virus surveillance: the co-circulation of DENV-1, DENV-2 and DENV-3 in the State of Rio de Janeiro, Brazil. *Transactions of The Royal Society of Tropical Medicine and Hygiene*, 98(9):553–562, Sept. 2004. ISSN 0035-9203. doi: 10.1016/j.trstmh.2003.09.003. URL https://academic-oup-com.insb.bib.cnrs.fr/trstmh/article/98/9/553/1917650.

N. B. DeFelice, E. Little, S. R. Campbell, and J. Shaman. Ensemble forecast of human West Nile virus cases and mosquito infection rates. *Nature Communications*, 8:14592, Feb. 2017. ISSN 2041-1723. doi: 10.1038/ncomms14592. URL https://www.nature.com/articles/ncomms14592.

W. Dejnirattisai, P. Supasa, W. Wongwiwat, A. Rouvinski, G. Barba-Spaeth, T. Duangchinda, A. Sakuntabhai, V.-M. Cao-Lormeau, P. Malasit, F. A. Rey, J. Mongkolsapaya, and G. R. Screaton. Dengue virus sero-cross-reactivity drives antibody-dependent enhancement of infection with zika virus. *Nature Immunology*, 17(9):1102–1108, Sept. 2016. ISSN 1529-2916. doi: 10.1038/ni.3515. URL https://www.nature.com/articles/ni.3515.

E. Delisle, C. Rousseau, B. Broche, I. Leparc-Goffart, G. L'Ambert, A. Cochet, C. Prat, V. Foulongne, J. B. Ferre, O. Catelinois, O. Flusin, E. Tchernonog, I. E. Moussion, A. Wiegandt, A. Septfons, A. Mendy, M. B. Moyano, L. Laporte, J. Maurel, F. Jourdain, J. Reynes, M. C. Paty, and F. Golliot. Chikungunya outbreak in Montpellier, France, September to October 2014. *Euro Surveillance: Bulletin Europeen Sur Les Maladies Transmissibles = European Communicable Disease Bulletin*, 20(17), Apr. 2015. ISSN 1560-7917.

O. Diekmann, J. a. P. Heesterbeek, and M. G. Roberts. The construction of next-generation matrices for compartmental epidemic models. *Journal of The Royal Society Interface*, 7(47):873–885, June 2010. ISSN 1742-5689, 1742-5662. doi: 10.1098/rsif.2009.0386. URL http://rsif.royalsocietypublishing.org/content/7/47/873.

K. Dietz. The estimation of the basic reproduction number for infectious diseases. *Statistical methods in medical research*, 2(1):23–41, 1993. URL http://smm.sagepub.com/content/2/1/23.short.

A. Doucet and A. M. Johansen. A tutorial on particle filtering and smoothing: fifteen years later, 2011.

A. Doucet, N. Freitas, and N. Gordon, editors. *Sequential Monte Carlo Methods in Practice*. Springer New York, New York, NY, 2001a. ISBN 978-1-4419-2887-0 978-1-4757-3437-9. URL http://link.springer.com/10.1007/978-1-4757-3437-9.

A. Doucet, N. d. Freitas, and N. Gordon. An Introduction to Sequential Monte Carlo Methods. In *Sequential Monte Carlo Methods in Practice*, Statistics for Engineering and Information Science, pages 3–14. Springer, New York, NY, 2001b. ISBN 978-1-4419-2887-0 978-1-4757-3437-9. doi: 10.1007/978-1-4757-3437-9_1. URL https://link.springer.com/chapter/10.1007/978-1-4757-3437-9_1.

M. R. Duffy, T.-H. Chen, W. T. Hancock, A. M. Powers, J. L. Kool, R. S. Lanciotti, M. Pretrick, M. Marfel, S. Holzbauer, C. Dubray, and others. Zika virus outbreak on Yap Island, federated states of Micronesia. *New England Journal of Medicine*, 360(24):2536–2543, 2009. URL http://www.nejm.org/doi/full/10.1056/NEJMoa0805715.

V. Duong, L. Lambrechts, R. E. Paul, S. Ly, R. S. Lay, K. C. Long, R. Huy, A. Tarantola, T. W. Scott, A. Sakuntabhai, and P. Buchy. Asymptomatic humans transmit dengue virus to mosquitoes. *Proceedings of the National Academy of Sciences*, Nov. 2015. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1508114112. URL http://www.pnas.org/cgi/doi/10.1073/pnas.1508114112.

M. Dupont-Rouzeyrol, O. O'Connor, E. Calvez, M. Daurès, M. John, J.-P. Grangeon, and A.-C. Gourinat. Co-infection with Zika and Dengue Viruses in 2 Patients, New Caledonia, 2014. *Emerging Infectious Diseases*, 21(2):381–382, Feb. 2015. ISSN 1080-6040. doi: 10.3201/eid2102.141553. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4313662/.

A. P. Durbin. Dengue Antibody and Zika: Friend or Foe? *Trends in Immunology*, 37(10):635–636, Oct. 2016. ISSN 1471-4906, 1471-4981. doi: 10.1016/j.it.2016.08.006. URL http://www.cell.com/trends/immunology/abstract/S1471-4906(16)30096-5.

J. Dureau, S. Ballesteros, and T. Bogich. ssm: Inference for State Space Models like playing with duplo blocks, Nov. 2013a. URL https://github.com/JDureau/ssm. original-date: 2014-08-29T13:15:10Z.

J. Dureau, K. Kalogeropoulos, and M. Baguelin. Capturing the time-varying drivers of an epidemic using stochastic dynamical systems. *Biostatistics*, 14(3):541–555, July 2013b. ISSN 1465-4644. doi: 10.1093/biostatistics/kxs052. URL https://academic.oup.com/biostatistics/article/14/3/541/259859.

C. Dye and G. Hasibeder. Population dynamics of mosquito-borne disease: effects of flies which bite some people more frequently than others. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 80(1):69–77, 1986. URL http://trstmh.oxfordjournals.org/content/80/1/69.short.

C. Dye and B. Williams. Nonlinearities in the Dynamics of Indirectly-Transmitted Infections (or, does having a Vector make a Difference?). In *Ecology of Infectious Diseases in Natural Populations*, pages 260–279. Grenfell, B.T. and Dobson, A.P., cambridge university press edition, 1995. URL /core/books/ecology-of-infectious-diseases-in-natural-populations/nonlinearities-in-the-dynamics-of-indirectly-transmitted-infections-or-does-h1E571DC1187028C2262A1D6EFE5A73A4.

O. Dyer. Philippines halts dengue immunisation campaign owing to safety risk. *BMJ*, 359: j5759, Dec. 2017. ISSN 0959-8138, 1756-1833. doi: 10.1136/bmj.j5759. URL https://www.bmj.com/content/359/bmj.j5759.

M. C. Eisenberg, S. L. Robertson, and J. H. Tien. Identifiability and estimation of multiple transmission pathways in cholera and waterborne disease. *Journal of Theoretical Biology*, 324: 84–102, May 2013. ISSN 1095-8541. doi: 10.1016/j.jtbi.2012.12.021.

S. P. Ellner, B. A. Bailey, G. V. Bobashev, A. R. Gallant, B. T. Grenfell, and D. W. Nychka. Noise and Nonlinearity in Measles Epidemics: Combining Mechanistic and Statistical Approaches to Population Modeling. *The American Naturalist*, 151(5):425–440, 1998. ISSN 0003-0147. doi: 10.1086/286130. URL http://www.jstor.org/stable/10.1086/286130.

R. A. Erickson, S. M. Presley, L. J. Allen, K. R. Long, and S. B. Cox. A dengue model with a dynamic Aedes albopictus vector population. *Ecological Modelling*, 221(24):2899–2908, Dec. 2010. ISSN 03043800. doi: 10.1016/j.ecolmodel.2010.08.036. URL http://linkinghub.elsevier.com/retrieve/pii/S0304380010004497.

N. Evans, M. Chapman, M. Chappell, and K. Godfrey. The structural identifiability of a general epidemic (SIR) model with seasonal forcing. *IFAC Proceedings Volumes*, 35(1):109–114, 2002. ISSN 14746670. doi: 10.3182/20020721-6-ES-1901.01327. URL http://linkinghub.elsevier.com/retrieve/pii/S1474667015397482.

G. Evensen. The Ensemble Kalman Filter: theoretical formulation and practical implementation. *Ocean Dynamics*, 53(4):343–367, Nov. 2003. ISSN 1616-7341, 1616-7228. doi: 10.1007/s10236-003-0036-9. URL http://link.springer.com/10.1007/s10236-003-0036-9.

R. C. G. Fares, K. P. R. Souza, G. Añez, and M. Rios. Epidemiological Scenario of Dengue in Brazil. *BioMed Research International*, 2015:e321873, Aug. 2015. ISSN 2314-6133. doi: 10.1155/2015/321873. URL https://www.hindawi.com/journals/bmri/2015/321873/abs/.

N. R. Faria, R. d. S. d. S. Azevedo, M. U. G. Kraemer, R. Souza, M. S. Cunha, S. C. Hill, J. Thézé, M. B. Bonsall, T. A. Bowden, I. Rissanen, I. M. Rocco, J. S. Nogueira, A. Y. Maeda, F. G. d. S. Vasami, F. L. d. L. Macedo, A. Suzuki, S. G. Rodrigues, A. C. R. Cruz, B. T. Nunes, D. B. d. A. Medeiros, D. S. G. Rodrigues, A. L. N. Queiroz, E. V. P. da Silva, D. F. Henriques, E. S. T. da Rosa, C. S. de Oliveira, L. C. Martins, H. B. Vasconcelos, L. M. N. Casseb, D. d. B. Simith, J. P. Messina, L. Abade, J. Lourenço, L. C. J. Alcantara, M. M. de Lima, M. Giovanetti, S. I. Hay, R. S. de Oliveira, P. d. S. Lemos, L. F. de Oliveira, C. P. S. de Lima, S. P. da Silva, J. M. de Vasconcelos, L. Franco, J. F. Cardoso, J. L. d. S. G. Vianez-Júnior, D. Mir, G. Bello, E. Delatorre, K. Khan, M. Creatore, G. E. Coelho, W. K. de Oliveira, R. Tesh, O. G. Pybus, M. R. T. Nunes, and P. F. C. Vasconcelos. Zika virus in the Americas: Early epidemiological and genetic findings. *Science (New York, N.Y.)*, 352(6283):345–349, Apr. 2016. ISSN 1095-9203. doi: 10.1126/science.aaf5036.

M. Fasiolo, N. Pya, and S. N. Wood. A Comparison of Inferential Methods for Highly Nonlinear State Space Models in Ecology and Epidemiology. *Statistical Science*, 31(1):96–118, Feb. 2016. ISSN 0883-4237, 2168-8745. doi: 10.1214/15-STS534. URL https://projecteuclid.org/euclid.ss/1455115916.

C. Favier, D. Schmit, C. D. Müller-Graf, B. Cazelles, N. Degallier, B. Mondet, and M. A. Dubois. Influence of spatial heterogeneity on an emerging infectious disease: the case of dengue epidemics. *Proceedings of the Royal Society B: Biological Sciences*, 272(1568):

1171–1177, June 2005. ISSN 0962-8452. doi: 10.1098/rspb.2004.3020. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1559809/.

C. Favier, N. Degallier, M. G. Rosa-Freitas, J. P. Boulanger, J. R. Costa Lima, J. F. Luitgards-Moura, C. E. Menkes, B. Mondet, C. Oliveira, E. T. S. Weimann, and P. Tsouris. Early determination of the reproductive number for vector-borne diseases: the case of dengue in Brazil. *Tropical Medicine and International Health*, 11(3):332–340, Mar. 2006. ISSN 1360-2276, 1365-3156. doi: 10.1111/j.1365-3156.2006.01560.x. URL http://doi.wiley.com/10.1111/j.1365-3156.2006.01560.x.

N. Ferguson, R. Anderson, and S. Gupta. The effect of antibody-dependent enhancement on the transmission dynamics and persistence of multiple-strain pathogens. *Proceedings of the National Academy of Sciences of the United States of America*, 96(2):790–794, Jan. 1999. ISSN 0027-8424. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC15215/.

N. M. Ferguson, Z. M. Cucunubá, I. Dorigatti, G. L. Nedjati-Gilani, C. A. Donnelly, M.-G. Basanez, P. Nouvellet, and J. Lessler. Countering Zika in Latin America. *Science*, page aag0219, July 2016. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aag0219. URL http://science.sciencemag.org/content/early/2016/07/13/science.aag0219.

E. Fernandez, W. Dejnirattisai, B. Cao, S. M. Scheaffer, P. Supasa, W. Wongwiwat, P. Esakky, A. Drury, J. Mongkolsapaya, K. H. Moley, I. U. Mysorekar, G. R. Screaton, and M. S. Diamond. Human antibodies to the dengue virus E-dimer epitope have therapeutic activity against Zika virus infection. *Nature Immunology*, 18(11):1261–1269, Nov. 2017. ISSN 1529-2916. doi: 10.1038/ni.3849.

B. F. Finkenstädt and B. T. Grenfell. Time series modelling of childhood diseases: a dynamical systems approach. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 49 (2):187–205, Jan. 2002. ISSN 1467-9876. doi: 10.1111/1467-9876.00187. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/1467-9876.00187.

W. E. Fitzgibbon, J. J. Morgan, and G. F. Webb. An outbreak vector-host epidemic model with spatial structure: the 2015-2016 Zika outbreak in Rio De Janeiro. *Theoretical Biology & Medical Modelling*, 14, Mar. 2017. ISSN 1742-4682. doi: 10.1186/s12976-017-0051-z. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5368945/.

S. Flasche, M. Jit, I. Rodríguez-Barraquer, L. Coudeville, M. Recker, K. Koelle, G. Milne, T. J. Hladish, T. A. Perkins, D. A. T. Cummings, I. Dorigatti, D. J. Laydon, G. Espana, J. Kelso, I. Longini, J. Lourenco, C. A. B. Pearson, R. C. Reiner, L. Mier-y Terán-Romero, K. Vannice, and N. Ferguson. The Long-Term Safety, Public Health Impact, and Cost-Effectiveness of Routine Vaccination with a Recombinant, Live-Attenuated Dengue Vaccine (Dengvaxia): A Model Comparison Study. *PLOS Medicine*, 13(11):e1002181, Nov. 2016. ISSN 1549-1676. doi: 10.1371/journal.pmed.1002181. URL http://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1002181.

S. Funk, M. Salathé, and V. A. A. Jansen. Modelling the influence of human behaviour on the spread of infectious diseases: a review. *Journal of The Royal Society Interface*, 7(50): 1247–1256, Sept. 2010. ISSN 1742-5689, 1742-5662. doi: 10.1098/rsif.2010.0142. URL http://rsif.royalsocietypublishing.org/content/7/50/1247.

S. Funk, S. Bansal, C. T. Bauch, K. T. D. Eames, W. J. Edmunds, A. P. Galvani, and P. Klepac. Nine challenges in incorporating the dynamics of behaviour in infectious diseases models. *Epidemics*, 10:21–25, Mar. 2015. ISSN 1755-4365. doi: 10.1016/j.epidem.2014.09.005. URL http://www.sciencedirect.com/science/article/pii/S1755436514000541.

S. Funk, A. J. Kucharski, A. Camacho, R. M. Eggo, L. Yakob, and W. J. Edmunds. Comparative analysis of dengue and Zika outbreaks reveals differences by setting and virus. Technical Report biorxiv;043265v3, Mar. 2016. URL http://biorxiv.org/lookup/doi/10.1101/043265.

D. Gao, Y. Lou, D. He, T. C. Porco, Y. Kuang, G. Chowell, and S. Ruan. Prevention and Control of Zika as a Mosquito-Borne and Sexually Transmitted Disease: A Mathematical Modeling Analysis. *Scientific Reports*, 6:28070, June 2016. ISSN 2045-2322. doi: 10.1038/srep28070. URL http://www.nature.com/articles/srep28070.

A. Gelman. How to think about "identifiability" in Bayesian inference?, Feb. 2014. URL http://andrewgelman.com/2014/02/12/think-identifiability-bayesian-inference/.

A. Gelman and D. B. Rubin. Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, 7(4):457–472, 1992. ISSN 0883-4237. URL http://www.jstor.org/stable/2246093.

A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian Data Analysis, Third Edition*. Chapman and Hall/CRC, Boca Raton, 3 edition edition, Nov. 2013. ISBN 978-1-4398-4095-5.

J. George, W. G. Valiant, M. J. Mattapallil, M. Walker, Y.-J. S. Huang, D. L. Vanlandingham, J. Misamore, J. Greenhouse, D. E. Weiss, D. Verthelyi, S. Higgs, H. Andersen, M. G. Lewis, and J. J. Mattapallil. Prior Exposure to Zika Virus Significantly Enhances Peak Dengue-2 Viremia in Rhesus Macaques. *Scientific Reports*, 7(1):10498, Sept. 2017. ISSN 2045-2322. doi: 10.1038/s41598-017-10901-1. URL https://www.nature.com/articles/s41598-017-10901-1.

J. Geweke. Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments. In *In Bayesian Statistics*, pages 169–193. University Press, 1992.

G. J. Gibson and E. Renshaw. Estimating parameters in stochastic compartmental models using Markov chain methods. *Mathematical Medicine and Biology: A Journal of the IMA*, 15(1):19–40, Mar. 1998. ISSN 1477-8599. doi: 10.1093/imammb/15.1.19. URL https://academic.oup.com/imammb/article/15/1/19/683936.

G. J. Gibson, G. Streftaris, and D. Thong. Comparison and Assessment of Epidemic Models. *Statistical Science*, 33(1):19–33, Feb. 2018. ISSN 0883-4237, 2168-8745. doi: 10.1214/17-STS615. URL https://projecteuclid.org/euclid.ss/1517562022.

D. T. Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of Computational Physics*, 22(4):403–434, Dec. 1976. ISSN 0021-9991. doi: 10.1016/0021-9991(76)90041-3. URL http://www.sciencedirect.com/science/article/pii/0021999176900413.

D. T. Gillespie. Approximate accelerated stochastic simulation of chemically reacting systems. *The Journal of Chemical Physics*, 115(4):1716–1733, July 2001. ISSN 0021-9606. doi: 10.1063/1.1378322. URL https://aip.scitation.org/doi/abs/10.1063/1.1378322.

J. R. Gog, S. Ballesteros, C. Viboud, L. Simonsen, O. N. Bjornstad, J. Shaman, D. L. Chao, F. Khan, and B. T. Grenfell. Spatial Transmission of 2009 Pandemic Influenza in the US. *PLoS Comput Biol*, 10(6):e1003635, June 2014. doi: 10.1371/journal.pcbi.1003635. URL http://dx.doi.org/10.1371/journal.pcbi.1003635.

E. Gould, J. Pettersson, S. Higgs, R. Charrel, and X. de Lamballerie. Emerging arboviruses: Why today? *One Health*, 4:1–13, July 2017. ISSN 2352-7714. doi: 10.1016/j.onehlt.2017.06.001. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5501887/.

R. R. Graham, M. Juffrie, R. Tan, C. G. Hayes, I. Laksono, C. Ma'roef, n. Erlin, n. Sutaryo, K. R. Porter, and S. B. Halstead. A prospective seroepidemiologic study on dengue in children four to nine years of age in Yogyakarta, Indonesia I. studies in 1995-1996. *The American Journal of Tropical Medicine and Hygiene*, 61(3):412–419, Sept. 1999. ISSN 0002-9637.

L. Grange, E. Simon-Loriere, A. Sakuntabhai, L. Gresh, R. Paul, and E. Harris. Epidemiological Risk Factors Associated with High Global Frequency of Inapparent Dengue Virus Infections. *Frontiers in Immunology*, 5, June 2014. ISSN 1664-3224. doi: 10.3389/fimmu.2014.00280. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4052743/.

D. Guha-Sapir and B. Schimmer. Dengue fever: new paradigms for a changing epidemiology. *Emerging Themes in Epidemiology*, 2:1, 2005. ISSN 1742-7622. doi: 10.1186/1742-7622-2-1. URL http://dx.doi.org/10.1186/1742-7622-2-1.

M. G. Guzman, S. B. Halstead, H. Artsob, P. Buchy, J. Farrar, D. J. Gubler, E. Hunsperger, A. Kroeger, H. S. Margolis, E. Martínez, M. B. Nathan, J. L. Pelegrino, C. Simmons, S. Yoksan, and R. W. Peeling. Dengue: a continuing global threat. *Nature Reviews Microbiology*, 8:S7–S16, Jan. 2010. ISSN 1740-1526. doi: 10.1038/nrmicro2460. URL http://www.nature.com/nrmicro/journal/v8/n12_supp/full/nrmicro2460.html.

H. Haario, E. Saksman, and J. Tamminen. An Adaptive Metropolis Algorithm. *Bernoulli*, 7(2): 223, Apr. 2001. ISSN 13507265. doi: 10.2307/3318737. URL https://www.jstor.org/stable/3318737?origin=crossref.

S. B. Halstead. Dengue. *Lancet (London, England)*, 370(9599):1644–1652, Nov. 2007. ISSN 1474-547X. doi: 10.1016/S0140-6736(07)61687-0.

S. B. Halstead. Dengue Virus–Mosquito Interactions. *Annual Review of Entomology*, 53(1): 273–291, 2008. doi: 10.1146/annurev.ento.53.103106.093326. URL http://dx.doi.org/10.1146/annurev.ento.53.103106.093326.

S. B. Halstead. Biologic Evidence Required for Zika Disease Enhancement by Dengue Antibodies. *Emerging Infectious Diseases*, 23(4):569–573, 2017. ISSN 1080-6059. doi: 10.3201/eid2304.161879.

S. N. Hammond, A. Balmaseda, L. Pérez, Y. Tellez, S. I. Saborío, J. C. Mercado, E. Videa, Y. Rodriguez, M. A. Pérez, R. Cuadra, S. Solano, J. Rocha, W. Idiaquez, A. Gonzalez, and E. Harris. Differences in dengue severity in infants, children, and adults in a 3-year hospital-based study in Nicaragua. *The American Journal of Tropical Medicine and Hygiene*, 73(6):1063–1070, Dec. 2005. ISSN 0002-9637.

L. C. Harrington, A. Fleisher, D. Ruiz-Moreno, F. Vermeylen, C. V. Wa, R. L. Poulson, J. D. Edman, J. M. Clark, J. W. Jones, S. Kitthawee, and T. W. Scott. Heterogeneous feeding patterns of the dengue vector, Aedes aegypti, on individual human hosts in rural Thailand. *PLoS neglected tropical diseases*, 8(8):e3048, Aug. 2014. ISSN 1935-2735. doi: 10.1371/journal.pntd.0003048.

I. Harris and P. Jones. CRU TS4.01: Climatic Research Unit (CRU) Time-Series (TS) version 4.01 of high-resolution gridded data of month-by-month variation in climate (Jan. 1901- Dec. 2016). *University of East Anglia Climatic Research Unit, Centre for Environmental Data Analysis*, Dec. 2017. doi: doi:10.5285/58a8802721c94c66ae45c3baa4d814d0. URL http://dx.doi.org/10.5285/58a8802721c94c66ae45c3baa4d814d0.

I. Harris, P. Jones, T. Osborn, and D. Lister. Updated high-resolution grids of monthly climatic observations—The CRU TS3.10 Dataset. *International Journal of Climatology*, 34:n/a–n/a, Mar. 2014. doi: 10.1002/joc.3711.

E. B. Hayes. Zika Virus Outside Africa. *Emerging Infectious Diseases*, 15(9):1347–1350, Sept. 2009. ISSN 1080-6040, 1080-6059. doi: 10.3201/eid1509.090442. URL http://wwwnc.cdc.gov/eid/article/15/9/09-0442_article.htm.

D. He, E. L. Ionides, and A. A. King. Plug-and-play inference for disease dynamics: measles in large and small populations as a case study. *Journal of The Royal Society Interface*, 7 (43):271–283, Feb. 2010. ISSN 1742-5689, 1742-5662. doi: 10.1098/rsif.2009.0151. URL http://rsif.royalsocietypublishing.org/content/7/43/271.

D. He, D. Gao, Y. Lou, S. Zhao, and S. Ruan. A comparison study of Zika virus outbreaks in French Polynesia, Colombia and the State of Bahia in Brazil. *Scientific Reports*, 7(1):273, Mar. 2017. ISSN 2045-2322. doi: 10.1038/s41598-017-00253-1. URL https://www.nature.com/articles/s41598-017-00253-1.

H. Heesterbeek, R. M. Anderson, V. Andreasen, S. Bansal, D. De Angelis, C. Dye, K. T. D. Eames, W. J. Edmunds, S. D. W. Frost, S. Funk, T. D. Hollingsworth, T. House, V. Isham, P. Klepac, J. Lessler, J. O. Lloyd-Smith, C. J. E. Metcalf, D. Mollison, L. Pellis, J. R. C. Pulliam, M. G. Roberts, C. Viboud, and Isaac Newton Institute IDD Collaboration. Modeling infectious disease dynamics in the complex landscape of global health. *Science*, 347(6227): aaa4339–aaa4339, Mar. 2015. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aaa4339. URL http://www.sciencemag.org/cgi/doi/10.1126/science.aaa4339.

M. Heringer, T. M. A. Souza, M. d. R. Q. Lima, P. C. G. Nunes, N. R. d. C. Faria, F. de Bruycker-Nogueira, T. Chouin-Carneiro, R. M. R. Nogueira, and F. B. dos Santos. Dengue type 4 in Rio de Janeiro, Brazil: case characterization following its introduction in an endemic region. *BMC Infectious Diseases*, 17(1):410, June 2017. ISSN 1471-2334. doi: 10.1186/s12879-017-2488-4. URL https://doi.org/10.1186/s12879-017-2488-4.

K. S. Hickmann, G. Fairchild, R. Priedhorsky, N. Generous, J. M. Hyman, A. Deshpande, and S. Y. D. Valle. Forecasting the 2013–2014 Influenza Season Using Wikipedia. *PLOS Computational Biology*, 11(5):e1004239, May 2015. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1004239. URL http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004239.

N. A. Honório, R. M. R. Nogueira, C. T. Codeço, M. S. Carvalho, O. G. Cruz, M. d. A. F. M. Magalhães, J. M. G. d. Araújo, E. S. M. d. Araújo, M. Q. Gomes, L. S. Pinheiro, C. d. S. Pinel,

and R. Lourenço-de Oliveira. Spatial Evaluation and Modeling of Dengue Seroprevalence and Vector Density in Rio de Janeiro, Brazil. *PLOS Neglected Tropical Diseases*, 3(11):e545, Nov. 2009. ISSN 1935-2735. doi: 10.1371/journal.pntd.0000545. URL http://journals.plos.org/plosntds/article?id=10.1371/journal.pntd.0000545.

R. Huy, P. Buchy, A. Conan, C. Ngan, S. Ong, R. Ali, V. Duong, S. Yit, S. Ung, V. Te, N. Chroeung, N. C. Pheaktra, V. Uok, and S. Vong. National dengue surveillance in Cambodia 1980–2008: epidemiological and virological trends and the impact of vector control. *Bulletin of the World Health Organization*, 88(9):650–657, Sept. 2010. ISSN 0042-9686. doi: 10.2471/BLT.09.073908. URL http://www.who.int/bulletin/volumes/88/9/09-073908.pdf.

N. Imai, I. Dorigatti, S. Cauchemez, and N. M. Ferguson. Estimating Dengue Transmission Intensity from Sero-Prevalence Surveys in Multiple Countries. *PLoS Neglected Tropical Diseases*, 9(4), Apr. 2015. ISSN 1935-2727. doi: 10.1371/journal.pntd.0003719. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4400108/.

N. Imai, I. Dorigatti, S. Cauchemez, and N. M. Ferguson. Estimating Dengue Transmission Intensity from Case-Notification Data from Multiple Countries. *PLOS Negl Trop Dis*, 10(7):e0004833, July 2016. ISSN 1935-2735. doi: 10.1371/journal.pntd.0004833. URL http://journals.plos.org/plosntds/article?id=10.1371/journal.pntd.0004833.

Insee. Populations légales dans les collectivités d'outre-mer et Mayotte - Populations légales au recensement de la population 2012 de Polynésie française, 2012. URL http://www.insee.fr/fr/themes/detail.asp?ref_id=populegalescom&page=recensement/populegalescom/polynesie.htm.

Insee. Populations légales dans les collectivités d'outre-mer et Mayotte - Populations légales au recensement de la population 2014 de Nouvelle-Calédonie, 2014. URL http://www.insee.fr/fr/themes/detail.asp?ref_id=populegalescom&page=recensement/populegalescom/nouvelle_caledonie.htm.

E. L. Ionides, C. Breto, and A. A. King. Inference for nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 103(49):18438–18443, 2006. URL http://www.pnas.org/content/103/49/18438.short.

E. L. Ionides, A. Bhadra, Y. Atchadé, and A. King. Iterated filtering. *The Annals of Statistics*, 39(3):1776–1802, June 2011. ISSN 0090-5364, 2168-8966. doi: 10.1214/11-AOS886. URL https://projecteuclid.org/euclid.aos/1311600283.

E. L. Ionides, D. Nguyen, Y. Atchadé, S. Stoev, and A. A. King. Inference for dynamic and latent variable models via iterated, perturbed Bayes maps. *Proceedings of the National Academy of Sciences*, 112(3):719–724, Jan. 2015. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1410597112. URL http://www.pnas.org/content/112/3/719.

ISPF. Institut de la Statistique de la Polynésie Française. Bilan démographique 2014. Technical report, Institut de la Statistique de la Polynésie Française, 2014.

P. E. Jacob, L. M. Murray, C. C. Holmes, and C. P. Robert. Better together? Statistical learning in models made of modules. *arXiv:1708.08719 [stat]*, Aug. 2017. URL http://arxiv.org/abs/1708.08719. arXiv: 1708.08719.

M. A. Johansson, J. Hombach, and D. A. Cummings. Models of the impact of dengue vaccines: a review of current research and potential approaches. *Vaccine*, 29(35):5860–5868, Aug. 2011. ISSN 0264-410X. doi: 10.1016/j.vaccine.2011.06.042. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4327892/.

T. Jombart, D. M. Aanensen, M. Baguelin, P. Birrell, S. Cauchemez, A. Camacho, C. Colijn, C. Collins, A. Cori, X. Didelot, C. Fraser, S. Frost, N. Hens, J. Hugues, M. Höhle, L. Opatowski, A. Rambaut, O. Ratmann, S. Soubeyrand, M. A. Suchard, J. Wallinga, R. Ypma, and N. Ferguson. OutbreakTools: A new platform for disease outbreak analysis using the R software. *Epidemics*, 7:28–34, June 2014a. ISSN 1755-4365. doi: 10.1016/j.epidem.2014.04.003. URL http://www.sciencedirect.com/science/article/pii/S1755436514000206.

T. Jombart, A. Cori, X. Didelot, S. Cauchemez, C. Fraser, and N. Ferguson. Bayesian Reconstruction of Disease Outbreaks by Combining Epidemiologic and Genomic Data. *PLoS Computational Biology*, 10(1):e1003457, Jan. 2014b. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1003457. URL http://dx.plos.org/10.1371/journal.pcbi.1003457.

K. E. Jones, N. G. Patel, M. A. Levy, A. Storeygard, D. Balk, J. L. Gittleman, and P. Daszak. Global trends in emerging infectious diseases. *Nature*, 451(7181):990–993, Feb. 2008. ISSN 1476-4687. doi: 10.1038/nature06536. URL https://www.nature.com/articles/nature06536.

J.-Y. Kang and J. Aldstadt. The Influence of Spatial Configuration of Residential Area and Vector Populations on Dengue Incidence Patterns in an Individual-Level Transmission Model. *International Journal of Environmental Research and Public Health*, 14(7):792, July 2017. doi: 10.3390/ijerph14070792. URL http://www.mdpi.com/1660-4601/14/7/792.

Y.-H. Kao and M. C. Eisenberg. Practical unidentifiability of a simple vector-borne disease model: implications for parameter estimation and intervention assessment. *Epidemics*, May 2018. ISSN 1755-4365. doi: 10.1016/j.epidem.2018.05.010. URL http://www.sciencedirect.com/science/article/pii/S1755436517301627.

S. Karl, N. Halder, J. K. Kelso, S. A. Ritchie, and G. J. Milne. A spatial simulation model for dengue virus infection in urban areas. *BMC Infectious Diseases*, 14:447, Aug. 2014. ISSN 1471-2334. doi: 10.1186/1471-2334-14-447. URL https://doi.org/10.1186/1471-2334-14-447.

R. E. Kass and A. E. Raftery. Bayes Factors. *Journal of the American Statistical Association*, 1995. URL https://amstat.tandfonline.com/doi/abs/10.1080/01621459.1995.10476572.

L. C. Katzelnick, L. Gresh, M. E. Halloran, J. C. Mercado, G. Kuan, A. Gordon, A. Balmaseda, and E. Harris. Antibody-dependent enhancement of severe dengue disease in humans. *Science*, page eaan6836, Nov. 2017. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aan6836. URL http://science.sciencemag.org/content/early/2017/11/01/science.aan6836.

A. B. Kawiecki and R. C. Christofferson. Zika Virus-Induced Antibody Response Enhances Dengue Virus Serotype 2 Replication In Vitro. *The Journal of Infectious Diseases*, 214(9):1357–1360, Nov. 2016. ISSN 1537-6613. doi: 10.1093/infdis/jiw377.

M. J. Keeling and P. Rohani. *Modelling infectious diseases in humans and animals*. Princeton University Press, 2008. URL https://docs.google.com/file/d/0BxtzprbWr-K_WXlPZHhMNmdITGM/edit?usp=drive_web&pli=1&usp=embed_facebook.

W. O. Kermack and A. G. McKendrick. A contribution to the mathematical theory of epidemics. *Proc. R. Soc. Lond. A*, 115(772):700–721, Aug. 1927. ISSN 0950-1207, 2053-9150. doi: 10.1098/rspa.1927.0118. URL http://rspa.royalsocietypublishing.org/content/115/772/700.

A. A. King, E. L. Ionides, M. Pascual, and M. J. Bouma. Inapparent infections and cholera dynamics. *Nature*, 454(7206):877–880, Aug. 2008. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature07084. URL http://www.nature.com/doifinder/10.1038/nature07084.

A. A. King, M. Domenech de Celles, F. M. G. Magpantay, and P. Rohani. Avoidable errors in the modelling of outbreaks of emerging pathogens, with special reference to Ebola. *Proceedings of the Royal Society B: Biological Sciences*, 282(1806):20150347–20150347, Apr. 2015. ISSN 0962-8452, 1471-2954. doi: 10.1098/rspb.2015.0347. URL http://rspb.royalsocietypublishing.org/cgi/doi/10.1098/rspb.2015.0347.

A. A. King, D. Nguyen, and E. L. Ionides. Statistical Inference for Partially Observed Markov Processes via the R Package pomp. *Journal of Statistical Software*, 2016. doi: 10.18637/jss.v069.i12. URL https://www.jstatsoft.org/article/view/v069i12.

J. Knape, N. Jonzén, and M. Sköld. On observation distributions for state space models of population survey data: Observation models for population data. *Journal of Animal Ecology*, 80(6):1269–1277, Nov. 2011. ISSN 00218790. doi: 10.1111/j.1365-2656.2011.01868.x. URL http://doi.wiley.com/10.1111/j.1365-2656.2011.01868.x.

J. Koopman. Modeling Infection Transmission. *Annual Review of Public Health*, 25 (1):303–326, Mar. 2004. ISSN 0163-7525. doi: 10.1146/annurev.publhealth.25.102802.124353. URL https://www.annualreviews.org/doi/10.1146/annurev.publhealth.25.102802.124353.

G. P. Kouri, M. G. Guzmán, J. R. Bravo, and C. Triana. Dengue haemorrhagic fever/dengue shock syndrome: lessons from the Cuban epidemic, 1981. *Bulletin of the World Health Organization*, 67(4):375–380, 1989. ISSN 0042-9686. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2491263/.

M. U. Kraemer, M. E. Sinka, K. A. Duda, A. Q. Mylne, F. M. Shearer, C. M. Barker, C. G. Moore, R. G. Carvalho, G. E. Coelho, W. V. Bortel, G. Hendrickx, F. Schaffner, I. R. Elyazar, H.-J. Teng, O. J. Brady, J. P. Messina, D. M. Pigott, T. W. Scott, D. L. Smith, G. W. Wint, N. Golding, and S. I. Hay. The global distribution of the arbovirus vectors Aedes aegypti and Ae. albopictus. *eLife*, 4:e08347, June 2015. ISSN 2050-084X. doi: 10.7554/eLife.08347. URL https://elifesciences.org/articles/08347.

M. U. G. Kraemer, D. Bisanzio, R. C. Reiner, R. Zakar, J. B. Hawkins, C. C. Freifeld, D. L. Smith, S. I. Hay, J. S. Brownstein, and T. A. Perkins. Inferences about spatiotemporal variation in dengue virus transmission are sensitive to assumptions about human mobility: a case study using geolocated tweets from Lahore, Pakistan. *EPJ Data Science*, 7(1), Dec. 2018. ISSN 2193-1127. doi: 10.1140/epjds/s13688-018-0144-x. URL https://epjdatascience.springeropen.com/articles/10.1140/epjds/s13688-018-0144-x.

A. J. Kucharski, S. Funk, R. M. Eggo, H.-P. Mallet, W. J. Edmunds, and E. J. Nilles. Transmission Dynamics of Zika Virus in Island Populations: A Modelling Analysis of the 2013-14 French Polynesia Outbreak. *PLOS Neglected Tropical Diseases*, 10(5):e0004726, May 2016. ISSN 1935-2735. doi: 10.1371/journal.pntd.0004726. URL http://dx.plos.org/10.1371/journal.pntd.0004726.

A. J. Kucharski, M. Kama, C. H. Watson, M. Aubry, S. Funk, A. D. Henderson, O. Brady, J. Vanhomwegen, J.-C. Manuguerra, C. L. Lau, J. Edmunds, J. Aaskov, E. J. Nilles, V.-M. Cao-Lormeau, S. Hue, and M. Hibberd. Using paired serology and surveillance data to quantify dengue transmission and control during a large outbreak in Fiji. *bioRxiv*, page 246116, Jan. 2018. doi: 10.1101/246116. URL https://www.biorxiv.org/content/early/2018/01/10/246116.

G. Kuno. Review of the factors modulating dengue transmission. *Epidemiologic Reviews*, 17(2): 321–335, 1995. ISSN 0193-936X.

L. Lambrechts, T. W. Scott, and D. J. Gubler. Consequences of the Expanding Global Distribution of Aedes albopictus for Dengue Virus Transmission. *PLOS Neglected Tropical Diseases*, 4(5):e646, May 2010. ISSN 1935-2735. doi: 10.1371/journal.pntd.0000646. URL http://journals.plos.org/plosntds/article?id=10.1371/journal.pntd.0000646.

L. Lambrechts, K. P. Paaijmans, T. Fansiri, L. B. Carrington, L. D. Kramer, M. B. Thomas, and T. W. Scott. Impact of daily temperature fluctuations on dengue virus transmission by Aedes aegypti. *Proceedings of the National Academy of Sciences*, 108(18):7460–7465, May 2011. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1101377108. URL http://www.pnas.org/content/108/18/7460.

K. Laneri, A. Bhadra, E. L. Ionides, M. Bouma, R. C. Dhiman, R. S. Yadav, and M. Pascual. Forcing Versus Feedback: Epidemic Malaria and Monsoon Rains in Northwest India. *PLoS Computational Biology*, 6(9):e1000898, Sept. 2010. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1000898. URL http://dx.plos.org/10.1371/journal.pcbi.1000898.

M. S. Y. Lau, G. Marion, G. Streftaris, and G. J. Gibson. New model diagnostics for spatio-temporal systems in epidemiology and ecology. *Journal of The Royal Society Interface*, 11(93): 20131093–20131093, Feb. 2014. ISSN 1742-5689, 1742-5662. doi: 10.1098/rsif.2013.1093. URL http://rsif.royalsocietypublishing.org/cgi/doi/10.1098/rsif.2013.1093.

M. L'Azou, A. Moureau, E. Sarti, J. Nealon, B. Zambrano, T. A. Wartel, L. Villar, M. R. Capeding, and R. L. Ochiai. Symptomatic Dengue in Children in 10 Asian and Latin American Countries. *New England Journal of Medicine*, 374(12):1155–1166, Mar. 2016. ISSN 0028-4793, 1533-4406. doi: 10.1056/NEJMoa1503877. URL http://www.nejm.org/doi/10.1056/NEJMoa1503877.

M. L'Azou, J. Assoukpa, K. Fanouillere, E. Plennevaux, M. Bonaparte, A. Bouckenooghe, C. Frago, F. Noriega, B. Zambrano, R. L. Ochiai, B. Guy, and N. Jackson. Dengue seroprevalence: data from the clinical development of a tetravalent dengue vaccine in 14 countries (2005-2014). *Transactions of the Royal Society of Tropical Medicine and Hygiene*, May 2018. ISSN 1878-3503. doi: 10.1093/trstmh/try037.

J. P. Ledermann, L. Guillaumot, L. Yug, S. C. Saweyog, M. Tided, P. Machieng, M. Pretrick, M. Marfel, A. Griggs, M. Bel, M. R. Duffy, W. T. Hancock, T. Ho-Chen, and A. M. Powers. Aedes hensilli as a Potential Vector of Chikungunya and Zika Viruses. *PLOS Negl Trop Dis*, 8(10):e3188, Oct. 2014. ISSN 1935-2735. doi: 10.1371/journal.pntd.0003188. URL http://journals.plos.org/plosntds/article?id=10.1371/journal.pntd.0003188.

E. C. Lee, M. R. Kelly, B. M. Ochocki, S. M. Akinwumi, K. E. S. Hamre, J. H. Tien, and M. C. Eisenberg. Model distinguishability and inference robustness in mechanisms of cholera transmission and loss of immunity. *Journal of Theoretical Biology*, 420:68–81, May 2017. ISSN 0022-5193. doi: 10.1016/j.jtbi.2017.01.032. URL http://www.sciencedirect.com/science/article/pii/S0022519317300322.

J. Legrand, R. Grais, P.-Y. Boelle, A. Valleron, and A. Flahault. Understanding the dynamics of Ebola epidemics. *Epidemiology and Infection*, 135(4):610–621, May 2007. ISSN 0950-2688. doi: 10.1017/S0950268806007217. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2870608/.

S. Lequime and L. Lambrechts. Vertical transmission of arboviruses in mosquitoes: A historical perspective. *Infection, Genetics and Evolution*, 28:681–690, Dec. 2014. ISSN 15671348. doi: 10.1016/j.meegid.2014.07.025. URL http://linkinghub.elsevier.com/retrieve/pii/S1567134814002779.

J. Lessler, C. T. Ott, A. C. Carcelen, J. M. Konikoff, J. Williamson, Q. Bi, L. M. Kucirka, D. A. Cummings, N. G. Reich, and L. H. Chaisson. Times to key events in Zika virus infection and implications for blood donation: a systematic review. *Bulletin of the World Health Organization*, 94(11):841–849, Nov. 2016. ISSN 0042-9686. doi: 10.2471/BLT.16.174540. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5096355/.

J. Liu-Helmersson, H. Stenlund, A. Wilder-Smith, and J. Rocklöv. Vectorial Capacity of Aedes aegypti: Effects of Temperature and Implications for Global Dengue Epidemic Potential. *PLoS ONE*, 9(3):e89783, Mar. 2014. ISSN 1932-6203. doi: 10.1371/journal.pone.0089783. URL http://dx.plos.org/10.1371/journal.pone.0089783.

A. L. Lloyd. Destabilization of epidemic models with the inclusion of realistic distributions of infectious periods. *Proceedings. Biological Sciences / The Royal Society*, 268(1470):985–993, May 2001. ISSN 0962-8452. doi: 10.1098/rspb.2001.1599.

A. L. Lloyd, J. Zhang, and A. M. Root. Stochasticity and heterogeneity in host–vector models. *Journal of The Royal Society Interface*, 4(16):851–863, Oct. 2007. ISSN 1742-5689, 1742-5662. doi: 10.1098/rsif.2007.1064. URL http://rsif.royalsocietypublishing.org/content/4/16/851.

W. P. London and J. A. Yorke. Recurrent outbreaks of measles, chickenpox and mumps. I. Seasonal variation in contact rates. *American Journal of Epidemiology*, 98(6):453–468, Dec. 1973. ISSN 0002-9262.

J. Lourenço and M. Recker. Viral and Epidemiological Determinants of the Invasion Dynamics of Novel Dengue Genotypes. *PLOS Neglected Tropical Diseases*, 4(11):e894, Nov. 2010. ISSN 1935-2735. doi: 10.1371/journal.pntd.0000894. URL http://journals.plos.org/plosntds/article?id=10.1371/journal.pntd.0000894.

J. Lourenço and M. Recker. Natural, Persistent Oscillations in a Spatial Multi-Strain Disease System with Application to Dengue. *PLoS Comput Biol*, 9(10):e1003308, Oct. 2013. doi: 10.1371/journal.pcbi.1003308. URL http://dx.doi.org/10.1371/journal.pcbi.1003308.

J. Lourenço and M. Recker. The 2012 Madeira Dengue Outbreak: Epidemiological Determinants and Future Epidemic Potential. *PLoS Neglected Tropical Diseases*, 8(8):e3083, Aug. 2014. ISSN 1935-2735. doi: 10.1371/journal.pntd.0003083. URL http://dx.plos.org/10.1371/journal.pntd.0003083.

J. Lourenço, M. M. d. Lima, N. R. Faria, A. Walker, M. U. Kraemer, C. J. Villabona-Arenas, B. Lambert, E. M. d. Cerqueira, O. G. Pybus, L. C. Alcantara, and M. Recker. Epidemiological and ecological determinants of Zika virus transmission in an urban setting. *eLife*, 6:e29820, Sept. 2017. ISSN 2050-084X. doi: 10.7554/eLife.29820. URL https://elifesciences.org/articles/29820.

R. Lowe, B. Cazelles, R. Paul, and X. Rodó. Quantifying the added value of climate information in a spatio-temporal dengue model. *Stochastic Environmental Research and Risk Assessment*, Mar. 2015. ISSN 1436-3240, 1436-3259. doi: 10.1007/s00477-015-1053-1. URL http://link.springer.com/10.1007/s00477-015-1053-1.

P. M. Luz, C. T. Codeço, E. Massad, and C. J. Struchiner. Uncertainties regarding dengue modeling in Rio de Janeiro, Brazil. *Memórias do Instituto Oswaldo Cruz*, 98(7):871–878, Oct. 2003. ISSN 0074-0276. doi: 10.1590/S0074-02762003000700002. URL http://www.scielo.br/scielo.php?script=sci_abstract&pid=S0074-02762003000700002&lng=en&nrm=iso&tlng=pt.

H. P. Mallet, A. L. Vial, and D. Musso. Bilan de l'épidémie à virus Zika en Polynésie française. *Bulletin d'Informations Sanitaires, Epidemiologiques et Statistiques*, (13), 2015. URL http://www.hygiene-publique.gov.pf/IMG/pdf/no13_-_mai_2015_-_zika.pdf.

M. P. Mammen Jr, C. Pimgate, C. J. Koenraadt, A. L. Rothman, J. Aldstadt, A. Nisalak, R. G. Jarman, J. W. Jones, A. Srikiatkhachorn, C. A. Ypil-Butac, and others. Spatial and temporal clustering of dengue virus transmission in Thai villages. *PLoS Med*, 5(11):e205, 2008. URL http://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.0050205.

J.-M. Marin, P. Pudlo, C. P. Robert, and R. J. Ryder. Approximate Bayesian computational methods. *Statistics and Computing*, 22(6):1167–1180, Nov. 2012. ISSN 0960-3174, 1573-1375. doi: 10.1007/s11222-011-9288-2. URL http://link.springer.com/10.1007/s11222-011-9288-2.

M. A. Martín-Acebes, J.-C. Saiz, and N. Jiménez de Oya. Antibody-Dependent Enhancement and Zika: Real Threat or Phantom Menace? *Frontiers in Cellular and Infection Microbiology*, 8, Feb. 2018. ISSN 2235-2988. doi: 10.3389/fcimb.2018.00044. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5818409/.

E. Massad, M. N. Burattini, F. A. B. Coutinho, and L. F. Lopez. Dengue and the risk of urban yellow fever reintroduction in Sao Paulo State, Brazil. *Revista de Saúde Pública*, 37(4):477–484, 2003. URL http://www.scielo.br/scielo.php?pid=S0034-89102003000400013&script=sci_arttext&tlng=es.

A. P. Masucci, J. Serras, A. Johansson, and M. Batty. Gravity versus radiation models: On the importance of scale and heterogeneity in commuting flows. *Physical Review E*, 88(2):022812, Aug. 2013. doi: 10.1103/PhysRevE.88.022812. URL https://link.aps.org/doi/10.1103/PhysRevE.88.022812.

T. McKinley, A. R. Cook, and R. Deardon. Inference in Epidemic Models without Likelihoods. *The International Journal of Biostatistics*, 5(1), 2009. ISSN 1557-4679. doi: 10.2202/1557-4679.1171. URL https://www.degruyter.com/view/j/ijb.2009.5.1/ijb.2009.5.1.1171/ijb.2009.5.1.1171.xml.

J. P. Messina, O. J. Brady, T. W. Scott, C. Zou, D. M. Pigott, K. A. Duda, S. Bhatt, L. Katzelnick, R. E. Howes, K. E. Battle, C. P. Simmons, and S. I. Hay. Global spread of dengue virus types: mapping the 70 year history. *Trends in Microbiology*, 22(3):138–146, Mar. 2014. ISSN 1878-4380. doi: 10.1016/j.tim.2013.12.011.

J. P. Messina, O. J. Brady, D. M. Pigott, N. Golding, M. U. G. Kraemer, T. W. Scott, G. R. W. Wint, D. L. Smith, and S. I. Hay. The many projected futures of dengue. *Nature Reviews. Microbiology*, 13(4):230–239, Apr. 2015. ISSN 1740-1534. doi: 10.1038/nrmicro3430.

C. J. E. Metcalf, J. Farrar, F. T. Cutts, N. E. Basta, A. L. Graham, J. Lessler, N. M. Ferguson, D. S. Burke, and B. T. Grenfell. Use of serological surveys to generate key insights into the changing global landscape of infectious disease. *The Lancet*, Apr. 2016. ISSN 01406736. doi: 10.1016/S0140-6736(16)30164-7. URL http://linkinghub.elsevier.com/retrieve/pii/S0140673616301647.

H. Miao, X. Xia, A. S. Perelson, and H. Wu. On identifiability of nonlinear ODE models and applications in viral dynamics. *SIAM review. Society for Industrial and Applied Mathematics*, 53(1):3–39, Jan. 2011. ISSN 0036-1445. doi: 10.1137/090757009. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3140286/.

H. L. Mills and S. Riley. The Spatial Resolution of Epidemic Peaks. *PLOS Computational Biology*, 10(4):e1003561, Apr. 2014. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1003561. URL http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003561.

D. Moulay, N. Verdière, and L. Denis-Vidal. Identifiablility of parameters in an epidemiologic model modeling the transmission of the Chikungunya. 2012. URL https://hal.archives-ouvertes.fr/hal-00699172.

N. A. Muhammad Azami, S. A. Salleh, H.-m. Neoh, S. Z. Syed Zakaria, and R. Jamal. Dengue epidemic in Malaysia: Not a predominantly urban disease anymore. *BMC Research Notes*, 4:216, 2011. ISSN 1756-0500. doi: 10.1186/1756-0500-4-216. URL http://dx.doi.org/10.1186/1756-0500-4-216.

L. M. Murray. Bayesian State-Space Modelling on High-Performance Hardware Using LibBi. *arXiv:1306.3277 [stat]*, June 2013. URL http://arxiv.org/abs/1306.3277. arXiv: 1306.3277.

D. Musso and D. J. Gubler. Zika Virus. *Clinical Microbiology Reviews*, 29(3):487–524, Jan. 2016. ISSN 0893-8512, 1098-6618. doi: 10.1128/CMR.00072-15. URL http://cmr.asm.org/content/29/3/487.

Y. Nagao and K. Koelle. Decreases in dengue transmission may act to increase the incidence of dengue hemorrhagic fever. *Proceedings of the National Academy of Sciences of the United States of America*, 105(6):2238–2243, Feb. 2008. ISSN 1091-6490. doi: 10.1073/pnas.0709029105.

S. Naish, P. Dale, J. S. Mackenzie, J. McBride, K. Mengersen, and S. Tong. Climate change and dengue: a critical and systematic review of quantitative modelling approaches. *BMC Infectious Diseases*, 14:167, Mar. 2014. ISSN 1471-2334. doi: 10.1186/1471-2334-14-167. URL https://doi.org/10.1186/1471-2334-14-167.

I. Nasell. Stochastic models of some endemic infections. *Mathematical Biosciences*, 2002.

P. P. C. National Institute of Statistics, Ministry of Planning. General Population Census of Cambodia 2008. National report on final census results, Aug. 2009.

E. M. Netto, A. Moreira-Soto, C. Pedroso, C. Höser, S. Funk, A. J. Kucharski, A. Rockstroh, B. M. Kümmerer, G. S. Sampaio, E. Luz, S. N. Vaz, J. P. Dias, F. A. Bastos, R. Cabral, T. Kistemann, S. Ulbert, X. d. Lamballerie, T. Jaenisch, O. J. Brady, C. Drosten, M. Sarno, C. Brites, and J. F. Drexler. High Zika Virus Seroprevalence in Salvador, Northeastern Brazil Limits the Potential for Further Outbreaks. *mBio*, 8(6):e01390–17, Dec. 2017. ISSN , 2150-7511. doi: 10.1128/mBio.01390-17. URL http://mbio.asm.org/content/8/6/e01390-17.

A. L. Nevai and E. Soewono. A model for the spatial transmission of dengue with daily movement between villages and a city. *Mathematical medicine and biology: a journal of the IMA*, 31(2): 150–178, June 2014. ISSN 1477-8602. doi: 10.1093/imammb/dqt002.

N. M. Nguyen, D. T. H. Kien, T. V. Tuan, N. T. H. Quyen, C. N. B. Tran, L. V. Thi, D. L. Thi, H. L. Nguyen, J. J. Farrar, E. C. Holmes, M. A. Rabaa, J. E. Bryant, T. T. Nguyen, H. T. C. Nguyen, L. T. H. Nguyen, M. P. Pham, H. T. Nguyen, T. T. H. Luong, B. Wills, C. V. V. Nguyen, M. Wolbers, and C. P. Simmons. Host and viral features of human dengue cases shape the population of infected and infectious Aedes aegypti mosquitoes. *Proceedings of the National Academy of Sciences*, 110(22):9072–9077, May 2013. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1303395110. URL http://www.pnas.org/content/110/22/9072.

A. Nisalak, H. E. Clapham, S. Kalayanarooj, C. Klungthong, B. Thaisomboonsuk, S. Fernandez, J. Reiser, A. Srikiatkhachorn, L. R. Macareo, J. T. Lessler, D. A. T. Cummings, and I.-K. Yoon. Forty Years of Dengue Surveillance at a Tertiary Pediatric Hospital in Bangkok, Thailand, 1973-2012. *The American Journal of Tropical Medicine and Hygiene*, 94(6):1342–1347, June 2016. ISSN 1476-1645. doi: 10.4269/ajtmh.15-0337.

H. Nishiura, R. Kinoshita, K. Mizumoto, Y. Yasuda, and K. Nah. Transmission potential of Zika virus infection in the South Pacific. *International Journal of Infectious Diseases*, 45: 95–97, Apr. 2016a. ISSN 1201-9712. doi: 10.1016/j.ijid.2016.02.017. URL http://www.ijidonline.com/article/S1201971216000370/abstract.

H. Nishiura, K. Mizumoto, W. E. Villamil-Gómez, and A. J. Rodríguez-Morales. Preliminary estimation of the basic reproduction number of Zika virus infection during Colombia epidemic, 2015–2016. *Travel Medicine and Infectious Disease*, 0(0), Apr. 2016b. ISSN 1477-8939, 1873-0442. doi: 10.1016/j.tmaid.2016.03.016. URL http://www.travelmedicinejournal.com/article/S1477893916300084/abstract.

R. M. Nogueira and A. L. Eppinghaus. Dengue virus type 4 arrives in the state of Rio de Janeiro: a challenge for epidemiological surveillance and control. *Memórias do Instituto Oswaldo Cruz*,

106(3):255–256, May 2011. ISSN 0074-0276. doi: 10.1590/S0074-02762011000300001. URL http://www.scielo.br/scielo.php?script=sci_abstract&pid=S0074-02762011000300001&lng=en&nrm=iso&tlng=en.

R. M. R. Nogueira, H. G. Schatzmayr, A. M. Bispo de Filippis, F. Barreto dos Santos, R. Venâncio da Cunha, J. O. Coelho, L. José de Souza, F. R. Guimarães, E. S. Machado de Araújo, T. S. De Simone, M. Baran, G. Teixeira, and M. P. Miagostovich. Dengue Virus Type 3, Brazil, 2002. *Emerging Infectious Diseases*, 11(9):1376–1381, Sept. 2005. ISSN 1080-6040. doi: 10.3201/eid1109.041043. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3310608/.

M. R. T. Nunes, N. R. Faria, J. M. de Vasconcelos, N. Golding, M. U. Kraemer, L. F. de Oliveira, R. d. S. d. S. Azevedo, D. E. A. da Silva, E. V. P. da Silva, S. P. da Silva, V. L. Carvalho, G. E. Coelho, A. C. R. Cruz, S. G. Rodrigues, J. L. da Silva Gonçalves Vianez, B. T. D. Nunes, J. F. Cardoso, R. B. Tesh, S. I. Hay, O. G. Pybus, and P. F. da Costa Vasconcelos. Emergence and potential for spread of Chikungunya virus in Brazil. *BMC Medicine*, 13(1):102, Apr. 2015. ISSN 1741-7015. doi: 10.1186/s12916-015-0348-x. URL https://doi.org/10.1186/s12916-015-0348-x.

P. D. O'Neill. Introduction and snapshot review: Relating infectious disease transmission models to data. *Statistics in Medicine*, 29(20):2069–2077, Sept. 2010. ISSN 1097-0258. doi: 10.1002/sim.3968. URL http://onlinelibrary.wiley.com/doi/10.1002/sim.3968/abstract.

P. D. O'Neill and G. O. Roberts. Bayesian Inference for Partially Observed Stochastic Epidemics. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 162(1):121–129, 1999. ISSN 0964-1998. URL http://www.jstor.org/stable/2680471.

J. Ong, G. Yap, C. Hapuarachchi, C. S. Tang, S. Liang, J. Rajarethinam, S. Tan, C. Koo, K. Zeng, J. Ycasas, L. Tay, C.-S. Chong, M. A. Abdul Razak, C. Lee, S. Chiang, Y.-L. Lai, and L.-C. Ng. Dengue review of 2016 and outlook for 2017. *Singapore Epidemiol News Bull*, 43 (2):47–56, Apr. 2017. URL https://www.moh.gov.sg/content/dam/moh_web/Publications/Epidemiological%20News%20Bulletin/ENB%20Quarterly_Apr%202017%20Vol%2043%20No%202%20(Final).pdf.

L. Opatowski, E. Varon, C. Dupont, L. Temime, S. v. d. Werf, L. Gutmann, P.-Y. Boëlle, L. Watier, and D. Guillemot. Assessing pneumococcal meningitis association with viral respiratory infections and antibiotics: insights from statistical and mathematical models. *Proceedings of the Royal Society of London B: Biological Sciences*, 280(1764):20130519, Aug. 2013. ISSN 0962-8452, 1471-2954. doi: 10.1098/rspb.2013.0519. URL http://rspb.royalsocietypublishing.org/content/280/1764/20130519.

L. Opatowski, M. Baguelin, and R. M. Eggo. Influenza interaction with cocirculating pathogens and its impact on surveillance, pathogenesis, and epidemic profile: A key role for mathematical modelling. *PLOS Pathogens*, 14(2):e1006770, Feb. 2018. ISSN 1553-7374. doi: 10.1371/journal.ppat.1006770. URL http://journals.plos.org/plospathogens/article?id=10.1371/journal.ppat.1006770.

K. O'Reilly, R. Lowe, J. Edmunds, P. Mayaud, A. Kucharski, R. M. Eggo, S. Funk, D. Bhatia, K. Khan, M. Kramer, A. Wilder-Smith, L. Rodrigues, P. Brasil, E. Massad, T. Jaenisch, S. Cauchemez, O. Brady, and L. Yakob. Projecting the end of the Zika virus epidemic

in Latin America: a modelling analysis. May 2018. doi: 10.1101/323915. URL http://biorxiv.org/lookup/doi/10.1101/323915.

J. Owen, D. J. Wilkinson, and C. S. Gillespie. Likelihood free inference for Markov processes: a comparison. *Statistical Applications in Genetics and Molecular Biology*, 14(2), Jan. 2015a. ISSN 1544-6115, 2194-6302. doi: 10.1515/sagmb-2014-0072. URL http://www.degruyter.com/view/j/sagmb.2015.14.issue-2/sagmb-2014-0072/sagmb-2014-0072.xml.

J. Owen, D. J. Wilkinson, and C. S. Gillespie. Scalable inference for Markov processes with intractable likelihoods. *Statistics and Computing*, 25(1):145–156, Jan. 2015b. ISSN 0960-3174, 1573-1375. doi: 10.1007/s11222-014-9524-7. URL https://link.springer.com/article/10.1007/s11222-014-9524-7.

A. Pandey, A. Mubayi, and J. Medlock. Comparing vector-host and SIR models for dengue transmission. *Mathematical Biosciences*, 246(2):252–259, Dec. 2013. ISSN 00255564. doi: 10.1016/j.mbs.2013.10.007. URL http://linkinghub.elsevier.com/retrieve/pii/S0025556413002435.

S. Pant. Information sensitivity functions to assess parameter information gain and identifiability of dynamical systems. *Journal of The Royal Society Interface*, 15(142):20170871, May 2018. ISSN 1742-5689, 1742-5662. doi: 10.1098/rsif.2017.0871. URL http://rsif.royalsocietypublishing.org/content/15/142/20170871.

P. Pantoja, E. X. Pérez-Guzmán, I. V. Rodríguez, L. J. White, O. González, C. Serrano, L. Giavedoni, V. Hodara, L. Cruz, T. Arana, M. I. Martínez, M. A. Hassert, J. D. Brien, A. K. Pinto, A. de Silva, and C. A. Sariol. Zika virus pathogenesis in rhesus macaques is unaffected by pre-existing immunity to dengue virus. *Nature Communications*, 8:15674, June 2017. ISSN 2041-1723. doi: 10.1038/ncomms15674.

M. Pascual and A. Dobson. Seasonal Patterns of Infectious Diseases. *PLOS Medicine*, 2(1):e5, Jan. 2005. ISSN 1549-1676. doi: 10.1371/journal.pmed.0020005. URL http://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.0020005.

S. R. L. Passos, B. D. Santos, M. A, J. Cerbino-Neto, S. N. Buonora, T. M. L. Souza, D. Oliveira, R. V. C, A. Vizzoni, G. Barbosa-Lima, Y. R. Vieira, M. Silva de Lima, and Y. H. M. Hökerberg. Detection of Zika Virus in April 2013 Patient Samples, Rio de Janeiro, Brazil. *Emerg Infect Dis*, 23, 2017. URL http://dx.doi.org/10.3201/eid2312.171375.

T. A. Perkins, R. C. Reiner, I. Rodriguez-Barraquer, D. L. Smith, T. W. Scott, and D. A. T. Cummings. A review of transmission models of dengue: a quantitative and qualitative analysis of model features. In D. J. Gubler, E. E. Ooi, S. Vasudevan, and J. Farrar, editors, *Dengue and dengue hemorrhagic fever*, pages 99–114. CABI, Wallingford, 2 edition, 2014. ISBN 978-1-84593-964-9. doi: 10.1079/9781845939649.0099. URL http://www.cabi.org/cabebooks/ebook/20143309285.

T. A. Perkins, V. A. Paz-Soldan, S. T. Stoddard, A. C. Morrison, B. M. Forshey, K. C. Long, E. S. Halsey, T. J. Kochel, J. P. Elder, U. Kitron, T. W. Scott, and G. M. Vazquez-Prokopec. Calling in sick: impacts of fever on intra-urban human mobility. *Proc. R. Soc. B*, 283(1834):20160390, July 2016a. ISSN 0962-8452, 1471-2954. doi: 10.1098/rspb.2016.0390. URL http://rspb.royalsocietypublishing.org/content/283/1834/20160390.

T. A. Perkins, A. S. Siraj, C. W. Ruktanonchai, M. U. G. Kraemer, and A. J. Tatem. Model-based projections of Zika virus infections in childbearing women in the Americas. *Nature Microbiology*, 1(9):16126, 2016b. ISSN 2058-5276. doi: 10.1038/nmicrobiol.2016.126.

L. R. Petersen, D. J. Jamieson, A. M. Powers, and M. A. Honein. Zika Virus. *New England Journal of Medicine*, 374(16):1552–1563, Apr. 2016. ISSN 0028-4793. doi: 10.1056/NEJMra1602113. URL http://dx.doi.org/10.1056/NEJMra1602113.

P. Pongsumpun and I. Tang. Transmission of dengue hemorrhagic fever in an age structured population. *Mathematical and Computer Modelling*, 37(9-10):949–961, May 2003. ISSN 08957177. doi: 10.1016/S0895-7177(03)00111-0. URL http://linkinghub.elsevier.com/retrieve/pii/S0895717703001110.

A. Prayitno, A.-F. Taurel, J. Nealon, H. I. Satari, M. R. Karyanti, R. Sekartini, S. Soedjatmiko, H. Gunardi, B. E. Medise, R. T. Sasmono, J. M. Simmerman, A. Bouckenooghe, and S. R. Hadinegoro. Dengue seroprevalence and force of primary infection in a representative population of urban dwelling Indonesian children. *PLOS Neglected Tropical Diseases*, 11(6):e0005621, June 2017. ISSN 1935-2735. doi: 10.1371/journal.pntd.0005621. URL http://journals.plos.org/plosntds/article?id=10.1371/journal.pntd.0005621.

PrefeituraRio. Cobertura Vegetal e Uso da Terra 2010, 2010. URL https://www.arcgis.com/home/item.html?id=cf1b862c8e014e8d9b02b9822f0ec3ff.

PrefeituraRio. Limite Bairro - Prefeitura da Cidade do Rio de Janeiro - IPP, 2016. URL https://portalgeo-pcrj.opendata.arcgis.com/datasets/8454eb0454b7424d89c61b67742286a1_15.

P. Pudlo, J.-M. Marin, A. Estoup, J.-M. Cornuet, M. Gautier, and C. P. Robert. Reliable ABC model choice via random forests. *Bioinformatics (Oxford, England)*, 32(6):859–866, 2016. ISSN 1367-4811. doi: 10.1093/bioinformatics/btv684.

G. Pujol, B. Iooss, and A. Janon. Package 'sensitivity', 2016. URL https://cran.r-project.org/web/packages/sensitivity/sensitivity.pdf.

H. Putter, S. H. Heisterkamp, J. M. A. Lange, and F. d. Wolf. A Bayesian approach to parameter estimation in HIV dynamical models. *Statistics in Medicine*, 21(15):2199–2214, Aug. 2002. ISSN 1097-0258. doi: 10.1002/sim.1211. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.1211.

A. E. Raftery, T. Gneiting, F. Balabdaoui, and M. Polakowski. Using Bayesian Model Averaging to Calibrate Forecast Ensembles. *Monthly Weather Review*, 133(5):1155–1174, May 2005. ISSN 0027-0644. doi: 10.1175/MWR2906.1. URL https://journals.ametsoc.org/doi/full/10.1175/MWR2906.1.

B. Randuineau. *Interactions between pathgogens*. PhD thesis, Université Pierre et Marie Curie-Paris VI, Paris, 2015.

D. A. Rasmussen, O. Ratmann, and K. Koelle. Inference for Nonlinear Epidemiological Models Using Genealogies and Time Series. *PLoS Computational Biology*, 7(8):e1002136, Aug. 2011. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1002136. URL http://dx.plos.org/10.1371/journal.pcbi.1002136.

D. A. Rasmussen, M. F. Boni, and K. Koelle. Reconciling Phylodynamics with Epidemiology: The Case of Dengue Virus in Southern Vietnam. *Molecular Biology and Evolution*, 31(2):258–271, Feb. 2014. ISSN 0737-4038. doi: 10.1093/molbev/mst203. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3907054/.

S. A. Rasmussen, D. J. Jamieson, M. A. Honein, and L. R. Petersen. Zika Virus and Birth Defects–Reviewing the Evidence for Causality. *The New England Journal of Medicine*, 374 (20):1981–1987, May 2016. ISSN 1533-4406. doi: 10.1056/NEJMsr1604338.

A. Raue, C. Kreutz, T. Maiwald, J. Bachmann, M. Schilling, U. Klingmüller, and J. Timmer. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics (Oxford, England)*, 25(15):1923–1929, Aug. 2009. ISSN 1367-4811. doi: 10.1093/bioinformatics/btp358.

L. Raynal, J.-M. Marin, P. Pudlo, M. Ribatet, C. P. Robert, and A. Estoup. ABC random forests for Bayesian parameter inference. *Peer Community in Evolutionary Biology*, page 100036, Nov. 2017. ISSN 2551668X. doi: 10.24072/pci.evolbiol.100036. URL http://arxiv.org/abs/1605.05537. arXiv: 1605.05537.

M. Recker, K. B. Blyuss, C. P. Simmons, T. T. Hien, B. Wills, J. Farrar, and S. Gupta. Immunological serotype interactions and their effect on the epidemiological pattern of dengue. *Proceedings. Biological Sciences*, 276(1667):2541–2548, July 2009. ISSN 0962-8452. doi: 10.1098/rspb.2009.0331.

N. G. Reich, S. Shrestha, A. A. King, P. Rohani, J. Lessler, S. Kalayanarooj, I.-K. Yoon, R. V. Gibbons, D. S. Burke, and D. A. T. Cummings. Interactions between serotypes of dengue highlight epidemiological impact of cross-immunity. *Journal of The Royal Society Interface*, 10(86):20130414–20130414, July 2013. ISSN 1742-5689, 1742-5662. doi: 10.1098/rsif.2013.0414. URL http://rsif.royalsocietypublishing.org/cgi/doi/10.1098/rsif.2013.0414.

N. G. Reich, S. A. Lauer, K. Sakrejda, S. Iamsirithaworn, S. Hinjoy, P. Suangtho, S. Suthachana, H. E. Clapham, H. Salje, D. A. T. Cummings, and J. Lessler. Challenges in Real-Time Prediction of Infectious Disease: A Case Study of Dengue in Thailand. *PLOS Neglected Tropical Diseases*, 10(6):e0004761, June 2016. ISSN 1935-2735. doi: 10.1371/journal.pntd.0004761. URL http://journals.plos.org/plosntds/article?id=10.1371/journal.pntd.0004761.

R. C. Reiner, T. A. Perkins, C. M. Barker, T. Niu, L. F. Chaves, A. M. Ellis, D. B. George, A. Le Menach, J. R. C. Pulliam, D. Bisanzio, C. Buckee, C. Chiyaka, D. A. T. Cummings, A. J. Garcia, M. L. Gatton, P. W. Gething, D. M. Hartley, G. Johnston, E. Y. Klein, E. Michael, S. W. Lindsay, A. L. Lloyd, D. M. Pigott, W. K. Reisen, N. Ruktanonchai, B. K. Singh, A. J. Tatem, U. Kitron, S. I. Hay, T. W. Scott, and D. L. Smith. A systematic review of mathematical models of mosquito-borne pathogen transmission: 1970-2010. *Journal of The Royal Society Interface*, 10(81):20120921–20120921, Feb. 2013. ISSN 1742-5689, 1742-5662. doi: 10.1098/rsif.2012.0921. URL http://rsif.royalsocietypublishing.org/cgi/doi/10.1098/rsif.2012.0921.

R. C. Reiner, S. T. Stoddard, and T. W. Scott. Socially structured human movement shapes dengue transmission despite the diffusive effect of mosquito dispersal. *Epidemics*, 6:30–36, Mar. 2014.

ISSN 1755-4365. doi: 10.1016/j.epidem.2013.12.003. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3971836/.

P. Reiter, S. Lathrop, M. Bunning, B. Biggerstaff, D. Singer, T. Tiwari, L. Baber, M. Amador, J. Thirion, J. Hayes, C. Seca, J. Mendez, B. Ramirez, J. Robinson, J. Rawlings, V. Vorndam, S. Waterman, D. Gubler, G. Clark, and E. Hayes. Texas Lifestyle Limits Transmission of Dengue Virus. *Emerging Infectious Diseases*, 9(1):86–89, Jan. 2003. ISSN 1080-6040. doi: 10.3201/eid0901.020220. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2873752/.

M. E. Reller, C. Bodinayake, A. Nagahawatte, V. Devasiri, W. Kodikara-Arachichi, J. J. Strouse, A. Broadwater, T. Østbye, A. de Silva, and C. W. Woods. Unsuspected Dengue and Acute Febrile Illness in Rural and Semi-Urban Southern Sri Lanka. *Emerging Infectious Diseases*, 18(2):256–263, Feb. 2012. ISSN 1080-6040. doi: 10.3201/eid1802.110962. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3310451/.

P. Renault, J.-L. Solet, D. Sissoko, E. Balleydier, S. Larrieu, L. Filleul, C. Lassalle, J. Thiria, E. Rachou, H. de Valk, D. Ilef, M. Ledrans, I. Quatresous, P. Quenel, and V. Pierre. A major epidemic of chikungunya virus infection on Reunion Island, France, 2005-2006. *The American Journal of Tropical Medicine and Hygiene*, 77(4):727–731, Oct. 2007. ISSN 0002-9637.

F. A. Rey, K. Stiasny, M.-C. Vaney, M. Dellarole, and F. X. Heinz. The bright and the dark side of human antibody responses to flaviviruses: lessons for vaccine design. *EMBO reports*, 19(2):206–224, Feb. 2018. ISSN 1469-3178. doi: 10.15252/embr.201745302.

G. Rezza. Dengue and other *Aedes* -borne viruses: a threat to Europe? *Eurosurveillance*, 21(21), May 2016. ISSN 1025-496X, 1560-7917. doi: 10.2807/1560-7917.ES.2016.21.21.30238. URL http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=22486.

G. Rezza, L. Nicoletti, R. Angelini, R. Romi, A. C. Finarelli, M. Panning, P. Cordioli, C. Fortuna, S. Boros, F. Magurano, G. Silvi, P. Angelini, M. Dottori, M. G. Ciufolini, G. C. Majori, A. Cassone, and CHIKV study group. Infection with chikungunya virus in Italy: an outbreak in a temperate region. *Lancet (London, England)*, 370(9602):1840–1846, Dec. 2007. ISSN 1474-547X. doi: 10.1016/S0140-6736(07)61779-6.

G. S. Ribeiro, M. Kikuti, L. B. Tauro, L. C. J. Nascimento, C. W. Cardoso, G. S. Campos, A. I. Ko, S. C. Weaver, M. G. Reis, U. Kitron, I. A. D. Paploski, M. M. O. Silva, A. M. Kasper, A. S. Tavares, J. S. Cruz, P. S. S. Moreira, R. O. Anjos, J. M. G. Araújo, R. Khouri, and S. I. Sardi. Does immunity after Zika virus infection cross-protect against dengue? *The Lancet Global Health*, 6(2):e140–e141, Feb. 2018. ISSN 2214-109X. doi: 10.1016/S2214-109X(17)30496-5. URL http://www.thelancet.com/journals/langlo/article/PIIS2214-109X(17)30496-5/abstract.

J. Riou, C. Poletto, and P.-Y. Boëlle. A comparative analysis of Chikungunya and Zika transmission. *Epidemics*, 2017. ISSN 1755-4365. doi: 10.1016/j.epidem.2017.01.001. URL http://www.sciencedirect.com/science/article/pii/S1755436517300014.

J. Riou, C. Poletto, and P.-Y. Boëlle. Improving early epidemiological assessment of emerging Aedes-transmitted epidemics using historical data. *PLOS Neglected Tropical Diseases*, 12(6):e0006526, June 2018. ISSN 1935-2735. doi: 10.1371/journal.pntd.

0006526. URL http://journals.plos.org/plosntds/article?id=10.1371/journal.pntd.0006526.

C. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer Texts in Statistics. Springer-Verlag, New York, 2 edition, 2004. ISBN 978-0-387-21239-5. URL //www.springer.com/la/book/9780387212395.

G. Roberts and R. Tweedie. Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika*, 83(1):95–110, Mar. 1996. ISSN 0006-3444, 1464-3510. doi: 10.1093/biomet/83.1.95. URL https://academic.oup.com/biomet/article-lookup/doi/10.1093/biomet/83.1.95.

G. O. Roberts and J. S. Rosenthal. Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science*, 16(4):351–367, Nov. 2001. ISSN 0883-4237, 2168-8745. doi: 10.1214/ss/1015346320. URL https://projecteuclid.org/euclid.ss/1015346320.

G. O. Roberts and J. S. Rosenthal. Examples of Adaptive MCMC. *Journal of Computational and Graphical Statistics*, 18(2):349–367, Jan. 2009. ISSN 1061-8600. doi: 10.1198/jcgs.2009.06134. URL http://dx.doi.org/10.1198/jcgs.2009.06134.

M. G. Roberts and J. A. P. Heesterbeek. A new method for estimating the effort required to control an infectious disease. *Proceedings of the Royal Society B: Biological Sciences*, 270(1522):1359–1364, July 2003. ISSN 0962-8452, 1471-2954. doi: 10.1098/rspb.2003.2339. URL http://rspb.royalsocietypublishing.org/cgi/doi/10.1098/rspb.2003.2339.

F. Rocha, M. Aguiar, M. Souza, and N. Stollenwerk. Time-scale separation and centre manifold analysis describing vector-borne disease dynamics. *International Journal of Computer Mathematics*, 90(10):2105–2125, Oct. 2013. ISSN 0020-7160, 1029-0265. doi: 10.1080/00207160.2013.783208. URL http://www.tandfonline.com/doi/abs/10.1080/00207160.2013.783208.

F. Rocha, L. Mateus, U. Skwara, M. Aguiar, and N. Stollenwerk. Understanding dengue fever dynamics: a study of seasonality in vector-borne disease models. *International Journal of Computer Mathematics*, 93(8):1405–1422, Aug. 2016. ISSN 0020-7160. doi: 10.1080/00207160.2015.1050961. URL http://dx.doi.org/10.1080/00207160.2015.1050961.

B. Roche, B. Gaillard, L. Léger, R. Pélagie-Moutenda, T. Sochacki, B. Cazelles, M. Ledrans, A. Blateau, D. Fontenille, M. Etienne, F. Simard, M. Salathé, and A. Yébakima. An ecological and digital epidemiology analysis on the role of human behavior on the 2014 Chikungunya outbreak in Martinique. *Scientific Reports*, 7(1):5967, July 2017. ISSN 2045-2322. doi: 10.1038/s41598-017-05957-y. URL https://www.nature.com/articles/s41598-017-05957-y.

I. Rodriguez-Barraquer, M. T. Cordeiro, C. Braga, W. V. d. Souza, E. T. Marques, and D. A. T. Cummings. From Re-Emergence to Hyperendemicity: The Natural History of the Dengue Epidemic in Brazil. *PLOS Neglected Tropical Diseases*, 5(1):e935, Jan. 2011. ISSN 1935-2735. doi: 10.1371/journal.pntd.0000935. URL http://journals.plos.org/plosntds/article?id=10.1371/journal.pntd.0000935.

A. J. Rodriguez-Morales, W. E. Villamil-Gómez, and C. Franco-Paredes. The arboviral burden of disease caused by co-circulation and co-infection of dengue, chikungunya and Zika in the Americas. *Travel Medicine and Infectious Disease*, 14(3):177–179, May 2016. ISSN 1477-8939, 1873-0442. doi: 10.1016/j.tmaid.2016.05.004. URL http://www.travelmedicinejournal.com/article/S1477-8939(16)30037-0/fulltext.

P. Rohani, D. J. Earn, B. Finkenstadt, and B. T. Grenfell. Population dynamic interference among childhood diseases. *Proceedings of the Royal Society B: Biological Sciences*, 265 (1410):2033–2041, Nov. 1998. ISSN 0962-8452, 1471-2954. doi: 10.1098/rspb.1998.0537. URL http://rspb.royalsocietypublishing.org/cgi/doi/10.1098/rspb.1998.0537.

P. Rohani, M. J. Keeling, and B. T. Grenfell. The interplay between determinism and stochasticity in childhood diseases. *The American Naturalist*, 159(5):469–481, May 2002. ISSN 1537-5323. doi: 10.1086/339467.

D. Rojas, N. Dean, Y. Yang, E. Kenah, J. Quintero, S. Tomasi, E. Ramirez, Y. Kelly, C. Castro, G. Carrasquilla, M. Halloran, and I. Longini. The epidemiology and transmissibility of Zika virus in Girardot and San Andres island, Colombia, September 2015 to January 2016. *Euro surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin*, 21(28), July 2016. ISSN 1025-496X. doi: 10.2807/1560-7917.ES.2016.21.28.30283. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5124348/.

V. Romeo-Aznar, R. Paul, O. Telle, and M. Pascual. Mosquito-borne transmission in urban landscapes: the missing link between vector abundance and human density. *Proc. R. Soc. B*, 285(1884):20180826, Aug. 2018. ISSN 0962-8452, 1471-2954. doi: 10.1098/rspb.2018.0826. URL http://rspb.royalsocietypublishing.org/content/285/1884/20180826.

M. G. Rosa-Freitas, P. Tsouris, I. C. Reis, M. d. A. F. M. Magalhães, T. F. S. d. Nascimento, and N. A. Honório. Dengue and land cover heterogeneity in Rio de Janeiro. *Oecologia Australis*, 14 (3):641–667, Sept. 2010. ISSN 2177-6199. URL https://revistas.ufrj.br/index.php/oa/article/view/7103.

M. Roy, M. J. Bouma, E. L. Ionides, R. C. Dhiman, and M. Pascual. The Potential Elimination of Plasmodium vivax Malaria by Relapse Treatment: Insights from a Transmission Model and Surveillance Data from NW India. *PLOS Negl Trop Dis*, 7(1):e1979, Jan. 2013. ISSN 1935-2735. doi: 10.1371/journal.pntd.0001979. URL http://journals.plos.org/plosntds/article?id=10.1371/journal.pntd.0001979.

H. Salje, J. Lessler, T. P. Endy, F. C. Curriero, R. V. Gibbons, A. Nisalak, S. Nimmannitya, S. Kalayanarooj, R. G. Jarman, S. J. Thomas, D. S. Burke, and D. A. T. Cummings. Revealing the microscale spatial signature of dengue transmission and immunity in an urban population. *Proceedings of the National Academy of Sciences of the United States of America*, 109(24): 9535–9538, June 2012. ISSN 1091-6490. doi: 10.1073/pnas.1120621109.

H. Salje, J. Lessler, I. M. Berry, M. C. Melendrez, T. Endy, S. Kalayanarooj, A. A-Nuegoonpipat, S. Chanama, S. Sangkijporn, C. Klungthong, B. Thaisomboonsuk, A. Nisalak, R. V. Gibbons, S. Iamsirithaworn, L. R. Macareo, I.-K. Yoon, A. Sangarsang, R. G. Jarman, and D. A. T. Cummings. Dengue diversity across spatial and temporal scales: Local structure and the

effect of host population size. *Science*, 355(6331):1302–1306, Mar. 2017. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aaj9384. URL http://science.sciencemag.org/content/355/6331/1302.

K. Sallah, R. Giorgi, L. Bengtsson, X. Lu, E. Wetter, P. Adrien, S. Rebaudet, R. Piarroux, and J. Gaudart. Mathematical models for predicting human mobility in the context of infectious disease spread: introducing the impedance model. *International Journal of Health Geographics*, 16(1):42, Nov. 2017. ISSN 1476-072X. doi: 10.1186/s12942-017-0115-7.

N. Sangkawibha, S. Rojanasuphot, S. Ahandrik, S. Viriyapongse, S. Janatasen, V. Salitul, B. Phanthumachinda, and S. B. Halstead. Risk factors in dengue shock syndrome: a prospective epidemiologic study in Rayong, Thailand I. The 1980 outbreak. *American journal of epidemiology*, 120(5):653–669, 1984. URL http://aje.oxfordjournals.org/content/120/5/653.short.

S. Sarkka. *Bayesian Filtering and Smoothing*. Cambridge University Press, Cambridge, 2013. ISBN 978-1-139-34420-3. doi: 10.1017/CBO9781139344203. URL http://ebooks.cambridge.org/ref/id/CBO9781139344203.

M. Sarzynska, O. Udiani, and N. Zhang. A study of gravity-linked metapopulation models for the spatial spread of dengue fever. Aug. 2013. URL https://arxiv.org/abs/1308.4589.

E. Saulnier, O. Gascuel, and S. Alizon. Inferring epidemiological parameters from phylogenies using regression-ABC: A comparative study. *PLOS Computational Biology*, 13(3):e1005416, Mar. 2017. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1005416. URL http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005416.

L. Schuler-Faccini. Possible Association Between Zika Virus Infection and Microcephaly - Brazil, 2015. *MMWR. Morbidity and Mortality Weekly Report*, 65, 2016. ISSN 0149-21951545-861X. doi: 10.15585/mmwr.mm6503e2. URL https://www.cdc.gov/mmwr/volumes/65/wr/mm6503e2.htm.

T. M. Sharp, K. M. Tomashek, J. S. Read, H. S. Margolis, and S. H. Waterman. A New Look at an Old Disease: Recent Insights into the Global Epidemiology of Dengue. *Current Epidemiology Reports*, 4(1):11–21, 2017. doi: 10.1007/s40471-017-0095-y.

D. S. Shepard, E. A. Undurraga, Y. A. Halasa, and J. D. Stanaway. The global economic burden of dengue: a systematic analysis. *The Lancet. Infectious Diseases*, 16(8):935–941, Aug. 2016. ISSN 1474-4457. doi: 10.1016/S1473-3099(16)00146-8.

Z. Shousheng, L. Denis-Vidal, and N. Verdière. Identifiability study in a model describing the propagation of the chikungunya to the human population. In *MOSIM 2014, 10ème Conférence Francophone de Modélisation, Optimisation et Simulation*, Nancy, France, Nov. 2014. URL https://hal.archives-ouvertes.fr/hal-01166654.

S. Shrestha, A. A. King, and P. Rohani. Statistical Inference for Multi-Pathogen Systems. *PLOS Computational Biology*, 7(8):e1002135, Aug. 2011. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1002135. URL http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002135.

S. Shrestha, B. Foxman, D. M. Weinberger, C. Steiner, C. Viboud, and P. Rohani. Identifying the interaction between influenza and pneumococcal pneumonia using incidence data. *Science*

translational medicine, 5(191):191ra84, June 2013. ISSN 1946-6234. doi: 10.1126/scitranslmed.3005982. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4178309/.

D. P. Shutt, C. A. Manore, S. Pankavich, A. T. Porter, and S. Y. Del Valle. Estimating the reproductive number, total outbreak size, and reporting rates for Zika epidemics in South and Central America. *Epidemics*, 21:63–79, Dec. 2017. ISSN 1755-4365. doi: 10.1016/j.epidem.2017.06.005. URL http://www.sciencedirect.com/science/article/pii/S1755436517300257.

F. Simini, M. C. González, A. Maritan, and A.-L. Barabási. A universal model for mobility and migration patterns. *Nature*, 484(7392):96, Apr. 2012. ISSN 1476-4687. doi: 10.1038/nature10856. URL https://www.nature.com/articles/nature10856.

D. L. Smith, K. E. Battle, S. I. Hay, C. M. Barker, T. W. Scott, and F. E. McKenzie. Ross, Macdonald, and a Theory for the Dynamics and Control of Mosquito-Transmitted Pathogens. *PLoS Pathogens*, 8(4), Apr. 2012. ISSN 1553-7366. doi: 10.1371/journal.ppat.1002588. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3320609/.

D. L. Smith, T. A. Perkins, R. C. Reiner, C. M. Barker, T. Niu, L. F. Chaves, A. M. Ellis, D. B. George, A. L. Menach, J. R. C. Pulliam, D. Bisanzio, C. Buckee, C. Chiyaka, D. A. T. Cummings, A. J. Garcia, M. L. Gatton, P. W. Gething, D. M. Hartley, G. Johnston, E. Y. Klein, E. Michael, A. L. Lloyd, D. M. Pigott, W. K. Reisen, N. Ruktanonchai, B. K. Singh, J. Stoller, A. J. Tatem, U. Kitron, H. C. J. Godfray, J. M. Cohen, S. I. Hay, and T. W. Scott. Recasting the theory of mosquito-borne pathogen transmission dynamics and control. *Transactions of The Royal Society of Tropical Medicine and Hygiene*, 108(4):185–197, Jan. 2014. ISSN 0035-9203, 1878-3503. doi: 10.1093/trstmh/tru026. URL http://trstmh.oxfordjournals.org/content/108/4/185.

SMS. Secretaria Municipal de Saúde, Prefeitura do Rio de Janeiro, Infected cases of Dengue per neighborhood and period (Casos de Dengue por bairro e período), 2016a. URL http://www.rio.rj.gov.br/web/sms/exibeConteudo?id=5953773.

SMS. Secretaria Municipal de Saúde, Prefeitura do Rio de Janeiro, Zika, Epidemiological data (Dado Epidemiológicos), 2016b. URL http://www.rio.rj.gov.br/web/sms/exibeConteudo?id=5953773.

H. E. Soper. The Interpretation of Periodicity in Disease Prevalence. *Journal of the Royal Statistical Society*, 92(1):34–73, 1929. ISSN 0952-8385. doi: 10.2307/2341437. URL http://www.jstor.org/stable/2341437.

D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. Van Der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639, Oct. 2002. ISSN 1467-9868. doi: 10.1111/1467-9868.00353. URL http://onlinelibrary.wiley.com/doi/10.1111/1467-9868.00353/abstract.

{Stan Development Team}. Stan Modeling Language: User's Guide and Reference Manual, 2017. URL http://mc-stan.org/users/documentation/.

M. Stefanouli and S. Polyzos. Gravity vs radiation model: two approaches on commuting in Greece. *Transportation Research Procedia*, 24(Supplement C):65–72, Jan. 2017. ISSN 2352-1465.

doi: 10.1016/j.trpro.2017.05.069. URL http://www.sciencedirect.com/science/article/pii/S2352146517303502.

T. Stocks, T. Britton, and M. Höhle. pomp-Astic Inference For Epidemic Models: Simple Vs. Complex. *bioRxiv*, page 125880, Apr. 2017. doi: 10.1101/125880. URL http://biorxiv.org/content/early/2017/04/13/125880.

T. Stocks, T. Britton, and M. Höhle. Model selection and parameter estimation for dynamic epidemic models via iterated filtering: application to rotavirus in Germany. *Biostatistics*, 2018. doi: 10.1093/biostatistics/kxy057. URL https://academic.oup.com/biostatistics/advance-article/doi/10.1093/biostatistics/kxy057/5108499.

S. T. Stoddard, A. C. Morrison, G. M. Vazquez-Prokopec, V. P. Soldan, T. J. Kochel, U. Kitron, J. P. Elder, and T. W. Scott. The Role of Human Movement in the Transmission of Vector-Borne Pathogens. *PLOS Neglected Tropical Diseases*, 3(7):e481, July 2009. ISSN 1935-2735. doi: 10.1371/journal.pntd.0000481. URL http://journals.plos.org/plosntds/article?id=10.1371/journal.pntd.0000481.

S. T. Stoddard, B. M. Forshey, A. C. Morrison, V. A. Paz-Soldan, G. M. Vazquez-Prokopec, H. Astete, R. C. Reiner, S. Vilcarromero, J. P. Elder, E. S. Halsey, T. J. Kochel, U. Kitron, and T. W. Scott. House-to-house human movement drives dengue virus transmission. *Proceedings of the National Academy of Sciences of the United States of America*, 110(3):994–999, Jan. 2013. ISSN 0027-8424. doi: 10.1073/pnas.1213349110. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3549073/.

L. M. Stolerman, D. Coombs, and S. Boatto. SIR-Network Model and Its Application to Dengue Fever. *SIAM Journal on Applied Mathematics*, Dec. 2015. doi: 10.1137/140996148. URL https://epubs.siam.org/doi/abs/10.1137/140996148.

D. Strickman, R. Sithiprasasna, P. Kittayapong, and B. L. Innis. Distribution of dengue and Japanese encephalitis among children in rural and suburban Thai villages. *The American Journal of Tropical Medicine and Hygiene*, 63(1-2):27–35, Aug. 2000. ISSN 0002-9637.

J. A. Swanstrom, J. A. Plante, K. S. Plante, E. F. Young, E. McGowan, E. N. Gallichotte, D. G. Widman, M. T. Heise, A. M. d. Silva, and R. S. Baric. Dengue Virus Envelope Dimer Epitope Monoclonal Antibodies Isolated from Dengue Patients Are Protective against Zika Virus. *mBio*, 7(4):e01123–16, July 2016. ISSN , 2150-7511. doi: 10.1128/mBio.01123-16. URL http://mbio.asm.org/content/7/4/e01123-16.

R. D. C. Team. R: A Language and Environment for Statistical Computing, 2015. URL http://www.R-project.org.

M. G. Teixeira, J. B. Siqueira, G. L. C. Ferreira, L. Bricks, and G. Joint. Epidemiological Trends of Dengue Disease in Brazil (2000–2010): A Systematic Literature Search and Analysis. *PLoS Neglected Tropical Diseases*, 7(12), Dec. 2013. ISSN 1935-2727. doi: 10.1371/journal.pntd.0002520. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3871634/.

O. Telle, A. Vaguet, N. K. Yadav, B. Lefebvre, E. Daudé, R. E. Paul, A. Cebeillac, and B. N. Nagpal. The Spread of Dengue in an Endemic Urban Milieu - The Case of Delhi, India. *PLOS ONE*, 11(1):e0146539, Jan. 2016. ISSN 1932-6203. doi: 10.1371/journal.pone.

0146539. URL http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0146539.

Q. A. ten Bosch, B. K. Singh, M. R. A. Hassan, D. D. Chadee, and E. Michael. The Role of Serotype Interactions and Seasonality in Dengue Model Selection and Control: Insights from a Pattern Matching Approach. *PLOS Neglected Tropical Diseases*, 10(5):e0004680, May 2016. ISSN 1935-2735. doi: 10.1371/journal.pntd.0004680. URL http://journals.plos.org/plosntds/article?id=10.1371/journal.pntd.0004680.

Q. A. ten Bosch, H. E. Clapham, L. Lambrechts, V. Duong, P. Buchy, B. M. Althouse, A. L. Lloyd, L. A. Waller, A. C. Morrison, U. Kitron, G. M. Vazquez-Prokopec, T. W. Scott, and T. A. Perkins. Contributions from the silent majority dominate dengue virus transmission. *PLoS Pathogens*, 14(5), May 2018. ISSN 1553-7366. doi: 10.1371/journal.ppat.1006965. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5933708/.

A. C. B. Terzian, A. S. Schanoski, M. T. d. O. Mota, R. A. da Silva, C. F. Estofolete, T. E. Colombo, P. Rahal, K. A. Hanley, N. Vasilakis, J. Kalil, and M. L. Nogueira. Viral Load and Cytokine Response Profile Does Not Support Antibody-Dependent Enhancement in Dengue-Primed Zika Virus-Infected Patients. *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America*, 65(8):1260–1265, 2017. ISSN 1537-6591. doi: 10.1093/cid/cix558.

M. Teurlai, R. Huy, B. Cazelles, R. Duboz, C. Baehr, and S. Vong. Can Human Movements Explain Heterogeneous Propagation of Dengue Fever in Cambodia? *PLOS Neglected Tropical Diseases*, 6(12):e1957, Dec. 2012. ISSN 1935-2735. doi: 10.1371/journal.pntd.0001957. URL http://journals.plos.org/plosntds/article?id=10.1371/journal.pntd.0001957.

K. T. Thai, T. Q. Binh, P. T. Giao, H. L. Phuong, L. Q. Hung, N. V. Nam, T. T. Nga, J. Groen, N. Nagelkerke, and P. J. Vries. Seroprevalence of dengue antibodies, annual incidence and risk factors among children in southern Vietnam. *Tropical Medicine & International Health*, 10 (4):379–386, 2005. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1365-3156.2005.01388.x/full.

K. T. D. Thai, T. T. T. Nga, N. Van Nam, H. L. Phuong, P. T. Giao, L. Q. Hung, T. Q. Binh, G. J. J. Van Doornum, and P. J. De Vries. Incidence of primary dengue virus infections in Southern Vietnamese children and reactivity against other flaviviruses. *Tropical Medicine & International Health*, 12(12):1553–1557, Dec. 2007. ISSN 1365-3156. doi: 10.1111/j.1365-3156.2007.01964.x. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1365-3156.2007.01964.x/abstract.

R. N. Thompson, C. A. Gilligan, and N. J. Cunniffe. Detecting Presymptomatic Infection Is Necessary to Forecast Major Epidemics in the Earliest Stages of Infectious Disease Outbreaks. *PLOS Comput Biol*, 12(4):e1004836, Apr. 2016. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1004836. URL http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004836.

N. T. K. Tien, C. Luxemburger, N. T. Toan, L. Pollissard-Gadroy, V. T. Q. Huong, P. Van Be, N. N. Rang, T.-A. Wartel, and J. Lang. A prospective cohort study of dengue infection in schoolchildren in Long Xuyen, Viet Nam. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 104(9):592–600, Sept. 2010. ISSN 00359203. doi: 10.1016/j.trstmh.2010.

06.003. URL https://academic.oup.com/trstmh/article-lookup/doi/10.1016/j.trstmh.2010.06.003.

T. Toni, D. Welch, N. Strelkowa, A. Ipsen, and M. P. H. Stumpf. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of The Royal Society Interface*, 6(31):187–202, Feb. 2009. ISSN 1742-5689, 1742-5662. doi: 10.1098/rsif.2008.0172. URL http://rsif.royalsocietypublishing.org/content/6/31/187.

P. Touloupou, N. Alzahrani, P. Neal, S. E. F. Spencer, and T. J. McKinley. Efficient Model Comparison Techniques for Models Requiring Large Scale Data Augmentation. *Bayesian Analysis*, 13(2):437–459, June 2018. ISSN 1936-0975, 1931-6690. doi: 10.1214/17-BA1057. URL https://projecteuclid.org/euclid.ba/1493431262.

S. Towers, F. Brauer, C. Castillo-Chavez, A. K. I. Falconar, A. Mubayi, and C. M. E. Romero-Vivas. Estimate of the reproduction number of the 2015 Zika virus outbreak in Barranquilla, Colombia, and estimation of the relative role of sexual transmission. *Epidemics*, 17:50–55, Dec. 2016a. ISSN 1755-4365. doi: 10.1016/j.epidem.2016.10.003. URL http://www.sciencedirect.com/science/article/pii/S1755436516300330.

S. Towers, F. Brauer, C. Castillo-Chavez, A. K. I. Falconar, A. Mubayi, and C. M. E. Romero-Vivas. Estimation of the reproduction number of the 2015 Zika virus outbreak in Barranquilla, Colombia, and a first estimate of the relative role of sexual transmission. *arXiv:1606.01422 [q-bio]*, June 2016b. URL http://arxiv.org/abs/1606.01422. arXiv: 1606.01422.

J. Truscott and N. M. Ferguson. Evaluating the Adequacy of Gravity Models as a Description of Human Mobility for Epidemic Modelling. *PLOS Computational Biology*, 8(10):e1002699, Oct. 2012. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1002699. URL http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002699.

P. van den Driessche and J. Watmough. Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission. *Mathematical Biosciences*, 180:29–48, Dec. 2002. ISSN 0025-5564.

D. A. Vasco, H. J. Wearing, and P. Rohani. Tracking the dynamics of pathogen interactions: Modeling ecological and immune-mediated processes in a two-pathogen single-host system. *Journal of Theoretical Biology*, 245(1):9–25, Mar. 2007. ISSN 0022-5193. doi: 10.1016/j.jtbi.2006.08.015. URL http://www.sciencedirect.com/science/article/pii/S0022519306003638.

F. Verelst, L. Willem, and P. Beutels. Behavioural change models for infectious disease transmission: a systematic review (2010–2015). *Journal of the Royal Society Interface*, 13(125), Dec. 2016. ISSN 1742-5689. doi: 10.1098/rsif.2016.0820. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5221530/.

C. Viboud, O. N. Bjørnstad, D. L. Smith, L. Simonsen, M. A. Miller, and B. T. Grenfell. Synchrony, Waves, and Spatial Hierarchies in the Spread of Influenza. *Science*, 312(5772):447–451, Apr. 2006. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1125237. URL http://www.sciencemag.org/content/312/5772/447.

C. Viboud, K. Sun, R. Gaffey, M. Ajelli, L. Fumanelli, S. Merler, Q. Zhang, G. Chowell, L. Simonsen, and A. Vespignani. The RAPIDD ebola forecasting challenge: Synthesis and

lessons learnt. *Epidemics*, 22:13–21, Mar. 2018. ISSN 1755-4365. doi: 10.1016/j.epidem.2017.08.002. URL http://www.sciencedirect.com/science/article/pii/S1755436517301275.

D. a. M. Villela, L. S. Bastos, L. M. DE Carvalho, O. G. Cruz, M. F. C. Gomes, B. Durovni, M. C. Lemos, V. Saraceni, F. C. Coelho, and C. T. Codeço. Zika in Rio de Janeiro: Assessment of basic reproduction number and comparison with dengue outbreaks. *Epidemiology and Infection*, 145(8):1649–1657, 2017. ISSN 1469-4409. doi: 10.1017/S0950268817000358.

S. Vong, V. Khieu, O. Glass, S. Ly, V. Duong, R. Huy, C. Ngan, O. Wichmann, G. W. Letson, H. S. Margolis, and P. Buchy. Dengue Incidence in Urban and Rural Cambodia: Results from Population-Based Active Fever Surveillance, 2006–2008. *PLoS Neglected Tropical Diseases*, 4(11):e903, Nov. 2010. ISSN 1935-2735. doi: 10.1371/journal.pntd.0000903. URL http://dx.plos.org/10.1371/journal.pntd.0000903.

S. Vongpunsawad, D. Intharasongkroh, T. Thongmee, and Y. Poovorawan. Seroprevalence of antibodies to dengue and chikungunya viruses in Thailand. *PLOS ONE*, 12(6):e0180560, June 2017. ISSN 1932-6203. doi: 10.1371/journal.pone.0180560. URL http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0180560.

E. Walter and L. Pronzato. On the identifiability and distinguishability of nonlinear parametric models. *Mathematics and Computers in Simulation*, 42(2):125–134, Oct. 1996. ISSN 0378-4754. doi: 10.1016/0378-4754(95)00123-9. URL http://www.sciencedirect.com/science/article/pii/0378475495001239.

L. Wang, S. Valderramos, A. Wu, S. Ouyang, C. Li, P. Brasil, M. Bonaldo, T. Coates, K. Nielsen-Saines, T. Jiang, R. Aliyari, and G. Cheng. From Mosquitos to Humans: Genetic Evolution of Zika Virus. *Cell Host & Microbe*, 19(5):561–565, May 2016. ISSN 19313128. doi: 10.1016/j.chom.2016.04.006. URL http://linkinghub.elsevier.com/retrieve/pii/S1931312816301421.

L. Wang, X. Zhang, and Z. Liu. An SEIR Epidemic Model with Relapse and General Nonlinear Incidence Rate with Application to Media Impact. *Qualitative Theory of Dynamical Systems*, pages 1–21, Feb. 2017. ISSN 1575-5460, 1662-3592. doi: 10.1007/s12346-017-0231-6. URL https://link.springer.com/article/10.1007/s12346-017-0231-6.

H. J. Wearing and P. Rohani. Ecological and immunological determinants of dengue epidemics. *Proceedings of the National Academy of Sciences*, 103(31):11802–11807, Aug. 2006. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.0602960103. URL http://www.pnas.org/content/103/31/11802.

WeatherUnderground. Weather Forecast & Reports - Long Range & Local, 2016. URL https://www.wunderground.com/.

R. A. Weiss and A. J. McMichael. Social and environmental risk factors in the emergence of infectious diseases. *Nature Medicine*, 10(12s):S70–S76, Dec. 2004. ISSN 1546-170X. doi: 10.1038/nm1150. URL https://www.nature.com/articles/nm1150.

A. Wesolowski, T. Qureshi, M. F. Boni, P. R. Sundsøy, M. A. Johansson, S. B. Rasheed, K. Engø-Monsen, and C. O. Buckee. Impact of human mobility on the emergence of dengue epidemics in Pakistan. *Proceedings of the National Academy of Sciences*, 112(38):

11887–11892, Sept. 2015. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1504964112. URL http://www.pnas.org/lookup/doi/10.1073/pnas.1504964112.

WHO. Dengue vaccine: WHO position paper. Weekly epidemiological record 30, World Health Organization, Geneva, Switzerland, Oct. 2016a. URL http://www.who.int/wer/2016/wer9130.pdf.

WHO. WHO statement on the first meeting of the International Health Regulations (2005) (IHR 2005) Emergency Committee on Zika virus and observed increase in neurological disorders and neonatal malformations, Feb. 2016b. URL http://www.who.int/mediacentre/news/statements/2016/1st-emergency-committee-zika/en/.

WHO. Dengue | World Health Organisation, 2017. URL http://www.who.int/denguecontrol/en/.

O. Wichmann, I.-K. Yoon, S. Vong, K. Limkittikul, R. V. Gibbons, M. P. Mammen, S. Ly, P. Buchy, C. Sirivichayakul, R. Buathong, R. Huy, G. W. Letson, and A. Sabchareon. Dengue in Thailand and Cambodia: An Assessment of the Degree of Underrecognized Disease Burden Based on Reported Cases. *PLoS Neglected Tropical Diseases*, 5(3):e996, Mar. 2011. ISSN 1935-2735. doi: 10.1371/journal.pntd.0000996. URL http://dx.plos.org/10.1371/journal.pntd.0000996.

H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2009. ISBN 978-0-387-98140-6. URL http://ggplot2.org.

H. Wu, H. Zhu, H. Miao, and A. S. Perelson. Parameter Identifiability and Estimation of HIV/AIDS Dynamic Models. *Bulletin of Mathematical Biology*, 70(3):785–799, Apr. 2008. ISSN 0092-8240, 1522-9602. doi: 10.1007/s11538-007-9279-9. URL https://link.springer.com/article/10.1007/s11538-007-9279-9.

D. R. Xavier, M. d. A. F. M. Magalhães, R. Gracie, I. C. D. Reis, V. P. d. Matos, and C. Barcellos. Spatial-temporal diffusion of dengue in the municipality of Rio de Janeiro, Brazil, 2000-2013. *Cadernos De Saude Publica*, 33(2):e00186615, Mar. 2017. ISSN 1678-4464. doi: 10.1590/0102-311X00186615.

Y. Xia, O. N. Bjørnstad, and B. T. Grenfell. Measles metapopulation dynamics: a gravity model for epidemiological coupling and dynamics. *The American Naturalist*, 164(2):267–281, Aug. 2004. ISSN 1537-5323. doi: 10.1086/422341.

X.-Y. Yan, C. Zhao, Y. Fan, Z. Di, and W.-X. Wang. Universal predictability of mobility patterns in cities. *Journal of the Royal Society Interface*, 11(100), Nov. 2014. doi: 10.1098/rsif.2014.0834. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4191116/.

J. Yang and J. S. Rosenthal. Automatically tuned general-purpose MCMC via new adaptive diagnostics. *Computational Statistics*, 32(1):315–348, Mar. 2017. ISSN 0943-4062, 1613-9658. doi: 10.1007/s00180-016-0682-2. URL http://link.springer.com/10.1007/s00180-016-0682-2.

W. Yang, A. Karspeck, and J. Shaman. Comparison of Filtering Methods for the Modeling and Retrospective Forecasting of Influenza Epidemics. *PLoS Computational Biology*, 10(4):e1003583, Apr. 2014. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1003583. URL http://dx.plos.org/10.1371/journal.pcbi.1003583.

G. Yap, S. Liang, C. Hapuarachchi, C.-S. Chong, L.-K. Tan, X. Liu, Y. Shi, K. Zeng, C. Koo, A. Cook, Y.-L. Lai, and L.-C. Ng. Dengue outlook for Singapore in 2016. *Singapore Epidemiol News Bull*, 42(1):3–10, Jan. 2016. URL https://www.moh.gov.sg/content/dam/moh_web/Statistics/Epidemiological_News_Bulletin/2016/ENB%20Quarterly_Jan%202016.pdf.

I.-K. Yoon, A. L. Rothman, D. Tannitisupawong, A. Srikiatkhachorn, R. G. Jarman, J. Aldstadt, A. Nisalak, M. P. Mammen, S. Thammapalo, S. Green, D. H. Libraty, R. V. Gibbons, A. Getis, T. Endy, J. W. Jones, C. J. M. Koenraadt, A. C. Morrison, T. Fansiri, C. Pimgate, and T. W. Scott. Underrecognized Mildly Symptomatic Viremic Dengue Virus Infections in Rural Thai Schools and Villages. *The Journal of Infectious Diseases*, 206(3):389–398, Aug. 2012. ISSN 0022-1899. doi: 10.1093/infdis/jis357. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3490697/.

D. Yu, Q. Lin, A. P. Chiu, and D. He. Effects of reactive social distancing on the 1918 influenza pandemic. *PLOS ONE*, 12(7):e0180545, July 2017. ISSN 1932-6203. doi: 10.1371/journal.pone.0180545. URL http://dx.plos.org/10.1371/journal.pone.0180545.

Q. Zhang, K. Sun, M. Chinazzi, A. P. y. Piontti, N. E. Dean, D. P. Rojas, S. Merler, D. Mistry, P. Poletti, L. Rossi, M. Bray, M. E. Halloran, I. M. Longini, and A. Vespignani. Spread of Zika virus in the Americas. *Proceedings of the National Academy of Sciences*, 114(22): E4334–E4343, May 2017. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1620161114. URL http://www.pnas.org/content/114/22/E4334.

université PARIS-SACLAY | École doctorale de mathématiques Hadamard (EDMH)

Titre : Modélisation mathématique et inférence statistique pour une meilleure compréhension des dynamiques des arboviroses

Mots Clefs : dengue, virus Zika, modèle de maladie infectieuse, MCMC, PMCMC, stochasticité

Résumé : L'importance et l'expansion des arboviroses comme la dengue ou le virus Zika nécessite des modèles pour mieux comprendre et prédire leurs dynamiques. La propagation vectorielle de ces maladies est influencée par de multiples facteurs humains et environnementaux qui rendent complexe la construction de modèles épidémiologiques parcimonieux. Parallèlement, de nombreux outils théoriques et computationnels existent désormais pour confronter ces modèles aux données observées. L'objectif de ce travail de thèse est donc d'apporter l'éclairage des données sur les modèles de propagation des arboviroses. Dans un premier temps, il s'agit d'identifier les éléments les plus importants à incorporer pour modéliser les dynamiques de la dengue en milieu rural, dans la région de Kampong Cham (Cambodge). Différents modèles sont comparés, complexifiant à la fois le détail de l'histoire de la maladie et la prise en compte des formes de stochasticité. Dans le cadre déterministe, on a pu souligner l'importance des interactions entre sérotypes, et le faible intérêt pour la représentation explicite des moustiques vecteurs et des individus asymptomatiques. Par ailleurs, la prise en compte des incertitudes indique qu'une large part de la dynamique est capturée seulement par la stochasticité et non par les éléments du squelette déterministe du modèle. Aussi étudie-t-on dans un second temps d'autres aspects de la transmission de la dengue, comme la saisonnalité et la structure spatiale, grâce à des données d'épidémies à Rio de Janeiro (Brésil). Dans un dernier temps, ces méthodes et modèles sont appliqués à l'étude d'un arbovirus émergent, le virus Zika. A partir de données d'épidémies survenues dans le Pacifique, les paramètres-clé de la propagation du virus sont estimés dans le cadre stochastique, et leur variabilité est envisagée à la fois en termes de contexte géographique et de modèle épidémiologique, par la comparaison de quatre îles et de deux modèles à transmission vectorielle. Par ailleurs, la question des interactions potentielles du virus Zika avec celui de la dengue est explorée.

Title : Mathematical modeling and statistical inference to better understand arbovirus dynamics

Key words : dengue, Zika virus, infectious disease model, MCMC, PMCMC, stochasticity

Abstract : Arboviruses such as the dengue and Zika viruses are expanding worldwide and modeling their dynamics can help to better understand and predict their propagation, as well as experiment control scenarios. These mosquito-borne diseases are influenced by a multiplicity of human and environmental factors that are complex to include in parsimonious epidemiological models. In parallel, statistical and computational tools are nowadays available to confront theoretical models to the observed data. The objective of this PhD work is therefore to study arbovirus propagation models in the light of data. Firstly, in order to identify the most important elements to incorporate in models for dengue dynamics in a rural setting, several dengue models are compared using data from the Kampong Cham region in Cambodia. Models incorporate increasing complexity both in the details of disease life history and in the account for several forms of stochasticity. In the deterministic framework, including serotype interactions proved decisive, whereas explicit modeling of mosquito vectors and asymptomatic infections had limited added value, when seasonality and underreporting are already accounted for. Moreover, including several forms of uncertainties highlighted that a large part of the disease dynamics is only captured by stochasticity and not by the elements of the deterministic skeleton. Therefore, secondly, we explore other aspects of transmission, such as seasonality and spatial structure, in the case of dengue epidemics in Rio de Janeiro (Brazil). Finally, the models and estimation methods are applied to study an emerging arbovirus, the Zika virus. Using data from epidemics in the Pacific, we estimate the key parameters of disease propagation in the stochastic framework and explore their variability in terms of geographic setting and model formulation by comparing four islands and two models with vector-borne transmission. In addition, potential interactions with the dengue virus are explored.