Enhancing the Reliability of Deep Learning Models to Improve the Observability of French Rooftop Photovoltaic Installations

Gabriel Kasmi

▶ **To cite this version:**

Gabriel Kasmi. Enhancing the Reliability of Deep Learning Models to Improve the Observability of French Rooftop Photovoltaic Installations. Chemical and Process Engineering. Université Paris sciences et lettres, 2024. English. NNT : 2024UPSLM027 . tel-04909303

HAL Id: tel-04909303

https://pastel.hal.science/tel-04909303v1

Submitted on 23 Jan 2025

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

PSL UNIVERSITÉ PARIS

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PSL

Préparée à Mines Paris - PSL

Enhancing the Reliability of Deep Learning Models to Improve the Observability of French Rooftop Photovoltaic Installations

Améliorer la fiabilité des modèles d'apprentissage profond en vue d'accroitre l'observabilité des installations photovoltaïques en toiture en France

Soutenue par
Gabriel KASMI
Le 5 avril 2024

Dirigée par
Philippe BLANC, co-encadrée par Yves-Marie SAINT-DRENAN et Laurent DUBUS

École doctorale n°621
Ingénierie des Systèmes, Matériaux, Mécanique, Énergétique

Spécialité
Energétique et génie des procédés

Composition du jury :

Stéphane MALLAT Professeur, Collège de France	*Président du jury* *Examinateur*
Angela MEYER Professeure, Berner Fachhochschule	*Rapporteure*
Mathieu SERRURIER Maitre de conférences (HDR), Université Toulouse 3	*Rapporteur*
Cristina CORNARO Professeure associée, Università degli studi di Roma Tor Vergata	*Examinatrice*
Laurent DUBUS Expert émérite, RTE	*Examinateur*
Yves-Marie SAINT-DRENAN Ingénieur de recherche, Mines Paris - PSL	*Examinateur*
Philippe BLANC Directeur de recherche, Mines Paris - PSL	*Directeur de thèse*

MINES PARIS | PSL

— Un livre comme ça est donc possible, alors ? fit Liza, toute contente.
— Il faut voir, réfléchir. Une chose comme ça, c'est énorme. On ne trouvera rien d'un seul coup. On a besoin d'expérience. Même quand on aura publié le livre, on ne saura toujours pas comment le publier. Rien que des expériences successives ; mais l'idée, elle a l'air de coller. L'idée, elle est utile.
Il finit enfin par lever les yeux, et ses yeux brillèrent même de plaisir, tant il était intéressé.
— C'est vous qui avez trouvé ça ? demanda-t-il à Liza d'une voix tendre et comme pudique.
— Mais le trouver, vous comprenez, ce n'est rien, ce qui n'est pas rien, c'est le plan, répondit Liza en souriant, je n'y comprends pas grand-chose, je ne suis pas très intelligente, et je ne poursuis que ce qui me semble clair...

Dostoievski, F. M. (1872/1995). *Les Démons* (A. Markovicz, Trad.). Actes Sud.

Acknowledgements

I would like to express my warmest thoughts to all of those who have contributed to the completion of this thesis.

I am deeply grateful to my jury members, *Stéphane Mallat* and *Cristina Cornaro* for dedicating their time to evaluate my work and to my reviewers *Angela Meyer* and *Mathieu Serrurier* for the the time spent in reviewing my thesis manuscript. I also have a thought for *Ronan Fablet* for accepting to be part of the thesis jury.

I would like to express my deepest gratitude to my supervisors, *Philippe*, *Yves-Marie* and *Laurent*. Thank you for your insightful, challenging, and stimulating guidance and unfailing support throughout this journey. It was a privilege to have you as supervisors, and this work owes you so much more than these few lines can express. Thank you for helping me grow academically and as a person.

I would like to thank my co-authors, *Raphaël Jolivet*, *David Trebosc*, *Jonathan Leloux* and *Babacar Sarr* for their work on BDAPPV. I would also like to thank *Maxime Fortin* and *Augustin Touron* for sharing their expertise on the estimation of PV power production. I would like to thank the *BDPV community*, thanks to whom I had access to invaluable data for completing this thesis work. I have a particular thought for *David Trebosc* and his tremendous work constructing the annotation platform and motivating the BDPV community to participate in the crowdsourcing campaigns. I also would like to acknowledge *Raphaël Jolivet* for his dedication to this project, helping me improve the core algorithm of DeepPVMapper, and supervising students with me. I would also like to thank *Olivier Bretteville* for his unfailing technical support, *Benjamin Donnot*, *Jerôme Picault* and *Guillaume Grosjean*, the users of groesplu0 for their technical guidance and advice.

During these 42.2 months[a], I benefited from a tremendous research environment in both Sophia-Antipolis and Paris. I would like to acknowledge the people at OIE, *Thierry*, *Sandra*, *Lionel*, *Rodrigo*, *Paula*, *Mathilde*, *Hadrien*, *Benoit*, *Joanna* and my fellow *doctorants*, *Sara*, *Alejandra*, *Mahefa*, *Vadim*, *Arthur*, *Fuqiang*, *Jérémie*. At RTE, I have a thought for the *Pôle CLER*, my desk-mates *Laurent*, *Valentin* (and formerly *Hélène*), *Anne-Claire*, *Olivier D*, *François*, and *Imane* for the enjoyable moments spent together. I would also like to acknowledge *Maxime C* and his innate ability for small talk, from movies to consumption outlooks to hiking. I also have a

a. It eventually turned out that the thesis *is* the metaphor.

thought for the former *Pôle EODCT*. Finally, I would like to acknowledge my fellow *doctorants* at RTE, *Marie*, *Florent*, *Maxime*, *Emily*, *Thomas V* and *Thomas H*, *Rémi*, *Marie-Alix*, *Benjamin*, *Marion*, *Yacine*, *Hugo C*, *Hugo H*, *Laurent*, *Antoine*, *Eva*, and *Amaury*. I have a special thought for *Maxime Laasri* and *Virginie Dussartre* for their trust in me, and to *Virgninie Dordonnat*.

I also have a thought for the running club *Sam Paris 12* for the countless laps at Léo Lagrange and aux Cyclistes, and for the many kudos. Thanks to the club, I've been able to go far beyond my limits, and the thesis work has undoubtedly benefited from it.

Going through this journey would not have been possible without my friends' and Family's support. I would like to thank *B*, *C*, *G*, *H* and *T*, for the countless moments together, for sharing the ups and downs of the doctoral life and growing from young students to jeunes actifs. I feel privileged to be surrounded by you. I have a special thought for *S*, for the pivotal moments we shared and to *M* and *B*.

Finally, I would like to thank my *Family*, for their unconditional love and especially to my *Mom* for her unconditional support. I am also profoundly grateful to *the L's*; *C*, *C*, *A*, *H*, *B* who took me as one of theirs and for their unwavering support and care. Thank you *S*; words cannot express how blessed I am to have you in my life and how significant your presence by my side has been throughout these years.

Abstract

In November 2023, the French photovoltaic (PV) installed capacity stood at 18.6 GW_p, and the French electricity transmission system operator (TSO) lacked power measurements for 20% of the fleet, which mostly corresponded to small-scale (rooftop) systems. In the context of decarbonizing the electric mix, the PV installed capacity will continue to experience sustained growth in the coming years, and the so-called problem of poor PV observability threatens its long-term integration into the grid due to the uncertainty it creates. A better knowledge of the rooftop PV fleet, embodied in a nationwide technical registry recording the localization and characteristics of the PV installations, is necessary to improve PV observability. This thesis proposes to assess whether deep learning-based remote sensing on orthoimagery is a suitable method for constructing this technical registry. The thesis first discusses the quality standards the technical registry should satisfy and introduces an unsupervised evaluation method to monitor the accuracy of the registry in the absence of ground truth labels. Second, the thesis introduces a new feature attribution method that enables the auditing of the model's decisions by decomposing its predictions into the space-scale domain. The thesis discusses the relevance of this decomposition for assessing what the model sees on the input image, understanding the model's sensitivity to varying acquisition conditions, which are found to affect the model's accuracy and reliability, and introducing a robust and reliable algorithm for mapping rooftop PV installations. Finally, the relevance of the registry for improving rooftop PV observability is established by showing that accurate and scalable estimations of the rooftop PV power production can be derived from the registry and weather data. This thesis features contributions in power systems by showing how to effectively improve rooftop PV observability and in deep learning by improving the interpretability of deep learning models thanks to a new feature attribution method. More generally, this thesis underlines the necessary conditions for using deep learning in critical industrial contexts.

Résumé

En novembre 2023, la puissance photovoltaïque (PV) installée en France s'élevait à 18,6 GW_c, et le gestionnaire du réseau de transport d'électricité (GRT) français ne disposait pas de mesures de production pour 20% du parc, correspondant principalement à des systèmes de petite taille sur toitures. Dans le contexte de décarbonisation du mix électrique, la puissance installée PV continuera de croître rapidement, aussi le manque d'observabilité du PV risque-t-il compromettre l'intégration du PV dans le système électrique en raison des incertitudes qu'il engendre. Une meilleure connaissance du parc photovoltaïque en toiture, matérialisée par un registre technique national contenant la localisation et les caractéristiques des installations photovoltaïques, est nécessaire pour améliorer l'observabilité du PV. Cette thèse evalue si l'utilisation d'algorithmes d'apprentissage profond et d'orthoimages est une méthode adaptée à la construction d'un registre technique national d'installations photovoltaiques (PV) sur toiture destiné à améliorer l'observabilité de la production PV en France. La thèse discute d'abord des normes de qualité que le registre technique doit satisfaire et introduit une méthode d'évaluation non supervisée pour contrôler l'exactitude du registre en l'absence de données de référence. Deuxièmement, la thèse introduit une nouvelle méthode d'attribution qui permet d'analyser des décisions du modèle en décomposant ses prédictions dans l'espace des ondelettes. La thèse discute de la pertinence de cette décomposition pour évaluer ce que le modèle voit sur l'image d'entrée, comprendre la sensibilité du modèle à des conditions d'acquisition variables, qui affectent la précision et la fiabilité du modèle, et introduire un algorithme robuste et fiable pour cartographier les installations PV sur toiture. Enfin, la pertinence du registre pour améliorer l'observabilité des installations photovoltaïques sur les toits est établie en montrant que des estimations précises et réplicables à grande échelle de la production issue des installations PV sur toiture peuvent être construites à partir du registre et de données météorologiques. Cette thèse apporte des contributions en énergétique et procédés, en montrant comment améliorer l'observabilité du PV toiture et en apprentissage statistique, en améliorant l'interprétabilité des modèles d'apprentissage profond grâce à une nouvelle méthode d'attribution. Plus généralement, cette thèse souligne les conditions nécessaires à l'utilisation de modèles d'apprentissage profond dans des contextes industriels critiques.

Contents

Acknowledgements iii

Abstract v

Résumé vii

Contents ix

List of figures xiii

List of Tables xxi

Acronyms and abbreviations xxv

Résumé étendu xxix

1 Introduction 1
1 Context . . . 1
1.1 Curbing anthropogenic CO_2 emissions through electrification and decarbonization . . . 1
1.2 Conditions for integrating high shares of wind and solar energy into the grid . . . 3
1.3 Overcoming the poor PV observability . . . 6
2 Literature review . . . 7
2.1 Earth observation data for large-scale mapping of rooftop PV installations . . . 7
2.2 Current methods are not reliable enough to be integrated into critical industrial processes . . . 9
2.3 On the limits of deep learning in applied settings, beyond the case of the detection of rooftop PV installations . . . 10
3 Scientific questions and outline . . . 11
3.1 Scientific questions . . . 11
3.2 Approach and outline . . . 12

2 Characterization and evaluation of a rooftop PV registry in the absence of ground truth labels 15
1 Overview of the existing and missing data sources . . . 16
1.1 Geographical information system (GIS) data . . . 16
1.2 PV registries and databases for France . . . 20

1.3 Specific requirements of the PV registry . . . 21
1.4 Training data: the BDAPPV training dataset . . . 24
2 Monitoring the accuracy of the registry without ground truth labels . . 26
2.1 Defining the evaluation criteria . . . 26
2.2 Leveraging existing data to meet these criteria: the downstream task accuracy . . . 27
3 How does the accuracy of state-of-the-art models vary over the mapping area? . . . 29
3.1 Evaluation framework . . . 30
3.2 Results: detection inconsistencies during deployment . . . 35
3.3 Understanding these shifts . . . 37

3 Assessing and improving the reliability of the model's decision process 47
1 Characterizing a model's decision in the space-scale domain . . . 48
1.1 Towards assessing what model sees on images . . . 48
1.2 Feature importance quantification in the scale-space domain . . 51
1.3 The Wavelet sCale Attribution Method (WCAM) . . . 55
2 Assessing the reliability of a model's decision process . . . 59
2.1 Relevance of the decision process: understanding the model's decision process through the lenses of the space-scale decomposition . . . 59
2.2 Robustness of the decision process: the impact of acquisition conditions on false negatives . . . 64
2.3 Acquisition conditions as image corruptions . . . 68
3 Reliably improving the robustness of the model's decision process . . 74
3.1 Robustness to image corruptions: review of existing works . . . 75
3.2 A benchmark of existing approaches . . . 77
3.3 A novel data augmentation technique for improving the robustness to acquisition conditions . . . 79

4 Constructing a reliable and scalable algorithm for mapping rooftop PV installations 85
1 Identifying the limitations of the current approaches for PV systems mapping . . . 86
1.1 Where should we focus on ? . . . 86
1.2 Standardized characteristics extraction: the `PyPVRoof` library . . 90
2 From DeepSolar to DeepPVMapper: how to make state-of-the-art more reliable? . . . 96
2.1 Evaluation on metrics that are more representative of the operational conditions . . . 97
2.2 Reducing the occurrence of false negatives through overlapping the thumbnails . . . 101
2.3 Reducing the occurrence of false positives by focusing on relevant areas . . . 103
3 Results . . . 106
3.1 The DTA reveals accuracy differences among models and the better performance brought by the sampling . . . 106
3.2 Building and deploying DeepPVMapper . . . 109
3.3 Broader impact . . . 112

5 Assessing the gains of the registry for improving PV observability 115
1 Additional requirements for improving rooftop PV observability 116
1.1 BDPV ground measurement data 116
1.2 Solar radiation and temperature data 118
1.3 PV registry 121
2 Assessing the relevance of our approach 121
2.1 Evaluation metrics 121
2.2 Proposed approach for improving rooftop PV observability 122
2.3 Assessing the relevance for improving PV observability 125
3 Our approach paves the way towards better PV observability 130
3.1 Improving rooftop PV observability 130
3.2 The impact of characteristics estimation biases on accuracy . . 133
3.3 Broader impact: closing the gap with the TSO's aggregated approaches 137

6 Conclusion and discussion 139
1 Conclusion 139
1.1 Answer to the scientific question 139
1.2 Discussion 143
1.3 Contributions 144
2 Limitations 145
2.1 On the power system's side... 145
2.2 ... and on the deep learning side 146
3 Perspectives 147
3.1 Power system perspectives 147
3.2 Deep learning perspectives 148

References 151

Appendices 173

Appendix A Discussion of the environmental impact 175
1 Literature and proposed approach 175
2 Results 176
2.1 Energy consumption 176
2.2 Environmental impact 177

Appendix B The BDAPPV training dataset 179
1 Additional details on the data extraction and the raw data records . . 179
1.1 Extraction of the raw data 179
1.2 Data and quality checks 182
2 Crowdsourcing campaign analysis 187

Appendix C Supplementary materials 191
1 Supplementary material to chapter 2 191
1.1 Additional plots of the distribution of the tilt angle 191
1.2 Complementary regressions of the city's coordinates on its error 192
2 Supplementary material to chapter 3 193
2.1 Additional visualization of the dissappearance of important components 193

2.2 Additional examples of the extraction of the critical component 195
2.3 Details on the data augmentation techniques 196
2.4 Training results 196
3 Supplementary material to chapter 4 197
3.1 Methods evaluated for constructing `PyPVRoof` 197
3.2 Example of shifted thumbnails generated from a larger image . 203
4 Supplementary material to chapter 5 203
4.1 Examples of reports generated to inspect the quality of the PV power measurements from BDPV 204
4.2 Using the LiDAR and the ground truth power measurements to evaluate the accuracy of the tilt angles reported in BDPV 205

Appendix D Introduction to machine learning **209**
1 Notations and definitions 209
2 Empirical risk minimization 211
3 Excess risk bounds 212
4 From the excess risk to generalization bounds 212

Appendix E Publications associated with this thesis **215**
1 Peer-reviewed journal papers 215
2 Conference and workshop papers (peer-reviewed) 216
3 Communications in conferences 217
4 Submitted works 219
5 Preprints 219
6 Communication in expert groups 220
7 Posters 220

List of figures

1 Croissance attendue du solaire photovoltaïque (et part dans le mix électrique) en 2035 (PPE, 2020) et en 2050 (RTE France, 2022). xxxi
2 Exemples d'orthoimages de l'IGN. xxxv
3 Histogramme des puissances installées des installations répertoriées dans BDPV. xxxvii
4 Diagramme de Venn résumant les caractéristiques attendues et satisfaites par les différentes sources de données sur le PV. xxxviii
5 Diagramme présentant le principe de fonctionnement de la précision aval (DTA), notre méthode pour surveiller les prédictions du modèle. D'après Kasmi et al. (2022a). xxxix
6 Explications du modèle de classification générées en utilisant la méthode GradCAM (Selvaraju et al., 2020) pour quelques vrais positifs, faux positifs, vrais négatifs et faux négatifs. Plus une zone est rouge, plus elle contribue à la prédiction du modèle pour la classe considérée. xlii
7 Décomposition en différentes échelles d'un panneau PV vu sur une orthoimage. Adapté de Kasmi et al. (2023b). xliv
8 Image et transformée en ondelettes dyadique à deux niveaux associée, avec des indications pour interpréter la transformée en ondelettes de l'image. Les termes "horizontaux", "diagonaux" et "verticaux" indiquent la direction des coefficients de détail. La direction est la même à tous les niveaux. xlv
9 Diagramme de notre méthode d'attribution, la *Wavelet sCale Attribution Method* (WCAM). Adapté de Kasmi et al. (2023a). xlvi
10 Décomposition dans l'espace des échelles des prédictions d'un modèle. Adapté de Kasmi et al. (2023b). xlvii
11 Image originale, information suffisante, dernier "faux négatif" et composant critique, dans l'espace des images (haut) et dans l'espace des échelles (bas). xlviii
12 Exemple d'images de BDAPPV (Kasmi et al., 2023d) issues de Google (gauche) et de l'IGN (droite). xlix
13 Prédictions sur l'image Google (gauche, rangée supérieure) et l'image IGN (droite, rangée supérieure) et WCAMs associés (rangée inférieure). Plus la région en surbrillance est claire, plus la prédiction est importante. La colonne la plus à droite présente les composants les plus importants de l'image Google et les composants critiques. Adapté de Kasmi et al. (2023b). l
14 Diagramme de DeepPVMapper. liv

15 Comportement de l'erreur d'agrégation des estimations des courbes de production PV dans le cas d'une inclinaison et d'une orientation estimées avec DeepPVMapper. lix

1.1 Overview of mitigation options and their estimated ranges of costs and potentials in 2030. Source: IPCC (2021b). 2
1.2 Typology of PV installations. Rows correspond to classes of installed capacities and columns to classes of tilt angles (in degrees). Adapted from Saint-Drenan et al. (2015). 4
1.3 Expected PV share growth according to the PPE and RTE's Energy Pathways 2050. 6

2.1 Examples of different types of orthoimagery. Sources: USGS (2024), IGN (2024a), and the ESA (2024). USGS and IGN are updated every three years on a rolling basis, and Pléiades from ESA is updated twice a day. 16
2.2 A) An orthophoto rectified over a terrain model. The church is not moved to its correct position. B) Orthophoto based on a city model. The church is rectified to its correct location, but a "ghost image" is left on the terrain. C) Same as B, but the obscured area has been detected. D) True orthophoto, where the obscured area has been replaced with imagery from other images. Taken from Nielsen (2004). 17
2.3 Screenshot of the QGIS software displaying building layers from the BD TOPO. 18
2.4 Examples of rasters extracted from the photogrammetry DSM (bottom row) and the LiDAR DSM (upper row) The middle row presents the RGB image associated with these digital surface models. We can see that despite the same GSD, the LiDAR data is more accurate due to its finer effective resolution. Source: IGN. 19
2.5 Distribution of the installed capacities registered in BDPV. Source: BDPV. 21
2.6 Comparison of the distribution of the installed capacity at the scale of the departements reported in BDPV (upper left) the RNI (lower left), and the relative spread of the two (upper right). On the relative spread plot, the red means that the reported installed capacity at the size of the departement is higher in BDPV than in the RNI. Blue means that the reported installed capacity is lower in BDPV than in the RNI. 23
2.7 Flowchart of the BDAPPV dataset construction workflow. Source: Kasmi et al. (2023d). 25
2.8 Overview of the principle of the DTA with the RNI as an example. GIS, "Geographical Information System, " corresponds to georeferenced PV data such as the RNI, RTE's registry, or BDPV. 28
2.9 Flowchart of our algorithm based on the literature for mapping rooftop PV. Source: Kasmi et al. (2022a). 30
2.10 Visualization of the lookup table used in this study to estimate the tilt angle of the installations. Taken from Kasmi et al. (2022a). 32
2.11 Mapping area over which we deployed our model. The numbers correspond to the number of the departements used to identify them in Table 2.5. 34

2.12 Comparison of the distribution of the azimuth (left) and tilt (right) angles obtained from our mapping algorithm and BDPV. The azimuth angle distribution is obtained for the 11 departements and the tilt angle distribution is computed for the departement Loire-Atlantique (44). . . 36
2.13 Comparison of the distribution of the installed capacity at the scale of the cities reported from our registry (upper left), the RNI (lower left), and the relative spread of the two (upper right) in the departement Isère (38). On the relative spread plot, the red means that the reported installed capacity at the size of the city is higher in our registry than in the RNI. Blue means that the reported installed capacity is lower in our registry than in the RNI. 38
2.14 Boxplots of the distributions of the APE metrics across four departement. 39
2.15 Model explanations using the GradCAM (Selvaraju et al., 2020) for some true positives, false positives, true negatives and false negatives. The redder, the higher the contribution of an image region to the predicted class (1 for true and false positives, 0 for true and false negatives). 42
2.16 Predicted probabilities of the true positives, false positives, true negatives, and false negatives on the BDAPPV test dataset. 43
2.17 False detection identification with an example for the city of Cobrieux (Nord). 44

3.1 Decomposition of the scales of a PV panel. Source: Kasmi et al. (2023b). 49
3.2 Examples of saliency maps, computed by Simonyan et al. (2014). Taken from Simonyan et al. (2014). 50
3.3 Correspondence between the scales in the wavelet domain and the frequency ranges in the Fourier domain. In this diagram, f_s corresponds to the highest frequency contained in the signal. Inspired by Chen et al. (2019). 52
3.4 Image and associated two-level dyadic wavelet transform with indications to interpret the wavelet transform of the image. "Horizontal," "diagonal," and "vertical" indicate the direction of the detail coefficients. The direction is the same at all levels. 53
3.5 Flowchart of the wavelet scale attribution method (WCAM). Source: Kasmi et al. (2023a). 55
3.6 Workflow on a grayscale image and for a 2-level wavelet transform. We first compute the discrete wavelet transform of the image and then apply a mask on the discrete wavelet transform (DWT). It yields the perturbed DWT, which we invert to generate the perturbed image. We evaluate the model on the perturbed image. 56
3.7 Decomposition of a prediction from the pixel domain (i) into the wavelet domain (ii) with the WCAM. Source: Kasmi et al. (2023a). 57

3.8 Representation of the scales of the WCAM as frequencies. Levels (numbered from (0 to 4) indicate the scales, from the coarser (i.e., lowest frequencies) to the finest (i.e., highest frequencies. The level 0 or "a" corresponds to the approximation coefficients. Labels "h," "v," and "d" correspond to the horizontal, vertical, and diagonal details, respectively. The rightmost index plots the cumulative curve. "AT," "RT," and "ST" stand for adversarial, robust, and standard training, respectively. The dotted line indicates the concentration towards coarser scales associated with better robustness. Adapted from Kasmi et al. (2023a). . 58
3.9 Decomposition in the space-scale domain of PV panel predictions (true positives). Adapted from Kasmi et al. (2023b). 59
3.10 Examples of false positives on IGN and corresponding WCAM. Adapted from Kasmi et al. (2023b). 60
3.11 Sufficient images reconstructed from the WCAM. 62
3.12 Identification of the critical component (highlighted in white on the "Critical component" plot on the bottom right of the image. Without this component, the model does not predict the PV panel. The sufficient image is the image reconstructed with the minimal set of components. 63
3.13 Comparison of model explanations using the GradCAM (Selvaraju et al., 2020) and the WCAM (Kasmi et al., 2023a) correct and incorrect predictions. The WCAM shows that different scales contribute to the prediction and that when focusing on single scales, shortcuts such as gridded-like patterns can arise. 64
3.14 Examples of images used in this experiment. 65
3.15 Evolution of the predicted probabilities for images depicting a PV panel on the Google test set and the corresponding images on the IGN test set. The predicted probability completely flips over when the model no longer recognizes the PV panel. Source: Kasmi et al. (2023b). 67
3.16 Predictions on Google image (left, upper row) and IGN image (middle, upper row) and associated WCAMs (bottom row). The brighter the highlighted region for the prediction, the more important it is. The rightmost column plots the most important components of Google Images and the critical components. Adapted from Kasmi et al. (2023b). 68
3.17 Image acquisition process. SN: Shannon-Nyquist. 69
3.18 Examples of varying acquisition conditions modeled after the Gaussian blur and noise model. We assume that the object O corresponds to the image without alteration (leftmost column). 71
3.19 Evolution of the F1 score on the test set depending on the noise and blur levels of the test set. Each letter marks a combination whose confusion matrix is unwrapped in Table 3.2. 72
3.20 Accuracy of a model's predictions under different levels of blur and noise and plot of some corrupted images and their associated WCAMs. We can see that for the same result at the macroscopic scale (a lower F1-score caused by a rise in the false negatives), the model behaves in two different ways at the microscopic level. If blurring increases, it tends to look for new components. If the noise increases, it tends to be disrupted by this noise and to focus on higher frequencies than if there were no noise. 74

3.21 Fast adversarial sample generation . . . 75
3.22 Examples of images coming from the ImageNet-C dataset of Hendrycks and Dietterich (2019). Each corruption has different levels of severity. Source: Hendrycks and Dietterich (2019). . . . 76
3.23 Illustration of augmented images with the selected data augmentation techniques. . . . 78
3.24 Illustration of augmented images with our data augmentation techniques. The colored pixels that appear with the Blurring + WP augmentation are a consequence of the fact that we hide some information in channels and not others. . . . 80
3.25 WCAMs on IGN of models trained on Google with different augmentation techniques. . . . 82

4.1 Original DeepSolar pipeline. Source: Yu et al. (2018). . . . 87
4.2 Semi-supervised approach of DeepSolar. The left column contains original images. The middle column contains the original images' Class Activation Maps (CAMs) without greedy layer-wise training. The right column is the CAMs of the original images with greedy layer-wise training. Taken from Yu et al. (2018). . . . 88
4.3 Flowchart of the proposed method to extract installations' characteristics. The installed capacity depends on the surface of the PV system, as the installed capacity is equal to the surface area multiplied by the efficiency of the PV modules. Source: Trémenbert et al. (2023). . . . 91
4.4 `PyPVRoof` flowchart when only auxiliary data is available. Adapted from Trémenbert et al. (2023). . . . 92
4.5 Perimeter of the evaluations carried out in this study. Headers in the black boxes correspond to the different modules of the mapping algorithm. . . . 100
4.6 Illustration of the cropping of thumbnails of a size of 224×224 pixels from the raw BDAPPV image with a size of 400×400 pixels to simulate various locations of the PV panel on the image. The black square is always included in the smaller thumbnails, the red dashed lines indicate the boundaries of the thumbnails, such that this black square is contained in the thumbnail. . . . 101
4.7 Average predicted probability (a) and true positives and false positives domains (b) as the center of the thumbnail moves away from the location of the panel. . . . 102
4.8 Illustration of how sampling strategies would cover a tile. All strategies represent $n = 100$ points. . . . 105
4.9 Increase in the number of points to reach a distance of at most d^* between two thumbnail centers according to the deterministic and random (Sobol and Uniform) sampling strategies. . . . 105
4.10 Flowchart of DeepPVMapper. . . . 109
4.11 Comparison of the behavior of DeepSolar (left) and DeepPVMapper (right) on the barn (false positive) that was detected by DeepSolar and avoided by DeepPVMapper. . . . 111

5.1 Localizations and example of power curves contained in our PV measurements dataset. . . . 117

5.2 Example of solar irradiation time series provided by CAMS for an installation located near Toulouse, France. The first day is cloudy, and the second day is sunny. . . . 119
5.3 Sample of 2m air temperature from ERA5. . . . 120
5.4 Illustration of the POA irradiation modeled with our approach. θ indicates angles, "AOI": "angle of incidence" and "SZA": solar zenith angle. The light gray surface is flat, and the dark grey surface is tilted with tilt angle θ. . . . 124
5.5 Power curves generated with our conversion model plotted against the corresponding ground truth measurement for one installation. . . . 125
5.6 Examples of individual characterization error. The red star shows the localization of the true parameter values, where the characterization error is equal to 0 by construction. . . . 129
5.7 30-minute pRMSE of the estimation of the PV power production using DeepPVMapper and the conversion model. The interquantile range plots the range between the 5th and 95th percentile. . . . 132
5.8 Geographical variability of the pRMSE [%] of the PV power estimation depending on the localization of the installation. . . . 133
5.9 Characterization errors as a function of the number of installations for the unbiased case and a set of worst (**(I)** and **(VIII)**) and best cases (**(II)** and **(V)**). . . . 135
5.10 Behavior of the error of the aggregation of the estimations of the PV power curves in the case of tilt and azimuth angles estimated with DeepPVMapper. . . . 136

B.1 Flowchart of the training dataset generation based on the BDPV PV data and crowdsourcing. "GSD" stands for the ground sampling distance, i.e., the distance between the centers of two adjacent pixels measured on the ground. Taken from Kasmi et al. (2023d). . . . 180
B.2 Examples of images from the BDPV training database. . . . 184
B.3 Validation by comparison of the surface estimated from the masks and the surface reported in the PV installations' metadata. Taken from Kasmi et al. (2023d). . . . 186
B.4 Number of installations filtered through the different filtering steps during the association between the masks and the installations' metadata. *During the mask uniqueness step, we account for the fact that (a) not all BDPV installations were identified on images (13,303 were identified on Google images and 7,686 on IGN images) and (b) among these identified installations, some of the masks contained more than one polygon. Adapted from Kasmi et al. (2023d). . . . 187
B.5 Screenshot of the BDAPPV crowdsourcing platform (first phase). . . . 187
B.6 Number of annotations per user. . . . 188
B.7 Number of daily annotations (log scale) during the first crowdsourcing campaign. Light blue indicates the second phase, and dark blue the first phase. . . . 189

C.1 Comparison of the distribution of the azimuth (left) and tilt (right) angles obtained from our mapping algorithm and BDPV. . . . 191
C.2 Comparison of the distribution of the azimuth (left) and tilt (right) angles obtained from our mapping algorithm and BDPV. . . . 192

C.3 Comparison of the distribution of the azimuth (left) and tilt (right) angles obtained from our mapping algorithm and BDPV. 192
C.4 Predictions on Google image (left, upper row) and IGN image (right, upper row) and associated WCAMs (bottom row, displayed in logscale and with the same color scale). The brighter the highlighted region for the prediction, the more important it is. Taken from: Kasmi et al. (2023b). 194
C.5 Identification of the critical component (highlighted in white on the "Critical component" plot on the bottom right of the image. Without this component, the model does not predict the PV panel. The sufficient image is the image reconstructed with the minimal set of components. 195
C.6 Identification of the critical component (highlighted in white on the "Critical component" plot on the bottom right of the image. Without this component, the model does not predict the PV panel. The sufficient image is the image reconstructed with the minimal set of components. 195
C.7 Example of DSM: the rasterization of the LiDAR from the IGN. 198
C.8 Theil-Sen method principle. The plane is deduced from the raster and is parameterized as $z(x, y) = ax + by + c$. φ corresponds to the azimuth angle and θ to the tilt angle. $\overrightarrow{e_x}$, $\overrightarrow{e_y}$, $\overrightarrow{e_z}$ correspond to the canonical basis of $\mathbb{R}^3$. Taken from Trémenbert et al. (2023). 199
C.9 Lookup table for 50×50 grid-points and four surface categories computed for the PV mapping algorithm of Kasmi et al. (2022a). Surface categories correspond to quartiles of the distribution of the surface in the auxiliary data. 200
C.10 Left: clustered linear regression. Right: linear regression with a single coefficient. 202
C.11 Example of thumbnails generated from a 400×400 image with the panel's position on the image being shifted. The coordinates indicate the position of the center of the thumbnail (in pixels relative to the upper-left corner of the image). 203
C.12 Geographical variability of the pRMSE [%] of the PV power estimation depending on the localization of the installation (Oracle). 204
C.13 Example of an installation that passed the QC. 205
C.14 Example of an installation that failed the QC. 205
C.15 Fit of the PV power estimation with different model parameterization. 206
C.16 Report comparing the estimation using BDPV parameters and LiDAR parameters. The pRMSEs are reported in %, and the tilt and azimuth angles are reported in degrees. 208

List of Tables

1 Performance en classification et en segmentation. Plus la GSD est basse, plus l'image est détaillée. Les meilleurs résultats sont en **gras**. xl
2 Précision aval (DTA) sur la zone cartographiée. Les valeurs entre parenthèses correspondent aux résultats sans filtrage par bâtiments. La ligne "Test" considère les images de test de l'ensemble de données d'entraînement comme une seule ville. k_i et C_i indiquent respectivement le nombre d'installations et la puissance installée. Un chapeau indique l'estimation par notre algorithme. Les valeurs entre parenthèses indiquent la précision avec et sans filtrage par bâtiment. Source : Kasmi et al. (2022a). xli
3 **Score F1** et décomposition en vrai positifs, vrais négatifs, faux positifs et faux négatifs des prédictions d'un modèle entraîné sur des images Google et déployé sur des images IGN (même scène, même résolution, mais différentes conditions d'acquisition). Adapté de Kasmi et al. (2023b). l
4 **Score F1** et décomposition en vrais positifs, vrais négatifs, faux positifs et faux négatifs pour des modèles entraînés sur Google avec différentes techniques d'atténuation. L'évaluation est conduite sur les images IGN. L'Oracle désigne un modèle entraîné sur des images IGN sans technique d'atténuation. Le meilleur résultat est en **gras**, le deuxième meilleur souligné. lii
5 Performances mesurées avec les métriques DTA de différentes configurations de l'algorithme. Les meilleurs résultats sont en **gras** et les seconds meilleurs soulignés. lv
6 Comparaison de la RMSE [W] et de la pRMSE [%] (entre parenthèses) de l'estimation à l'échelle de l'installation individuelle avec les paramètres de DeepPVMapper. Les meilleurs résultats sont en **gras** et les seconds meilleurs soulignés. n indique le nombre d'installations utilisées dans cette étude. lviii

1.1 Overview of the PV observability in France 5
1.2 Outline of the chapters of this thesis and the scientific sub-questions (SQ) they address. 13
1.3 Summary of the publications associated with this thesis. 14

2.1 Data requirements applicable for the technical registry and accessibility from existing sources for PV installations below 36 kW_p. 24
2.2 Overview of the data records of the training dataset BDAPPV. 26

2.3 Training dataset characteristics. 33
2.4 Classification and segmentation accuracy. The lower the GSD, the more detailed the image. Best results are **bolded**. 34
2.5 Downstream task accuracy across the mapping area. Values in parentheses correspond to the results without filtering by buildings. The line "Test" considers the test images of the training dataset as one city. k_i and C_i denote the count of installations and the installed capacity, respectively. The hat indicates the estimation by our algorithm. Source: Kasmi et al. (2022a). 36
2.6 Results of estimating the linear model defined in Equation 2.1 for the dependent variables APE, AIPE, and ratio. 40
2.7 Extract of the registry generated by DeepSolar for the city of Cobrieux (Nord). 44

3.1 **F1 Score** and decomposition in true positives, true negatives, false positives, and false negatives rates of the disentanglement of the distribution shift between the GSD (Google 10 cm/px), the geographical variability (Google OOD) and the acquisition conditions (IGN). Taken from Kasmi et al. (2023b). 66
3.2 **F1 score** and decomposition in terms of true positives, true negatives, false negatives, and false positives. Each line corresponds to a given level of corruption of the dataset, parameterized by the noise and blur level, σ_n and σ_b. 72
3.3 **F1 Score** and decomposition in true positives, true negatives, false positives, and false negatives for models trained on Google with different mitigation strategies. Evaluation of IGN images. The Oracle corresponds to a model trained on IGN images with standard augmentations. The best results are **bolded** and second best underlined. . . . 79
3.4 **F1 Score** and decomposition in true positives, true negatives, false positives, and false negatives for models trained on Google with different mitigation strategies. Evaluation on IGN images. The Oracle corresponds to a model trained on IGN images with standard augmentations. The best results are **bolded** and second best underlined. . . . 81

4.1 Overview of the methods considered for building `PyPVRoof`. The column "data requirements" indicates the additional requirements besides the geolocalized polygon. 92
4.2 Performance metrics for the estimation of the tilt angle. The two lines for the Theil-Sen method report the accuracy results whether photogrammetry DSM or LiDAR DSM are passed as inputs. The best results are **bolded** and second best underlined. 94
4.3 Performance metrics for the estimation of the azimuth angle. The two lines for the Theil-Sen method report the accuracy results whether photogrammetry DSM or LiDAR DSM are passed as inputs. The best results are **bolded** and second best underlined. 94
4.4 Performance metrics for the estimation of the installed capacity. Column θ indicates the method used to derive the tilt necessary to compute the estimated surface S_{est}, taken as input to estimate the installed capacity. "RF" indicates random forest, and "TS" Theil-Sen. The best results are **bolded** and second best underlined. 95

4.5 Accuracy results of PyPVRoof's methods for PV panels characteristics extraction. . . . 96
4.6 Benchmark of various model architectures on BDAPPV (Kasmi et al., 2023d). The best results are **bolded** and the second best results underlined. . . . 107
4.7 Benchmark of various model architectures for the segmentation branch. The best results are **bolded** and second best underlined. . . . 107
4.8 Accuracy results (DTA) on a 120km² area around Lyon, representative of the operational conditions. The best results are **bolded** and the second best results underlined. . . . 108
4.9 Computational gains brought by the sampling. The best results are **bolded**. . . . 109
4.10 Extract of the registry generated by DeepSolar and DeepPVMapper for the city of Cobrieux (Nord). . . . 110

5.1 Summary statistics on the PV power measurements. The load factor is the ratio between the PV power production at time t and the installed capacity of the installation. . . . 118
5.2 Description of the main variables included in CAMS. . . . 119
5.3 Descriptive statistics of the PV systems' characteristics extracted from the PV registry for the systems used in this study. . . . 121
5.4 Set of minimal PV system characteristics for the conversion model Parameters that we input are **bolded**, advanced parameters are in *italics*. Source: Dobos (2014). . . . 122
5.5 Accuracy (RMSE in [W] and pRMSE in [%] in parenthesis) of the statistical models considered in this study on their test dataset. n indicates the number of samples (here: number of power measurements in the test set). . . . 127
5.6 Summary of the cases for aggregating the characterization errors from the installation level to the representative cell level. . . . 130
5.7 Comparison of the RMSE [W] and pRMSE [%] (in parenthesis) of the estimation at the individual installation scale with parameters from DeepPVMapper. Best results are **bolded** and second best underlined. n indicates the number of installations used in this study. . . . 131
5.8 pRMSE [%] of the aggregated rooftop PV power production estimation under different estimation biases of the PV systems' characteristics. . 133

A.1 Computation of C_{train} and $C_{inference}$ for some selected variants of the mapping algorithm. . . . 176
A.2 Total energy consumption in Wh of deploying variants of our pipeline. Best results are **bolded**. . . . 177
A.3 Carbon intensity of DeepPVMapper (ResNet + Sampling). Source of the carbon intensities: Our World in Data (2024). . . . 177

B.1 Data attributes and description of the metadata.csv data file. Taken from Kasmi et al. (2023d). . . . 183
B.2 Summary statistics of the contributions during the crowdsourcing campaigns. Source: Kasmi et al. (2023d). . . . 188

C.1 Results of estimating the linear model defined in Equation 2.1 for the dependent variables APE, AIPE, and ratio. 193
C.2 Results of estimating the linear model defined in Equation 2.1 for the dependent variables APE, AIPE, and ratio. 193
C.3 **F1 Score** and decomposition in true positives, true negatives, false positives, and false negatives for models trained on Google images with different strategies to mitigate the sensitivity to acquisition conditions. Evaluation computed on the Google (source) dataset. 197
C.4 Accuracy of the PV power production estimation using parameterizations from the LiDAR data. Best results are **bolded**. 206
C.5 Average pRMSE for different configurations of parameterizations of the conversion model. 207

Acronyms and abbreviations

CO_2 Carbon dioxyde

CO_2e Carbon dioxyde equivalent

GW_p Gigawatt peak

MW_p Megawatt peak

W_p Watt peak

kW_p Kilowatt peak

AIPE Average installation percentage error

AL Angular losses

APE Average percentage error

API Application programming interface

BCE Binary Cross Entropy

BDAPPV *Base de données d'apprentissage profond photovoltaïque* (PV database for deep learning)

BDPV *Base de données photovoltaïque* (PV database)

BHI Direct (beam) horizontal irradiance or irradiation

BNI Direct (beam) normal irradiance or irradiation

C3S Copernicus Climat Change Service

CAM Class attribution map or class attribution mapping

CAMS Copernicus Atmospheric Monitoring Service

CCD Charge-coupled device

CMOS Complementary metal oxide semiconductor

CNN Convolutional neural network

CRAFT Concept Recursive Activation FacTorization for Explainability

CWT Continuous wavelet transform

DC Direct current

DHI Diffuse horizontal irradiance or irradiation

DRI Detection-Recognition-Identification

DSM Digital Surface Model

DSO Distribution system operator

DTA Downstream Task Accuracy

DWT Discrete wavelet transform

ECMWF European Centre for Medium-Range Weather Forecasts

ERA5 ECMWF reanalysis v5

ERM Empirical risk minimization

EU European Union

EVA Explaining using Verified perturbation Analysis

GAP Global average pooling

GEE Google Earth Engine

GHI Global horizontal irradiance or irradiation

GIS Geographical Information System

GradCAM Gradient class activation map or gradient class activation mapping

GSD Ground Sampling Distance

HSIC Hilbert-Schmidt Independence Criterion

IAM Incident angle modifier

IEA International Energy Agency

IGN *Institution Géographique National* (French National Institute of Geographical and Forest Information)

IoU Intersection-over-Union

IPCC Intergovernmental Panel on Climate Change

IS Importance sampling

KPI Key performance index

kV Kilovolt

LiDAR Light Detection and Ranging

LIME Local Interpretable Model-agnostic Explanations

LUT Look-up table

MAE Mean absolute error

MAPE Mean average percentage error or mean absolute percentile error

MCC Matthews correlation coefficient

ME Mean error

MSE Mean squared error

MSG Meteosat Second Generation

ODRE *Open Data Réseau Energies*

OSM OpenStreetMap

PCG permuted congruential generator

PIL Python imaging library

POA Plane-of-array

PPE *Programmation pluriannuelle de l'énergie* (pluriannual energy plan)

pRMSE Percentage Root Mean Square Error (or normalized RMSE)
PSF Point spread function
PV Photovoltaic
QGIS Quantum Geographic Information System
QMC Quasi-Monte Carlo
RGB Red-Green-Blue
RISE Randomized Input Sampling for Explanation of Black-box Models
RMSE Root Mean Square Error
RNI *Registre National d'Installations* (national registry of installations)
ROI Region of interest
RTE *Réseau de transport d'électricité*
SAR Synthetic Aperture Radar
SNBC *Stratégie nationale bas carbone* (national low carbon strategy)
SOTA State-Of-The-Art
STC Standard test conditions
SVM Support vector machine
SZA Solar zenith angle
TOA Top of atmosphere
TOD Time-of-the-day
TSI Total Sobol indices
TSO Transmission system operator
TWh Terawatt hour
USGS United States Geological Survey
WCAM Wavelet sCale Attribution Method

Résumé étendu

1 Introduction

1.1 Contexte et motivation

Changement climatique et électrification D'après le Groupe d'experts intergouvernemental sur l'évolution du climat (GIEC), la lutte contre le changement climatique nécessite une baisse drastique des émissions de gaz à effet de serre, et en particulier des émissions de dioxyde de carbone (IPCC, 2021a). La réduction des emissions de dioxyde de carbone (CO_2) peut se faire selon deux leviers : les économies d'énergie (sobriété) et la décarbonation. Les mesures de sobriété correspondent par exemple à l'amélioration de l'isolation des bâtiments, ou le fait de privilégier les transports en commun pour les déplacements du quotidien. La décarbonation implique quant à elle d'électrifier massivement certains usages, en particulier les transports. Par conséquent, la transition énergétique entraînera une augmentation de la consommation d'électricité. L'Agence internationale de l'énergie (AIE) prédit une augmentation de la part de l'électricité dans la demande d'énergie finale de 4% par an pour atteindre les objectifs de décarbonation (IEA, 2023). En France, la consommation d'électricité pourrait passer de 459,3 TWh en 2022 à 580 à 640 TWh en 2035 (RTE France, 2023). La décarbonation suppose donc une hausse de la production d'électricité.

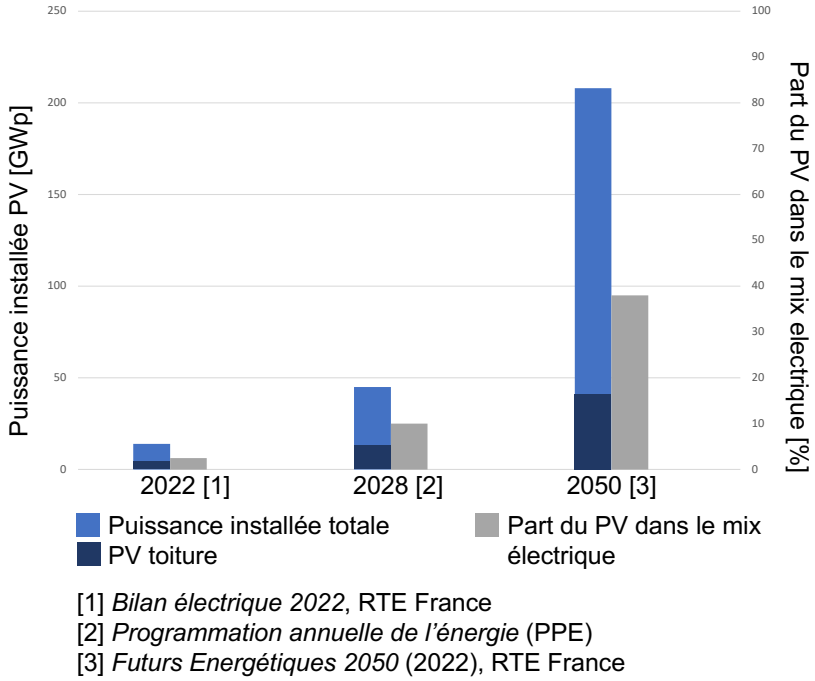
L'électricité est un vecteur énergétique résultant de la conversion d'une source d'énergie primaire en énergie électrique. Les sources primaires peuvent être le gaz, le fioul, le nucléaire ou les renouvelables ; ces dernières différant selon leur intensité en carbone. La décarbonation du secteur électrique implique donc de favoriser le déploiment d'énergies bas carbone telles que les renouvelables (hydraulique, solaire, éolien, biomasse). Selon le GIEC, l'énergie éolienne et le solaire photovoltaïque (PV) constituent les deux leviers les plus puissants pour réduire les émissions de CO_2 d'ici à 2030 (IPCC, 2021b).

Solaire et éolien : deux sources météo-dépendantes La production d'électricité éolienne et PV dépend des conditions météorologiques. Elle varie donc de manière importante, quelle que soit l'échelle spatiale ou temporelle considérée. Un système électrique comprenant une large part de PV et d'éolien est ainsi plus sensible au climat et sujet aux incertitudes de production. Afin de limiter les consé-

quences de ces incertitudes sur le système électrique, telles que les congestions ou la hausse du niveau de marges requises (Pierro et al., 2022; RTE France and IEA, 2021), il est nécessaire de mesurer ou d'estimer avec précision la production d'électricité solaire et éolienne. Je définis l'observabilité comme la capacité du gestionnaire du réseau de transport d'électricité (GRT) à estimer avec précision la production en temps réel et future d'une unité de production. En pratique, le GRT mesure en temps réel la production (il dispose de télémesures), ou a accès à des relevés *a posteriori* (la télérelève, généralement dans le mois suivant le temps réel). La télémesure ou la télérelève permettent de calibrer des modèles d'estimation et de prévision de production.

Disposer *a minima* de la télérelève de la production d'électricité issue des centrales éoliennes et solaires est au fondement de l'observabilité de ces sources de production. Si le parc éolien est observé de manière homogène, ce n'est pas le cas du parc photovoltaïque. Les installations PV sont caractérisées par une très grande variabilité de puissances installées, allant de quelques kW_c, à plusieurs centaines de MW_c. Actuellement, le GRT ne dispose pas de télérelève pour les installations de moins de 36 kW_c, qui représentaient 94% des installations et 22% de la puissance installée PV en France en 2023 : le PV sous 36 kW_c souffre donc d'un manque d'observabilité.

Objectif industriel : améliorer l'observabilité du PV en toiture Le manque d'observabilité sera de plus en plus préoccupant dans le contexte de la croissance rapide de la capacité photovoltaïque installée. Comme illustré sur la figure 1, nous pouvons voir que la puissance installée PV pourrait atteindre jusqu'à 200 GW_c en 2050 (RTE France, 2022). La Programmation Pluriannuelle de l'Energie (PPE) vise déjà à atteindre 35 et 45 GW_c de capacité photovoltaïque installée en 2029. Ces scénarios et objectifs supposent un taux de déploiement constant pour toutes les typologies d'installations PV, ce qui signifie que jusqu'à 40 GW_c (c'est-à-dire les deux tiers du parc nucléaire français actuel) pourraient ne pas être observés d'ici 2050 si l'observabilité du PV reste constante. L'objectif industriel de cette thèse est d'introduire une méthode permettant d'améliorer l'observabilité des petites installations PV en toiture (c'est-à-dire les installations photovoltaïques d'une puissance installée inférieure à 36 kW_c).



[1] *Bilan électrique 2022*, RTE France
[2] *Programmation annuelle de l'énergie* (PPE)
[3] *Futurs Energétiques 2050* (2022), RTE France

Figure 1 – Croissance attendue du solaire photovoltaïque (et part dans le mix électrique) en 2035 (PPE, 2020) et en 2050 (RTE France, 2022).

Les données disponibles sur le parc photovoltaïque en toiture sont incomplètes. Au mieux, seule la puissance installée de chaque installation dans chaque commune est connue. Or, pour améliorer l'observabilité du PV sur toiture, il est nécessaire de disposer d'estimations précises de la production issue de ces systèmes. Un des pré-requis nécessaires à l'estimation de production est l'acquisition d'informations sur les caractéristiques techniques et la localisation ponctuelle (latitude et longitude) des systèmes PV, en plus de leur puissance installée. Ainsi, l'amélioration de l'observabilité PV implique la constitution d'un **registre technique** (ou registre) répertoriant la localisation, l'inclinaison, l'azimuth, et la puissance installée du plus grand nombre d'installations photovoltaïques en toiture possible.

1.2 Revue de littérature et question scientifique

Télédétection d'installations PV sur des orthoimages Le manque d'informations sur les instalations PV de petite taille est un problème récurrent dans de nombreux pays (Malof et al., 2015, 2019). Je renvoie le lecteur aux revues de littérature de Puttemans et al. (2016), de Hoog et al. (2020) et Arnaudo et al. (2023) pour une présentation exhaustive des travaux sur le sujet. Les premiers travaux portant sur la télédétection d'installations PV remontent à 2015 avec Malof et al.. Dans un premier temps, des bases de données contenant des annotations d'installations sur

des orthoimages[1] (Bradbury et al., 2016) ont été constituées et différentes méthodes pour identifier quels pixels de correspondaient à des panneaux solaires ont été testées. Certains auteurs ont étudié des méthodes consistant à calculer manuellement des statistiques pour chaque pixel et à classifier les pixels en fonction de ces statistiques (Malof et al., 2016a,c; Li et al., 2020; Wang et al., 2018; Devarajan et al., 2016), tandis que d'autres ont utilisé des réseaux convolutifs profonds (Golovko et al., 2017, 2018; Yuan et al., 2016; Huang et al., 2018; Camilo et al., 2018; Malof et al., 2019). La performance de ces algorithmes est évaluée avec le score F1[2], et les études comparatives ont montré la supériorité des performances des algorithmes utilisant des réseaux convolutifs sur les algorithmes se fondant sur une extraction manuelle des statistiques de l'image. Ainsi par exemple, sur une même base de données, la méthode "manuelle" de Malof et al. (2016a) atteint un score F1 d'à peine 0.6 tandis que le réseau convolutif de Malof et al. (2019) dépasse 0.8.

Le projet DeepSolar (Yu et al., 2018) constitue une étape importante dans la cartographie d'installations PV. Ce projet utilise des réseaux convolutifs profonds pour détecter les installations et estimer leur surface. Ce modèle a été déployé à l'échelle des Etats-Unis et les auteurs ont rapporté une erreur dans l'estimation de la superficie inférieure à 5%[3]. De nombreux travaux ont réutilisé les modèles de DeepSolar pour cartographier des régions ou des pays, notamment en Europe : Kausika et al. (2021) aux Pays-Bas, Mayer et al. (2020, 2022) en Rhénanie du Nord-Westphalie (Allemagne), Frimane et al. (2023); Lindahl et al. (2023) en Suède, Arnaudo et al. (2023) pour le Nord de l'Italie.

Limites des approches existantes et question scientifique En dépit de leurs performances, les méthodes actuelles d'apprentissage profond ne peuvent pas être appliquées directement sur une nouvelle région ni être appliquées sur de nouvelles données pour mettre à jour un registre existant (De Jong et al., 2020; Arnaudo et al., 2023). Cette faible capacité de généralisation sur des nouvelles données est identifiée par De Jong et al. (2020) comme le principal frein à l'utilisation de ces méthodes pour constituer des statistiques officielles.

La mauvaise capacité de généralisation est connue dans la littérature en ap-

1. Les orthoimages sont des images aériennes ou satellitaires dont la géométrie a été redressée de sorte que chaque point soit superposable à une carte plane qui lui correspond. En d'autres termes, une orthophotographie semble être prise à la verticale de tous les points qu'elle figure, ces points étant situés sur un terrain parfaitement plat

2. Le score F1 mesure la performance d'un classifieur binaire. Un classifieur parfait a un score de 1. Il correspond à la moyenne harmonique entre la précision et le rappel d'un classifieur. Les métriques de performances sont définies au chapitre 4.

3. La métrique utilisée par Yu et al. (2018) est l'erreur relative moyenne (MRE), définie comme suit :

$$MRE = \frac{\sum_{i=1}^{\#\text{vrais positifs}} \text{vraie aire}_i - \text{aire estimée}_i}{\sum_{i=1}^{\#\text{vrais positifs}} \text{vraie aire}_i}$$

prentissage profond comme la sensibilité aux *variations statistiques*[4] (**distribution shifts**), liée au fait que les statistiques des données d'apprentissage diffèrent des statistiques des données de test ou de déploiement (Koh and Liang, 2017). Cette sensibilité entraîne des baisses de performances imprévisibles ; ce qui implique que les mesures de performances sur les données d'entraînement ne sont pas représentatives des performances en conditions réelles. Ce problème est critique au-delà de la télédétection d'installations et se pose dans tous les domaines applicatifs des méthodes d'apprentissage profond, comme par exemple le diagnostic médical (Pooch et al., 2020) ou la conduite autonome (Sun et al., 2022b).

De nombreuses méthodes ont été introduites pour atténuer les effets des variations statistiques. Je renvoie le lecteur à Zhou et al. (2023); Tuia et al. (2016); Guan and Liu (2022); Csurka (2017); Csurka et al. (2021) pour des revues de ces méthodes dans différents contextes. Je désigne les méthodes visant à atténuer la sensibilité aux changements de distribution comme des méthodes de *réalignement de domaine* (**domain adaptation**, Saenko et al., 2010). L'idée générale est qu'un modèle est entraîné sur un jeu de données d'entraînement source (par exemple, des images d'installations photovoltaïques en France) et est déployé sur une ou plusieurs bases de données cibles (par exemple, une nouvelle livrée d'images ou un autre pays). Gulrajani and Lopez-Paz (2021) ont montré que la méthode standard de minimisation empirique du risque (ERM, Vapnik,1999) pouvait être aussi performante dans ce contexte que des méthodes de réalignement de domaine, tout en étant beaucoup plus simple à implémenter en pratique.

Par conséquent, plutôt que d'introduire une nouvelle méthode de réalignement de domaine, je m'attacherai à étudier pourquoi les variations statistiques affectent les performances des modèles d'apprentissage profond en vue d'évaluer si leurs prédictions sont *fiables*. Je définis la **fiablité** d'un modèle comme comme étant la résultante de trois facteurs :

— La pertinence par l'auditabilité de son processus de décision : on veut pouvoir savoir si une prédiction ponctuelle (Schulam and Saria, 2019) se fonde sur de bonnes raisons (Ross et al., 2017) ou non ;

— La robustesse de ce processus : on veut s'assurer que ce processus de décision est invariant à des variations statistiques, étant donné que ces dernières arrivent nécessairement dans un contexte opérationnel (Peng et al., 2017) ;

— Le contrôle ou le suivi de la qualité des prédictions, afin d'identifier les cas où le modèle fait de mauvaises prédictions (Schulam and Saria, 2019). Par contrôle, j'entends la capacité de soumettre à un examen systématique de l'utilisateur la qualité des données produites par le registre.

4. Par variations statistiques, j'entends à la fois des différences dans les statistiques des distributions, leur domain de définition, ou bien des différences dans la distribution des différentes classes entre les différentes sources de données.

L'utilisation d'orthoimages et d'algorithmes d'apprentissage profond est une approche prometteuse pour cartographier les installations photovoltaïques sur toiture. Cependant, ces données et cette méthodologie doivent répondre à des normes précises concernant leur qualité et leur fiabilité ; normes qui ne sont actuellement pas satisfaites. Cette thèse vise à définir quelles normes doivent être respectées et à introduire une méthodologie pour évaluer si les systèmes de cartographie fondés sur l'apprentissage profond peuvent les respecter. À cette fin, j'aborderai la question scientifique générale suivante :

L'utilisation d'algorithmes d'apprentissage profond et d'orthoimages est-elle une méthode adaptée à la construction d'un registre technique national d'installations photovoltaïques (PV) sur toiture destiné à améliorer l'observabilité de la production PV en France ? Pour répondre à cette question, je traite trois sous-questions dont j'expose dans les sections ci-après les enjeux, la méthodologie et les principaux résultats obtenus.

2 Première sous-question : quels standards de qualité le registre technique d'installations doit-il satisfaire et comment vérifier qu'il les remplit ?

Cette première sous-question est traitée dans le chapitre 2. Nous passons tout d'abord en revue les données disponibles sur le parc PV diffus et les données nécessaires pour le cartographier à partir d'orthoimages et d'algorithmes d'apprentissage profond. Je définis ensuite les standards de qualité et notre méthode pour évaluer nos données à l'aune de ces critères et ainsi en assurer le contrôle. Je présente enfin les résultats de l'évaluation des données issues d'un modèle état-de-l'art selon ces critères. Nous analysons ensuite empiriquement les décisions du modèle afin de motiver le besoin d'auditabilité, objet de la deuxième sous-question. Les contributions de ce chapitre sont la précision aval (*downstream task accuracy*, Kasmi et al., 2022a), qui permet de contrôler la qualité des prédictions d'un modèle d'apprentissage profond au cours de son déploiement opérationnel et la base de données BDAPPV (Base de données d'apprentissage profond photovoltaïque, Kasmi et al., 2023d), sur laquelle nous avons entrainé nos modèles de classification et de segmentation destinés à être déployés en France.

2.1 Données disponibles et manquantes

2.1.1 Données géographiques

La principale source de données utilisée pour cette étude est la base de données BD ORTHO[5] de l'IGN (IGN, 2024a). La base BD ORTHO contient des orthoimages aériennes couvrant l'ensemble du territoire français. Les données sont livrées pour chaque département. La résolution (GSD, *ground sampling distance*) de ces images est de 20 cm/pixel et la fréquence de rafraîchissement est de trois ans. Cependant, les images sont mises à jour "au fil de l'eau", de sorte que de nouvelles images sont disponibles tous les mois. Ces données sont accessibles sous licence ouverte. La figure 2 présente des échantillons d'orthoimages issues de la BD ORTHO. Nous utilisons aussi la BD TOPO[6] (IGN, 2023) de l'IGN. Cette base de donnée répertorie tout le bati français (bâtiments, routes, lignes électriques, etc) sous forme de polygones géolocalisés. Cette information est utile pour regrouper des panneaux PV localisés sur le même bâtiment. Enfin, nous avons à notre disposition des modèles numériques de surface (MNS). Ces modèles indiquent le relief du sol et du sursol, permettant ainsi de calculer la pente des toits des habitations. L'IGN met à disposition les données LiDAR HD[7] (IGN, 2024b), suffisamment précises pour estimer l'inclinaison des toits des maisons individuelles. A l'heure où cette thèse est rédigée, le LiDAR ne couvre pas pas tout le territoire français.

Figure 2 – Exemples d'orthoimages de l'IGN.

2.1.2 Revue des données disponibles sur photovoltaïque en France

Pour la France, nous disposons de trois sources d'informations principales sur le PV en général (centrales au sol et PV toiture) : le registre national d'installations

5. Les données BD ORTHO sont accessibles à l'adresse suivante : https://geoservices.ign.fr/bdortho.
6. Les données BD TOPO sont accessibles à l'adresse suivante : https://geoservices.ign.fr/bdtopo.
7. Les données LiDAR HD sont accessibles à l'adresse suivante : https://geoservices.ign.fr/lidarhd.

(RNI), les données internes de RTE et les données participatives, issues de BDPV et d'OpenStreetMap.

Le registre national d'installations Le registre national d'installations contient l'ensemble des installations de production et de stockage d'électricité raccordées au réseau français. Ces données sont collectées par RTE et accessible sur le portail Open Data Réseaux Energie[8] (ODRE), une plateforme portée par les gestionnaires du réseau de transport d'électricité, de gaz et des réseaux de distributions en France. Pour des raisons de confidentialité, les données sur les petites installations photovoltaïques (moins de 36 kW_c) sont agrégées à la maille de la commune. Ainsi, le RNI nous indique au mieux, pour chaque commune, le nombre d'installations leur puissance installée cumulée. Le RNI est mis à jour tous les trois mois.

Les données internes de RTE Les données internes de RTE répertorient l'ensemble des installations raccordées au réseau de transport et au sein du réseau de distribution d'Enedis, qui couvre 95% du territoire français. Par rapport au RNI, RTE a accès aux données désagrégées pour les installations de moins de 36 kW_c. Cependant, ces données ne répertorient que la puissance installée des systèmes PV et pas leurs charactéristiques techniques.

Les données participatives Nous nous sommes appuyés dans le cadre de cette thèse sur les données de l'association *Base de données photovoltaïque* (BDPV)[9]. Cette association propose aux propriétaires de systèmes PV sur toiture de renseigner les caractéristiques de leur installation afin de savoir si cette dernière fonctionne correctement. Ainsi, la base de données de BDPV contient les caractéristiques détaillées et la localisation de 28 000 installations PV de petite taille, dont 24 000 en France. La figure 3 présente la distribution des puissances installées des petites installations répertoriées dans BDPV. Nous avons également accès aux historiques de production de 1700 installations. Nous pouvons enfin mentionner les données issues d'OpenStreetMap (OSM). OSM est un projet collaboratif visant à constituer une base de donnée géographique en libre accès. La plupart des centrales PV et quelques petites installations PV sont répertoriées dans OSM sous la forme de polygones géolocalisés.

8. Cette plateforme est accessible ici : https://opendata.reseaux-energies.fr/.
9. Site internet de l'association : https://asso.bdpv.fr/.

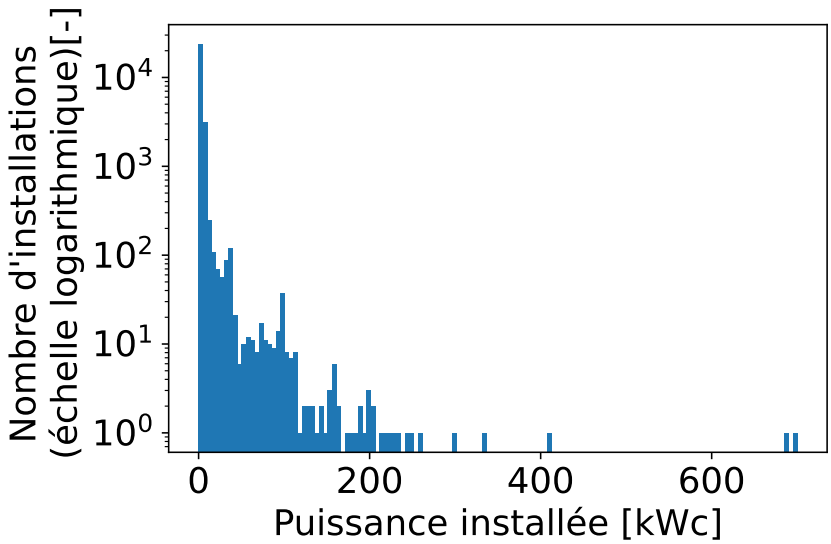


Figure 3 – Histogramme des puissances installées des installations répertoriées dans BDPV.

2.1.3 Pré-requis pour le registre PV

Nous avons mentionné en introduction que la finalité principale de notre registre est d'estimer la production PV issue des petites installations. Cette estimation se fonde sur un modèle physique de l'installation PV et requiert par conséquent un certain nombre de paramètres. Saint-Drenan et al. (2015) ont montré que l'inclinaison, l'orientation et la puissance installée d'une installation sont des caractéristiques suffisantes pour estimer de manière satisfaisante la production d'un système PV. Notre registre s'attache donc à collecter ces données à l'échelle de la France. Ce registre doit satisfaire trois critères principaux : il doit être aussi complet que possible (c'est-à-dire répertorier autant d'installations que possible afin notamment de refléter la vraie répartition spatiale des installations) ; il doit être désagrégé et doit contenir les caractéristiques techniques des installations.

Aucune des données disponibles antérieurement à cette thèse ne répond simultanément à ces trois critères. La figure 4 illustre les conditions satisfaites et manquantes des données disponibles. Notre registre doit répondre aux trois critères simultanément.

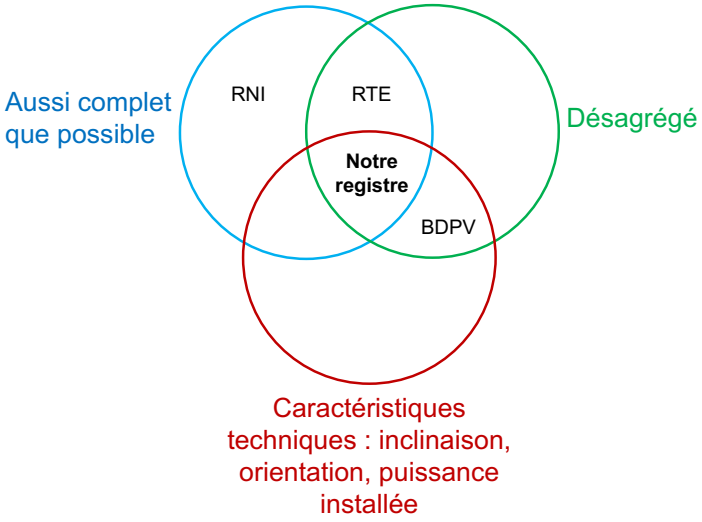


Figure 4 – Diagramme de Venn résumant les caractéristiques attendues et satisfaites par les différentes sources de données sur le PV.

2.2 Controle non supervisé de la précision du registre : la précision aval (*downstream task accuracy*, DTA)

2.2.1 Définition des standards de qualité

Un registre représentatif Notre registre vise à refléter fidèlement les caractéristiques du PV toiture en France. Il existera nécessairement un décalage du fait de la fréquence de rafraîchissement des images. Ce décalage est acceptable tant que les installations répertoriées sont représentatives des installations raccordées dans les trois années qui suivent. En faisant cette hypothèse de "stationnarité" sur trois ans des caractéristiques des installations, il faut alors surtout s'assurer que la puissance installée estimée (c'est-à-dire répertoriée dans notre registre) reflète la distribution spatiale du PV ainsi que sa répartition en termes de typologie d'installations. L'estimation de l'inclinaison et de l'orientation doit être représentative des distributions réelles pour éviter des biais systématiques entraînant des surestimations ou des sous-estimations de la production d'énergie à différents moments de la journée. Enfin, puisque l'inclinaison dépend de la latitude (Killinger et al., 2018), le registre doit refléter cette propriété.

Approche proposée Le principal problème pour assurer le suivi de la performance du modèle est que nous ne disposons pas de données de référence sur tout le territoire. Dans le cas contraire, il ne serait pas nécessaire de recourir à une cartographie par télédétection. Cependant, même si nous ne disposons pas de données de références à l'échelle de l'installation, nous disposons d'aggrégations auxquelles nous pouvons nous référer pour comparer les données générées par l'algorithme de cartographie du PV. Il nous suffit pour cela d'agréger les sorties de notre modèle

de cartographie et de les comparer avec une donnée existante, qu'il s'agisse du RNI, du registre de RTE ou de BDPV. J'appelle "précision aval" (ou *downstream task accuracy*, DTA) l'ensemble de mesures permettant d'évaluer la précision du registre selon une caractéristique donnée et sur un territoire donné. Cette méthode est une méthode non supervisée d'évaluation (Zhang et al., 2008; Chen et al., 2021a) en ce qu'elle ne requiert pas d'intervention humaine pour être calculée. Ainsi, elle permet de contrôler la précision du modèle sur toute une zone et l'utilisateur peut directement identifier les localisations où la DTA indique une moindre performance. La figure 5 présente le principe de fonctionnement de la précision aval.

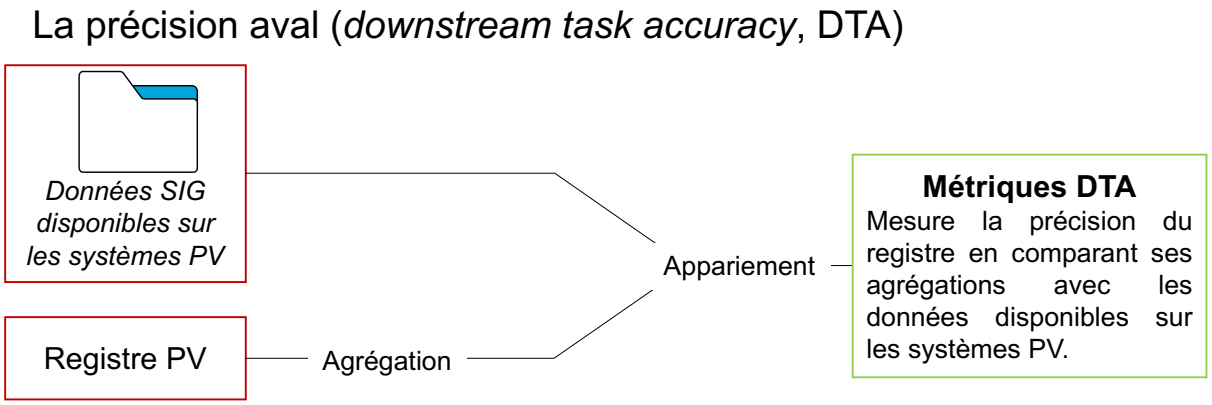


Figure 5 – Diagramme présentant le principe de fonctionnement de la précision aval (DTA), notre méthode pour surveiller les prédictions du modèle. D'après Kasmi et al. (2022a).

2.3 Quantifier et comprendre les limites des algorithmes de cartographie existants

2.3.1 Quantification de la sensibilité aux variations statistiques grâce à la DTA

Réplication de la littérature Nous évaluons les algorithmes existants de cartographie du PV toiture avec la DTA. A cette fin, nous répliquons un modèle s'inspirant de Mayer et al. (2022), qui est fondé sur DeepSolar (Yu et al., 2018; Mayer et al., 2020) et est l'un des meilleurs algorithmes de cartographie existants. Notre adaptation de cet algorithme prend en entrée des orthoimages et renvoie en sortie un registre répertoriant la localisation, l'inclinaison, l'orientation, la puissance installée et la surface des installations. Ce modèle requiert les données de la BD TOPO pour filtrer les détections du modèle et les données de BDPV pour inférer l'inclinaison des installations à partir de leur localisation, et calibrer un paramètre d'efficacité permettant de déduire de la surface estimée de l'installation sa puissance installée (à la suite de So et al. (2017), nous supposons que la relation entre la puissance installée et la surface est linéaire).

Nous ré-entraînons ce modèle sur des données d'entraînement collectées pour la France (Kasmi et al., 2023d). Ce jeu de données d'entraînement contient 17 000 images de l'IGN annotées de panneaux PV en France. Les performances obtenues sur les données d'entraînement sont comparables avec la littérature existante, comme le montre le tableau 1.

Table 1 – Performance en classification et en segmentation. Plus la GSD est basse, plus l'image est détaillée. Les meilleurs résultats sont en **gras**.

	Classification	**Segmentation**	
Méthode	Score F1 (↑)	IoU (↑)	GSD (cm/pixel)
Mayer et al. (2022)	0.87	0.74	10
Malof et al. (2019)	-	0.67	30
Zech and Ranalli (2020)	0.82	-	10
Parhar et al. (2021)	**0.97**	0.86	10
Notre approche	0.84	**0.86**	20

Déploiement à grande échelle Nous déployons ensuite ce modèle sur 11 départements français et évaluons le registre produit par le modèle avec la DTA. Nous comparons la distribution des inclinaisons et des orientations estimées par l'algorithme avec les distributions issues de BDPV. Nous nous intéressons ensuite à l'estimation de la puissance installée et nous intéressons en particulier au pourcentage d'erreur moyen, médian, au ratio de détections (c'est-à-dire le nombre d'installations détectées sur le nombre d'installations répertoriées dans le RNI) et l'erreur moyenne par installation, qui évalue l'écart en pourcentage entre l'installation moyenne estimée et l'installation moyenne répertoriée. Notre analyse révèle que l'estimation de l'inclinaison et de l'orientation est satisfaisante.

Le tableau 2 présente l'évaluation du modèle selon la DTA, pour l'estimation de la puissance installée. Nous pouvons voir une baisse importante de la performance, de l'ordre de 30 points de pourcentage. Ces résultats permettent de voir comment se transcrit en pratique une précision mesurée par le score F1 ou l'intersection-sur-union [10].

10. L'intersection-sur-union (IoU) ou indice de Jaccard calcule le degré de superposition entre deux ensembles. Il vaut 1 si les deux ensembles sont confondus et 0 s'ils sont parfaitement disjoints. Dans un contexte de segmentation, l'indice de Jaccard permet d'indiquer si la prédiction du modèle est conforme au masque de référence.

Table 2 – Précision aval (DTA) sur la zone cartographiée. Les valeurs entre parenthèses correspondent aux résultats sans filtrage par bâtiments. La ligne "Test" considère les images de test de l'ensemble de données d'entraînement comme une seule ville. k_i et C_i indiquent respectivement le nombre d'installations et la puissance installée. Un chapeau indique l'estimation par notre algorithme. Les valeurs entre parenthèses indiquent la précision avec et sans filtrage par bâtiment. Source : Kasmi et al. (2022a).

Echantillon	MAPE [%]	APE médiane [%]	ratio moyen [-]	AIPE moyenne [%]	k_i [-]	$\hat{k}_i$ [-]	C_i [kW$_p$]	$\hat{C}_i$ [kW$_p$]
Test	17.61	-	0.92	-0.10	1485	1362	6473.8	5334
Zone cartographiée	47.45 *(66.20)*	32.81 *(30.66)*	1.03 *(1.46)*	16.33 *(12.03)*	72595	58818 *(84015)*	293120.7	280807.9 *(382967.8)*

2.3.2 Comment expliquer les variations de performances sur la zone cartographiée ?

La surveillance de l'estimation de la puissance installée par commune avec la DTA montre que la précision varie grandement d'un département à l'autre. La question dès lors est de savoir *pourquoi* une telle variabilité émerge. Les travaux existants mettent en avant le rôle de la variabilité géographique dans la perte de précision du modèle (Wang et al., 2017; Malof et al., 2019). Nos analyses (voir la section 3.3.1 du chapitre 2 pour plus de détails) montrent que dans notre cas, des facteurs géographiques évidents n'expliquent pas la variabilité des performances du modèle. Nous analysons dans le détail les prédictions du modèle afin d'identifier une tendance et de formuler une hypothèse que nous étudierons dans le cadre de la deuxième sous-question.

Analyse des décisions du modèle Nous utilisons la méthode GradCAM (Selvaraju et al., 2020) pour analyser les décisions du modèle. Les *class activation maps* (CAM) sont une méthode d'interprétabilité consistant à mettre en evidence les zones images qui contribuent le plus à la décision du modèle. C'est un moyen courant pour analyser et interpréter les décisions d'un modèle (Lapuschkin et al., 2019; Zhang et al., 2021b). Dans notre cas, il s'agit de savoir "ce que voit" le modèle lorsqu'il prédit que l'image contient un panneau solaire, ou au contraire lorsqu'il prédit qu'il n'y a pas de panneau solaire sur l'image. La figure 6 présente des illustrations de cette analyse.

L'inspection des échantillons de la figure 6 révèle que cette région de l'image représente des éléments qui *ressemblent* à des panneaux photovoltaïques. Sur l'image de la première ligne (deuxième colonne) de la figure 6, nous pouvons voir que le modèle confond une ombrière qui a la même couleur et la même forme

générale qu'un panneau photovoltaïque avec un panneau réel. Dans l'image de la deuxième ligne, les vérandas avec des stries trompent le modèle.

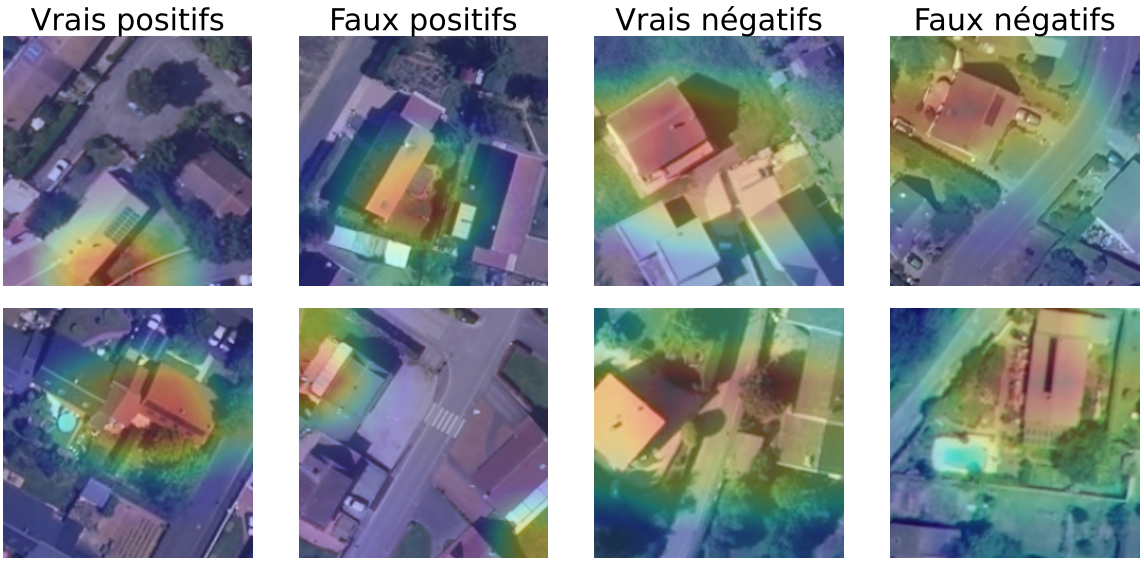


Figure 6 – Explications du modèle de classification générées en utilisant la méthode GradCAM (Selvaraju et al., 2020) pour quelques vrais positifs, faux positifs, vrais négatifs et faux négatifs. Plus une zone est rouge, plus elle contribue à la prédiction du modèle pour la classe considérée.

Hypothèse de travail Suite à cette analyse, nous formulons l'hypothèse de travail suivante. Nous supposons que, pendant l'apprentissage, le modèle extrait différentes caractéristiques corrélées à un panneau photovoltaïque sur l'image. Ces caractéristiques peuvent correspondre à des textures à différentes échelles, des composants tels que des lignes horizontales ou verticales, des couleurs ou, dans certains cas, à la forme générale du panneau photovoltaïque. Nous avons un contrôle limité sur ce que le modèle apprend des données, car cela dépend de la qualité des données, de l'initialisation du modèle et des hyperparamètres. Un modèle entraîné prédit, pendant le déploiement, la présence d'un panneau photovoltaïque si l'une de ses caractéristiques est identifiée sur l'image d'entrée. Il se peut qu'un motif ne soit pas évident sur l'image brute, mais plus facile à identifier à une échelle donnée.

Pour vérifier cette hypothèse et comprendre pourquoi le modèle fait de fausses prédictions, nous devons décomposer son processus de décision et comprendre comment ce dernier est affecté le cas échéant par des variations statisiques. La décomposition du processus de décision du modèle vise à mettre en évidence les *éléments* sur lesquels il s'appuie. D'autre part, l'évaluation de la robustesse de la prédiction nous permettra de voir si le modèle peut s'appuyer sur des facteurs moins susceptibles d'être perturbés par des variations statistiques.

3 Deuxième sous-question : comment s’assurer qu’un modèle d’apprentissage profond détecte de manière fiable des installations PV ?

Le chapitre 3 traite cette sous-question. Nous introduisons dans un premier temps une nouvelle méthode d’explicabilité permettant de localiser spatialement *et dans les échelles* les éléments importants pour la prédiction d’un modèle. Nous appelons cette méthode *Wavelet sCale Attribution Method* (WCAM, Kasmi et al., 2023a). Nous nous appuyons ensuite sur cette méthode d’explicabilité pour analyser et auditer les décisions du modèle. A partir de cette analyse, nous introduisons une méthode permettant d’améliorer la robustesse du modèle à des variations dans les conditions d’acquisition, variations identifiées comme étant un facteur important de variabilité statistique (Kasmi et al., 2023b).

3.1 Décomposer la décision d’un modèle dans l’espace des échelles

Motivation et littérature La méthode GradCAM que nous avons utilisée au chapitre précédent pour analyser le processus de décision est représentative des approches actuelles les plus répandues en explicabilité. Appliquée à la vision par ordinateur, l’explicabilité consiste à identifier les zones de l’image qui sont importantes pour la décision du modèle. On parle généralement de méthode d’attribution (*feature attribution method*). La GradCAM appartient à une famille de méthodes dites "boîte blanche" : il est nécessaire d’avoir accès au modèle et à ses gradients afin de pouvoir calculer l’explication. Cette classe de méthode produit les meilleures explications, que ce soit en termes de fidélité (Bhatt et al., 2020) ou de stabilité (Crabbé and van der Schaar, 2023). Cependant, il n’est pas toujours possible en pratique d’avoir accès au modèle : dans un contexte opérationnel, les modèles sont souvent appelés via une interface de programmation d’application (API). Dans ce cas, on peut s’appuyer sur une autre classe d’approches, dite "boîte noire" (Fel et al., 2021). Avec ces approches, les explications sont calculées en perturbant les données d’entrée (Zeiler and Fergus, 2014; Ribeiro et al., 2016; Petsiuk et al., 2018; Fel et al., 2021) et requièrent seulement d’avoir accès aux prédictions du modèle. Le principe des méthodes de perturbation est d’occulter certaines parties de l’image de manière aléatoire. Si la prédiction du modèle varie fortement lorsqu’une zone est fortement occultée, alors on peut en déduire que cette zone contribue à la prédiction du modèle.

Qu’elles soient "boîte noire" ou "boîte blanche", les méthodes d’explications existantes se limitent à indiquer *où* est-ce que le modèle voit et non *ce qu’il* voit sur les images. Or fournir seulement une indication spatiale n’est pas suffisant pour de nombreux cas pratiques (Achtibat et al., 2022). Pour commencer à aborder la question du "quoi", Fel et al. (2023b) a récemment introduit une méthode consis-

tant à calculer des concepts à partir des données d'entraînement puis à corréler les zones importantes sur une image avec ces différents concepts. Cependant, cette méthode suppose d'avoir accès au modèle et aux données d'entraînement, ce qui la rend difficile à appliquer dans un contexte opérationnel.

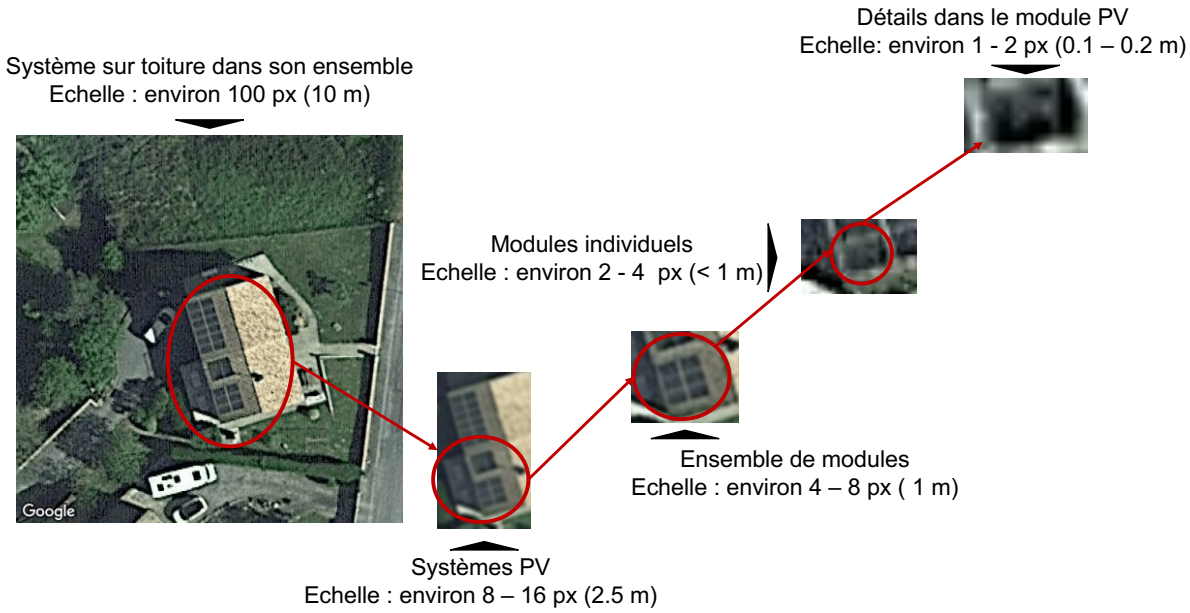


Figure 7 – Décomposition en différentes échelles d'un panneau PV vu sur une orthoimage. Adapté de Kasmi et al. (2023b).

Dans notre cas – la télédétection d'installations PV – nous pouvons remarquer qu'une installation photovoltaïque peut se décomposer en plusieurs échelles, pouvant être aisément interprétées en termes de concepts ou composantes. La figure 7 montre un exemple d'une telle décomposition dans les échelles. Selon l'échelle considérée, le panneau photovoltaïque présente des caractéristiques différentes. Aux plus petites échelles (20 à 40 cm/pixel), le panneau se manifeste par les détails au sein des modules PV. A l'inverse au plus grandes échelles (de l'ordre du mètre et au delà), il se manifeste par le cadre, soit sous une forme rectangulaire.

Ces échelles sont aussi localisées en termes de fréquences, ce qui peut constituer une propriété intéressante en vue d'étudier la robustesse d'un modèle ou d'une prédiction à une perturbation donnée. De nombreux travaux ont en effet utilisé l'espace des fréquences pour caractériser la robustesse d'un modèle (Yin et al., 2019; Chen et al., 2022). Nous utilisons cette propriété dans la section 3.3.

La transformée en ondelettes La transformée en ondelettes décompose un signal (par exemple, une image) en composantes élémentaires, les coefficients d'ondelettes, qui sont localisées spatialement et dans les échelles. La transformée en ondelettes est donc un moyen naturel pour décomposer une image en différentes échelles, ce qui est utile pour analyser des caractéristiques spécifiques de l'image

à ces différentes échelles. La transformée en ondelettes *dyadique* (Mallat, 1989) décompose les différentes échelles d'une image I de manière récursive. Le signal d'entrée f_{i-1} à l'étape i est filtré par un filtre passe-haut pour obtenir les coefficients de détail à l'échelle i et un filtre passe-bas [11] pour obtenir les coefficients d'approximation f_i à l'étape i. Pour une image, qui est un signal en deux dimensions, on distingue les coefficients de détail horizontaux, verticaux et diagonaux. Ces coefficients "résument" les changements brusques visibles à une échelle comprise entre 2^i et 2^{i+1} pixels. A l'étape $i+1$, on répète ce processus sur les coefficients d'approximation f_i obtenus. La figure 8 présente la transformée en ondelettes dyadique à deux niveaux d'une image en niveaux de gris.

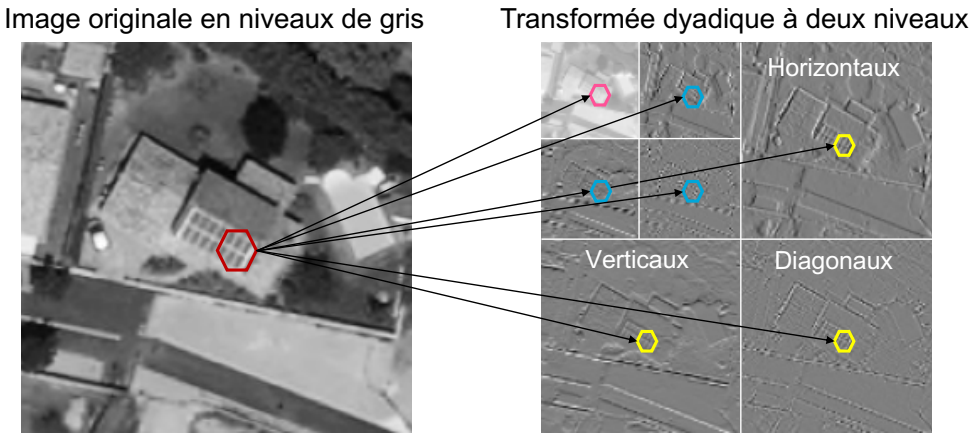


Figure 8 – Image et transformée en ondelettes dyadique à deux niveaux associée, avec des indications pour interpréter la transformée en ondelettes de l'image. Les termes "horizontaux", "diagonaux" et "verticaux" indiquent la direction des coefficients de détail. La direction est la même à tous les niveaux.

Cette représentation nous permet d'isoler les différentes échelles pour une localisation donnée. Ainsi pour la localisation donnée par l'hexagone rouge sur l'image de gauche, on peut distinguer les coefficients de détail aux échelles comprises entre 1 et 2 pixels (hexagones jaunes), 2 et 4 pixels (hexagones bleus) et les coefficients d'approximation restants (dont l'échelle est supérieure à 4 pixels, hexagone rose). Si les méthodes d'attribution traditionnelles nous permettent de savoir si la zone encadrée par l'hexagone rouge contribue à la prédiction, notre méthode vise à savoir quelle échelle (illustrée par les hexagones jaunes, bleus ou roses), à cette localisation, contribue à la prédiction du modèle, et donc quelle partie du système PV est importante pour la prédiction (nous rappelons que ces échelles correspondent à différentes composantes d'un système PV).

11. Un filtre passe-haut est un filtre qui laisse passer les hautes fréquences et qui atténue les basses fréquences. A l'inverse, un Un filtre passe-bas est un filtre qui laisse passer les basses fréquences et qui atténue les hautes fréquences.

Approche proposée Afin de quantifier la contribution des différentes échelles – interprétées comme des concepts ou des éléments structurels de l'image – dans la prédiction d'un modèle, nous introduisons la *Wavelet sCale Attribution Method* (WCAM, Kasmi et al., 2023a). Cette méthode consiste à perturber la transformée en ondelettes d'une image afin de générer une image perturbée, puis à évaluer la sensibilité du modèle à ces perturbations. Etant donné que nous pouvons relier chaque perturbation *dans les ondelettes* à une réponse d'un modèle, nous pouvons identifier les régions de la transformée en ondelettes qui contribuent le plus à la prédiction du modèle. La méthode de perturbation que nous utilisons est issue de Fel et al. (2021).

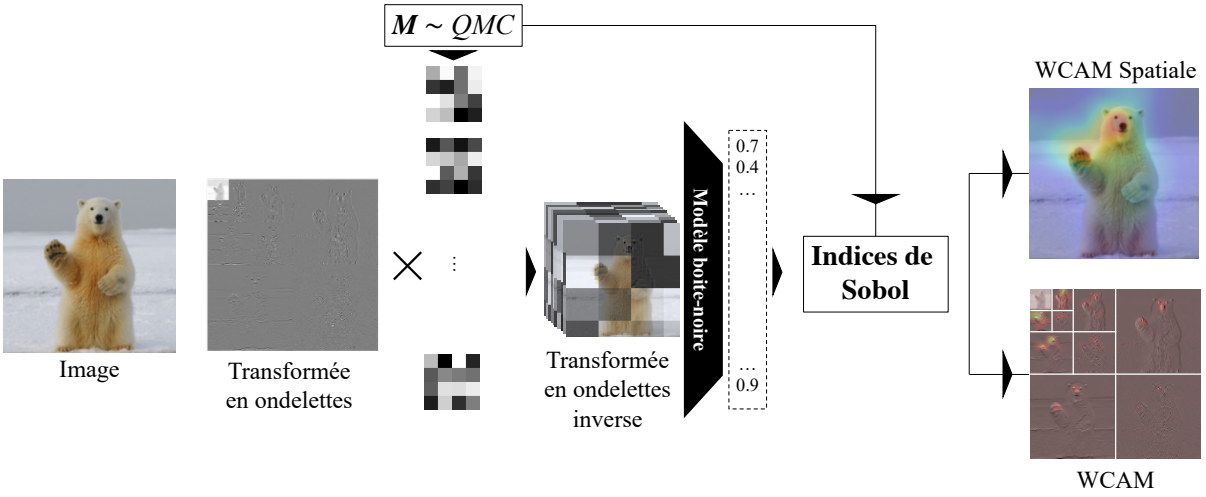


Figure 9 – Diagramme de notre méthode d'attribution, la *Wavelet sCale Attribution Method* (WCAM). Adapté de Kasmi et al. (2023a).

Le principe de fonctionnement de la WCAM est décrit sur la figure 9. L'idée est de générer une suite de n masques $\mathbf{M} = (M_1, \ldots, M_n)$ pseudo-aléatoirement (on la génère à partir d'une séquence de Quasi Monte-Carlo, QMC). On applique ensuite ces masques à la transformée en ondelettes dyadique de l'image pour la perturber, puis on reconstruit l'image à partir de cette transformée en ondelettes perturbée, pour obtenir n images perturbées. On évalue ensuite ces images. Le modèle renvoie la probabilité que l'image perturbée contienne un panneau solaire. En utilisant la probabilité prédite initiale (c'est-à-dire correspondant à l'image non perturbée), les masques et les probablités prédites pour les images perturbées, on calcule la sensibilité du modèle aux perturbations (en calculant les indices de Sobol correspondants à ces perturbations). L'analyse de sensibilité nous renvoie une *heatmap* indiquant la sensibilité du modèle à l'altération des coefficients en ondelettes correspondants.

3.2 Evaluer la fiabilité du processus de décision d’un modèle

3.2.1 Auditer les décisions du modèle et isoler les composants critiques

Analyser des prédictions du modèle Grâce à la WCAM, nous pouvons identifier quelles sont les échelles qui contribuent à la décision du modèle. Sur la figure 10, nous pouvons par exemple voir que dans le cas **A**, la zone importante sur l’image n’est importante qu’à une seule échelle (identifiée par le cercle blanc sur la WCAM, sous l’image **A**). En revanche, Dans les cas **B** et **C**, plusieurs échelles, localisées spatialement au même endroit, contribuent à la prédiction du modèle. En se référant aux composants structurels identifiés sur la figure 7, il est possible d’affirmer grâce à la WCAM que sur ces exemples, le modèle *voit* surtout des modules ou des groupes de modules. Nous pouvons remarquer que c’est à l’échelle des groupes de modules que se trouvent les grilles archétypales des panneaux solaires installés au tournant des années 2010.

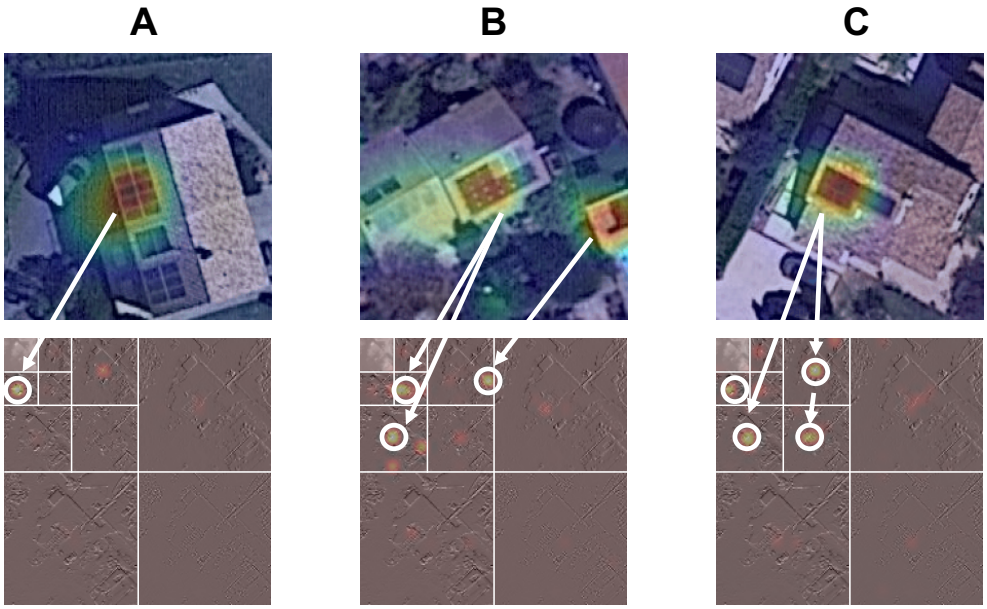


Figure 10 – Décomposition dans l’espace des échelles des prédictions d’un modèle. Adapté de Kasmi et al. (2023b).

Identifier les composants critiques Maintenant que nous avons vu comment se décomposait une prédiction dans les échelles, nous pouvons nous demander quelle est l’information *suffisante* pour prédire un panneau et où est-ce qu’elle est localisée dans les échelles. Pour identifier cette information suffisante, nous reconstruisons l’image des composants les plus importants aux composants les moins importants. La dernière image correspond à l’image initiale. Nous considérons que le modèle dispose de l’information suffisante dès lors qu’il parvient à identifier le panneau sur l’image. En faisant la différence entre cette image et la précédente

(la dernière image incorrecte) nous pouvons identifier le *composant critique*. La figure 11 présente un exemple d'une image reconstruite à partir des composantes les plus importantes de l'image. Nous pouvons voir que le motif grillagé est prépondérant dans la prédiction du modèle. Nous pouvons aussi voir que sans l'ajout du composant critique, qui correspond à des grilles verticales, le modèle ne parvient pas à prédire la présence du panneau.

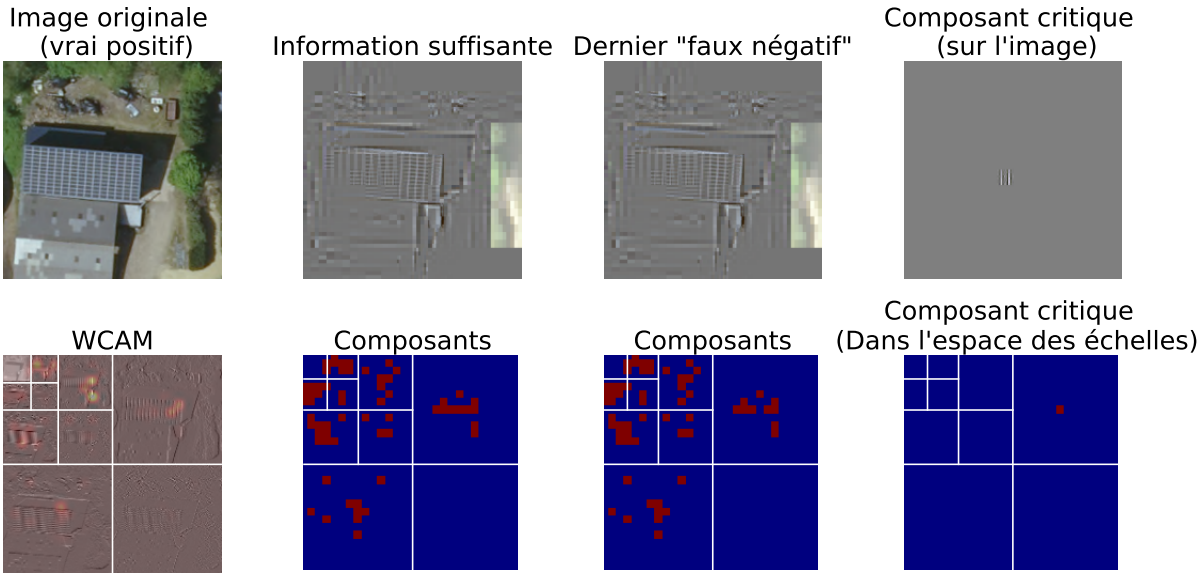


Figure 11 – Image originale, information suffisante, dernier "faux négatif" et composant critique, dans l'espace des images (haut) et dans l'espace des échelles (bas).

Ainsi, la WCAM étend les méthodes d'attribution existantes et permet de décomposer, pour une localisation donnée, les échelles qui sont importantes pour la prédiction. Mais il est également possible d'utiliser cette information pour extraire de l'image l'information suffisante à la prédiction, traduisant ce que le modèle "a besoin" de voir pour faire une prédiction correcte. Ces composantes étant interprétables, la WCAM permet ainsi d'auditer les prédictions du modèle : nous avons mis en avant l'importance des échelles, qui correspondent aux modules ou aux groupes de modules PV. Mais cette information suffisante nous permet également de comprendre pourquoi le modèle peut manquer de robustesse face aux variations des conditions d'acquisition.

3.2.2 Etudier la sensibilité aux conditions d'acquisition

Définition Les conditions d'acquisition désignent la chaîne de traitement allant de la captation d'une scène à sa conversion sous forme d'image numérique. Les conditions d'acquisition varient en fonction de l'heure de la journée et des conditions atmosphériques, du type de capteur optique, de sa sensibilité, de son réglage. On peut considérer que les perturbations induites par la conversion d'un signal op-

tique en un signal numérique (quantification du signal) sont négligeables. La variation des conditions d’acquisition se traduit par une perturbation des hautes fréquences de l’image, ce qui correspond aux échelles les plus fines (jusqu’à 4, voire 8 pixels). Ainsi, si le modèle dépend cruciallement de l’information localisée dans ces échelles, une perturbation des conditions d’acquisition peut perturber cette information et ainsi perturber la prédiction du modèle.

Etude de l’effet des conditions d’acquisition Notre base de données BDAPPV contient des annotations doubles pour environ 8 000 systèmes PV. La figure 12 présente des exemples de prises de vues aériennes avec deux sources différentes : Google (Gorelick et al., 2017) et IGN (IGN, 2024a).

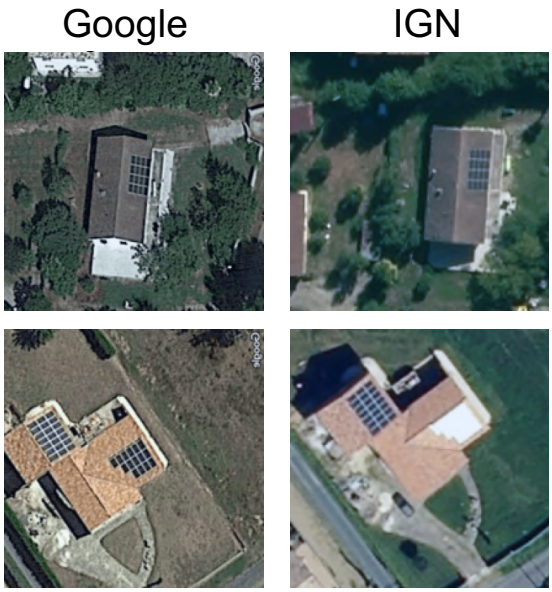


Figure 12 – Exemple d’images de BDAPPV (Kasmi et al., 2023d) issues de Google (gauche) et de l’IGN (droite).

Nous pouvons donc étudier l’effet des conditions d’acquisition sur la précision des prédictions d’un modèle, indépendamment d’autres facteurs de variabilité (résolution de l’image ou nature de la scène observée). Nous entraînons un modèle sur les images Google et nous mesurons sa précision sur les images IGN. Comme le montre le tableau 3, la performance s’effondre, à cause d’une hausse du nombre de faux négatifs : le modèle ne parvient plus à reconnaître les panneaux PV sur les images issues de l’IGN.

Table 3 – **Score F1** et décomposition en vrai positifs, vrais négatifs, faux positifs et faux négatifs des prédictions d'un modèle entraîné sur des images Google et déployé sur des images IGN (même scène, même résolution, mais différentes conditions d'acquisition). Adapté de Kasmi et al. (2023b).

	Score F1 (↑)	Vrais positifs (VP)	Vrai négatifs (VN)	Faux positifs (FP)	Faux négatifs (FN)
Google	0.98	1891	2355	36	39
IGN	0.46	566	2321	99	1335

3.3 Améliorer la robustesse des modèles à la variabilité des conditions d'acquisition

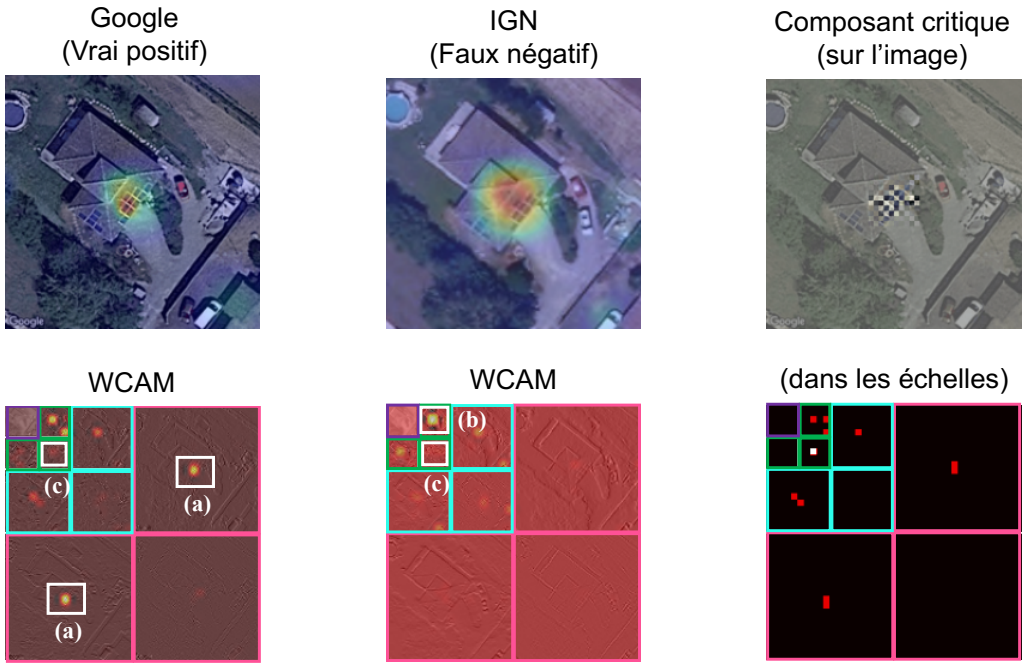


Figure 13 – Prédictions sur l'image Google (gauche, rangée supérieure) et l'image IGN (droite, rangée supérieure) et WCAMs associés (rangée inférieure). Plus la région en surbrillance est claire, plus la prédiction est importante. La colonne la plus à droite présente les composants les plus importants de l'image Google et les composants critiques. Adapté de Kasmi et al. (2023b).

Mettre en avant la disparition de composantes importantes La WCAM nous permet d'étudier la sensibilité du modèle aux conditions d'acquisitions. Nous pouvons calculer la WCAM pour une prédiction altérée par les conditions d'acquisitions, comme par exemple sur la figure 13. Nous pouvons voir que les composants **(a)** contribuaient à la prédiction du modèle. Or ces composants sont moins présents sur l'image IGN, ce qui peut expliquer pourquoi le modèle ne parvient plus à identifier le panneau solaire. Le composant **(c)** est le composant critique, qui contribue à

la prédiction dans le cas de l’image Google. Il n’y a pas d’équivalent sur l’image IGN. Enfin, sur l’image IGN, nous pouvons voir que le modèle s’appuie essentiellement sur les composants **(b)** pour faire sa prédiction.

Améliorer la robustesse aux conditions d’acquisition La WCAM nous a permis de confirmer le mécanisme selon lequel le fait d’identifier un panneau en recourant à des composants situés dans des échelles fines (c’est-à-dire des hautes fréquences) entraîne une sensibilité aux conditions d’acquisition ; ces dernières perturbant essentiellement l’information située dans les hautes fréquences de l’image.

Pour améliorer la robustesse du modèle aux conditions d’acquisition, nous mobilisons la littérature sur les corruptions naturelles (Hendrycks and Dietterich, 2019). Ce champ de la littérature propose un ensemble de méthodes visant à améliorer la robustesse des réseaux de neurones à des perturbations pouvant affecter les images : flou, bruit, pixelisation, etc. Les conditions d’acquisition faisant partie de ce type de perturbations, nous implémentons des méthodes populaires issues de cette littérature. Ces méthodes consistent en des augmentations de la base d’apprentissage : durant l’entraînement, on génère aléatoirement des copies altérées des images et on ajoute ces copies aux données d’entraînement. L’idée est que si le modèle est entraîné sur suffisamment d’images altérées, alors il apprendra à être invariant à ces perturbations.

Nous évaluons plusieurs méthodes répandues (Cubuk et al., 2019, 2020; Hendrycks et al., 2020) de génération de telles perturbations. Nous implémentons également une méthode visant à "reproduire" les conditions d’acquisition en bruitant et flouttant les images durant l’entraînement ("Noise and blur"), et une méthode qui floutte les images pour supprimer l’information contenue dans les hautes fréquences ("Blurring"). Enfin, nous introduisons une méthode visant à perturber les différentes échelles de l’image (en perturbant la transformée en ondelettes) en plus de la flouter ("Blurring + Wavelet Perturbation"). L’intuition derrière cette transformation est d’apprendre au modèle à être capable de se reposer sur différentes échelles pour prédire un panneau. Tous nos résultats sont comparés avec la minimisation du risque empirique (ERM), c’est-à-dire un modèle entraîné sans augmentations, et avec l’Oracle, c’est-à-dire un modèle ERM entraîné sur les images IGN. l’ERM et l’Oracle permettent de borner la performance.

Le tableau 4 présente nos résultats. Nous pouvons voir que les méthodes existantes sont à peine meilleures que l’ERM. Par ailleurs, le fait de flouter l’image restaure la performance mais au détriment du nombre de faux positifs : le modèle prédit presque toujours que l’image contient un panneau solaire. Finalement, nous pouvons voir que notre méthode consistant à perturber la transformée en ondelettes atteint des résultats honorables. Elle surclasse toutes les méthodes existantes tout en ayant un comportement cohérent : il n’y a pas d’explosion du nombre de faux positifs ou de faux négatifs. Ces résultats mettent en avant le fait qu’il est

possible d'améliorer la robustesse des modèles aux conditions d'acquisition.

Table 4 – **Score F1** et décomposition en vrais positifs, vrais négatifs, faux positifs et faux négatifs pour des modèles entraînés sur Google avec différentes techniques d'atténuation. L'évaluation est conduite sur les images IGN. L'Oracle désigne un modèle entraîné sur des images IGN sans technique d'atténuation. Le meilleur résultat est en **gras**, le deuxième meilleur souligné.

	Score F1 (↑)	VP	VN	FP	FN
Oracle	0.88	1818	1992	428	83
ERM (Vapnik, 1999)	0.44	566	2321	99	1335
AutoAugment (Cubuk et al., 2019)	0.46	598	2318	102	1303
AugMix (Hendrycks et al., 2020)	0.48	624	2318	102	1277
RandAugment (Cubuk et al., 2020)	0.51	707	2280	140	1194
Noise and blur	0.48	636	2287	133	1265
Blurring	**0.74**	1855	1196	1224	46
Blurring + WP	0.58	896	2114	306	1005

4 Troisième sous-question : comment construire un registre répondant aux standards de fiabilité et comment ce registre peut-il améliorer l'observabilité du PV en France ?

Dans les chapitres 2 et 3, nous traitons des trois aspects de la fiabilité définis en introduction : l'auditabilité avec la WCAM, la robustesse, avec l'augmentation de la taille de l'échantillon d'apprentissage et le contrôle avec la DTA. Ces éléments sont autant d'outils nous permettant de construire un modèle fiable pour cartographier le PV en France. La question qui se pose à présent est de savoir comment construire un tel algorithme et comment évaluer son apport pour l'amélioration de l'observabilité du PV. Le chapitre 4 se focalise sur la construction d'un algorithme fiable de cartographie du parc PV en toiture, DeepPVMapper. Le chapitre 5 étudie comment évaluer la pertinence de la cartographie réalisée par DeepPVMapper pour améliorer l'observabilité du PV diffus. Les contributions issues de ce chapitre sont la librarie Python `PyPVRoof` qui permet d'extraire les caractéristiques techniques d'une installation PV à partir de son polygône géolocalisé, l'algorithme DeepPVMapper (Kasmi et al., 2023c) et une étude montrant qu'il est possible d'améliorer l'observabilité de la production PV sur toiture en combinant dans un modèle de conversion physique simplifié les données issues de DeepPVMapper avec des données de rayonnement solaire et de température (Kasmi et al., 2024).

4.1 DeepPVMapper : un algorithme fiable pour cartographier le parc PV diffus en France

Critères d'évaluation En amont de l'implémentation d'un nouvel algorithme, nous nous proposons d'introduire de nouvelles métriques de performance, plus représentatives de conditions réelles de fonctionnement de cet algorithme. Ces métriques sont fondées sur la précision aval définie au chapitre 2. Nous rappelons les trois mesures permettant d'évaluer la précision de l'estimation de la puissance installée : la MAPE, qui compare l'estimation globale de la capacité installée avec la référence ; le ratio de détection, qui compare le nombre de détections avec le nombre réel d'installations et l'AIPE, qui compare la taille moyenne estimée de l'installation avec la taille moyenne réelle de l'installation. La MAPE mesure le décalage entre la capacité installée enregistrée et la capacité installée estimée au niveau de la ville. Le taux de détection garantit que l'algorithme détecte le nombre correct d'installations. L'AIPE indique si nous sous-estimons ou surestimons la taille des installations. Par construction, un AIPE négatif (resp. positif) indique qu'en moyenne, nous sous-estimons (resp. surestimons) la taille des installations.

Nous évaluons notre algorithme avec les métriques DTA sur une zone de 120 km² près de Lyon. Cette zone est suffisante pour évaluer les avantages de notre approche et suffisamment petite pour permettre plusieurs évaluations de variantes de l'algorithme de cartographie en un temps limité. Nous avons choisi cette zone parmi plusieurs autres en France car les types géographiques varient avec une zone urbaine densément peuplée et une campagne environnante. La densité des installations photovoltaïques est également assez inhomogène, ce qui rend la zone assez difficile pour l'algorithme. Ce benchmark est entièrement réplicable en suivant les instructions de notre dépôt public Kasmi et al. (2023c).

Approche proposée Notre approche améliore l'approche de Mayer et al. (2022), que nous avons adaptée au chapitre 2. La figure 14 présente notre approche. Cette dernière se décompose en deux étapes. Premièrement, on extrait des polygones d'installations PV géolocalisés à partir d'orthoimages et en utilisant un modèle de classification et un modèle de segmentation d'images. Ensuite, on utilise ces polygones et des données auxiliaires (BDPV dans notre cas, mais notre approche peut être utilisée avec des données LiDAR) pour estimer les caractéristiques des installations PV.

Par rapport à l'état-de-l'art, les améliorations portent sur trois aspects principaux :

1. Filtrage des zones à cartographier en amont avec la BD TOPO. Cela permet de réduire les faux positifs en ne se focalisant que sur des zones anthropisées ;
2. Introduction d'une superposition entre les vignettes extraites des orthoimages brutes et présentées au modèle de classification afin de réduire les faux nega-

tifs. Nous avons constaté que le modèle est plus susceptible de faire un faux négatif si l'installation est située sur un coin de l'image. Avec la superposition, nous nous assurons qu'une installation n'est jamais dans un coin de l'image ;

3. Introduction d'un module standardisé pour extraire les caractéristiques. Dans Trémenbert et al. (2023), nous comparons les méthodes d'extractions de caractéristiques existantes et développons un package Python standardisé, `PyPVRoof`. Ce package est conçu pour fonctionner dans différents cas de disponibilité de données et n'intègre que les méthodes les plus précises et rapides pour calculer les caractéristiques d'installations PV.

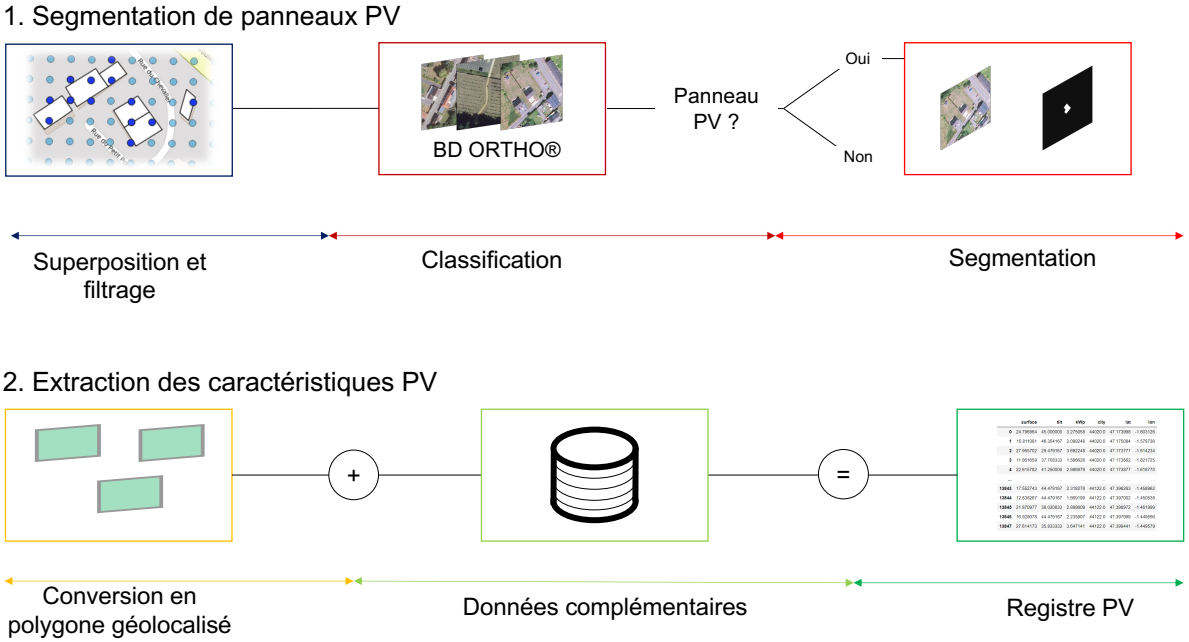


Figure 14 – Diagramme de DeepPVMapper.

Performances Nous évaluons la plus-value en termes de performance des améliorations intégrées dans DeepPVMapper avec la littérature existante. Nous testons également différentes configurations. Ces configurations consistent en des modèles de classification plus simples (le ResNet-50, He et al., 2016) et plus complexes (le ConvNext, Liu et al., 2022) que le modèle utilisé dans DeepSolar et pour notre référence.

Nos résultats (tableau 5) montrent que le filtrage et la superposition permettent d'améliorer significativement la performance. En revanche, la meilleure robustesse aux conditions d'acquisition ne se traduit pas par un gain de performances en termes de DTA. Cela peut être lié au fait que la variation des conditions d'acquisition au sein des images IGN est moins forte qu'entre IGN et Google. Outre ces résultats quantitatifs, nous discutons dans la section 3.2.2 du chapitre 4 comment

DeepPVMapper répond aux limites qualitatives de l'algorithme que nous avions implémenté dans le chapitre 2.

Table 5 – Performances mesurées avec les métriques DTA de différentes configurations de l'algorithme. Les meilleurs résultats sont en **gras** et les seconds meilleurs soulignés.

Configuration	DTA		
	MAPE [%]	Ratio [-]	AIPE [%]
Référence (Kasmi et al., 2022a)	55.7	1.29	15.4
ResNet-50	46.9	1.09	15.5
ConvNext	45.5	1.11	15.3
ResNet + Superposition/filtrage	39.5	**0.91**	**12.6**
ConvNext + Superposition/filtrage	**38.8**	0.84	18.2
ResNet + WP	48.22	0.57	16.97
ResNet + Sampling + WP	40.62	0.82	21.48

4.2 Critères d'évaluation de l'amélioration de l'observabilité du PV en France

4.2.1 Du registre aux courbes de production PV

Données supplémentaires Le registre technique ne suffit pas à l'estimation de la production PV diffuse. Nous avons besoin en plus des données de rayonnement solaire et de température. Les données de rayonnement solaire proviennent du Service Européen de surveillance de l'atmosphère Copernicus (CAMS, Qu et al., 2017) et les données de température proviennent de la 5ème réanalyse météorologique du Centre européen pour les prévisions météorologiques à moyen terme (ERA5, Hersbach et al., 2020).

Grâce à l'association *Asso BDPV*, nous avons eu accès à des mesures de la production d'énergie photovoltaïque de 1 793 systèmes photovoltaïques individuels (906 après contrôle qualité). Ces mesures couvrent toute la France et ont une granularité de 30 minutes. Cet ensemble de données est donc l'un des plus importants disponibles en termes de nombre de systèmes, d'emprise géographique et de niveau de détail associé sur les systèmes PV, étant donné que nous avons également accès aux caractéristiques techniques des installations. La base de données la plus proche de la nôtre est celle fournie par IBW, un fournisseur d'électricité à Wohlen en Suisse et utilisée par Walch et al. (2021). Cette base de données contient les profils de production d'énergie photovoltaïque de 15 foyers et les caractéristiques

techniques des systèmes photovoltaïques.

Modèle de conversion Les caractéristiques issues du registre technique permettent de paramétriser un modèle de conversion PV, qui prend en entrée des données de rayonnement et de température et renvoie une production photovoltaïque exprimée en watts. Nous nous focalisons dans cette étude sur un modèle très simple, le modèle PVWatts de Dobos (2014). La raison de ce choix est que nous souhaitons faire aussi peu d'hypothèses que possible sur le système PV, étant donné que le registre ne nous donne pas accès à de nombreux paramètres (efficacité des convertisseurs, pertes du système, type de module). Le modèle de conversion est donné par l'équation (1) :

$$p_{PV,t} = \frac{POA_{eff,t}}{G_{stc}} \times P_{PV} \times (1 + \gamma_{pdc} (T_{module,t} - T_{stc})) \quad (1)$$

Où $POA_{eff} = POA_{eff}(\theta, \phi)$ désigne le rayonnement solaire effectif (c'est-à-dire issu des trois formes de rayonnement, direct, diffus et réfléchi, et après prise en compte de la réfraction des modules) en fonction de l'inclinaison et de l'orientation du système PV. P_{PV} est la puissance installée de l'installation et γ_{pdc} est un facteur d'efficacité qui reflète la diminution de la performance du module avec la température. La température de référence T_{stc} est de 25°C et γ_{pdc} exprimé en K^{-1} correspond à la perte d'efficience du module au delà de cette température de référence. Enfin, G_{stc} est l'irradiance de référence, exprimée en W/m^2 et valant 1000 W/m^2.

Modélisation proposée Notre modélisation consiste à estimer la production de chacun des systèmes PV. De cette manière, il est possible de retrouver tous les niveaux d'agrégation (commune, poste-source, département, région, pays) sans hypothèse sur la répartition géographique des installations. Par ailleurs, nous pouvons directement évaluer la précision de cette méthode avec les données de production PV. La principale question soulevée par cette approche est de savoir si les erreurs d'estimation de la production vont se compenser ("foisonner" dans le vocabulaire du système électrique) lorsqu'on va agréger les installations entre elles. Saint-Drenan et al. (2016) ont montré que les erreurs d'estimation de production se compensent si ces dernières sont indépendantes. Dans notre cas, l'estimation de production dépend de l'estimation des caractéristiques des installations. Nous étudions par conséquent si des biais systématiques dans l'estimation des caractéristiques conduisent à des biais systématiques dans l'estimation de la production.

4.2.2 Critères d'amélioration de l'observabilité

Rappel de la définition En introduction, j'ai défini l'observabilité comme la capacité du GRT à estimer avec précision la production en temps réel et future d'une

unité de production. Afin d'évaluer la pertinence de notre méthode pour améliorer l'observabilité du PV diffus, il faut donc :

1. Qu'elle soit précise (ce qui requiert des données de référence),
2. Qu'elle soit plus performante que des méthodes alternatives,
3. Qu'elle puisse être agrégée efficacement ; tant sur le plan computationnel qu'en termes de convergence statistique des estimations de production. Ce critère est propre à notre approche et au grand nombre d'installations PV diffuses (environ 600 000 en France en 2023).

Approche proposée Afin d'évaluer si notre approche est précise et pertinente pour améliorer l'observabilité des systèmes photovoltaïques, nous la comparons avec les mesures de référence et définissons un ensemble de références auxquelles nous comparons notre modèle. Nous comparons notre approche "explicite" (de conversion du rayonnement en production PV) avec des approches "statistiques". Ces approches implicites ne nécessitent pas de registre, seulement la puissance installée du système PV et ont été entraînées sur les données de production PV de BDPV. Nous implémentons deux approches implicites ou statistiques : une régression linéaire et un réseau de neurones à une couche cachée.

Pertinence et périmètre des modèles Nous ne pouvons pas utiliser nos données de production PV comme mesure de référence au niveau des agrégats géographiques utilisés dans la modélisation du PV par le GRT sans hypothèse supplémentaire. En effet, l'approche probabiliste, généralement utilisée, suppose qu'il y a une forte densité d'installations sur un territoire donné (de l'ordre du millier d'installations). Dans le meilleur des cas, sur des zones géographiques pertinentes, nous n'avons au plus qu'une vingtaine de télémesures. Ainsi, nous avons choisi d'évaluer la pertinence de notre approche indépendamment des méthodes pratiquées actuellement par le GRT. Ce choix se justifie par le fait que le GRT sait que ses méthodes manquent de précision. Par conséquent, si notre approche se révèle être précise, il sera très probable qu'elle soit meilleure que les méthodes actuellement déployées par le GRT.

4.3 Résultats : le PV toiture est observable

Résultat principal D'après la tableau 6, il est possible d'améliorer l'observabilité du PV diffus grâce à notre approche. En effet, l'erreur d'estimation à l'échelle d'une installation est de l'ordre de 10%. D'autre part, des tests d'aggrégation montrent que l'erreur dans l'estimation de production agrégée reste contenue. L'erreur de notre méthode à l'échelle de l'installation est du même ordre que l'erreur commise par une régression linéaire (entraînée sur des courbes de charges d'installations PV en toiture).

Table 6 – Comparaison de la RMSE [W] et de la pRMSE [%] (entre parenthèses) de l'estimation à l'échelle de l'installation individuelle avec les paramètres de DeepPV-Mapper. Les meilleurs résultats sont en **gras** et les seconds meilleurs soulignés. n indique le nombre d'installations utilisées dans cette étude.

	Cas	Min [W]	Max [W]	Moyenne [W]	Médiane [W]	n
	Oracle	114.61 (3.90)	2137.82 (26.49)	281.53 (8.36)	223.06 (7.66)	255 255
Explicite	DeepPVMapper	**119.56** **(4.15)**	3001.42 (43.39)	**332.57** **(10.10)**	**245.33** **(8.18)**	255 255
Statistiques	Régression linéaire	134.42 (4.67)	7663.42 (33.27)	392.97 (10.18)	257.21 (8.86)	255 255
	Réseau de neurones	245.93 (8.42)	9261.82 (29.31)	744.24 (20.70)	607.64 (20.74)	255 255

Ainsi, nos résultats montrent qu'il est possible d'approximer de manière satisfaisante la production d'une installation PV en toiture en ne connaissant que sa puissance installée, sa localisation, son inclinaison et son orientation. La performance de l'approche proposée est de l'ordre de la performance d'une régression linéaire. Ainsi, si des courbes de charges du PV toiture sont accessibles, la régression linéaire est préférable. Cependant, cette disponibilité n'allant pas de soi, notre approche est une première approximation simple à obtenir de la production PV en toiture.

Agrégation Nous évaluons par ailleurs le comportement de l'erreur d'estimation de la production lorsque l'on agrège un nombre croissant d'installations. La figure 15 présente les résultats. Nous pouvons voir que l'erreur agrégée sur une vingtaine d'installations reste contenue. Le chapitre 5 discute plus en détail ces résultats et montre que l'erreur peut ne pas décroître avec le nombre d'installations dans le cas où les erreurs d'estimations des caractéristiques des installations sont systématiques. Grâce à la DTA, nous identifions quelles erreurs systématiques DeepPVMapper commet et pouvons donc estimer le comportement de l'erreur d'estimation agrégée en fonction de ces biais.

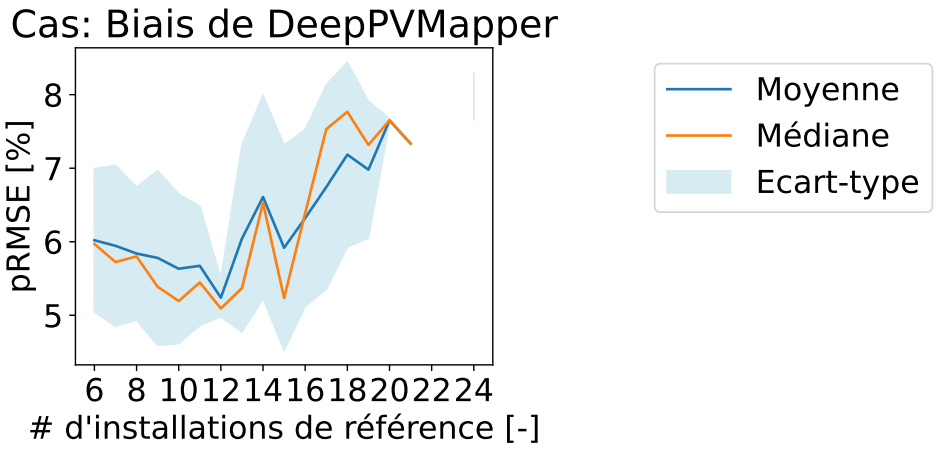


Figure 15 – Comportement de l'erreur d'agrégation des estimations des courbes de production PV dans le cas d'une inclinaison et d'une orientation estimées avec DeepPVMapper.

5 Conclusion

5.1 Réponses aux questions scientifique et industrielle

Réponse à la question scientifique La question scientifique était de savoir si la télédétection à partir d'orthoimages et utilisant des algorithmes d'apprentissage profond était une méthode appropriée pour construire un registre technique national du PV en toiture destiné à améliorer l'observabilité du PV. Plus généralement la question soulevée est celle de savoir si les algorithmes et pratiques actuelles sont suffisamment matures pour être utilisés dans un processus industriel plus large. La contribution centrale de ce travail est d'introduire une méthodologie permettant d'améliorer la fiabilité des algorithmes d'apprentissage statistique utilisés dans un contexte appliqué ; ainsi que la fiabilité des données qu'ils génèrent. A condition qu'il soit possible de croiser les données générées avec une source tierce afin d'en contrôler la précision, cette méthode montre que la fiabilité des algorithmes d'apprentissage profond est satisfaisante pour une application industrielle dès lors qu'il est possible d'auditer leur processus de décision de manière à s'assurer de sa pertinence et de sa robustesse. Dans ce contexte, il est en effet possible d'évaluer la précision des données et de pouvoir identifier et comprendre les erreurs commises par le modèle, améliorant ainsi la transparence de l'algorithme aux yeux de l'utilisateur final, ce qui renforce sa confiance et son recul critique vis-à-vis du modèle. Appliquée au cas de la télédétection d'installations PV en toiture, notre méthode montre que l'apprentissage profond et les données d'observation de la Terre sont une méthode pertinente pour construire un registre du PV en toiture : des données de contrôle sont disponibles et notre WCAM a mis en avant le fait que le processus de décision est pertinent, et nous a permis d'améliorer sa robustesse

à la variabilité des conditions d'acquisition.

Réponse à la question industrielle Ce travail contribue à améliorer l'observabilité des systèmes photovoltaïques dans la mesure où il fournit une information détaillée sur la distribution géographique et les caractéristiques du PV toiture. Il améliore également l'observabilité du PV en montrant qu'il est possible d'estimer précisément la production PV en toiture uniquement à partir d'un modèle simple de conversion et une information limitée (disponible dans le registre à grande échelle) sur les systèmes PV. Nous montrons que cette méthode d'estimation de la production PV est aussi précise que des méthodes alternatives calibrées sur des données de production PV en toiture. Ainsi la pertinence de notre méthode comparée à ces approches statistiques est discutable dès lors que des données de production pour le PV toiture sont disponibles ; ce qui n'est cependant pas évident. Ce travail suggère des pistes intéressantes pour construire des mesures de référence pour le PV toiture, lorsque la télérelève n'est pas disponible et mériterait d'être complété grâce à la collecte d'un grand nombre de télérelèves d'installations PV en toiture afin de démontrer empiriquement la plus grande précision de notre méthode par rapport aux pratiques actuelles du GRT.

5.2 Contributions

Contributions académiques Les contributions de cette thèse s'inscrivent dans deux domaines : l'apprentissage statistique et l'ingéniérie des systèmes électriques. J'ai tiré parti de l'étude de cas de la cartographie des installations photovoltaïques sur toitures pour étudier la question de la fiabilité des algorithmes d'apprentissage profond dans un contexte opérationnel. À cette fin, j'ai introduit une méthodologie de vérification de la précision des données générées par un algorithme fondée sur l'utilisation de mesures de référence indirectes (Kasmi et al., 2022a). J'ai également introduit une nouvelle méthode d'attribution qui identifie les échelles contribuant à la prédiction d'un modèle d'apprentissage profond (Kasmi et al., 2023a). Cette méthode d'attribution est fondée sur une analyse de sensibilité du modèle à la perturbation de la transformée en ondelettes de l'image d'entrée. Je montre que cette méthode permet d'améliorer la fiabilité des modèles en fournissant une information plus fine de leur processus de décision (Kasmi et al., 2023b). Enfin, grâce à des campagnes participatives, j'ai introduis une nouvelle base de données d'entraînement, BDAPPV (Kasmi et al., 2023d), contenant près de 50 000 images annotées et provenant de deux fournisseurs d'images. Cette base de données aide à la cartographie d'installations en France et dans des pays voisins (voir par exemple Freitas et al., 2023), mais permet également d'étudier la sensibilité des modèles à différentes conditions d'acquisitions (voir par exemple Kasmi et al. (2023b) ou Guo et al., 2024).

En ingéniérie des sytèmes électriques, ce travail améliore la connaissance du parc photovoltaïque français en cartographiant les installations PV en toiture sur 38 départements français (au moment de la rédaction de cette thèse, et à terme sur l'ensemble de la France métropolitaine). Le registre généré répertorie l'inclinaison, l'azimuth, la localisation ainsi que la puissance installée de chaque système PV. La zone cartographiée est actuellement la deuxième plus grande au monde en termes de superficie derrière DeepSolar, et la plus grande au monde avec ce niveau de détail en termes de caractérisation des systèmes PV. J'ai également montré qu'il était possible d'améliorer l'observabilité du PV toiture en utilisant les données de ce registre (Kasmi et al., 2024). Les outils introduits dans cette thèse peuvent être transposés dans d'autres pays ou régions, où le problème d'observabilité du PV toiture se pose. La seule condition est de disposer de données de référence agrégées (par exemple, à l'échelle des municipalités) concernant la puissance installée.

Les contributions complémentaires de cette thèse sont une librairie Python open-source permettant l'extraction des caractéristiques du PV en toiture à partir de polygones géolocalisés (Trémenbert et al., 2023), et DeepPVMapper, un algorithme de cartographie open-source qui peut être réutilisé et amélioré par la communauté (Kasmi et al., 2023c).

Applications pour RTE et au-delà Ce travail montre que l'amélioration l'observabilité de la production PV en toiture nécessite peu d'information sur ces systèmes. L'approche de modélisation choisie pour la production d'énergie photovoltaïque montre que l'inclinaison, l'orientation, la puissance installée et la localisation, couplées à des données de rayonnement et de température, suffisent pour obtenir une estimation précise de la production électrique d'un système PV. Cette approche ouvre la voie à l'amélioration de la précision de l'estimation de la production PV en toiture et, par conséquent, de l'estimation de la production photovoltaïque globale à différentes échelles spatiales et temporelles, des estimations individuelles aux agrégats nationaux et de la réanalyse aux prévisions.

Le registre fournit une vue actuelle du parc photovoltaïque. Le GRT peut donc l'utiliser pour calibrer les modèles de potentiel photovoltaïque utilisés dans les études prospectives. Il peut également être utilisé pour analyser les facteurs géographiques, sociaux et économiques à l'origine de l'adoption du photovoltaïque, comme le font des travaux tels que Wang et al. (2022) ou Freitas et al. (2023). Ces registres peuvent également être utiles aux pouvoirs publiques qui cherchent un moyen simple d'évaluer l'état actuel du déploiement de l'énergie photovoltaïque sur leur territoire.

5.3 Discussion et perspectives

Discussion Lorsque j'ai commencé le projet, j'avais à ma disposition de très nombreuses approches, dont certaines se focalisaient sur des problématiques très pointues, comme par exemple le fait que les polygones des installations soient le plus rectangulaires possibles. D'un autre côté, il n'y avait pas encore eu de publication au sujet de la cartographie d'installations PV (toiture ou non) en France, ni de base de données d'apprentissage disponible pour la France. Face à ce contraste, j'ai choisi dans un premier temps de collecter des données d'apprentissage et de déployer un prototype construit à partir de la littérature existante en France. J'ai ensuite amélioré ce prototype après en avoir identifié les principales limites. Les présentations des premiers résultats obtenus m'ont rapidement montré que les performances du modèle, mesurées avec le score F1 ou l'indice de Jaccard, ne trouvaient que peu d'écho auprès des utilisateurs potentiels. Leurs préoccupations portaient essentiellement sur la cohérence des données générées avec les données existantes (la standardisation de ce processus ayant conduit à la DTA) ou sur la question de savoir comment être sûr que le modèle détectait bien des panneaux solaires. Si la GradCAM était suffisante pour exclure des corrélations fallacieuses évidentes comme des piscines, elle ne permettait pas d'expliquer pourquoi le modèle confondait parfois une piste d'athlétisme avec un panneau PV.

Je pense que pour la plupart des projets concrets de *data science*, les modèles "sur étagère" sont suffisants pour répondre à la plupart des besoins, *a minima* pour construire un prototype. Par ailleurs, les travaux visant à introduire un nouveau modèle état-de-l'art connaissent souvent une postérité fugace étant donné la vitesse à laquelle l'apprentissage statistique progresse.

Cependant, répondre à la simple question "Le modèle fonctionne-t-il correctement ?" et élaborer un protocole de contrôle et d'audit des modèles d'apprentissage profond, permettant en particulier de détecter et analyser leurs erreurs soulève pléthore de questions. Je suis convaincu qu'il reste encore beaucoup à faire dans ce que j'appellerais l'audit de l'IA, en particulier dans contexte où des modèles d'IA sont utilisés quotidiennement et pour beaucoup de tâches différentes par des utilisateurs non spécialistes. Je pense que fournir de bons outils à ces utilisateurs, c'est-à-dire des outils leur permettant de comprendre concrètement comment fonctionnent les modèles et quelles sont leurs limites peut améliorer la confiance placée en ces outils et notre recul critique envers ces derniers. Je pense que la confiance et l'esprit critique sont deux ingrédients essentiels pour une utilisation saine des algorithmes d'apprentissage profond par le grand public.

Perspectives Concernant l'ingéniérie des sytèmes électriques, il serait intéressant de se pencher sur l'autoconsommation PV. Le contexte législatif Français favorise l'auto-consommation individuelle, qui est devenue début 2024 le mode ma-

joritaire de raccordement au réseau. Une bonne connaissance de la production PV en toiture restera importante, d'autant plus que l'on peut s'attendre à ce que la télérelève de courbes de productions du PV en toiture reste rare : la seule information disponible sera la demande nette (c'est-à-dire la différence entre la production d'énergie photovoltaïque et la consommation du ménage). Ainsi, les estimations de production d'énergie devront être intégrées dans des modèles plus larges qui prennent également en compte la consommation, à l'échelle des ménages ou bien de quartiers.

Du côté de l'apprentissage statistique, je pense que l'étude des biais inductifs des modèles pourrait contribuer à mieux comprendre les modèles et à les rendre plus fiables. Notre travail a montré que les fausses détections sont dues principalement au fait que le modèle prédit un panneau solaire lorsqu'il identifie un motif quadrillé sur l'image. Notre WCAM nous a permis d'atténuer ce phénomène, mais il serait intéressant de comprendre son émergence. Cette question est plus théorique et dépasse le cadre de la présente thèse. Selon ma compréhension actuelle de la question, pour comprendre pourquoi le motif quadrillé finit par être une caractéristique prédictive importante, il faut comprendre comment le modèle construit des caractéristiques prédictives à partir des données d'entrée durant l'entraînement.

Chapter 1

Introduction

1 Context

1.1 Curbing anthropogenic CO_2 emissions through electrification and decarbonization

As the Sixth Assessment Report of the Intergovernmental Panel on Climate Change (IPCC, 2021a) states, "*Human activities, principally through emissions of greenhouse gases, have unequivocally caused global warming, with global surface temperature reaching 1.1°C above 1850-1900 in 2011-2020. Global greenhouse gas emissions have continued to increase, with unequal historical and ongoing contributions arising from unsustainable energy use, land use and land-use change, lifestyles and patterns of consumption and production across regions, between and within countries, and among individuals*".

Mitigating global warming requires reducing greenhouse gas emissions, particularly CO_2 emissions. To this end, options exist, roughly summarized as energy savings (efficiency) and decarbonization. Better insulation of buildings and favoring public transportation over individual vehicles are examples of energy efficiency measures. On the other hand, decarbonizing uses requires electrifying them, especially in the transportation sector. As a result, the so-called energy transition will lead to an increase in electricity generation to face the increase in consumption. The International Energy Agency (IEA) expects an increase of the share of electricity in the final energy demand by 4% yearly to meet the decarbonization goals (IEA, 2023). In France, the electricity consumption could rise from 459.3 TWh in 2022 to 580 to 640 TWh in 2035 (RTE France, 2023).

To meet the decarbonization goals, the share of electricity in the energy supply needs to increase *and* to resort to low-carbon sources massively. Electricity is an energy vector that results from the conversion of a primary source of energy to electric energy. Primary sources differ in their carbon intensity, and the decarbonization of the electric sector necessitates favoring low-carbon sources such as

renewable energies. According to Figure 1.1, these energies, especially wind and solar energy, offer the highest potential for contributing to the reduction of CO_2 emissions by 2030.

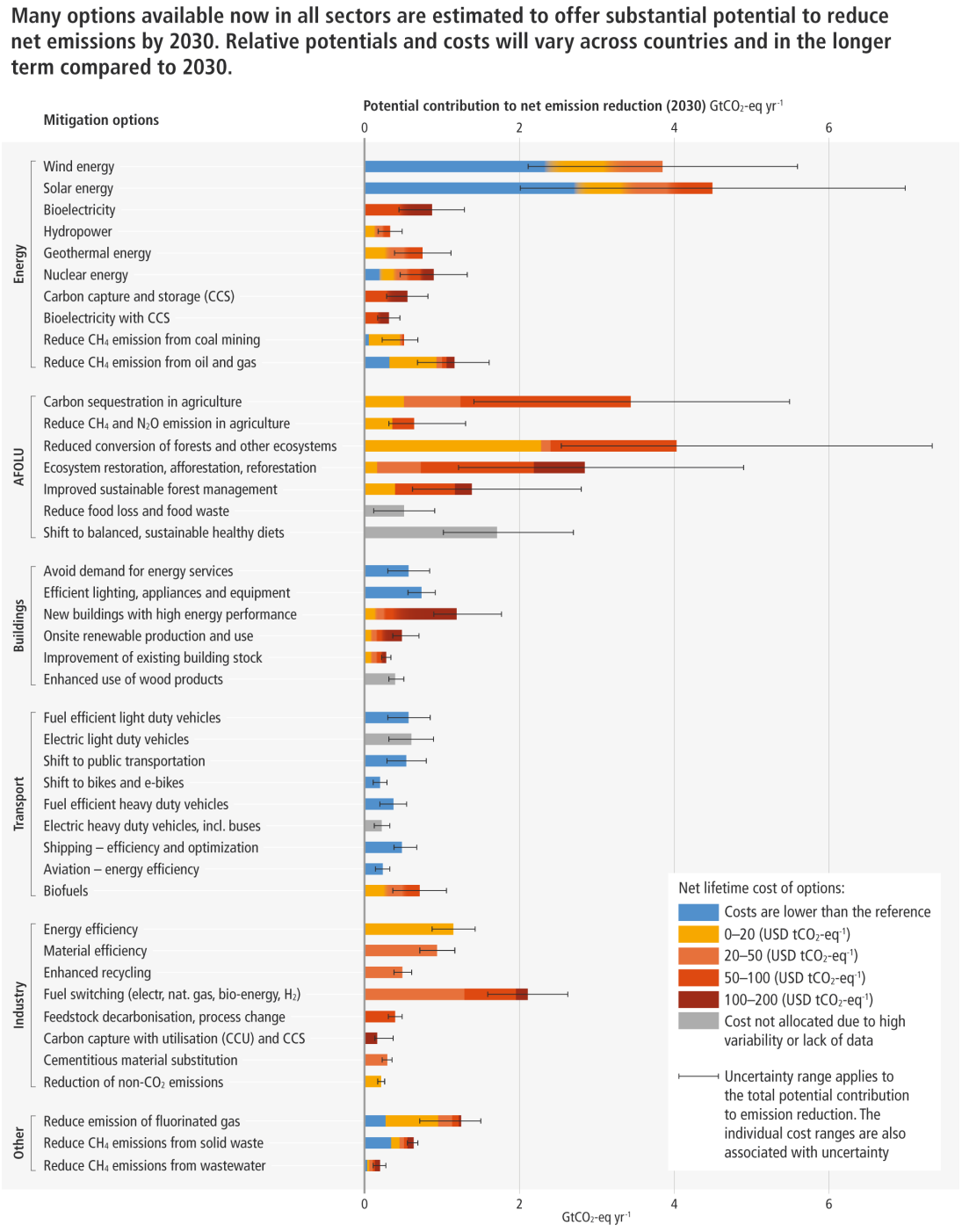


Figure 1.1 – Overview of mitigation options and their estimated ranges of costs and potentials in 2030. Source: IPCC (2021b).

Current climate action policies encourage the development of renewable energies. In the European Union (EU), the "Fit for 55" legislative bundle (European

Union, 2021) aims at cutting the EU's CO_2 emissions by 55% in 2030 compared to 1990, notably by setting the goal of having 40% of electric generation coming from renewable energies. In France, the *Programmation pluriannuelle de l'énergie* (PPE, 2020) and the *Stratégie nationale bas carbone* (SNBC, 2020) define the legislative framework for meeting the decarbonization goals of the country. The SNBC defines a trajectory for curbing CO_2 emissions and sets "carbon budgets" for each sector. The PPE focuses on the energy sector and defines energy consumption and generation goals. The PPE and SNBC are updated every five years. The current PPE, adopted in 2020, states that renewables must represent 40% of electricity generation by 2030. The development of renewables mainly focuses on wind and solar energy, as the potential of hydroelectricity is already at its peak, and the potential of biomass is negligible (about 2 GW or less than 1% of the current installed capacity).

1.2 Conditions for integrating high shares of wind and solar energy into the grid

In France, the transport network is a meshed grid of high-voltage lines, with voltages ranging from 63 kV to 400 kV. The operator responsible for the management and development of the transport network is the transmission system operator (TSO). The French TSO is the *Réseau de transport d'électricité*[1] (RTE), which is a public regulated monopoly. At all times, two constraints must be satisfied for the network to operate. Firstly, the sum of injections must equal the sum of the off-takes. Injections correspond to energy production and imports. The off-takes correspond to the exports, the consumption, and the network losses. This first constraint is a global constraint. Secondly, the intensity of the flows must not exceed the transit capacity of the lines. Otherwise, this could damage them, e.g., by overheating or present risks to the people and the environment. This second constraint is local, meaning that the localization of the injection and off-takes also needs to be considered.

Wind and solar photovoltaic (PV) electricity generation is weather-dependent and highly variable at different time and space scales. Variable power generation increases the grid's sensitivity to climate, and the uncertainties as wind and solar power generation are variable. To limit the uncertainties they bring to the grid, it is necessary to accurately observe (i.e., accurately measure or estimate) their power production. I define *observability* as the ability of the TSO to accurately estimate a power unit's real-time and future production. In practice, the TSO either measures in real time the production of the power unit (i.e., telemetry) or has access to *ex-post* measurements of the production (usually within one month). These *ex-post* measurements enable the calibration of power estimation models.

1. Website: https://www.rte-france.com/.

Accurate measurements of renewable power generation at the scale of the power units are the basis for validating short-term forecast models, which use weather forecasts to estimate future renewable power production and the power system's margins necessary to compensate for the variability of the production. Wind power generation is homogeneously measured (RTE France and IEA, 2021), but it is not the case for PV power production. A distinctive feature of PV energy is that PV installations vary in size. The smallest PV power installations have a few kW_p installed capacity, while the largest plants can have an installed capacity of up to dozens of kW_p or even several MW_p or GW_p. The variability in size results in a great diversity in terms of the installations' technical characteristics, as seen from Figure 1.2.

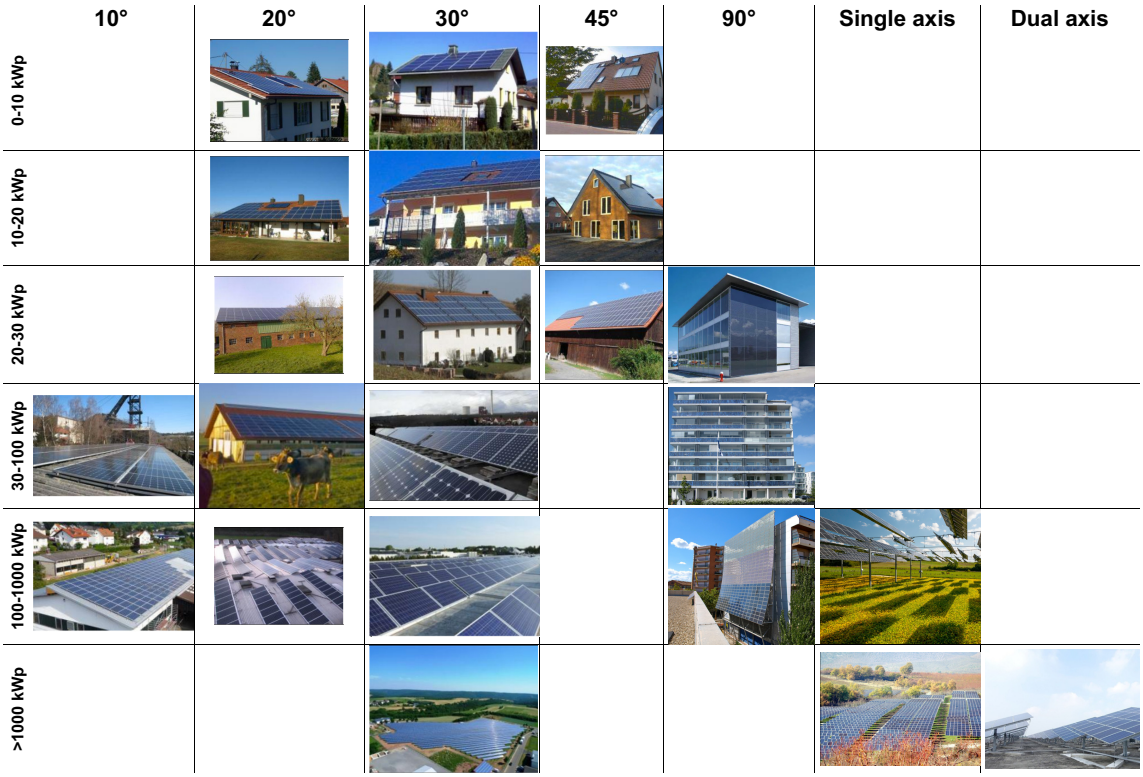


Figure 1.2 – Typology of PV installations. Rows correspond to classes of installed capacities and columns to classes of tilt angles (in degrees). Adapted from Saint-Drenan et al. (2015).

Table 1.1 summarizes the current state of PV observability, depending on the power class of the installation. Overall, about 94% of the PV fleet is not observable, corresponding to 22% of the installed capacity, equivalent to four nuclear units. The lack of observability primarily concerns small-scale installations with installed capacity below 36 kW_p.

Table 1.1 – Installed capacity, number of installations, and corresponding shares of observed PV installation by power class as of September 2023. TN: transport network. DN: distribution network. Source: RTE.

Power class [kW$_p$]	**Observed**		**Not observed**	
	Installed capacity [MW$_p$]	*Number of installations* [-]	*Installed capacity* [MW$_p$]	*Number of installations* [-]
0 - 3	0.006	3	1124	414579
(%)	*0.0*	*0.0*	*100.0*	*100.0*
3 - 6	0.05	9	872	171042
(%)	*0.0*	*0.0*	*100.0*	*100.0*
6 - 9	0.439	53	397	46704
(%)	*0.1*	*0.1*	*99.9*	*99.9*
9 - 36	14.3	340	637	25893
(%)	2.2	1.3	97.8	98.7
36 - 250	4438	40054	340	2897
(%)	92.9	93.3	7.1	6.7
250 - 1000	435	748	23.5	47
(%)	94.9	94.1	5.1	5.9
≥ 1000 (DN)	7586	1531	377	79
(%)	95.3	95.1	4.7	4.9
≥ 1000 (TN)	827	20	0	0
(%)	*100*	*100*	*0*	*0*
Total	13301	42758	3771	661241
(%)	*77.9*	*6.1*	***22.1***	***93.9***

The lack of observability will be increasingly concerning in the context of the quick growth of PV installed capacity. Following Figure 1.3, we can see that the overall PV installed capacity could reach up to 200 GW$_p$ in 2050 (RTE France, 2022). The PPE aims to reach 35 and 45 GW$_p$ of PV installed capacity by 2029. These scenarios and objectives assume a constant deployment rate for small-scale and large PV, meaning that up to 40 GW$_p$ (i.e., two-thirds of the current French nuclear park) could be unobserved by 2050 with the current practices. Therefore, **the industrial objective of this thesis is to seek methods for improving the observability of small-scale rooftop PV (i.e., PV installations with an installed capacity below 36 kW$_p$)**.

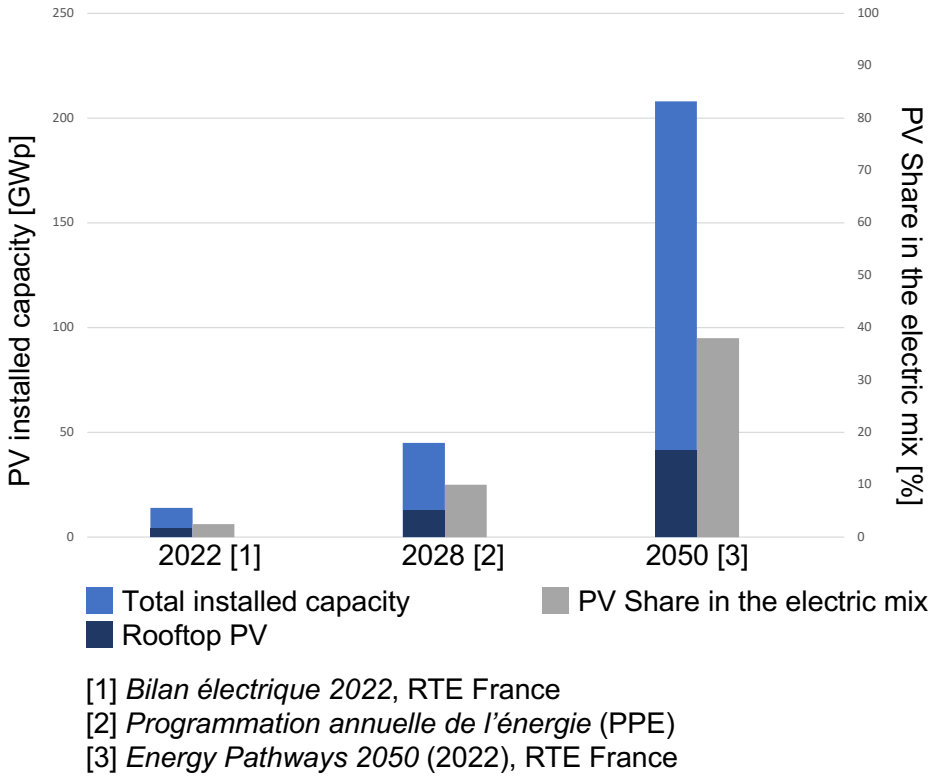


Figure 1.3 – Expected PV share growth according to the PPE and RTE's Energy Pathways 2050.

1.3 Overcoming the poor PV observability

As depicted on Table 1.1, in September 2023, 661,241 installations, amounting to 3.7 GW_p (i.e., 22% of the PV installed capacity) are currently not observed. Acquiring telemetry or *ex-post* measurements for all of these installations is computationally unbearable.

The current practice estimates the *regional* PV power production (i.e., the PV power production aggregated for areas up to a few hundred km^2). These areas are then summed to derive the PV power production at the scale of a country. My method will borrow from the probabilistic approach (Saint-Drenan et al., 2015), the idea of estimating the PV power production using a conversion model that requires a limited set of PV characteristics and solar irradiance and temperature data.

The probabilistic approach estimates the PV power production of a target set of systems for which no measurements are available. Traditional approaches (e.g., the upscaling method of Lorenz et al., 2011) rely on the production of metered neighboring plants. However, Saint-Drenan et al. (2016) showed that this method leads to *interpolation errors* when the neighboring plants are too far away. Estimation errors also arise when the reference set of installations is not representative of the target systems set. Instead of neighboring plants, Saint-Drenan et al. (2016) proposed using a physical PV system model and solar irradiance data. Indeed, only a few pa-

rameters are necessary to estimate the PV power production of a system. However, it is necessary to acquire these parameters. Saint-Drenan (2016) showed that in addition to its localization, the tilt, azimuth angles, and the nameplate capacity (or installed capacity) of the PV installation are sufficient. As existing data sources do not provide systematically this information for rooftop PV installations, one needs to acquire them through mapping small-scale rooftop PV installations over France. The outcome of this mapping is a **technical registry** (or registry) recording at least the localization, tilt, and azimuth angles and installed capacity of as many rooftop PV installations as possible. By rooftop PV installation, I refer to installations with an installed capacity lower than 36 kW_p. I may also refer to these installations as *small-scale* or *distributed*.

2 Literature review

2.1 Earth observation data for large-scale mapping of rooftop PV installations

Definitions Classification, segmentation, and the detection, recognition, and identification (DRI) framework: In remote sensing, one usually distinguishes between object detection, recognition, and identification. Object detection means that the goal is to know that *something* is here without knowing what instance it is. In our case, it corresponds to PV panel *classification*. The model predicts that the input image depicts a PV panel. Object recognition corresponds to classifying a detected instance into a given category. In our case, it would correspond to *segmentation*: we know where the PV panel is and its size. Finally, identification means one can derive specific information about the recognized instance. This corresponds to the characteristics extraction, where one identifies that the PV panel has a given nameplate capacity, tilt, and azimuth angles. For the remainder, I will use the terms image classification and segmentation to refer to the detection and recognition of PV panels.

Early works The field of remote sensing of PV installations using spaceborne or airborne orthoimagery [2] and computer vision techniques are now well established. I refer the reader to Puttemans et al. (2016), de Hoog et al. (2020) and Arnaudo et al. (2023) for comprehensive reviews on the topic. To our knowledge, the first work dealing with remote PV panel detection on orthoimagery is Malof et al. (2015). In this work, the authors leveraged aerial orthoimagery provided by the United States Geological Survey (USGS) to detect installations over Lemoore (California). The main question was whether aerial photography could enable automatic PV ar-

2. Orthoimagery is defined as overhead imagery geometrically corrected such that the scale is uniform: the image follows a map projection

ray detection. In an elementary setting with 50 images containing PV panels and 50 images without panels, the authors established that one could use aerial orthoimagery to map solar installations. The authors proposed a two-step method. First, a set of features is manually computed for each input image pixel. Then, they used a support vector machine (SVM) classifier to predict whether the pixel depicted a PV array. In Malof et al. (2016a,b), the same research team improved this work with better hand-crafted features (i.e., individual measurable properties) that include the surroundings of the pixel of interest and a random forest classifier to classify pixels of aerial images automatically. This method, deployed in a much more challenging and realistic setting than the first attempt, achieved an F1 score [3] of 0.65. Besides hand-crafted based detection methods, the use of convolutional neural networks (CNNs) has also been investigated by Yuan et al. (2016) and later by Malof et al. (2017) and Golovko et al. (2017). In the latter work, the F1 score reached 0.79. Finally, Bradbury et al. (2016) introduced the first publicly available dataset containing ground truth annotation of rooftop PV panels.

Deep learning-based detection on RGB images as a standard Several works (Czirjak, 2017; Karoui et al., 2018, 2019; Ji et al., 2021) attempted to use infrared and hyper-spectral imagery to detect PV panels. The rationale is that PV modules have an identifiable spectral signature, i.e., a variation of reflectance or emittance of a material concerning wavelengths, constant across manufacturers (Ji et al., 2021). The approach typically consists of (i) using portable spectrometers to determine the spectral signature of the solar array and construct a discriminative index and (ii) classifying pixels if the discriminative index exceeds a certain threshold. So far, these methods have neither been applied at a large scale nor demonstrated their benefit in terms of accuracy compared to CNN-based methods on orthoimagery. Moreover, measurements on the arrays replace the data-labeling step (e.g., in Czirjak, 2017), thus casting doubt on benefits in terms of data labeling can be discussed. Besides, the spectral signature of PV in the near-infrared domain can sometimes be mixed with unrelated components such as oil (Ji et al., 2021). Therefore, the literature overwhelmingly adopted methods based on RGB imagery and CNN-based models. Training data and pre-trained models are easily accessible, so even if a data labeling step is still required, it is easier to leverage RGB imagery and deep learning models for detecting PV panels.

Towards large scale mapping The DeepSolar project (Yu et al., 2018) introduced the first large-scale PV registry covering the continental US and reporting the surface area and the number of installations. DeepSolar leverages CNNs to ef-

3. The F1 score measures the accuracy of a binary classifier. A perfect classifier has a F1 score equal to 1. The F1 score is the harmonic mean between the precision and the recall of this classifier. The main accuracy metrics are defined in chapter 4.

ficiently detect PV installations from overhead imagery and to estimate the surface area they cover. Their surface area estimation only leveraged the horizontal projection of the array on the image and reached a mean relative error [4] almost always lower than 5%. Similar works such as SolarMapper (Malof et al., 2019) also leveraged deep learning-based methods to map the surface area of distributed installations. SolarMapper used a segmentation model, i.e., a deep learning model that identifies which input image pixels depict a PV panel. Subsequent works mapped numerous areas including North-Rhine Westphalia (Mayer et al., 2020), Switzerland (Casanova et al., 2021), Oldenburg in Germany (Zech and Ranalli, 2020), parts of Sweden (Lindahl et al., 2023; Frimane et al., 2023), Northern Italy (Arnaudo et al., 2023), the Netherlands (Kausika et al., 2021) or the surroundings of Berkeley in California (Parhar et al., 2021). Several works even included GIS data to construct registries of PV installations (Kausika et al., 2021; Mayer et al., 2022; Rausch et al., 2020). In the current context of rapid rooftop PV growth (Haegel et al., 2017), remote sensing of rooftop PV installations using deep learning and orthoimagery emerged as a promising solution to address the lack of systematic registration of small-scale PV installations (Kausika, 2022).

2.2 Current methods are not reliable enough to be integrated into critical industrial processes

The current methodology for mapping rooftop PV installations on orthoimagery consists of first training CNN models for image classification (matching an image to a pre-existing category) and segmentation (delineating the pixels on the image that correspond to the predicted category). These models are then integrated into a larger pipeline where they are used to extract polygons of rooftop PV installations from unlabelled orthoimagery. Depending on the works, the polygons are used to estimate the surface area or the installed capacity or combined with additional data such as 3D building data to derive the tilt and azimuth angles (Mayer et al., 2022). Some works (Hu et al., 2022; Malof et al., 2019) do not use the classification step.

The main limitation of current approaches is the spatial and temporal extent to which a trained model can generalize to (Tuia et al., 2016). Wang et al. (2017) showed in a small experiment that a classification model trained on a city generalizes poorly to another city unseen during training. Their explanation for this phenomenon was that the panels were more complex to recognize in one city than another. De Jong et al. (2020) and Arnaudo et al. (2023) underlined the fact that models trained over a region (e.g., Germany) cannot generalize to neighboring

4. Yu et al. (2018) define the mean relative error (MRE) as

$$MRE = \frac{\sum_{i=1}^{\#\text{true positives}} \text{true area}_i - \text{estimated area}_i}{\sum_{i=1}^{\#\text{true positives}} \text{true area}_i}$$

countries. On the temporal side, to build a historical database of rooftop PV adoption, Wang et al. (2019) had to construct a model based on Siamese networks [5] and cross-correlation modules to identify when a PV panel appeared on a satellite image. De Jong et al. (2020) identified the lack of generalizability to new regions and new images as the main limitation when using machine learning-based techniques to construct official statistics.

2.3 On the limits of deep learning in applied settings, beyond the case of the detection of rooftop PV installations

Deep learning suffers from a sensitivity to distribution shifts The uncertain ability to generalize to unseen settings is more broadly referred to as the sensitivity to **distribution shifts**, i.e., the sensitivity to the fact that *the training distribution differs from the test distribution* (Koh et al., 2021).

The sensitivity to distribution shifts causes unpredictable performance drops. As a result, the performance reported on the training dataset no longer represents the accuracy under real-life scenarios. These performance drops can have dire consequences as models are deployed in safety-critical settings, e.g., autonomous driving (Sun et al., 2022b) or medical diagnoses (Pooch et al., 2020). Numerous approaches have been introduced to mitigate the sensitivity to distribution shifts. I refer the reader to surveys such as Zhou et al. (2023); Tuia et al. (2016); Guan and Liu (2022); Csurka (2017); Csurka et al. (2021) for reviews of these methods in various settings. I broadly refer to the methods aiming at mitigating the sensitivity to distribution shifts as "domain adaptation" (Saenko et al., 2010) methods. The general idea is that a model is trained on a source training dataset S (e.g., labeled images of PV installations in France) and is deployed on one or several target datasets T. Gulrajani and Lopez-Paz (2021) showed that the standard empirical risk minimization [6] (ERM, Vapnik, 1999) method was a strong baseline for domain adaptation while being significantly more straightforward to use in practice. Besides, domain adaptation is a long-tail problem, meaning unseen situations eventually arise, and all situations cannot be accounted for (Torralba and Efros, 2011; Recht et al., 2019). Therefore, rather than introducing a new domain adaptation method, I will focus on understanding why distribution shifts affect a model's performance to assess whether its predictions are *reliable*.

Reliability in the context of deep learning To assess whether a *point-wise* decision (Schulam and Saria, 2019) is correct (as opposed to a decision that is correct on average), one needs to assess the reliability of individual predictions.

5. Siamese neural networks are a class of neural networks that leverage two identical networks with shared weights to learn the feature representation of different inputs. I refer the reader to Chicco (2021) for a review on the topic.

6. I provide a short introduction to the key notions of machine learning in appendix D.

I define the **reliability** of a deep learning algorithm as the combination of three factors:

— The *relevance* of its decision process: one wants to be able to evaluate whether the model relies on good features to make its prediction, i.e., is right for the right reasons (Ross et al., 2017),

— The *robustness* of the decision process: one wants the decision process to be invariant to distribution shifts, as these perturbations eventually arise when dealing with vast datasets (Peng et al., 2017),

— The *monitoring* of the output data: one wants to identify where the model fails (Schulam and Saria, 2019) by implementing a strategy to assess the model's accuracy indirectly. By monitoring, I mean the ability to keep the quality of the data produced by the registry under systematic review of the user.

3 Scientific questions and outline

3.1 Scientific questions

Mapping rooftop PV installations using deep learning and orthoimagery is feasible. However, industrial applications require quality standards regarding the data and the methods, which current works fail to meet. This thesis aims to define these standards and introduce a methodology to assess whether deep learning-based mapping systems can meet them. To this end, I will address the following general scientific question:

Is deep learning-based remote sensing on orthoimagery a suitable method for constructing a nationwide technical registry of rooftop photovoltaic (PV) installations intended to improve the observability of PV power production in France?

I shall address this question by tackling the following sub-questions (SQ):

(SQ1) What requirements should the registry have, and how can we check whether it meets these requirements? I first need to state what I need to know about the rooftop PV installations and how I can acquire such knowledge. I also need to assess the quality of the data contained in the registry and check its adequacy with the actual rooftop PV fleet.

(SQ2) How can we ensure deep learning models reliably map rooftop PV installations? I use deep learning models on vast amounts of data, which can be subject to unpredictable alterations. To ensure that the impact of these factors is as small as possible on the model's decision process, I need to unveil the model's

decision process and ensure that this process will not be impacted by the alterations that can occur to the data. I will identify these alterations first.

(SQ3) How to build and integrate the registry for rooftop PV power production estimation and evaluate its relevance for improving PV observability? Once I have designed the proper tools to ensure that a deep learning model can generate quality data reliably, I will have to deploy it over France, which means I will need adequate training data. As the registry is intended to improve PV observability, I will have to construct a benchmark using reference data to evaluate whether rooftop PV power estimations derived from the registry can contribute to improving the accuracy of the estimation of PV power production.

3.2 Approach and outline

Sparing the computational resources This thesis work is part of a more global commitment to mitigate and adapt to climate change. Considering the environmental impact of the tools used in this work is essential. Generally speaking, I preferred to reuse and discuss existing models rather than implement and train new models from scratch and argue in favor of taking into account the computational cost of methods alongside traditional performance metrics when evaluating them, as done, for instance, by Hugging Face (2023). In appendix A, I introduce a simple approach to take into account the computation cost of different methods and discuss how the energy cost of my approach can be translated into terms of environmental impact.

Thesis outline Table 1.2 summarizes the chapters of this thesis and outlines which research questions they address. In chapter 2, I review the existing data sources to map PV installations in France and define key performance indicators (KPIs) for our registry. I introduce a method to measure these KPIs without ground truth data. Using this method, called downstream task accuracy (DTA), I quantify the variation in accuracy that occurs during the large-scale deployment of the mapping algorithm. In chapter 3, I empirically show that *acquisition conditions* mainly contributes to the variations in accuracy. I assumed that PV panels could be decomposed into different scales on the input images to show this. As acquisition conditions perturb some scales, the prediction may be disrupted if the model relies on these scales. I introduced a new feature attribution method called the wavelet scale attribution method (WCAM) to identify which scales the model relies on. I derived from the WCAM a data augmentation technique to improve the model's robustness to acquisition conditions. Chapter 4 combined the methods of the two former chapters to introduce DeepPVMapper, our proposal for scalable and reliable mapping of rooftop PV installations. This algorithm introduces a new filtering

method to minimize the computational burden of large-scale deployment. I evaluated the reliability of our algorithm through extensive benchmarks and presented the results of this mapping approach. Finally, chapter 5 introduces a power production estimation model fitted for our registry to estimate the rooftop PV power production. I evaluate the accuracy of our method using ground truth measurements of rooftop PV installations and discuss its benefit for the TSO by comparing it with the current methods. Chapter 6 summarizes and discusses the present work.

Table 1.2 – Outline of the chapters of this thesis and the scientific sub-questions (SQ) they address.

Chapter	Title	SQ1	SQ2	SQ3
2	Characterization and evaluation of the quality of the rooftop PV registry in the absence of ground truth labels	✓		
3	Assessing the reliability of a model's decision process by generalizing attribution to the wavelet domain		✓	
4	Constructing a reliable and scalable algorithm for mapping rooftop PV installations in France			✓
5	Assessing the gains for rooftop PV observability of a physics-based method using a detailed registry of PV installations and solar irradiance data			✓

Associated publications Several research papers are associated with this thesis. The chapters of this manuscript consist of extended and enriched versions of material already published or under publication at the time of writing. This manuscript is self-contained, so it is unnecessary to read the associated publications to understand its content. I refer the interested reader to the appendix E for a more thorough presentation of these works and links to the documents. Table 1.3 lists for each chapter the associated research papers and the sub-questions they contribute to addressing.

Table 1.3 – Summary of the publications associated with this thesis.

Reference	Publication type	Contributes to ...		
		SQ1	SQ2	SQ3
Chapter 2				
A crowdsourced dataset of aerial images with annotated solar photovoltaic arrays and installation metadata (Kasmi et al., 2023d)	Journal	✓	✓	✓
Towards unsupervised assessment with open-source data of the accuracy of deep learning-based distributed PV mapping (Kasmi et al., 2022a)	Conference workshop	✓		
Chapter 3				
Can We Reliably Improve the Robustness to Image Acquisition of Remote Sensing of PV Systems? (Kasmi et al., 2023b)	Conference workshop		✓	
Assessment of the Reliability of a Model's Decision by Generalizing Attribution to the Wavelet Domain (Kasmi et al., 2023a)	Conference workshop		✓	
Chapter 4				
`PyPVRoof`: a Python package for extracting the characteristics of rooftop PV installations using remote sensing data (Trémenbert et al., 2023)	Preprint			✓
DeepPVMapper: reliable and scalable remote sensing of rooftop photovoltaic installations	Submitted work (journal)			✓
Chapter 5				
Remote Sensing-Based Estimation of Rooftop Photovoltaic Power Production Using Physical Conversion Models and Weather Data (Kasmi et al., 2024)	Journal			✓

Chapter 2

Characterization and evaluation of a rooftop PV registry in the absence of ground truth labels

Summary

This chapter discusses how to monitor the accuracy of a PV registry created with Earth observation data and deep learning. We review the data sources at our disposal for constructing and evaluating the registry, define key performance indicators (KPIs) for industrial use, and introduce an unsupervised evaluation method, the Downstream Task Accuracy (DTA). The DTA indirectly monitors the accuracy of the registry by aggregating its data, comparing it with available sources and deriving accuracy metrics. We assess tilt and azimuth angle estimations against the self-reported BDPV database and installed capacity estimation against the aggregation of the city-level *Registre national d'installations* (RNI).
We train a model on the training dataset BDAPPV and deploy it across 11 French departements. The DTA reveals satisfactory tilt and azimuth angle estimations but relatively poor installed capacity estimation accuracy. Investigating the performance drop, we find no geographical pattern impact. Analyzing how the model makes predictions, we uncover that it relies on features correlated with PV panels but not causally related. This can lead to confusions, such as mistaking a veranda for a PV panel. This explanation helps us understand some cases where the model performed very poorly and offers insights towards improving the reliability of the classification model, which will be the topic of the next chapter.

1 Overview of the existing and missing data sources

This section reviews the data for mapping rooftop PV installations from orthoimagery. We also review the existing PV data sources. We show that none meets all the criteria for constructing the technical registry of rooftop PV installations we defined in the introduction. We recall that this registry should record the localization, the tilt and azimuth angles, and the installed capacity of as many rooftop PV installations with an installed capacity of less than 36 kW_p as possible. Finally, we introduce our training database BDAPPV for training models to map PV installations over France.

1.1 Geographical information system (GIS) data

1.1.1 Orthoimagery and topological data

Different types of overhead imagery Overhead imagery encompasses satellite and airborne (or aerial) imagery. Figure 2.1 presents airborne and spaceborne imagery samples.

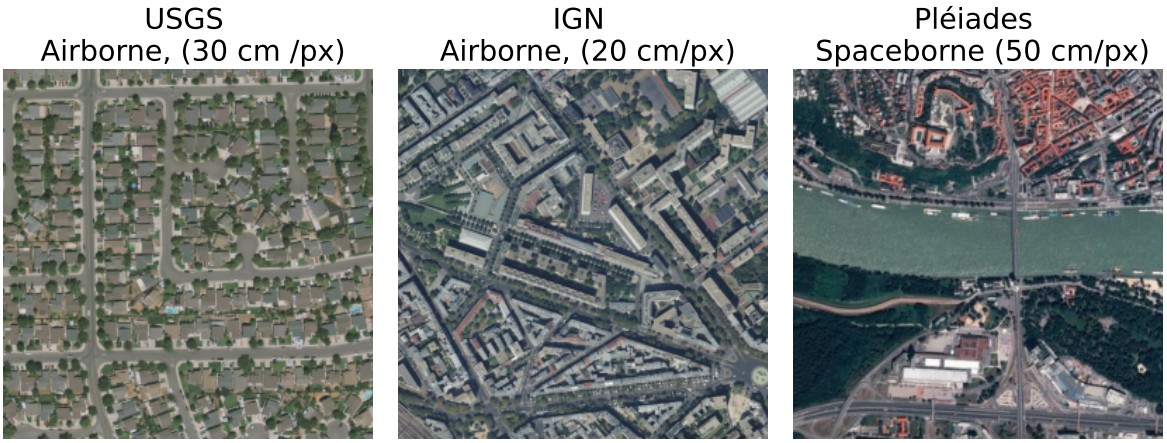


Figure 2.1 – Examples of different types of orthoimagery. Sources: USGS (2024), IGN (2024a), and the ESA (2024). USGS and IGN are updated every three years on a rolling basis, and Pléiades from ESA is updated twice a day.

For a fixed location, aerial imagery has three main characteristics. First, the ground sampling distance (GSD). The GSD corresponds to the distance between two consecutive pixels measured on the ground. The lower the GSD, the more details on the image. The GSD is expressed in meters per pixel.

Second, its effective resolution, which enables us to assess its quality. The effective resolution considers the distortions induced by the angle of incidence of the sensor (e.g., RGB camera). If the sensor is tilted, the effective resolution is larger than the GSD. For instance, if a sensor has a GSD of 50 cm/pixel but is very

tilted with respect to some places, it can have a higher effective resolution (e.g., 70 cm/pixel). In this case, the resulting image, despite theoretically having a GSD of 50 cm/pixel, will have a quality equivalent to an image with a GSD of 70 cm/pixel (with no distortion). The GSD gives us an upper bound on the image quality.

The third characteristic of overhead imagery is its revisit rate. Satellite images have a higher revisit rate at the expense of a higher GSD, while aerial images are usually more detailed but have a lower revisit rate.

Orthoimagery Orthoimagery is overhead imagery that is rectified for the angle of incidence of the sensor. Therefore, each pixel on the orthoimagery appears as if the sensor had been right above the localization. Orthoimagery is obtained after correcting the images and induces fewer geometric distortions than on non-rectified images. Figure 2.2 presents an image example before and after orthorectification.

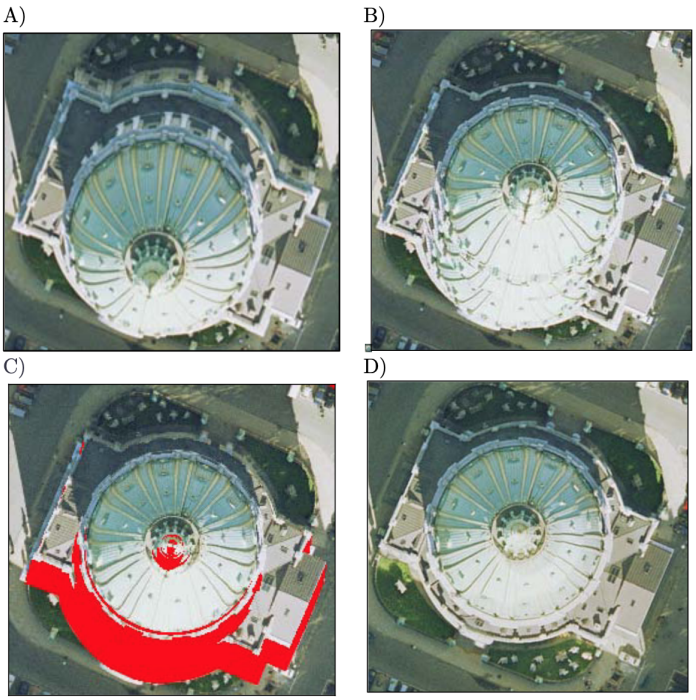


Figure 2.2 – A) An orthophoto rectified over a terrain model. The church is not moved to its correct position. B) Orthophoto based on a city model. The church is rectified to its correct location, but a "ghost image" is left on the terrain. C) Same as B, but the obscured area has been detected. D) True orthophoto, where the obscured area has been replaced with imagery from other images. Taken from Nielsen (2004).

We will use the the BD ORTHO database (IGN, 2024a), provided by the French *Institut national de l'information géographique et forestière* (IGN). These images are provided under an open license. The revisit rate for these images is three years, but updates are provided on a rolling basis, so updated images of departements are available every month. The images are classified by departements, a French

administrative unit between the city and the region[1]. There are 95 departements in metropolitan France. The ground sampling distance of these images is 20 cm/pixel, which is sufficient to detect PV systems (Li et al., 2021). The effective resolution is at most 30 cm/pixel. This effective resolution is derived by considering the expected GSD (20 cm/pixel) and the expected error between the true localization of the points (measured on the ground) and their localization derived from the BD ORTHO. See IGN (2024c) for more details on the quality controls of IGN images.

Topological data In addition to the orthoimagery, the IGN also provides (under open access) a register of all buildings and infrastructure in France. It should be noted that topological data is *not* a surface model. The buildings are 2D polygons, and we do not have information on the building or rooftop height from this database. Figure 2.3 presents an example of layers of the BD TOPO (IGN, 2023) opened in the GIS software QGIS. The BD TOPO is updated every three months.

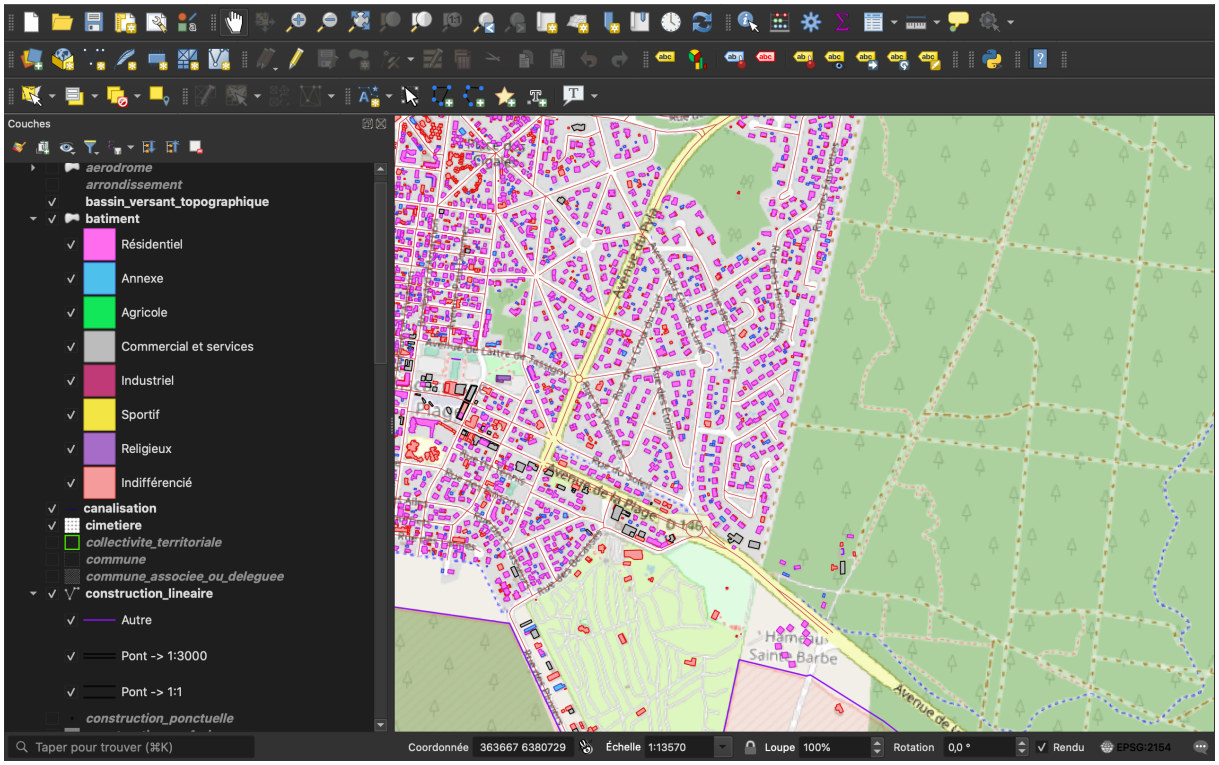


Figure 2.3 – Screenshot of the QGIS software displaying building layers from the BD TOPO.

1. A departement (*département* in French) corresponds to the NUTS-3 territorial subdivision level, according to the European Nomenclature of Territorial Units for Statistics (NUTS). See European Union (2024) for more details on the NUTS.

1.1.2 Digital surface models (DSMs)

Rasterization The rasterization is a process that converts an image in raw or vector format into a *raster image* described in pixels or dots. For digital surface models, rasterization involves converting a tri-dimensional points cloud into a two-dimensional image, where each pixel's value marks the height at the location (x, y).

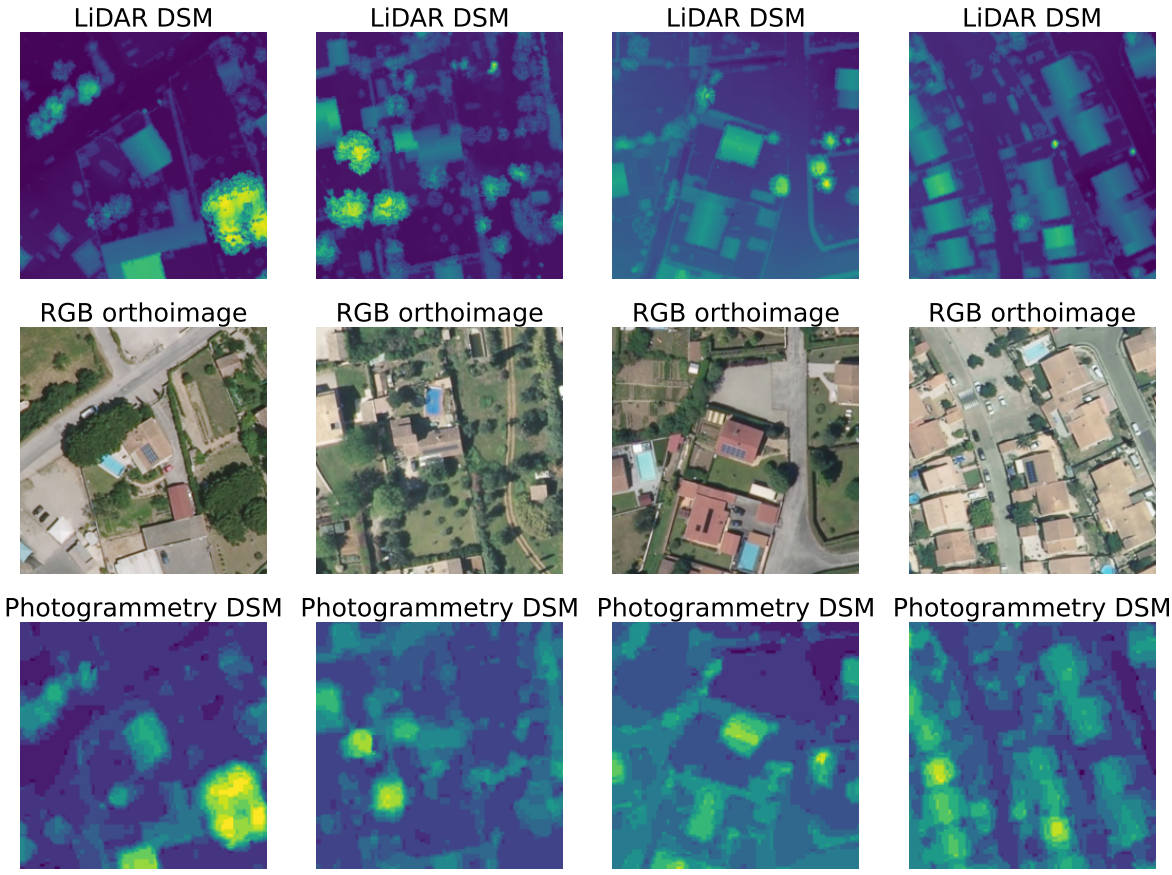


Figure 2.4 – Examples of rasters extracted from the photogrammetry DSM (bottom row) and the LiDAR DSM (upper row) The middle row presents the RGB image associated with these digital surface models. We can see that despite the same GSD, the LiDAR data is more accurate due to its finer effective resolution. Source: IGN.

Photogrammetry DSM Photogrammetry uses parallax to get the altitude points associated with each coordinate. Indeed, altitudes over an area are determined from different pictures from different points of view as nearer objects (from the aircraft carrying the aerial camera) move faster than distant objects. Such data is available almost everywhere in France, with a ground resolution of around 20 cm/pixel and an altimetric precision of around 150 cm.

LiDAR data Light Detection and Ranging (LiDAR) calculates distances from the reflection of a light beam on a surface. This technique has an altimetric resolution of

10 cm/pixel. Even though LiDAR raw data is composed of point clouds with around ten points/m², we have decided to interpolate and rasterize it to a 20 cm/pixel resolution to use the same developed methods to infer tilts and azimuths. By the time this thesis was written, the LiDAR HD DSM provided by the IGN did not cover all of France and is only accessible for demonstration purposes on the IGN's dedicated webpage (IGN, 2024b). Figure 2.4 presents rasters coming from the photogrammetry DSM and the LiDAR DSM.

1.2 PV registries and databases for France

1.2.1 The *Registre national d'installations* (RNI)

The RNI is an official registry that records all installations from all energy sources connected to the French grid (distribution and transportation networks). This data is openly accessible and updated every three months. The RNI aggregates information from the TSO and the distribution system operators (DSOs) for PV installations, as most PV installations are connected to the distribution network. The RNI provides a brief technical description of the PV installation (localization and installed capacity). Due to privacy constraints, installations below 36 kW_p are registered as aggregated installations in the registry. The aggregation is done at the city level. Therefore, the RNI indicates the number of installations below 36 kW_p for each city and their aggregated installed capacity. The RNI is accessible on the online portal *Open Data Réseaux Energies* (ODRE, 2024 [2]).

1.2.2 RTE internal data

RTE has access to the individual list of installations with their respective city and installed capacity for all installations connected to RTE's network and for installations connected to the network of the DSO Enedis. Enedis is the main DSO, covering 95% of the distribution network. The five missing percent correspond to the territory of the other French DSOs, which amount to about 160 in France (Commission de Régulation de l'Énergie, 2024). These DSOs most often cover cities and at most individual departements. RTE updates its internal database every three months. This dataset does not contain the technical characteristics of the installations. Besides, RTE cannot access real-time or past power measurements from small-scale PV installations.

1.2.3 The *Base de données photovoltaïque* (BDPV)

Asso BDPV is a non-profit association that enables individual owners of rooftop PV installation to monitor the PV power production of their PV system. This associ-

2. Website : https://opendata.reseaux-energies.fr/.

ation maintains a database of PV installations (BDPV), which records the technical characteristics, including the tilt and azimuth angles and the nameplate capacity, but also the surface area of the installation, the number, and manufacturer of PV modules, etc. This database is not exhaustive and is based on PV plant owners' declarations. The database contains over 28,000 installations, including more than 24,000 in metropolitan France. For approximately 2,000 installations, individual production records are also accessible. As individual owners of PV installations maintain it, the installations mostly have an installed capacity below 36 kW_p (see Figure 2.5) and records about 3.6% of the total number of rooftop PV installations (according to Table 1.1). Finally, for privacy reasons, the precise localization of the plants is private. We had access to the precise localization of the installations for constructing BDAPPV (which we introduce in section 1.4 of this chapter) but could not publish this information.

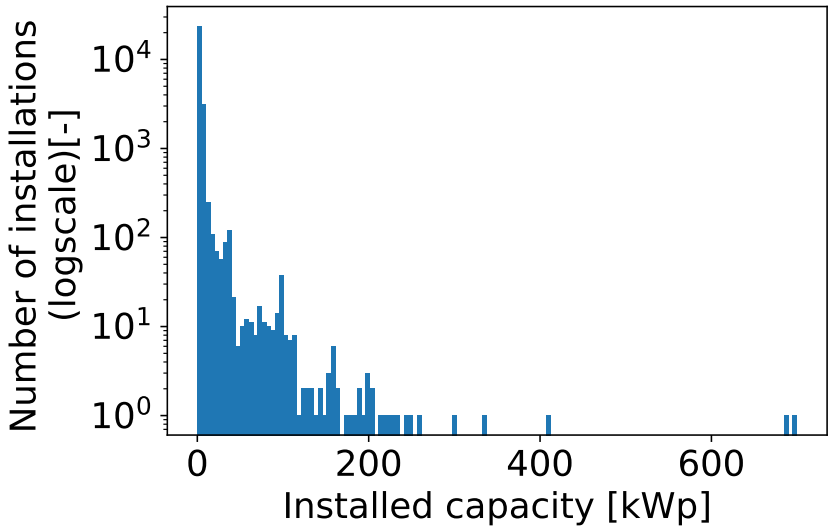


Figure 2.5 – Distribution of the installed capacities registered in BDPV. Source: BDPV.

1.3 Specific requirements of the PV registry

Now that we have introduced and reviewed the existing data sources for mapping PV installations, we present the requirements that our PV registry should meet and identify what is missing among existing data sources.

1.3.1 Overview of the criteria

The primary motivation for constructing our registry is to gather information on rooftop PV installations to estimate PV power production. For this task, our registry must satisfy four conditions: disaggregation, technical characteristics, representativeness, and updatability.

Disaggregation The registry must contain the individual list of installations. We need to identify individual installations because we need their precise location to match them with the correct weather data and cluster them in the correct cell around a power substation.

Technical characteristics Our registry must gather the minimal characteristics required to estimate PV power production using physical models. Killinger et al. (2018) showed that this set of characteristics is relatively limited: we need the tilt, azimuth angles, and nameplate capacity.

Representativeness Our registry must represent the actual installed capacity. Ideally, we wish that it covers all the territory. In practice, we require no sampling bias, such as those that occur in self-reported registries where the distribution of installations over the territory depends on the number of people who register their installation and not only on the number of installations. Figure 2.6 compares the coverage (in terms of distribution of the total installed capacity in each departement) reported in the BDPV database to the installed capacity per departement reported in the RNI, taken as the reference.

We can see that for some departements located in the West of France, BDPV overestimates the share of these departements in the national distribution of rooftop PV installations, compared to the RNI, which is considered as the reference. In other words, the distribution of the aggregated installed capacity at the size of the departement in BDPV is not representative of the true distribution considered to be given by the RNI. The relative spread indicates the direction of the biases. A value of 0 indicates that the share of installations in the departement corresponds to the reference. A negative value (highlighted in blue) indicates that the reported installed capacity is lower in BDPV than in the RNI. A positive value (highlighted in red) indicates that the reported installed capacity is larger in BDPV than in the RNI.

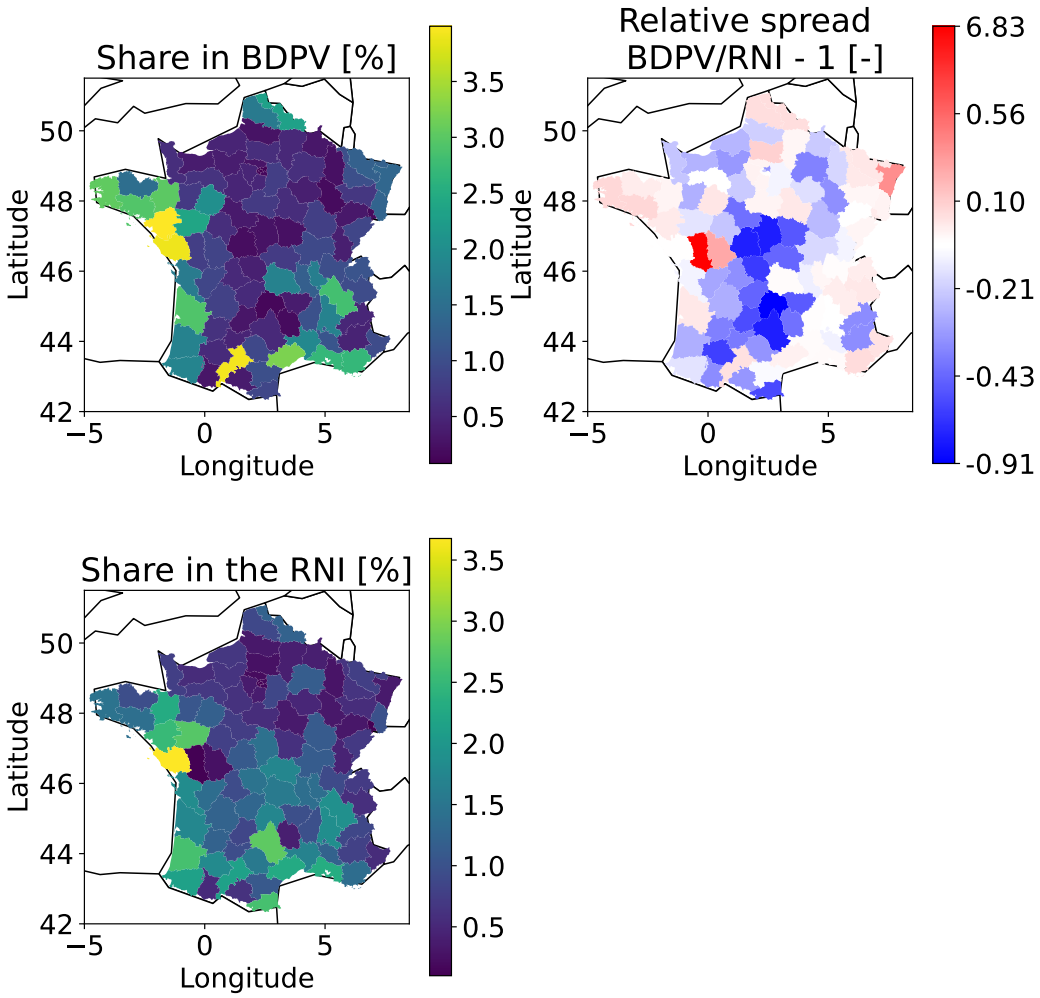


Figure 2.6 – Comparison of the distribution of the installed capacity at the scale of the departements reported in BDPV (upper left) the RNI (lower left), and the relative spread of the two (upper right). On the relative spread plot, the red means that the reported installed capacity at the size of the departement is higher in BDPV than in the RNI. Blue means that the reported installed capacity is lower in BDPV than in the RNI.

Updatability Finally, we want to be able to update the registry frequently to be able to cope with the quick growth of rooftop PV installations. The RNI and RTE's registry are updated every three months, but BDPV is updated only when a new user registers his installation, with no correlation with the rate of adoption of rooftop PV. A registry built from orthoimagery will depend on the images' revisit rate. For IGN, we said that it is three years for a departement. However, as these updates are carried out on a rolling basis, we can update the registry every month or three months for a set of departements and scale the installed capacity to the actual capacity in departements that have not been updated for a long time.

1.3.2 Summary: the need for a disaggregated, representative and complete registry

Table 2.1 compares the available data sources and evaluates them with respect to the criteria defined in section 1.3.1. We can see that all existing sources miss at least one criterion. Our registry aims to satisfy the four requirements (disaggregation, technical characteristics, representativeness, and updatability) simultaneously.

Table 2.1 – Data requirements applicable for the technical registry and accessibility from existing sources for PV installations below 36 kW_p.

	RTE's registry	RNI	BDPV	Technical registry (our target)
Disaggregation	✓	✗	✓	✓
Technical characteristics	✗	✗	✓	✓
Representativeness	✓	✓	✗	✓
Updatability	✓	✓	✗	✓

1.4 Training data: the BDAPPV training dataset

We need training data because we will use deep learning models in France, which has not been mapped prior to our work. This section briefly introduces our training dataset. We gathered training data through crowdsourcing campaigns using the PV registry of BDPV. We refer the reader to the appendix B, section 2 for more details regarding the crowdsourcing campaigns and to appendix B, section 1 or to Kasmi et al. (2023d) for more technical details and a comprehensive description of the data records. The BDAPPV training dataset is provided under open access at Kasmi et al. (2022b).

Overview Figure 2.7 summarizes the data collection process we used to construct BDAPPV. The source data comes from BDPV's registry, which contains localizations and PV installation characteristics. We then carried out a crowdsourcing campaign to annotate PV images. The crowdsourcing campaigns took place on a dedicated platform[3].

Overview of the crowdsourcing campaigns We extracted thumbnails based on the geolocation of the installations recorded in the BDPV dataset. However, this geolocation can be inaccurate, so before asking users to draw polygons of PV installations, we asked them to click on images if they depicted a PV panel. This first classification step corresponds to the first phase of the annotation campaign.

3. The platform is accessible here: https://www.bdpv.fr/_BDapPV/

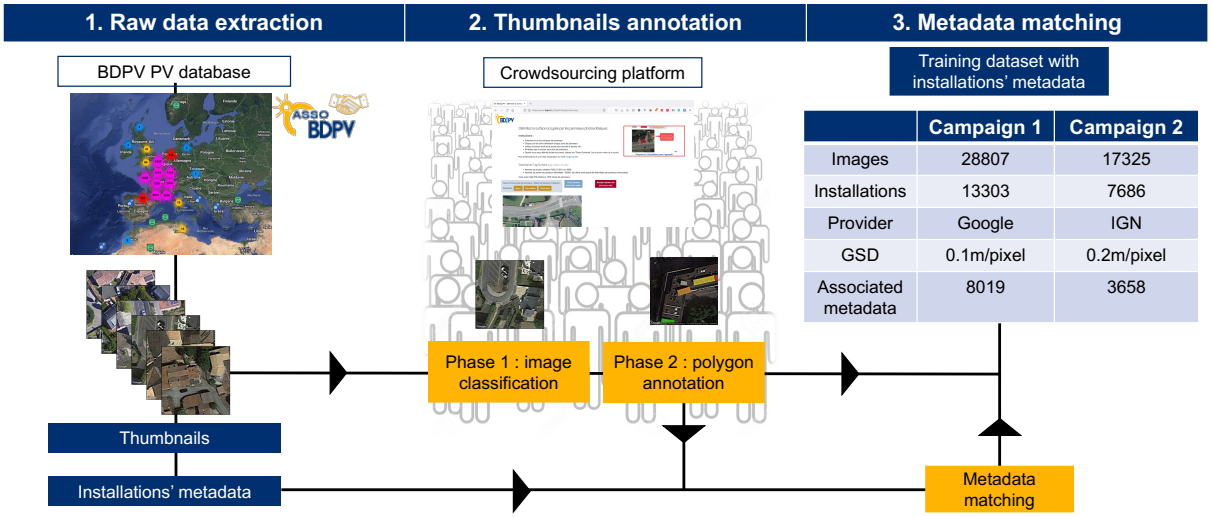


Figure 2.7 – Flowchart of the BDAPPV dataset construction workflow. Source: Kasmi et al. (2023d).

Once users classified images, we asked them to draw the PV polygons on the positively classified images. It corresponds to the second phase of the crowdsourcing campaign. We conducted crowdsourcing campaigns on two image sources: Google images from the Google Earth Engine (Gorelick et al., 2017) and IGN images from the BD ORTHO (IGN, 2024a).

During the first phase, the user clicks on an image if it depicts a PV panel. We recorded the localization of the user's click and instructed them to click *on* the PV panel if there was one. We collected an average of 10 actions (click with localization or no click) per image. After empirical investigation, we considered the image a true positive if at least three users clicked on it.

During the second phase, annotators delineated the PV panels on the images validated during phase 1. Users can draw as many polygons as they want. On average, we collected five annotations per image. We collect the coordinates of the polygons drawn by the annotators. We then averaged the different contributors' annotations to obtain a ground truth segmentation mask.

Matching with PV characteristics and data records After postprocessing the data from the crowdsourcing campaigns, we end up with images and associated ground truth segmentation masks. We matched the annotations for each image provider with the characteristics recorded in BDPV. This enables us to enrich the segmentation mask with PV characteristics, which is very useful as we ultimately wish to derive the PV characteristics from the segmentation mask generated by the model. Table 2.2 summarizes the data records of BDAPPV.

Table 2.2 – Overview of the data records of the training dataset BDAPPV.

Provider	Total number of samples	Positive samples (Share [%])	Number of installations linked with installation characteristics (Share [%])
Google	28 807	13 303 (46.18)	8019 (27.84)
IGN	17 325	7686 (44.36)	3658 (21.11)

2 Monitoring the accuracy of the registry without ground truth labels

In the previous section, we reviewed the available data sources and the requirements that our registry should meet to estimate rooftop PV power production. This section introduces our approach to ensure that the registry accurately maps PV installations. First, we define our criteria for evaluating the accuracy of the registry. Then, we introduce our approach: the downstream task accuracy (DTA). The DTA accounts for the fact that we do not have ground truth labels over the mapping area (i.e., the area over which the model is deployed). It relies on existing data sources to define a set of metrics that assess whether the registry is accurate and highlight where it is lacking accuracy. The DTA enables us to address the first condition for reliability: the ability to monitor (i.e., to keep under systematic review) the data produced by the model.

2.1 Defining the evaluation criteria

2.1.1 Overall criteria: representativeness and completeness

The registry's main requirement is to reflect the rooftop PV characteristics in France as closely as possible. However, we need to accommodate for the lag induced by the revisit rate of the images, which can be as high as three years for a single departement. The lag is not necessarily problematic if the registry is representative of the installations. In this case, by upscaling the estimated production from the registry, we can still estimate the PV power production fairly because the missing power plants are not systematically biased compared to the systems recorded in the registry. For instance, we wish to avoid situations where a set of installations popped up in a precise location of the departement and with characteristics that differ (e.g., larger installations) from those "historically" recorded in the registry.

To avoid such situations, we assume deploying PV installations in a departement is locally stationary (over three years). Drifts in the PV system's characteristics can be identified with updates made every three years. Under this assumption, we

have to ensure that the PV installed capacity registered in the model reflects the spatial distribution (and the distribution in terms of the system's sizes). We also have to ensure that the PV characteristics represent the PV characteristics in the departement or the city of interest.

2.1.2 Tilt and azimuth angles

We need to ensure that the estimation of the PV characteristics is representative of the actual distribution of the tilt and azimuth angles in a given locality. The worst case would be a systematic bias in estimating the azimuth angle, which could result in overestimating the PV power production at a moment in the day (and underestimating it at another moment). Another concern would be systematic bias in estimating the tilt angle. In this case, we could under or overestimate the PV power production. Since the tilt angle's value depends on the installation's latitude (Killinger et al., 2018), we want our registry to reflect this.

2.1.3 Installed capacity

We want to ensure that the distribution of the PV installations matches the geographical distribution of the actual installed capacity and avoid situations such as those reported in Figure 2.6. This enables the avoidance of local over- or underestimations of PV power production. We also want to represent the power class well within a departement.

2.2 Leveraging existing data to meet these criteria: the downstream task accuracy

An unsupervised evaluation metric The main issue for evaluating the accuracy of the registry during deployment is that we lack ground truth labels. Besides, we do not want to rely on a statistical extrapolation of the accuracy computed over a given area, as done by Mayer et al. (2022) because we want to monitor the accuracy of our registry over the complete mapping area. From Table 2.1 of section 1.3.2, we can see that existing data sources are not sufficient for PV power estimation but can assess at least at the aggregated level whether the registry is accurate or not. We can use the RNI and RTE's registry to evaluate the accuracy regarding the installed capacity and BDPV to assess whether the estimation of the technical characteristics is correct.

Figure 2.8 illustrates the procedure to derive metrics based on the DTA principle. The idea is to derive metrics that can be verified using an external source. The DTA is an unsupervised evaluation metric (Zhang et al., 2008; Chen et al., 2021b), i.e., it evaluates the accuracy of the registry without ground truth labels and without requiring human intervention in the evaluation. Unsupervised evaluation metrics are

initially intended to compare methods (Gao et al., 2017; Liu et al., 2016; Chabrier et al., 2006) according to a given set of criteria. In our case, we take advantage of the fact that we do not need human intervention and that the reference data is available throughout France to use the unsupervised evaluation metric as a monitoring tool. This enables us to inspect whether the registry is accurate over the mapping area.

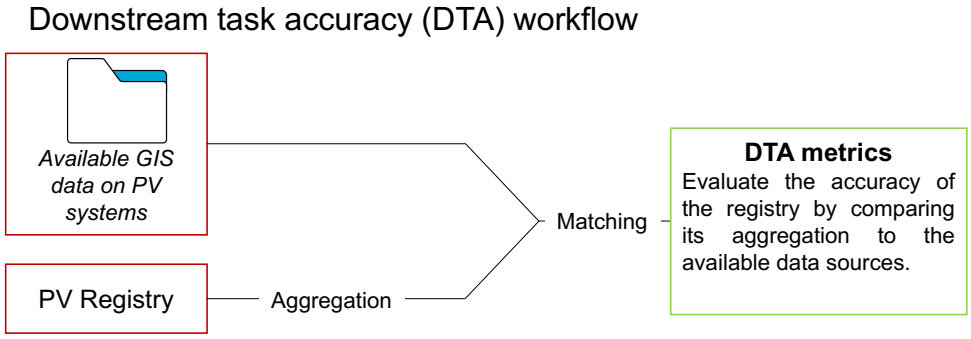


Figure 2.8 – Overview of the principle of the DTA with the RNI as an example. GIS, "Geographical Information System, " corresponds to georeferenced PV data such as the RNI, RTE's registry, or BDPV.

2.2.1 Metrics from the RNI for the installed capacity

Using the RNI dataset and the aggregations from our pipeline, we define three metrics to ensure that the model estimates the installed capacity correctly. Mayer et al. (2022) inspired these metrics, aiming to capture different properties of the estimation of the city-wise installed capacity.

We denote $\hat{C}_i$ our estimated installed capacity based on aggregating our detections in the city i[4], C_i the reference installed capacity from the RNI for city i. Similarly, $\hat{k}_i$ denotes our estimation of the number of installations in city i, and k_i is the reference number of installations recorded in the RNI.

— The **average percentage error (APE)**: $\frac{|C_i - \hat{C}_i|}{C_i}$ and the **mean APE (MAPE)** $\frac{1}{n}\sum_{i=1}^{n}\frac{|C_i - \hat{C}_i|}{C_i}$ computed over the whole departement. The APE and MAPE are computed for the installed capacity C_i in city i. The APE and MAPE indicate whether the estimated installed capacity at the aggregated level is correct. This is why we weigh each installation individually; we don't want to give more weight to large installations. The (M)APE reads as follows: if it is equal to 0,

4. Formally,

$$\hat{C}_i = \sum_{j=1}^{\hat{k}_i} \hat{C}_j^{(i)}$$

where $\hat{k}_i$ is the number of detected installations in city i and $\hat{C}_j^{(i)}$ the installed capacity of the j^{th} installation in city i, as recorded in our rooftop PV technical registry.

then the estimation and the reference are equal. If it is equal to 100, then the estimation is twice as large as the reference value.

— The **detection ratio** $\Delta := \frac{\hat{k}_i}{k_i}$, based on Mayer et al. (2022) computed at the city level and averaged over the departement. We compute this ratio for the number of installations k_i in city i. The ratio indicates if the model correctly detects the installations (irrespective of their installed capacity).

— The **average installation percentage error (AIPE)** $-\frac{C_i/k_i - \hat{C}_i/\hat{k}_i}{C_i/k_i}$ which is the APE computed for the *average* installation. By construction, a negative AIPE indicates that we underestimate the installations' size, and a positive AIPE indicates that we overestimate them. The AIPE assesses whether the average installation size derived from our registry represents the average installation size recorded in the RNI.

The MAPE ensures we do not overestimate or underestimate the overall installed capacity. The detection ratio ensures that we detect the correct number of installations, and the AIPE, which is a function of the installed capacity and the number of installations, ensures that, on average, we correctly estimate the size of the installations.

2.2.2 Evaluation of the accuracy of the tilt and azimuth angles

To assess whether the tilt and azimuth estimates are consistent with the actual underlying distribution, we use the data from BDPV as a reference and compare our estimations with the distribution from BDPV.

Distribution matching As done in previous works (e.g., Mayer et al., 2022), we can compute the distribution of the tilt and azimuth angles obtained from the reference data and our technical registry and then compare them. For the azimuth angle, we can compute the distribution over all departements. We compute the distributions in each departement for the tilt angle, as the mean tilt angle varies with the latitude (Killinger et al., 2018).

3 How does the accuracy of state-of-the-art models vary over the mapping area?

In this final section, we build a model for mapping PV installations based on the literature and train it on BDAPPV to deploy it over France. We evaluate the accuracy according to the DTA and show that this baseline mapping algorithm exhibits prediction inconsistencies over the mapping area. The mapping area corresponds to the area where we deploy the trained model. We conduct empirical experiments

to formulate a hypothesis regarding the origin of this prediction inconsistency. We will introduce the tools needed to verify this hypothesis in chapter 3.

3.1 Evaluation framework

3.1.1 An algorithm for mapping rooftop PV installations based on the literature

Figure 2.9 presents the model we used in this experiment. This pipeline follows Mayer et al. (2022) and consists of two main steps. Firstly, we extract geolocalized polygons of rooftop PV installations using a classification and a segmentation model. Then, we combine the polygons with additional data from the BDPV database to infer the characteristics of the rooftop PV installations. A key difference with Mayer et al. (2022) is that we do not use surface data to infer the tilt and azimuth angles of the installation. Instead, we use a lookup table (LUT), which indicates the average tilt angle in each grid point over France and an algorithm based on the bounding-box of the PV polygons.

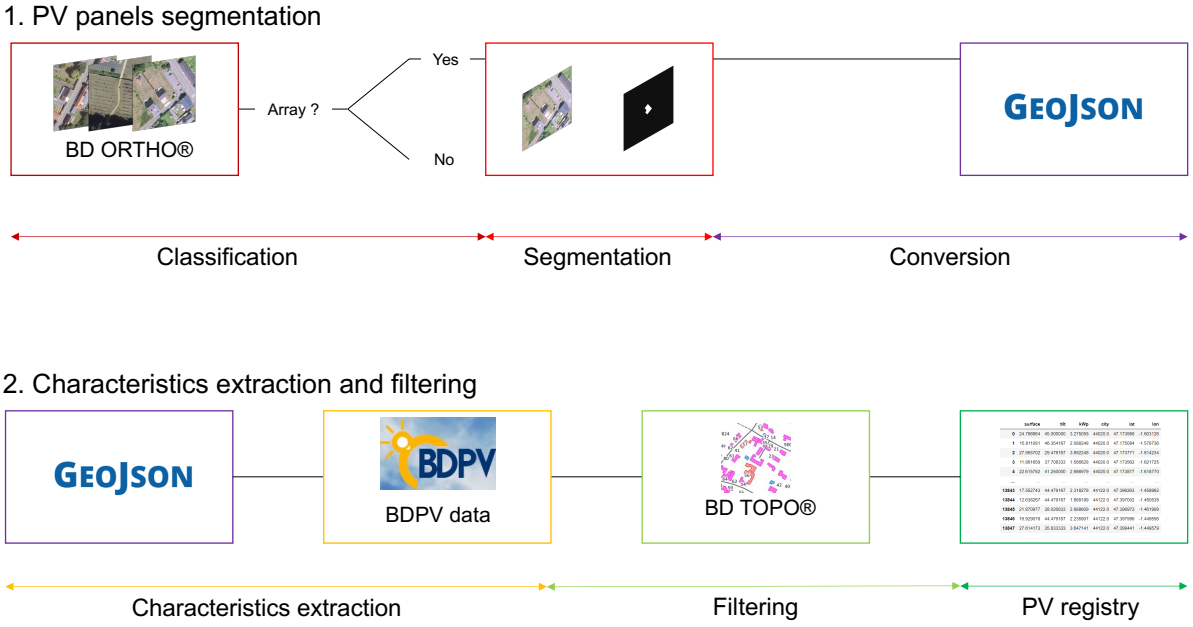


Figure 2.9 – Flowchart of our algorithm based on the literature for mapping rooftop PV. Source: Kasmi et al. (2022a).

Step 1: PV panels segmentation The first step of the pipeline consists in segmenting the PV panels from the orthoimages. Orthoimages consist of large image tiles with a size of 25,000×25,000 pixels. As these images are too large to be fed into the model, we cut the tiles into small thumbnails that have a size of 299×299 to fit the requirements of the Inception-v3 (Szegedy et al., 2016) classification model

that we use in this algorithm. The classification model identifies the thumbnails that contain a PV panel. These positive images are the input of a segmentation model (DeepLab-v3, Chen et al., 2018), which identifies which pixels correspond to the PV panel. The segmentation model returns a binary mask, where pixels equal to 1 depict a PV panel and pixels equal to 0 do not depict PV panels. We convert the binary masks into polygons of PV installations. This conversion step accommodates PV systems that may span across several thumbnails. In this case, we merge adjacent binary masks from neighboring thumbnails into a single polygon. As the thumbnails are geolocalized, our polygons are also geolocalized.

Step 2: Characteristics extraction and filtering We want to derive the tilt and azimuth angles and installed capacity from the polygons of the PV systems. We leveraged the dataset BDPV to estimate an average PV module efficiency, which enables us to relate an installation's surface with its installed capacity linearly, following the approach of So et al. (2017).

As surface models completely covering France are not yet available, we also used the BDPV database to construct a lookup table (LUT). To construct the LUT, we split France into 50×50 gridpoints and computed the average tilt angle in each gridpoint. We compute this average using the installations belonging to each gridpoint. To reflect the differences in tilt angles depending on the size of the installation, we computed these averages for four different systems' sizes classes, taken as the quantiles of the distribution of the installed capacities in the BDPV dataset. Finally, we converted these classes into projected surfaces to be able to input a tilt angle from the projected surface of the installation. Figure 2.10 presents the lookup table used in this work.

Using the lookup table and the efficiency parameter estimated from BDPV, we extract the characteristics of the PV systems as follows:

— Azimuth angle: computation of the angle of the bounding box of the PV polygon relative to the North.

— Tilt angle: imputation from the LUT, based on the localization of the PV system.

— Surface: computation based on the projected surface (taken as the input polygon's area) and the tilt angle's cosine.

— Installed capacity: product between the surface and the PV module efficiency derived from BDPV.

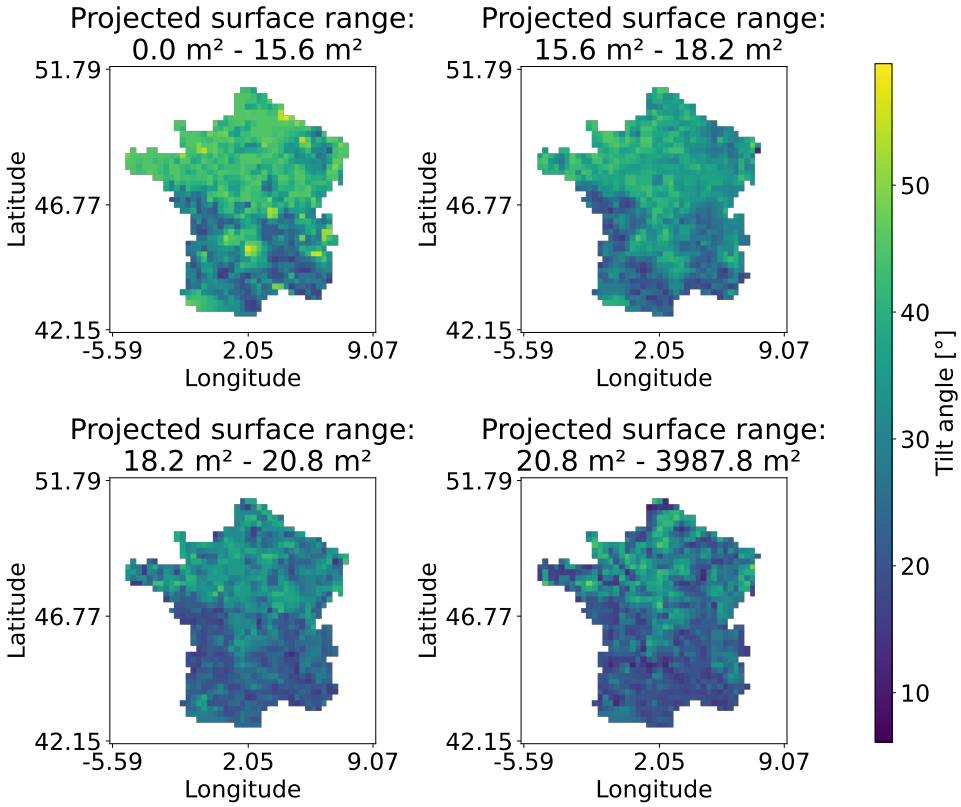


Figure 2.10 – Visualization of the lookup table used in this study to estimate the tilt angle of the installations. Taken from Kasmi et al. (2022a).

We expanded this methodology to create a Python package for extracting characteristics from PV polygons. We refer the reader to chapter 4, section 1.2, appendix C, section 3.1 or Trémenbert et al. (2023) for more details on our method for extracting characteristics of PV systems from PV polygons.

Filtering The final step of our algorithm consists in a filtering of the characteristics. As we focus on installation with an installed capacity lower than 36 kW_p, we remove all installations larger than 36 kW_p. We filter out the installations whose estimated installed capacity is lower than the installed capacity of a single PV module (i.e., 300 W_p). Finally, we use the BD TOPO to remove all installations that are not located on a rooftop.

3.1.2 Training details

Training dataset As we use IGN images to map PV installations in France, we trained our baseline models on the IGN images of BDAPPV (Kasmi et al., 2023d). Table 2.3 summarizes the characteristics of the data that we use for training our models. Our training dataset is nearly balanced. Following standard practice, we split this training dataset into training, validation and test images. All these images

have a resolution of 20 cm/pixels. Finally, we have the true installed capacity of the installations depicted in the test set.

Table 2.3 – Training dataset characteristics.

Dataset	Total number of samples	Positive samples (share in %)
Train	12,127	5,445 (44.90)
Validation	1,732	755 (43.59)
Test	3,466	1,485 (42.84)
Total	17,325	7,685 (44.36)

Images preprocessing Images are flipped following the vertical axis and rotated 90 degrees clockwise and counter-clockwise during training. We avoid rotations that lead to having arrays pointing north (i.e., upwards). Input images have a size of 400×400 pixels, so we randomly crop the images to generate input images whose size matches the requirements of the classification model (i.e., 299×299 pixels). Finally, images are also normalized with the ImageNet mean and standard deviation (i.e., a mean of (0.485, 0.456, 0.406) and a standard deviation of (0.229, 0.224, 0.225)). Validation and test images are only normalized[5].

Model training We use the model architecture and the model weights provided by Mayer et al. (2022). Their classification model is an Inception-v3 (Szegedy et al., 2016). We retrain all the layers of the classification model, using their weights as an initialization. We train the model for 25 epochs, enough for our model to converge on our training images. We picked the model that achieved the lowest accuracy on the validation set after the end of the training. We evaluate the final performance after threshold fine-tuning on the testing dataset. We use the binary cross entropy loss (BCE, without weighting) and Adam optimizer (Kingma and Ba, 2015) with a learning rate of 0.0001. We used a batch size of 128 images. The segmentation model of Mayer et al. (2022) is a Deeplab-v3 model (Chen et al., 2018). Like classification, we retrained all layers and picked the model parameterization that achieved the lowest validation accuracy after the end of the training. We also use BCE loss with a learning rate of 0.0001 and Adam optimizer. The batch size is 64.

3.1.3 Training results and deployment on the mapping area

Classification and segmentation accuracies Our fine-tuned model achieves competitive results compared to state-of-the-art models (see table 2.4). For the

5. This process corresponds to the usual data preprocessing approach when using pre-trained deep learning classification and segmentation models.

Table 2.4 – Classification and segmentation accuracy. The lower the GSD, the more detailed the image. Best results are **bolded**.

	Classification	**Segmentation**	
Work	F1 score (↑)	IoU (↑)	GSD (cm/pixel)
Mayer et al. (2022)	0.87	0.74	10
Malof et al. (2019)	-	0.67	30
Zech and Ranalli (2020)	0.82	-	10
Parhar et al. (2021)	**0.97**	0.86	10
Ours	0.84	**0.86**	20

classification branch, we reach an F1 score of 0.84. We reach an Intersection-over-Union (IoU) of 0.86 for the segmentation branch[6]. We aim not to establish a new state-of-the-art (SOTA) for classification or segmentation but to see how current performance translates to accuracy over the mapping area.

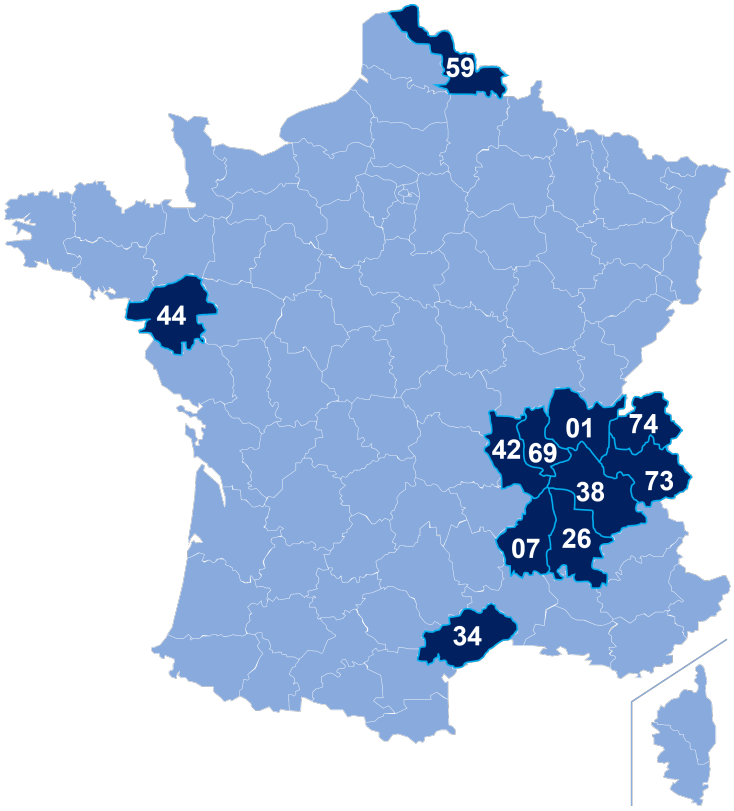


Figure 2.11 – Mapping area over which we deployed our model. The numbers correspond to the number of the departements used to identify them in Table 2.5.

6. The F1 score is the harmonic mean between the classifier's precision and recall. It varies between 0 and 1. The IoU evaluates how well the predicted and the true segmentation masks overlap. The IoU is equal to 0 if the masks are disjoint and equal to 1 if the masks perfectly overlap. We provide more detailed definitions of the F1 score and the IoU in chapter 4, section 2.1.1.

Mapping area We call the area over which we deployed our trained model the mapping area. Over the mapping area, we have no ground truth labels as in the test set to evaluate the model's accuracy. Figure 2.11 illustrates our mapping area, which corresponds to 11 departements in the North, South, West, and East of France, covering approximately 50,000 km², as depicted on Figure 2.11. As we do not have ground truth labels over this mapping area, we rely on the DTA metrics introduced in section 2.2 to monitor the accuracy of our registry.

3.2 Results: detection inconsistencies during deployment

3.2.1 Tilt and azimuth angles

Distribution matching Figure 2.12 compares the distribution of the azimuth and tilt angles coming from BDPV and our mapping algorithm. To reflect the geographical variability of the tilt angle, we compute its distribution within each departement. Figure 2.12 plots an example of the distribution of the tilt angle in the departement Loire-Atlantique (West of France). In appendix C, section 1.1, we supply more examples of the distribution of tilt angles coming from BDPV and our algorithm.

We can see that the estimation follows the same overall pattern as the source distribution. For the azimuth angle, we slightly underestimated the number of installations with a tilt oriented southwards (i.e., between 170 and 190 degrees) and slightly overestimated the eastward and westward azimuth angles. However, there is no evidence of a systematic bias compared to BDPV. Our average estimation of the tilt angle around a given localization is close to the reference for the tilt angle. The distribution of tilt angles is more concentrated towards the center than the actual distribution. This is a consequence of our imputation method, which tends to concentrate the tilt angles towards the mean. Finally, we can see that the tilt and azimuth angle estimations do not produce extreme values.

3.2.2 Installed capacity

Overall results Table 2.5 shows the accuracy results measured with the DTA for the installed capacity. As we have the installed capacities of the installations of the test set, we can compute the true aggregated installed capacity, as if the test set was one city, and compute the MAPE between the estimated installed capacity by the mapping algorithm and the true value. The MAPE is equal to 17.61%. When shifting to the mapping area, the MAPE increases to 47.45%, meaning that the accuracy drop between the test set and the mapping area is about 30 percentage points.

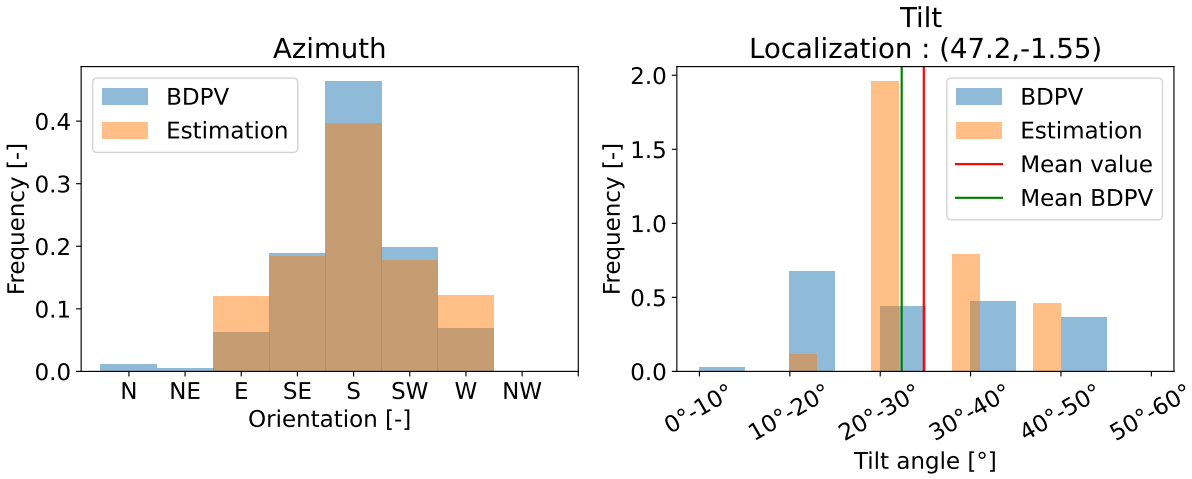


Figure 2.12 – Comparison of the distribution of the azimuth (left) and tilt (right) angles obtained from our mapping algorithm and BDPV. The azimuth angle distribution is obtained for the 11 departements and the tilt angle distribution is computed for the departement Loire-Atlantique (44).

Table 2.5 – Downstream task accuracy across the mapping area. Values in parentheses correspond to the results without filtering by buildings. The line "Test" considers the test images of the training dataset as one city. k_i and C_i denote the count of installations and the installed capacity, respectively. The hat indicates the estimation by our algorithm. Source: Kasmi et al. (2022a).

Departement	MAPE [%]	med. APE [%]	mean ratio [-]	mean AIPE [%]	k_i [-]	$\hat{k}_i$ [-]	C_i [kW$_p$]	$\hat{C}_i$ [kW$_p$]
Test	17.61	-	0.92	-0.10	1485	1362	6473.8	5334
44	39.09 *(33.20)*	38.05 *(26.69)*	0.67 *(0.91)*	22.83 *(18.54)*	12683	6838 *(9325)*	51955.2	34197.7 *(45206.6)*
69	31.99 *(39.45)*	28.91 *(23.29)*	0.83 *(1.26)*	12.18 *(10.33)*	8944	6508 *(9808)*	36500.6	31361.6 *(45434)*
59	130.06 *(224.22)*	88.13 *(168.29)*	2.23 *(3.22)*	61.16 *(41.56)*	6453	9524 *(15393)*	22083.8	52790.2 *(73697.4)*
34	26.80 *(45.57)*	17.78 *(30.05)*	1.01 *(1.33)*	6.82 *(4.29)*	9199	8408 *(11445)*	35398.4	39897.2 *(52841.8)*
01	35.90 *(38.77)*	35.35 *(27.07)*	0.77 *(1.13)*	6.18 *(7.34)*	4940	3654 *(5259)*	18433.2	14659.4 *(21372)*
38	33.41 *(31.29)*	31.15 *(21.80)*	0.81 *(1.07)*	9.18 *(6.20)*	10672	7835 *(10680)*	39691.4	32391.11 *(42617.4)*
42	29.12 *(46.30)*	23.81 *(28.66)*	1.00 *(1.41)*	15.42 *(11.06)*	6892	5831 *(8384)*	28594.1	27916.5 *(38222.4)*
26	30.46 *(55.35)*	23.51 *(25.80)*	0.92 *(1.43)*	3.74 *(6.15)*	5808	4933 *(7121)*	28262	25834 *(37645.8)*
74	44.08 *(41.45)*	41.18 *(28.53)*	0.67 *(0.93)*	-4.61 *(-6.10)*	7004	5287 *(6600)*	32202.1	21760.4 *(25931)*
Overall	47.45 *(66.20)*	32.81 *(30.66)*	1.03 *(1.46)*	16.33 *(12.03)*	72595	58818 *(84015)*	293120.7	280807.9 *(382967.8)*

We can see that the accuracy varies a lot across the departements: the MAPE reaches 130% in the Nord (59, North of France) while it is only 26% in Hérault (34, South of France). It means that our model estimates more than twice the actual installed capacity in the North.

The ratio and the AIPE help us understand the drivers behind the variability in estimating the city-wise installed capacity. In the North, we tend to detect too many installations (twice as many) and to overestimate their size individually. On the other hand, in the departement Loire-Atlantique (West of France), where the accuracy is better than on average, we tend to detect too few installations, compensated by an overestimation of the average size of the installation.

Visualization at the scale of the cities Figure 2.13 plots the spatial distribution of the installed capacity, aggregated at the size of the cities, referenced in the RNI, and estimated by our model.

The upper left plot presents the distribution of the installed estimated from our registry (the reference is the total installed capacity in the departement). The bottom left plot presents the recorded distribution from the RNI. The upper right plot presents the relative spread between the two. The interpretation is the same as in Figure 2.6: redder values indicate that our model overestimates the share of the installed capacity in these cities. Bluer (and negative) values indicate that the model underestimates the importance of the share of these cities.

The mean of the relative spread is equal to 0.05, and the median is -0.11. This means that overall, the estimation of the distribution of the aggregated installed capacities (at the scale of the cities) is close to the true distribution of the installed capacities. However, considering the qualitative pattern from Figure 2.13, we can see that the model tends to heavily overestimates the installed capacities in some cities (the estimation can be nearly seven times higher than in reality). In the next section, we will analyze how the detections of the model lead to such imbalances.

3.3 Understanding these shifts

In section 3.2, we evaluated the data produced by our registry according to the DTA metrics. These results showed that estimating the tilt and azimuth angles is satisfying, while estimating the installed capacity lacks precision. In particular, it is driven upwards by false positives. The analysis of the DTA for the installed capacity shows that the model detects PV panels inconsistently over the mapping area. In this section, we investigate why such inconsistencies occur. Since our images come from the same provider (the IGN) with the same ground sampling distance, we can rule out this factor as an explanation for the detection inconsistencies. Existing works in the literature often posit that geographical variability is an essential contributor to the loss of accuracy (Wang et al., 2017; Malof et al., 2019). In sec-

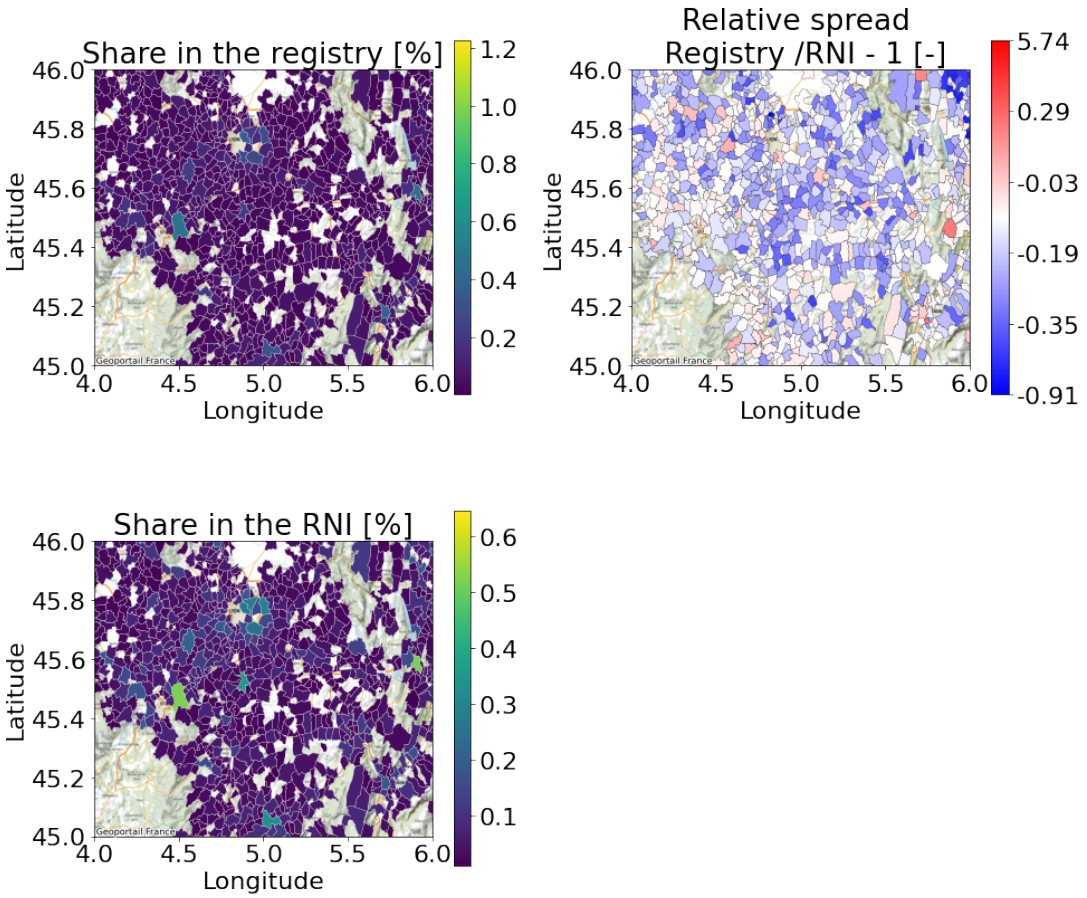


Figure 2.13 – Comparison of the distribution of the installed capacity at the scale of the cities reported from our registry (upper left), the RNI (lower left), and the relative spread of the two (upper right) in the departement Isère (38). On the relative spread plot, the red means that the reported installed capacity at the size of the city is higher in our registry than in the RNI. Blue means that the reported installed capacity is lower in our registry than in the RNI.

tion 3.3.1, we quantitatively review this hypothesis; in section 3.3.2, we conduct a qualitative analysis of the model's errors. Finally, in section 3.3.3, we present our hypothesis explaining the detection inconsistencies over the mapping area.

3.3.1 Geographical factors

Objective and approach Our goal is to look for a geographical pattern that could explain the differences in accuracies observed in table 2.5. We consider the accuracy measurements at the city level obtained according to the DTA, defined in section 2.2.1.

Qualitative analysis We first conduct the qualitative analysis by comparing the box plots of the distributions of interests. Figure 2.14 reports the boxplots of the city-wise MAPE (error in the estimation of the city-wise installed capacity) across four departement and for the three accuracy metrics (up: APE, middle: ratio, and down: AIPE).

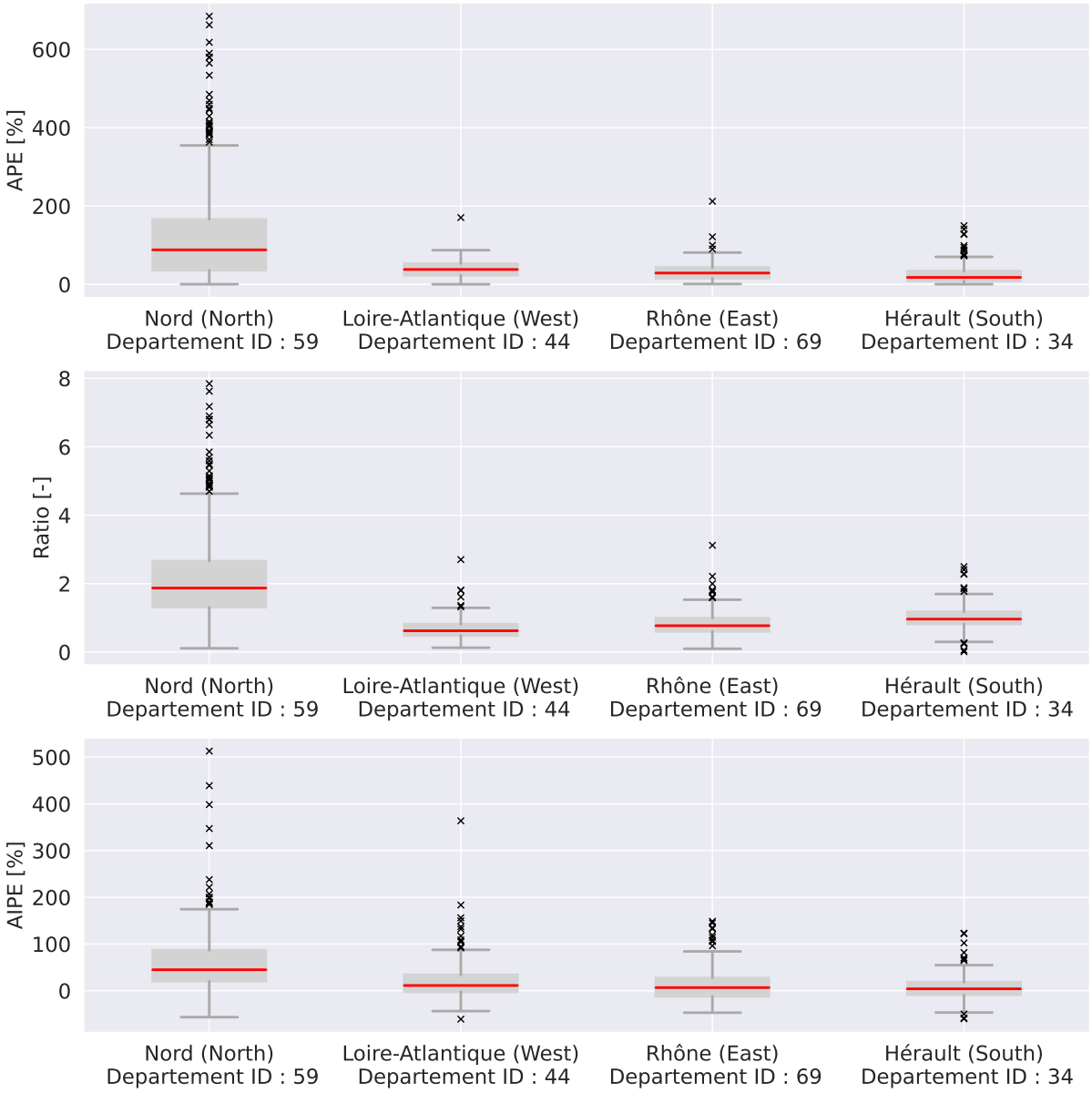


Figure 2.14 – Boxplots of the distributions of the APE metrics across four departement.

We can see that the variance of the distribution of the MAPE is much higher in the departement Nord (North of France), resulting in the higher mean and median error reported in Table 2.5. We investigate whether this variation is significantly correlated with the city coordinates. In this case, the city's geographical position would significantly impact the model's accuracy.

Is the observed difference significant? To study whether the effect of the geographical coordinates is significantly correlated with a variation in the model's accuracy, we set up a linear regression model. To test nonlinear relationships, we test alternative models, including polynomial and logarithmic transformations of the independent variables. The dependent variable of this model is the DTA metric of interest (i.e., the APE, the ratio, and the AIPE), and the independent variables of interest are the latitude and longitude. Our regression model is given by Equation 2.1

$$y_i = \alpha + \beta_1 \text{lat}_i + \beta_2 \text{lon}_i + \sum_{j=1}^{J} \gamma_j x_{i,j} + \varepsilon_i, \tag{2.1}$$

where y_i is the dependent variable (the accuracy in city i), lat_i and lon_i are our variables of interest, the latitude and longitude of the city and the x_j's are a set of control variables. These controls include the number of installations, the total PV installed capacity, the area of the city, and the number of buildings in the departement (a proxy for the urbanization level and complexity of the background). We do not include higher order and interaction terms to avoid multicollinearity issues that could artificially reduce the significance of the coefficients. ε is the remaining error term. We cluster the standard errors at the scale of the city. We evaluate the statistical significance of the coefficients β_1 and β_2 by performing a Student t-test on these coefficients.

Results Table 2.6 presents the results of our linear regression. We do not find evidence of a significant correlation between a city's APE and its latitude. In appendix C, section 1.2, we provide similar results for models that consider nonlinear and polynomial transformations of the coordinates.

Table 2.6 – Results of estimating the linear model defined in Equation 2.1 for the dependent variables APE, AIPE, and ratio.

	APE	**ratio**	**AIPE**
β_1 (Latitude)	6.5535	0.0329	11.3557*
	(5.902)	*(0.048)*	*(6.088)*
β_2 (Longitude)	-4.3700	-0.0034	5.0295
	(3.770)	*(0.026)*	*(3.304)*
α	-1.772e-10	-2.036e-13	-3.279e-10**
	(1.58e-10)	*(1.29e-12)*	*(1.67e-10)*
N	1839	1839	1839
Adjusted-R^2	0.382	0.572	0.412

Clustered standard errors in parenthesis.
$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Discussion In this section, we investigated whether the error of the model measured by the DTA was sensitive to the localization of the cities, defined as their coordinates. Comparing boxplots of the distributions of the APE at the city level across four departements, we found that in the departement Nord, the variance was much higher than in other departements, explaining the higher mean and median error reported in Table 2.5. We then investigated quantitatively whether the difference in mean and median accuracy was significantly correlated with the localization of the city and found no such evidence. A reason to explain the high variance could be that the departement Nord is, on average, more urbanized, thus with more complex scenes, than the other departements. Note that the higher prevalence of complex scenes drives the errors: if we considered only urbanized areas elsewhere, accuracy would also drop. We need to focus on how a model makes detections to verify this idea.

We warn the reader that these results do not state that the geographical localization of the city does not affect the model's accuracy. We wanted to see whether a straightforward relationship emerged and if this relationship was statistically significant. In chapter 3, section 2.2.1, we show in a similar setting that geography is not the primary driver that affects the model's performance. In both cases, we do not state that these results generally hold. A plausible explanation is that our training dataset covers most of France; therefore, the model has learned the geographical variability in France.

3.3.2 Feature analysis

Overview In this section, we analyze the model's prediction to formulate a hypothesis that could explain why false detections and, more precisely, false positives arise. A popular way to analyze the model's behavior is using feature attribution methods such as class activation maps (CAMs, Zhang et al., 2018). A CAM is a heatmap that highlights discriminative regions in an input image used by a CNN for predicting a specific class. The CAM visually indicates which parts of the image contribute the most to the network's decision.

Using attribution methods to explain models' decisions in remote sensing applications is a well-established approach. For instance, Dardouillet et al. (2023) leveraged Shapley values (Shapley, 1952), which quantify the marginal contribution of each feature to the prediction performance, to analyze the decision process of segmentation models designed to identify offshore oil slick on SAR images.

Our approach considers the popular class activation mapping method of Selvaraju et al. (2020), which has been applied in many settings for analyzing a models' decisions, such as Lapuschkin et al. (2019) or Zhang et al. (2021b). Lapuschkin et al. (2019) used CAMs to show that models can rely on spurious features to make predictions. We gather samples of the test dataset and analyze the behavior of

the model through the lenses of the class activation maps. We consider all types of detections (false positives, false negatives, true positives, and true negatives) to see whether a prediction pattern arises. These prediction patterns correspond to similarities observed through the visual inspection of the class activation maps. In addition to the qualitative investigation, we also analyze the model's predicted probabilities.

Results Figure 2.15 presents the explanations obtained using the GradCAM (Selvaraju et al., 2020). We can see two different prediction patterns depending on whether the model predicts a positive (true or false) or a negative (true or false). In the case of a true positive prediction, the model will focus on a specific, narrow region of the image, which indeed corresponds to a PV panel. However, for false positives, the model also focuses on a narrow image region. Inspecting the samples of Figure 2.15 reveals that this region of the image depicts items that *resemble* PV panels. On the image on the first row (second column) of Figure 2.15, we can see that the model confuses a shade house that shares the same color and overall shape of a PV panel with an actual panel. In the image on the second row, the verandas with groves fool the model.

On the other hand, when the model does not see a PV panel, it does not focus on a specific image region. This remains true for the false negatives, where we can see that the model does not see the panels on any of the images.

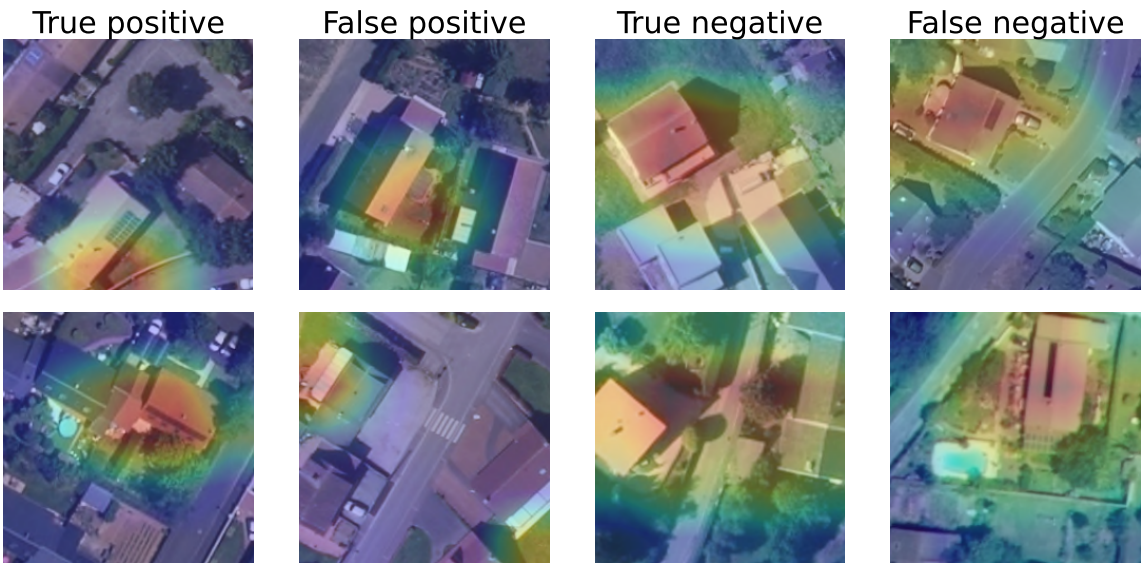


Figure 2.15 – Model explanations using the GradCAM (Selvaraju et al., 2020) for some true positives, false positives, true negatives and false negatives. The redder, the higher the contribution of an image region to the predicted class (1 for true and false positives, 0 for true and false negatives).

To further understand the model's behavior, we plot on Figure 2.16 the model's

predicted probabilities. We can see that except for false negatives, the distribution of the predicted probabilities is concentrated towards 0 or 1. This means the model predicts a PV panel (or no PV panel) with high confidence. This is particularly true for false positives, whose distribution closely follows the shape of the distribution of the true positives.

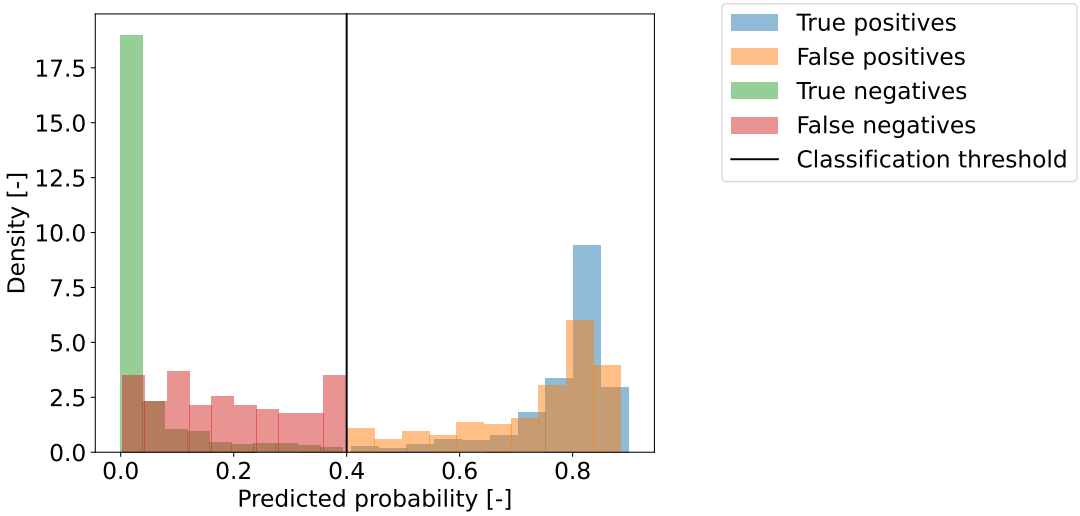


Figure 2.16 – Predicted probabilities of the true positives, false positives, true negatives, and false negatives on the BDAPPV test dataset.

False detections "in the wild" Now that we analyzed the model's behavior on the test set, let us focus on examples that occurred during deployment. Thanks to the DTA, we can target places where the model made significant mistakes and analyze them qualitatively. Table 2.7 unwraps the detections made in the city of Cobrieux (Nord) for which the APE is 96%, meaning that the model estimated the double of the true installed capacity.

The model detected 12 installations in this city for an aggregated installed capacity of 64 kW_p. The target, defined by the RNI, is ten installations for an installed capacity of 27.20 kW_p. Unwrapping the list of installations registered in these cities leads us to identify the culprit: an installation whose estimated installed capacity is 27.40 kW_p. If we remove this installation, we can recover an average installed capacity closer to the actual average size of the installations as given in the RNI.

Table 2.7 – Extract of the registry generated by DeepSolar for the city of Cobrieux (Nord).

	ID	Surface	Installed capacity
	29925	12.35	1.96
	29897	275.51	27.40
	29904	24.18	3.84
	29924	36.01	3.94
	29926	17.69	2.81
	29927	24.34	3.87
	29921	39.61	3.94
	29913	12.59	2.00
	29908	14.94	2.37
	29910	19.78	3.14
	29912	46.26	4.60
	29901	38.16	4.17
Average size			5.34
(curated)			3.33
Reference			*2.72*

Accessing this "installation"'s location on a GIS software as shown on Figure 2.17 reveals that the model confused the roof of a factory with a PV panel. The confusing factors were probably the lines on the roof of the farm.

Figure 2.17 – False detection identification with an example for the city of Cobrieux (Nord).

This example demonstrates the added value of the DTA for monitoring the model's output and identifying outliers such as those depicted on Figure 2.17. However, to improve the reliability of the mapping algorithms, we not only need to monitor the outputs and ensure that the decision process is relevant and robust. To address

these points, we formulate our work hypothesis for understanding the model's decision process in the next section.

3.3.3 Work hypothesis for understanding the model's decision process

PV panels as a set of features In section 3.3.1, we found no evidence of sensitivity to the geographical localization of the model's accuracy. We underlined that this result may be proper for our case study, as our training dataset spans France and thus captures the diversity of geographical characteristics we encounter in the mapping area. Nonetheless, this led us to analyze the model's prediction as documented in 3.3.2. The inquiries carried out after Figure 2.15 and Figure 2.16 lead us to assume that the model does not see the panel as a whole but as a set of characteristic features. These features comprise the groves, the color, and the overall shape. If the model matches at least one of these features on the image, then it predicts the presence of a PV panel with high confidence. On the other hand, if the model does not match such a pattern, it predicts – also with high confidence – that there is no PV panel on the image.

We suppose that, during training, the model extracts different features correlated with a PV panel on the image. These features can correspond to textures at different scales, components such as horizontal or vertical lines, colors, or the overall shape of the PV panel in some cases. We have limited control over what the model learns from the data as it depends on the data's quality, the model's initialization, and the hyperparameters. A trained model predicts during deployment the occurrence of a PV panel if one of these features is identified on the input image. However, it is possible to find patterns that resemble a PV panel but are not a panel, such as those presented on Figure 2.17.

False positives and negatives False positives can occur if the model sees a pattern resembling a PV panel. False negatives are more intricate since the panel *is* on the image, but the model does not recognize it. This could be because the image lacks the *decisive* feature, which is necessary for the model to recognize the panel *in this context*. We have limited control over what the model learns during training and the contexts seen during deployment. Therefore, we need to understand better what the model sees on the image to identify whether some factors are more likely to be context-dependent than others.

Assessing the relevance *and* robustness of the decision process To verify this hypothesis and understand why the model makes false predictions, we need to understand how it makes a prediction and whether it is robust (to distribution shifts). Understanding how the model makes a prediction will enable us to highlight the components it relies on. On the other hand, assessing whether the prediction

is robust will enable us to see whether the model can rely on factors more likely to be disrupted by the distribution shifts.

Conclusion of the chapter

In this chapter, we introduced our data sources, reviewed the existing data on rooftop PV panels, and introduced our training dataset. To address the first pillar of reliability, *monitoring* the data quality, we introduced the downstream task accuracy (DTA). The DTA is an unsupervised evaluation method that compares data aggregates generated from our model to existing data to measure the accuracy of the model's predictions indirectly. Using the DTA on our replication of existing works, we showed that existing models provide a satisfying estimation of the tilt and azimuth angles of the installations but show detection inconsistencies: they predict false positives that result in large, inexisting installations, which translates into an average estimation error of the city-wise installed capacity of 40%. We investigated common causes that could explain this performance drop and found that, in our case, it is driven by the fact that the model confidently predicts a PV panel when it sees a factor that resembles a PV panel or, on the other hand, completely misses the PV panel even if there is one. These false negatives appear because the image may not depict a *necessary* feature (in this context).

The next chapter will discuss this hypothesis to explain why a model makes false predictions. Addressing this question will enable us to tackle the two remaining pillars of reliability, as defined in chapter 1: assessing the decision process's relevance and robustness.

Chapter 3

Assessing the reliability of a model's decision process by generalizing attribution to the wavelet domain

Summary

This chapter introduces a new feature attribution method to assess the relevance and robustness of a model's decision process. This method highlights which regions of the space-scale domain are the most important for the model's decision. This feature attribution method is based on the efficient perturbation of the wavelet transform of the input image. It helps the user understand and disentangle the model's reliance on the structural components of the image. This method enables us to assess whether the classification model sees PV panels for the right reasons.
We then investigate the sensitivity to distribution shifts of PV detection models. We show that most of their sensitivity comes from the sensitivity to varying acquisition conditions. Thorough analysis using the WCAM reveals that this sensitivity can be explained by the fact that changes in the acquisition conditions can result in hiding a critical component the model needs to predict a PV panel. Finally, thanks to the proposed attribution method, we introduce a data augmentation technique aiming at reducing the sensitivity to varying acquisition conditions.

In the previous chapter, we characterized a PV panel as a set of features. To explain false detections, we hypothesize that false positives occur because the model sees a feature on the image that could be related to a PV panel (e.g., a grid pattern). On the other hand, in the case of a false negative, the feature that would have been predictive *in this context* (e.g., a color or lines in a given orientation) is absent for some reason. To verify this hypothesis, we first introduce a method that decomposes the model's prediction into a set of features. We chose the scales obtained from the wavelet decomposition as our set of features, as scales can be related to frequencies, thus enabling us to characterize the robustness of the prediction simultaneously.

1 Characterizing a model's decision in the space-scale domain: the Wavelet sCale Attribution Method (WCAM)

In chapter 2, section 3.3.2, we used the GradCAM method of Selvaraju et al. (2020) to analyze the model's decision. This approach is representative of usual practice in computer vision for interpreting a model's decision: we analyze which regions on the image contribute the most to the model's decision using a so-called feature attribution method. With methods such as the GradCAM, the features correspond to the regions of the image (i.e., in the pixel domain). However, we not only need the localization of the important features on the input image but also a characterization in terms of structural elements to assess whether the model sees the panels as a whole or as specific patterns on the panel. To this end, we introduce a new feature attribution method that expands existing works to characterize the detection of the model not only in the pixel domain but also in the scale domain.

1.1 Towards assessing what model sees on images

Decomposing a PV panel into scales A useful representation to understand how a model detects a PV panel on orthoimagery is to consider that the information leveraged by the model is located in space and in scale. Figure 3.1 shows an example of such a scale decomposition. Depending on the scale of interest, the PV panel will show different characteristics. The PV panel corresponds to small details within the individual PV modules at the smallest scales. On the other hand, if we consider the PV system as a whole, it is, in our case, about 10 meters long, i.e., 100 pixels, in this image.

Each scale can be related to semantic (i.e., meaningful) features or characteristics of the PV panel. The panel is a combination of all these features. We would like to know how to decompose the model prediction regarding these different scales. In section 3.3.2 of chapter 2, we leveraged a class of explainability methods called

feature attribution methods to identify the pixels that contributed the most to the model's prediction.

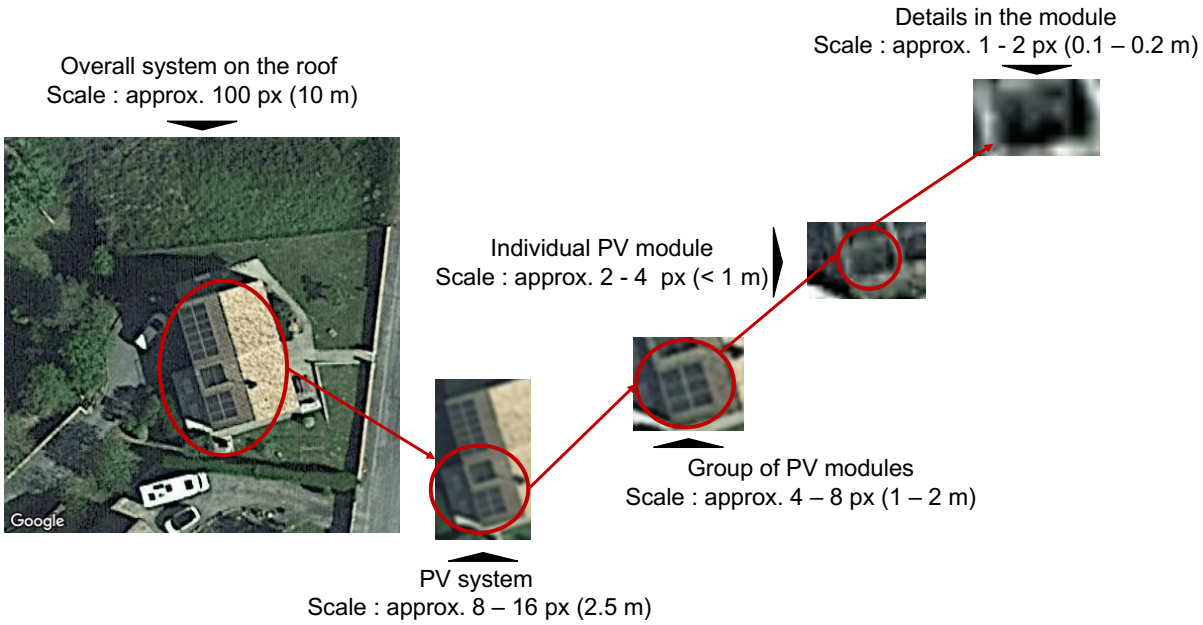


Figure 3.1 – Decomposition of the scales of a PV panel. Source: Kasmi et al. (2023b).

Feature attribution methods Explainability in computer vision typically highlights the regions on the image on which the model focuses, i.e., to construct saliency maps. Constructing such saliency maps can be done in two ways: using the model's gradients or activations or by quantifying the model's response to perturbations to the input image. We refer to the first class of feature attribution methods as white-box methods, as they require access to the model's weights and gradients, and the second class of methods as black-box methods, as they do not require access to the model's weights and gradients.

To our knowledge, the saliency maps of (Simonyan et al., 2014) were one of the earliest white-box feature attribution methods for deep neural networks. Given an image x and a model F, their approach consists in computing the gradient $\nabla_x F(x)$ with respect to the input and projecting it onto the input image. Figure 3.2 presents examples of saliency maps obtained following the approach of Simonyan et al. (2014).

This principle has been extended and refined in subsequent works (Shrikumar et al., 2017; Sundararajan et al., 2017). Similarly, Zhou et al. (2016) used the model's activations to compute class activation maps (CAMs). CAMS require the penultimate layer of a model to be a global average pooling (GAP) layer. Each neuron k of such a layer corresponds to the average of the coefficients of the preceding k_{th} convolutional layer C_k.The CAM is the average of the K convolutional

Figure 3.2 – Examples of saliency maps, computed by Simonyan et al. (2014). Taken from Simonyan et al. (2014).

layers $C_1, \ldots, C_K$ located before the penultimate GAP layer, weighted by their respective weights $w_1, \ldots, w_K$. The CAM takes advantage of the fact that the GAP is followed by a nonlinearity, meaning that some neurons in the penultimate layer of the model are not activated. CAMs require a particular architecture to be computed. To overcome this difficulty, the GradCAM, which computes the class activation maps using the model's gradients, has been introduced (Selvaraju et al., 2020). The main advantage of the white-box methods is that they are fast to compute, and gradient methods produce the best explanations whether in terms of faithfulness (Bhatt et al., 2020) or stability (Crabbé and van der Schaar, 2023).

In practical settings, models may be accessible only through an application programming interface (API), so their gradients are unavailable. In this case, black-box explainability methods can be preferred. Black box attribution methods compute explanations by perturbing (e.g., occluding parts of the image) the inputs and computing a score that reflects the model's sensitivity to the perturbation. The various proposed methods, e.g., Occlusion (Zeiler and Fergus, 2014), LIME (Ribeiro et al., 2016), RISE, (Petsiuk et al., 2018), Sobol (Fel et al., 2021), HSIC (Novello et al., 2022) or EVA (Fel et al., 2023a) differ in that they use different sampling strategies to explore the space of perturbations and can be seen as special cases of a more general approach based on Shapley values (Lundberg and Lee, 2017).

However, the main limitation of current explainability methods (both white-box and black-box) is that they only explain *where* the model focuses and are therefore not informative enough in many settings where one wants to assess *what* the model sees (Achtibat et al., 2022). To begin addressing the *what*, Fel et al. (2023b) recently introduced the Concept Recursive Activation FacTorization for Explainability (CRAFT). This method involves extracting concepts from the training dataset and attributing them to a prediction using a standard attribution method. CRAFT indi-

cates the most influential regions of the image and the concepts related to these influential regions. Other works focused on identifying the most significant points in the training dataset through influence functions (Koh and Liang, 2017). However, such approaches require access to the model and the training data and are, therefore, hard to implement in applied settings. These works enable understanding the model at the local level, i.e. for individual predictions.

Frequency-centric perspective on model robustness A line of works aimed at explaining the behavior of neural networks through the lenses of frequency analysis. Several works showed that convolutional neural networks (CNNs) are biased towards high frequencies (Wang et al., 2020; Yin et al., 2019) and that robust methods tend to limit this bias (Zhang et al., 2022; Chen et al., 2022). By "robust" or "robustness," we mean robustness to adversarial perturbations or natural image corruptions. We further discuss the robustness in section 3.1.1. Other works highlighted a so-called spectral bias (Rahaman et al., 2019; John Xu et al., 2020; Jo and Bengio, 2017), showing that CNNs learn the input image frequencies from the lowest to the highest. More recently, Wang et al. (2023) leveraged Fourier analysis to characterize shortcuts (Geirhos et al., 2020): this work showed that shortcuts are context-dependent as models tend to favor the most distinctive frequency to make a prediction. Decision shortcuts correspond to the fact that models rely on features highly predictive of the class of interest in a given setting at the expense of considering all the available information. Spurious correlations are an example of shortcuts. These methods enable us to understand the model at the global level.

1.2 Feature importance quantification in the scale-space domain

Overview Wavelets are a natural tool to decompose an image into scales while maintaining local analysis in space: they provide a single space-scale decomposition. We will, therefore, use this decomposition of the image. On the other hand, several attribution methods, Fel et al. (2021), for instance, leveraged Sobol indices to quantify the importance of features. Our contribution is to combine both tools to highlight the important contributors to the detection in the space-scale domain. The quantification of the importance relies on the Sobol indices. By highlighting the important areas in the space-scale domain, we can decompose the prediction in terms of scales (and thus assess what are the components highlighted in Figure 3.1 that are important for the model) *and* assess whether these components are robust or not from a frequency-centric perspective. Indeed, the finer scales correspond to the highest frequency ranges, which are most likely to be disrupted by the acquisition conditions.

Wavelet transform A wavelet is an integrable function $\psi \in L^2(\mathbb{R})$ with zero average, normalized and centered around 0. Unlike a sinewave, a wavelet is localized in space and in the Fourier domain. It implies that dilatations of this wavelet enable to scrutinize different scales while translations enable to scrutinize spatial location. In other words, scales correspond to different spatial frequency ranges or spectral domains (see Figure 3.3 for an illustration).

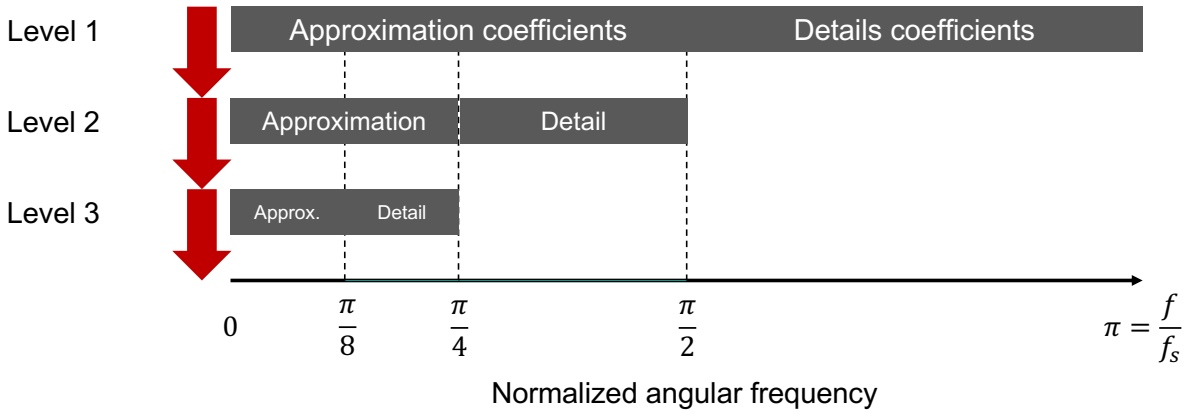


Figure 3.3 – Correspondence between the scales in the wavelet domain and the frequency ranges in the Fourier domain. In this diagram, f_s corresponds to the highest frequency contained in the signal. Inspired by Chen et al. (2019).

To compute an image's (continuous) wavelet transform (CWT), one first defines a filter bank $\mathcal{D}$ from the original wavelet ψ with the scale factor s and the 2D translation in space u. We have

$$\mathcal{D} = \left\{ \psi_{s,u}(x) = \frac{1}{\sqrt{s}} \psi \left(\frac{x-u}{s} \right) \right\}_{u \in \mathbb{R}^2,\, s \geq 0}, \tag{3.1}$$

where $|\mathcal{D}| = J$, and J denotes the number of levels. The computation of the wavelet transform of a function $f \in L^2(\mathbb{R})$ at location x and scale s is given by

$$\mathcal{W}(f)(x,s) = \int_{-\infty}^{+\infty} f(u) \frac{1}{\sqrt{s}} \psi^* \left(\frac{x-u}{s} \right) \mathrm{d}u, \tag{3.2}$$

which can be rewritten as a convolution (Mallat, 1999). Computing the multilevel decomposition of f requires applying Equation 3.2 J times with all dilated and translated wavelets of $\mathcal{D}$. Mallat (1989) showed that one could implement the multilevel dyadic decomposition of the discrete wavelet transform (DWT) by applying a high-pass filter H to the original signal f and subsampling by a factor of two to obtain the *detail* coefficients and applying a low-pass filter G and subsampling by a factor of two to obtain the *approximation* coefficients. Iterating on the approximation coefficients yields a multilevel transform where the j^{th} level extracts information at resolutions between 2^j and 2^{j-1} pixels. The detail coefficients can be decomposed into horizontal, vertical, and diagonal components when dealing with 2D signals

(e.g., images). Figure 3.3 illustrates the multilevel decomposition of a signal into approximation and detail coefficients. The detail coefficients at level k contain the frequencies located between $\frac{1}{2^{k+1}}\pi$ and $\frac{1}{2^k}\pi$ where $\pi = \frac{f}{f_s}$ and f_s is the highest frequency in the signal.

Interpreting the wavelet transform of an image Figure 3.4 illustrates how to interpret the (two-level) wavelet transform of an image. Reading is the same for any multilevel decomposition. The rightmost image plots the two-level dyadic decomposition of the leftmost image. Following this transform, the localization on the image highlighted by the red polygon can be decomposed into six detail components (marked yellow and blue) and one approximation component (marked pink). The yellow components correspond to the details at the 1-2 pixel scale, and the blue components to the details at the 2-4 pixel scale. For each location, the wavelet transform summarizes the information contained in the image at this scale and location.

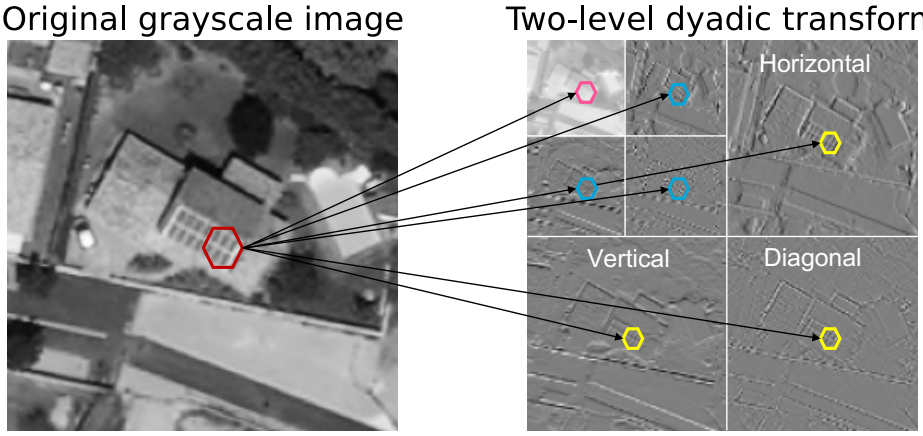


Figure 3.4 – Image and associated two-level dyadic wavelet transform with indications to interpret the wavelet transform of the image. "Horizontal," "diagonal," and "vertical" indicate the direction of the detail coefficients. The direction is the same at all levels.

Sobol sensitivity analysis The Sobol sensitivity analysis consists in decomposing the variance of the output of a model into fractions that can be attributed to a set of inputs. Let $(X_1, \ldots, X_K)$ be independent random variables and $\mathcal{K} = \{1, \ldots, K\}$ denote the set of indices. Let f be a model, X an input, and $f(X)$ the model's decision (e.g., the output probability). We denote $f_\kappa = f_\kappa(X_\kappa)$ the partial contributions of the variables $(X_k)_{k\in\kappa}$ to the score $f(X)$. The Sobol-Hoeffding decomposition (Hoeffding, 1992) decomposes the decision score $f(X)$ into summands of increasing

dimension

$$f(X) = f_{\emptyset} + \sum_{\kappa \in \mathcal{P}(\mathcal{K}) \setminus \{\emptyset\}} f_{\kappa}(X_{\kappa}), \tag{3.3}$$

Where $f_{\emptyset}$ denotes the prediction with no features (i.e., the average prediction), $\mathcal{P}(\mathcal{K})$ denotes the power set of $\mathcal{K}$ and $\emptyset$ the empty set. Then, $\forall (u, v) \in \mathcal{K}^2$ such that $u \neq v$, $\mathbb{E}[f_u(X_u) f_v(X_v)] = 0$, we derive from Equation 3.3 the variance of the model's score

$$Var(f(X)) = \sum_{\kappa \in \mathcal{P}(\mathcal{K})} Var(f_{\kappa}(X_{\kappa})), \tag{3.4}$$

Equation 3.4 enables us to describe the influence of a subset κ of features as the ratio between its own and the total variance. This corresponds to the first order **Sobol index** given by

$$S_{\kappa} = \frac{Var(f_{\kappa}(X_{\kappa}))}{Var(f(X))}. \tag{3.5}$$

S_{κ} measures the proportion of the output variance $Var(f(X))$ explained by the subset of variables X_{κ} (Sobol, 1990). In particular, the first order Sobol index S_k only captures the *direct* contribution of the feature X_k to the model's decision. To capture the indirect (or coupling) effect, due to the effect of X_k on the other variables, **total Sobol indices** S_{T_k} (Homma and Saltelli, 1996) can be computed as

$$S_{T_k} = \sum_{\kappa \in \mathcal{P}(\mathcal{K}), k \in \kappa} S_{\kappa}. \tag{3.6}$$

Total Sobol indices (TSIs) measure the contribution of the k^{th} feature, taking into account both its *direct* effect and its *indirect* effect through its interactions with the other features.

Efficient estimation of Sobol indices As seen from Equation 3.5, estimating the impact of a feature k on the model's decision requires recording the partial contribution $f_k(X_k)$. This partial contribution corresponds to a *forward*. Estimating Sobol indices requires computing variances by drawing at least N samples and computing N forwards to estimate a first-order Sobol index S_k of a single feature k. As we are interested in the TSI of a feature k, we need to estimate the Sobol index of all sets of features $\kappa \in \mathcal{K}$ such that $k \in \kappa$. To minimize the cost of this computation, Fel et al. (2021) leveraged the efficient estimator of Jansen (1999) based on Quasi-Monte Carlo (QMC) methods (Morokoff and Caflisch, 1995) to estimate the TSIs given the models' outputs and the perturbations. The N perturbations of dimension K are drawn from Sobol sequences (Sobol, 1967). Their approach requires $N(K+2)$ forwards (Fel et al., 2021).

To estimate the TSIs, they draw two matrices from a Quasi-Monte Carlo sequence of size $N \times K$ and convert them into perturbations, which they apply to X. The perturbated input yields two matrices, A and B. a_{jk} (resp. b_{kj}) is the element of A

(resp. B) corresponding to the k^{th} feature and the j^{th} sample. For the k^{th} feature, they define $C^{(k)}$ in the same way as A, except that the column corresponding to feature k is replaced by the column of B. They then derive an empirical estimator for the Sobol index and TSI as

$$\hat{S}_k = \frac{\hat{V} - \frac{1}{2N}\sum_{j=1}^{N}\left[f(B_j) - f\left(C_j^{(k)}\right)\right]^2}{\hat{V}}, \qquad \hat{S}_{T_k} = \frac{\frac{1}{2N}\sum_{j=1}^{N}\left[f(A_j) - f\left(C_j^{(k)}\right)\right]^2}{\hat{V}}, \tag{3.7}$$

where $f_{\varnothing} = \frac{1}{N}\sum_{j=1}^{N} f(A_j)$ and $\hat{V} = \frac{1}{N-1}\sum_{j=0}^{N}[f(A_j) - f_{\varnothing}]^2$. Further details on implementing the method can be found in Fel et al. (2021).

1.3 The Wavelet sCale Attribution Method (WCAM)

1.3.1 Expanding attribution in the space-scale domain

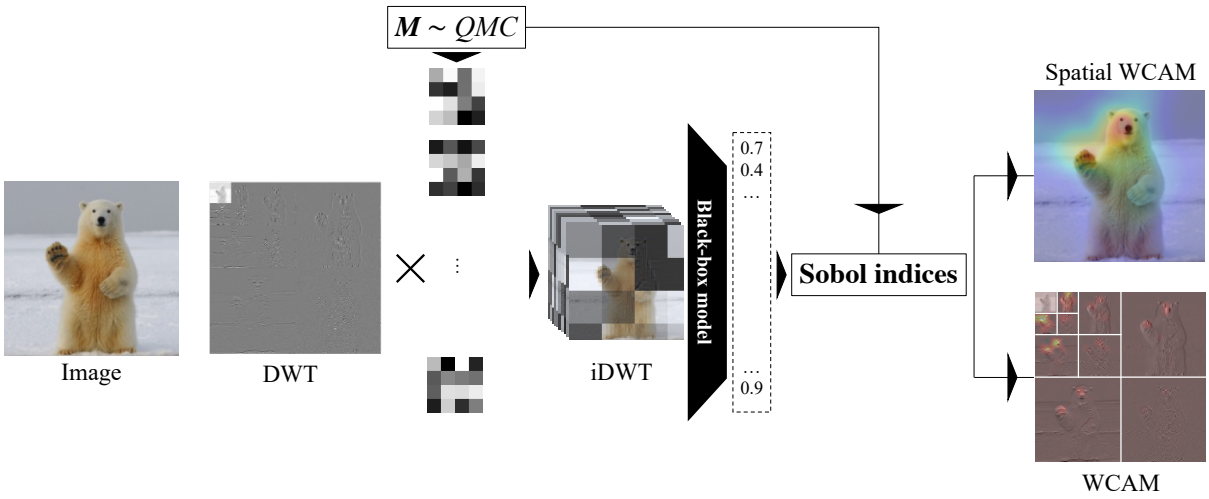


Figure 3.5 – Flowchart of the wavelet scale attribution method (WCAM). Source: Kasmi et al. (2023a).

Overview The **W**avelet s**C**ale **A**ttribution **M**ethod (**WCAM**) is an attribution method that quantifies the importance of the regions of the wavelet transform of an image to the predictions of a model. Figure 3.5 depicts the principle of the WCAM. The importance of the regions of the wavelet transform of the input image is estimated by **(1)** generating masks from a QMC sequence, **(2)** evaluating the model on perturbed images. We obtain these images by computing the DWT of the original image, applying the masks on the DWT to obtain perturbed DWT,[1] and inverting the perturbed DWT to generate perturbed images. We generate $N(K+2)$ perturbed images for a single image. **(3)** We estimate the total Sobol indices of the perturbed

1. On an RGB image, we apply the DWT channel-wise and apply the same perturbation to each channel.

regions of the wavelet transform using the masks and the model's outputs using Jansen's estimator (Jansen, 1999). Fel et al. (2021) introduced this approach to estimate the importance of image regions in the pixel space. We generalize it to the wavelet domain.

Computation of the perturbation masks We follow the sampling procedure introduced by Fel et al. (2021) to generate the masks. Their approach involves drawing two independent matrices of size $N \times K$ from a Sobol low discrepancy sequence. N is the number of designs necessary to estimate the variance, and K is the sequence dimension. We reshape this sequence as a two-dimensional mask to generate our perturbation. By default, we perturb the wavelet transform with a mask of size 28 $\times$ 28 to balance between the sequence's dimensionality and the perturbation's accuracy. We reshape the 784-dimensional sequence to a grid of 28 $\times$ 28 to define our perturbation masks. We tried alternative mapping from the unidimensional sequence to the mask. However, it had a limited effect on the dimensionality reduction and at the expense of the meaningfulness of the perturbation in the wavelet domain. Figure 3.6 illustrates our workflow for one mask to generate the images that are then passed to the model.

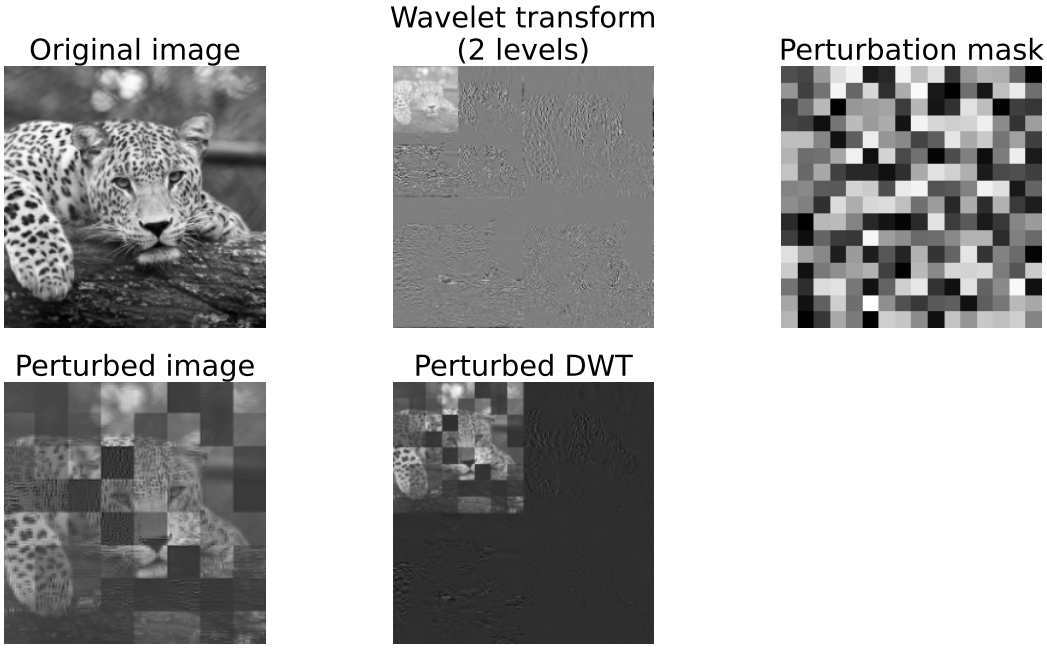


Figure 3.6 – Workflow on a grayscale image and for a 2-level wavelet transform. We first compute the discrete wavelet transform of the image and then apply a mask on the discrete wavelet transform (DWT). It yields the perturbed DWT, which we invert to generate the perturbed image. We evaluate the model on the perturbed image.

The WCAM expands attribution to the space-scale domain The WCAM decomposes a prediction into the wavelet domain. As Figure 3.7 depicts, highlighting an important area in the pixel domain (i) does not provide information on *what* the model sees. By decomposing the prediction into the wavelet domain (ii), the WCAM represents the important features of a prediction in terms of structural components. In the example of Figure 3.7, we can see two important areas for predicting the fox: the hind leg and the ear. We can see that three distinct components contribute to the prediction for the ear. Areas **(a)**, **(b)**, **(c)** and **(d)** highlight these components. **(a)** corresponds to details at the 1-2 pixel scale, i.e., fine-grained details such as the fur in the ear. **(b)** corresponds to details at the 2-4 pixel scale, i.e., larger details such as the shape of the ear. We can see that both vertical (**(b)**) and horizontal (**(c)**) components of the shape of the ear contribute to the prediction. On the other hand, for the hind leg, only the overall vertical shape (4-8 pixel size, **(d)**) contributes to the prediction.

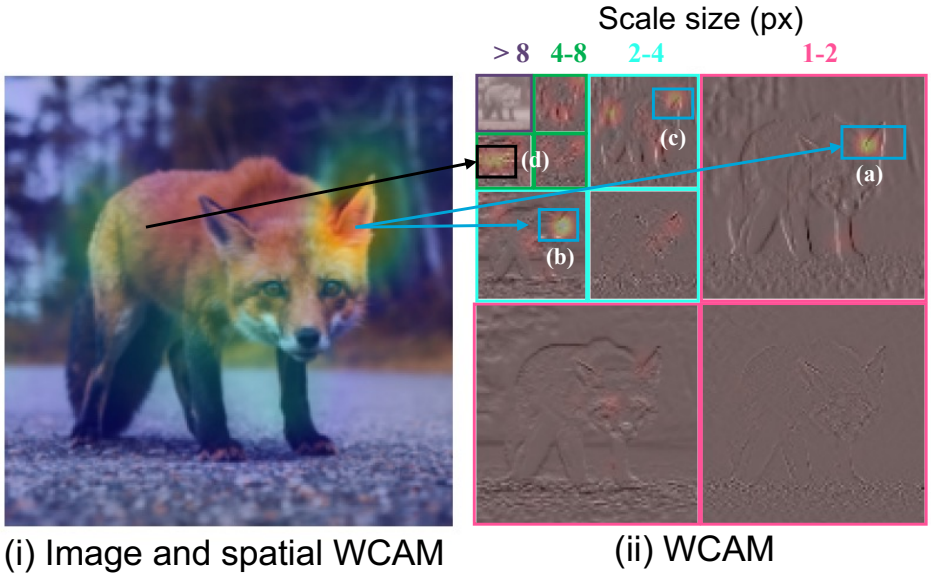


Figure 3.7 – Decomposition of a prediction from the pixel domain (i) into the wavelet domain (ii) with the WCAM. Source: Kasmi et al. (2023a).

1.3.2 Connecting attribution and robustness

Scales, frequencies, and robustness As stated in section 1.2, and illustrated in Figure 3.3 scales in the wavelet domain correspond to dyadic frequency ranges in the Fourier domain. The smallest scales correspond to the highest frequencies. Therefore, the WCAM connects attribution with frequency-centric approaches to model robustness. To show this connection, we replicate the Chen et al. (2022) experiment. In this work, the authors compared the standard model ("ST," i.e., a

ResNet-50 trained using the vanilla ERM) with adversarial models ("AT," i.e., models trained with adversarial training to improve their robustness to adversarial perturbations) and robust models ("RT," i.e., models trained to be robust against natural image corruptions). They showed that the ST model relies more on high-frequency components to make predictions than the RT and AT models.

In our case, we compute the importance of each detail ("h," "v," and "d") at each scale (1 to 4) and the importance of the approximation coefficients ("a"). The importance corresponds to the value of the Sobol indices within each level. We then average the importance across 500 randomly sampled images. Figure 3.8 shows the results. We see that robust models favor coarse scales (i.e., low-frequencies) over fine scales (i.e., high frequencies). The WCAM characterizes robust models by estimating the importance of each frequency component in the final prediction. We can see that the ordering from the detail coefficients corresponding to the largest scales from those corresponding to the highest remains the same.

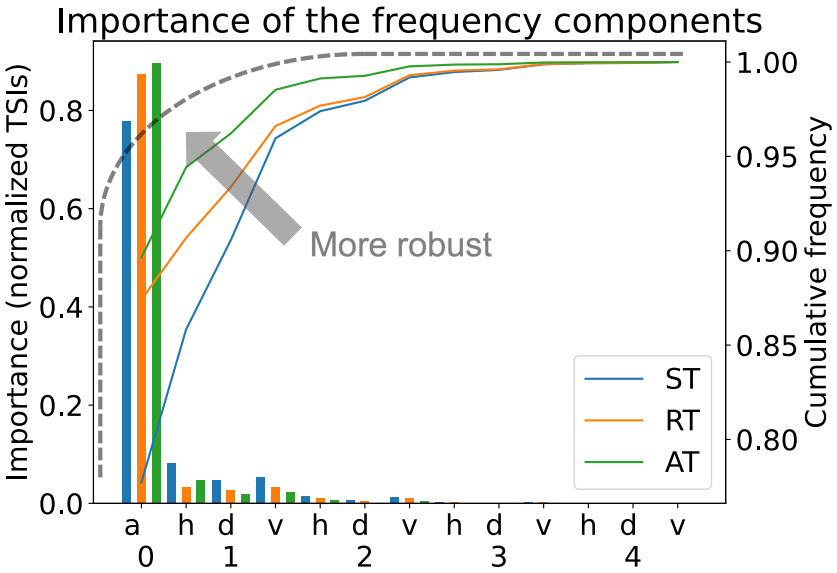


Figure 3.8 – Representation of the scales of the WCAM as frequencies. Levels (numbered from (0 to 4) indicate the scales, from the coarser (i.e., lowest frequencies) to the finest (i.e., highest frequencies. The level 0 or "a" corresponds to the approximation coefficients. Labels "h," "v," and "d" correspond to the horizontal, vertical, and diagonal details, respectively. The rightmost index plots the cumulative curve. "AT," "RT," and "ST" stand for adversarial, robust, and standard training, respectively. The dotted line indicates the concentration towards coarser scales associated with better robustness. Adapted from Kasmi et al. (2023a).

This can be further assessed by the cumulative frequencies (right axis), where the most robust models have a cumulative curve that is more concentrated than the cumulative frequency of nonrobust models. This indicates they rely more on the coarser scales (i.e., low-frequency components). These results are in line with existing works (Zhang et al., 2022; Chen et al., 2022; Wang et al., 2020; Yin et al.,

2019) and show that the WCAM correctly estimates the robustness of a model.

2 Assessing the reliability of a model's decision process through the lenses of the scale-space decomposition

In this section, we leverage the WCAM to assess the reliability of the model's decision process. We analyze the relevance, i.e., whether the model relies on semantically meaningful factors. This enables us to explain why false positives occur. To explain why false negatives occur, we set up a small motivating experiment to understand why the model no longer recognizes PV panels when we change the image provider. Using the WCAM, we show that the image no longer features an important factor and that this factor has disappeared due to the differences in acquisition conditions. We support this mechanism with a small empirical model that shows that we can alter the model's prediction by altering the frequency content of the image and that the resulting behavior, in this case, is the same as in our motivating experiment.

2.1 Relevance of the decision process: understanding the model's decision process through the lenses of the space-scale decomposition

2.1.1 Understanding predictions

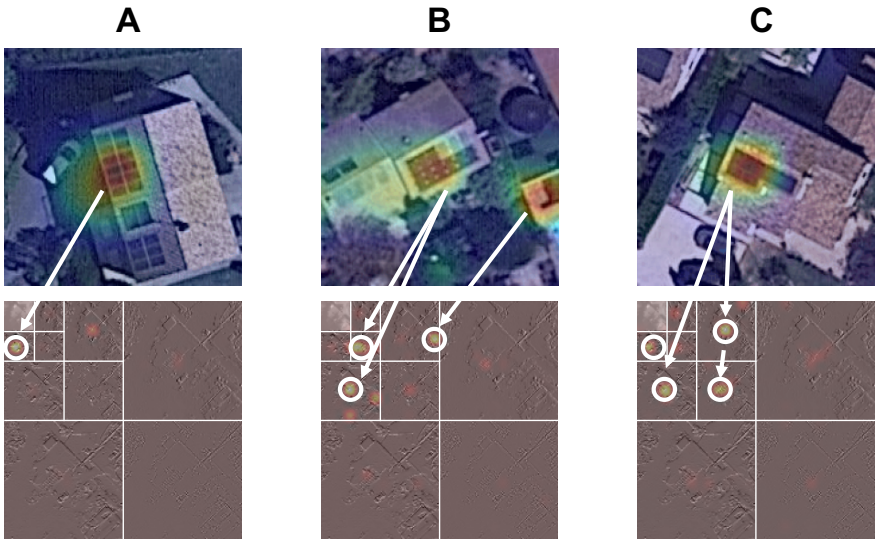


Figure 3.9 – Decomposition in the space-scale domain of PV panel predictions (true positives). Adapted from Kasmi et al. (2023b).

True positives Figure 3.9 presents some examples of predictions made by a model trained to classify images of PV systems. The upper row presents the image and the localization of the important regions in the space domain. The lower row represents the WCAM. We can see that for the image **A**, the important localization in the space domain mostly translates to a single position in the space-scale domain. On the other hand, different scales strongly contribute to the prediction for the images **B** and **C**. Overall, a single localization in the space domain translates into several localizations in the space-scale domain, meaning that information from different scales is important for the model to predict a PV panel correctly.

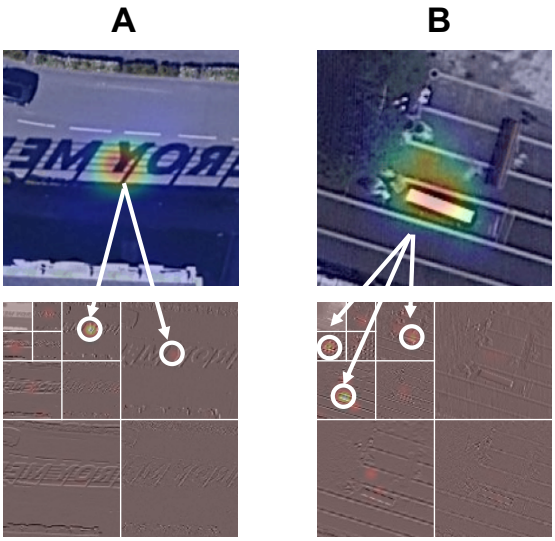


Figure 3.10 – Examples of false positives on IGN and corresponding WCAM. Adapted from Kasmi et al. (2023b).

False positives In line with the behavior highlighted by the GradCAM in chapter 2, section 3.3.2, Figure 2.15, the model's behavior is the same as for true positives when it comes to false positives. On the examples of Figure 3.10, we can see that the localization of the component that fools the model can also be decomposed into several localizations in the space-scale domain. Interestingly, for image **A** of figure Figure 3.10, we can see that the intersection between the grid and the "Y" on the shadow contributes the most to the model's wrong decision and that it contributes at two different scales. This evidence confirms the intuition that the model confuses the grids on the shadow with actual grid patterns typically found on PV panels. The same phenomenon occurs in image **B**, where the road lines and the parked vehicle are also seen as grid components.

Why do false positives arise? On Figure 3.10, we highlighted cases where the model predicts a PV panel because it sees a gridded pattern. This gridded pattern is not necessarily evident in the space domain (e.g., on image **B** of Figure 3.10)

but appears more clearly when considering the wavelet transform of the image and the WCAM. Therefore, we can suppose that the grid pattern, which can be found on many old PV panels, is a *learning shortcut*. Learning shortcuts (Geirhos et al., 2020) describe the tendency of neural networks to favor the most distinctive features to make a prediction. Whenever this feature arises on an image, it may tend to predict a PV panel.

2.1.2 Localizing the critical components in the space-scale domain

In the previous section, we leveraged the WCAM to understand the positive detections of a classification model. Following our hypothesis from chapter 2, section 3.3.3, the model predicts a PV panel when it identifies a component that can be semantically related to a PV panel on the image without necessarily taking into account additional components such as in the example of the sign shadow. In this section, we quantify this component as the *critical* component and show that once removed from the image, the model no longer sees the PV panel. In section 2.2, we will use the notion of a critical component to explain the sensitivity to acquisition conditions.

Definitions We call **sufficient image** the image reconstructed from the n first wavelet coefficients ranked according to their corresponding Sobol indices such that the model can correctly predict the image's label. Figure 3.11 displays examples of such images. In our examples, we can see that the model needs the information that is primarily located on the PV panels. On all images, we can see that it is not necessary to reconstruct the background for the model to correctly predict the PV panel. We can see that the finest information is required when the panel is smaller on the image. The number n_c of coefficients necessary to construct a sufficient image depends on the image.

If the n_c first components are sufficient to construct an image that the model correctly predicts, we call **critical component** the n_c^{th} coefficient. If we reconstruct an image with $n_c - 1$ components, the model does not have enough information. With $n_c + 1$ components, the information brought by the $(n_c + 1)^{th}$ component can be removed without changing the model's decision.

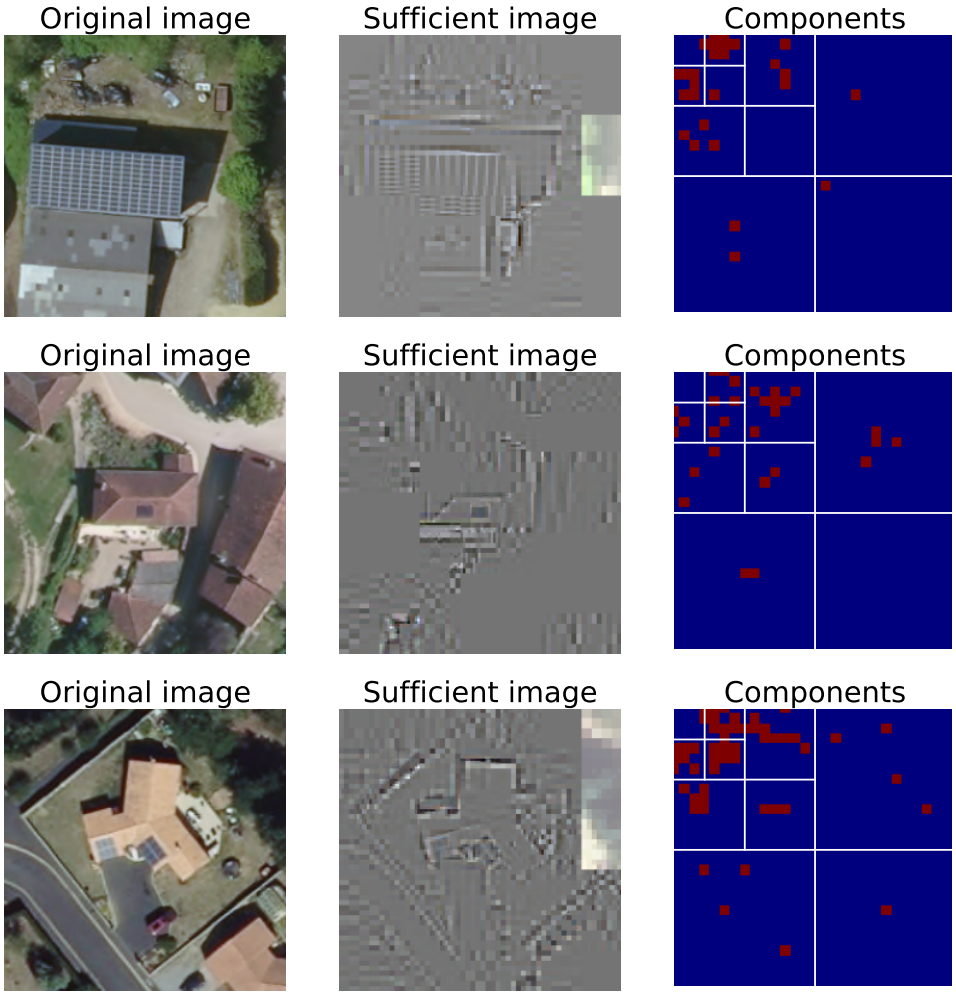


Figure 3.11 – Sufficient images reconstructed from the WCAM.

Shedding light on critical components On images of PV panels, the critical (or pivotal) component can be related to actual semantic content on the image, thus shedding light on *what* is important and constitutes a PV panel in the eyes of the model. On Figure 3.12, the model only needs n_c =17 coefficients to predict the PV panel correctly. Depicting these coefficients, we can see that they mainly capture the gridded structure of the PV panel. The critical information corresponds to the vertical lines at a precise location of the PV panel. Without this information, the model does not predict the PV panel. On the WCAM, we include a feature that enables us to see the critical component, as the red component on the bottom-right WCAM in figure 3.12. We provide additional examples on the model of Figure 3.12 in appendix C, section 2.2.

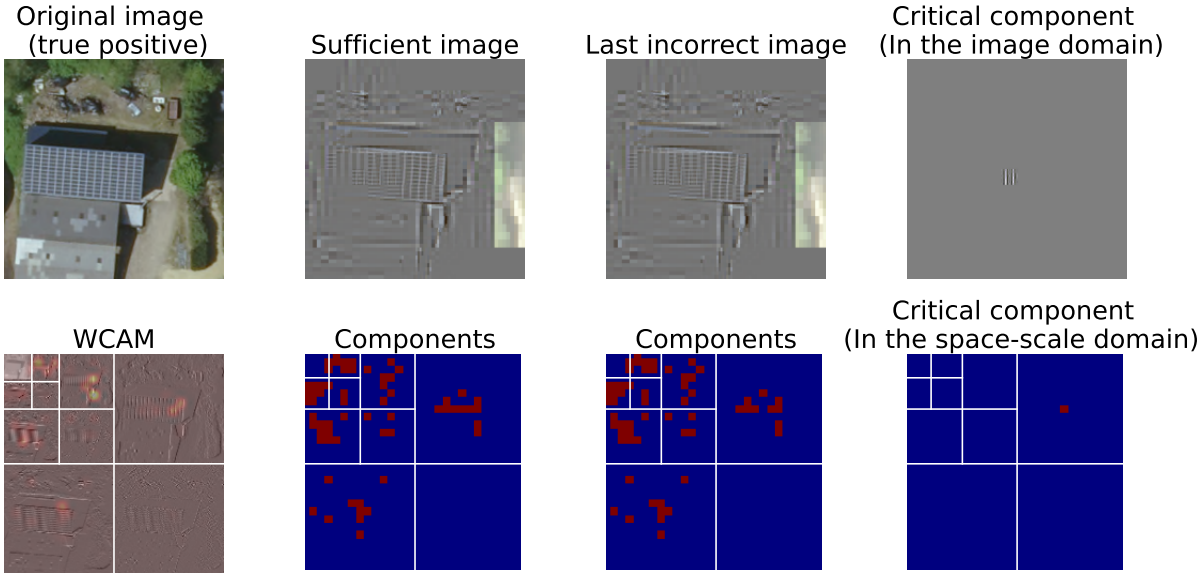


Figure 3.12 – Identification of the critical component (highlighted in white on the "Critical component" plot on the bottom right of the image. Without this component, the model does not predict the PV panel. The sufficient image is the image reconstructed with the minimal set of components.

Summary: assessing the relevance of the decision process In this section, we leveraged the WCAM to decompose the model's prediction in the space-scale domain. This enables us to understand *what* more precisely the model sees on the image when it predicts (or not) a PV panel. On Figure 3.13, we plot the WCAM alongside the GradCAM for the examples provided in Figure 2.15.

The WCAM enables us to gain intuition on the model's predictions by showing that a single localization in the space domain corresponds to different localizations in the space scale domain. In this domain, the WCAM enables us to see shortcuts that induce a false positive and that are less visible in the space domain only. For instance on Figure 3.13, on the false positive on the upper row, the important components are located in the 2-4 and 4-8 pixel scales, meaning that the shade roof and the veranda are confused with a PV panel because of their overall shape (and also probably because of their color as well).

Finally, we introduced the notion of *critical component* to refer to a component of the image necessary for the prediction. This component is localized in space and scale and can be isolated with the WCAM. This notion helps us understand what is critical for the model to make a decision. In the next section, we will discuss the robustness of this component to distribution shifts, as the robustness of the decision process is the last pillar for assessing its reliability.

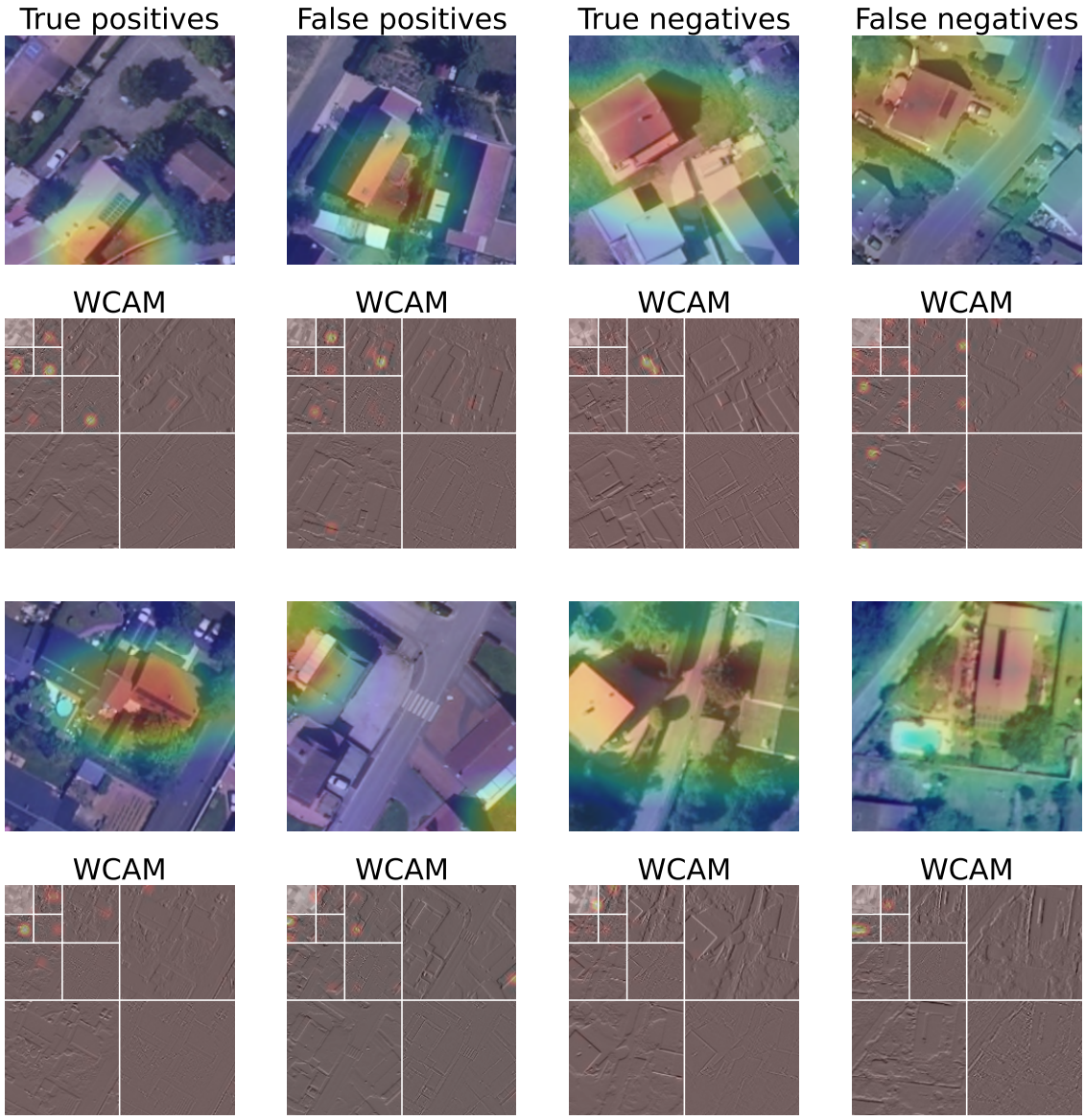


Figure 3.13 – Comparison of model explanations using the GradCAM (Selvaraju et al., 2020) and the WCAM (Kasmi et al., 2023a) correct and incorrect predictions. The WCAM shows that different scales contribute to the prediction and that when focusing on single scales, shortcuts such as gridded-like patterns can arise.

2.2 Robustness of the decision process: the impact of acquisition conditions on false negatives

2.2.1 Identifying the impact of acquisition conditions in the sensitivity to distribution shifts

The sources of distribution shifts in remote sensing The literature extensively discussed the sensitivity to distribution shifts of current models (see the discussion in chapter 1, section 2.2). Tuia et al. (2016) identified two primary distri-

bution shifts in remote sensing: the geographical variability and the heterogenous acquisition conditions. Following Murray et al. (2019), we can add the ground sampling distance, as the models are sensitive to zooming on images (Taesiri et al., 2023). However, to our knowledge, few works discussed this issue in the context of remote sensing of PV panels. Wang et al. (2017) argued that the ability to generalize depended on how hard to recognize the PV systems are. However, no work properly disentangled the effect of each source of variability. We propose to bridge this gap through a small experiment using the BDAPPV dataset (Kasmi et al., 2023d). Our goal is to evaluate the impact of the different sources of distribution shifts (i.e., acquisition conditions, geographical variability, and ground sampling distance) on the model's generalization ability.

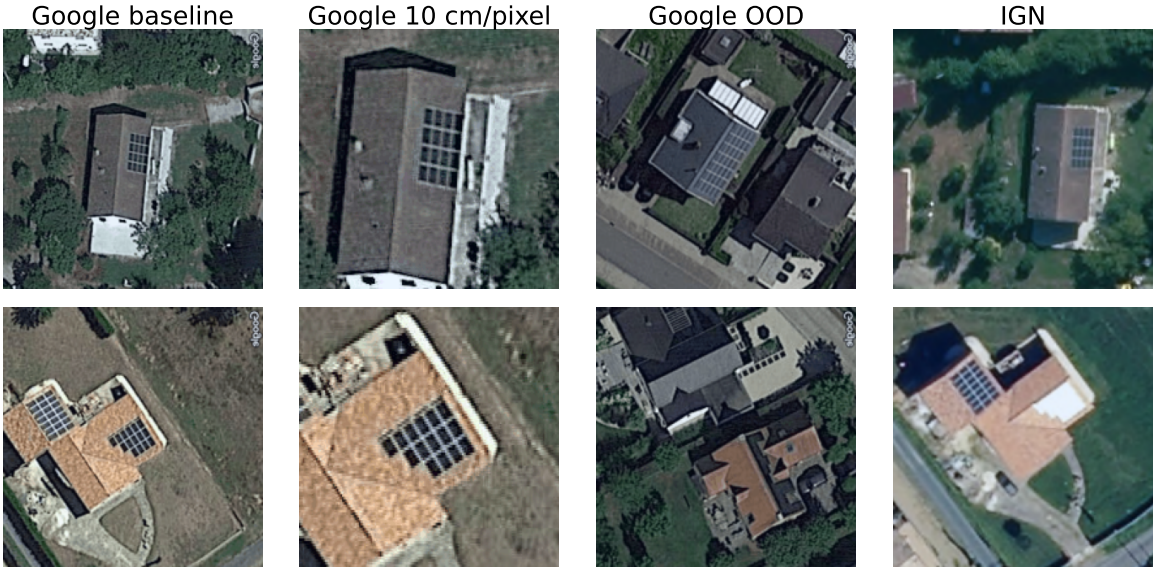


Figure 3.14 – Examples of images used in this experiment.

Disentangling the sources of distribution shifts BDAPPV features images of the same installations from two providers and records the crude location of the PV installations. Using this information, we can define three test cases to disentangle the distribution shifts that occur with remote sensing data: the resolution, the acquisition conditions, and the geographical variability. We train a ResNet-50 model (He et al., 2016) on Google images downsampled at 20cm/pixel of resolution and evaluate it on three datasets: a dataset with Google images at their native 10cm/pixel resolution ("Google 10 cm/pixel" on Figure 3.14), the IGN images with a native 20cm/pixel resolution ("IGN" on Figure 3.14) and Google images downsampled at 20 cm/pixel located outside of France ("Google OOD[2]" on Figure 3.14).

2. OOD: out-of-distribution. This corresponds in our framework to the geographical variability of the images

We add the test set to record the test accuracy without distribution shifts ("Google baseline" on Figure 3.14). We only do random crops, rotations, and ImageNet normalizations during training.

A strong impact of the acquisition conditions Table 3.1 shows the results of the disentanglement of distribution shifts into three components: resolution, acquisition conditions, and geographical shift. We can see that the F1 score[3] drops the most when the model faces new acquisition conditions (IGN in this experiment). The second most significant impact comes from the change in the ground sampling distance, but the performance drop remains relatively small compared to the effect of the acquisition conditions. In our framework, there is no evidence of an effect of the geographical variability once we isolate the effects of the acquisition conditions and ground sampling distance. This effect is probably underestimated, as images of our dataset that are not in France are near France. Nonetheless, we are primarily concerned with the acquisition conditions: the latter vary over France, and if we deploy the model on new images to update the registry, we will have new acquisition conditions. We want to know how acquisition conditions will impact the model's behavior.

Table 3.1 – **F1 Score** and decomposition in true positives, true negatives, false positives, and false negatives rates of the disentanglement of the distribution shift between the GSD (Google 10 cm/px), the geographical variability (Google OOD) and the acquisition conditions (IGN). Taken from Kasmi et al. (2023b).

	F1 Score (↑)	TPR	TNR	FPR	FNR
Google baseline	0.98	0.99	0.98	0.02	0.01
Google 10cm/px	0.89	0.81	1.00	0.00	0.19
Google OOD	0.98	0.99	0.98	0.02	0.01
IGN	0.46	0.32	0.95	0.03	0.68

Going further, we inspect on Figure 3.15 how varying the acquisition conditions affects the model's predicted probabilities (i.e., the probability assigned to a class after the softmax normalization and before applying the classification threshold). We can see that when the classifier no longer recognizes the PV panel, the probability shift is large, suggesting that the important factor for prediction disappeared from the image. In other words, if a critical component is no longer depicted due to the change in the acquisition condition, then the model no longer sees a PV panel (despite having other information about it). These results are in line with the hypothesis made in chapter 2, section 3.3.2, where on Figure 2.16, we witnessed that the model sees or does not see a PV panel with great confidence.

3. See chapter 4, section 2.1.1 for a definition of the F1 score.

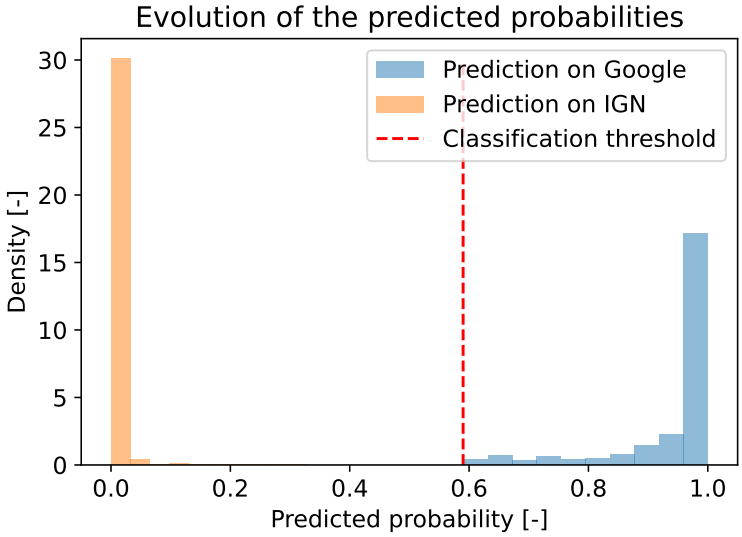


Figure 3.15 – Evolution of the predicted probabilities for images depicting a PV panel on the Google test set and the corresponding images on the IGN test set. The predicted probability completely flips over when the model no longer recognizes the PV panel. Source: Kasmi et al. (2023b).

2.2.2 Unveiling the impact of acquisition conditions on the prediction

When important factors disappear Plotting an explanation on the pixel domain does not help understand why acquisition conditions fooled the model (upper row of Figure 3.16). We can see that on Google and IGN, important areas for predicting a PV panel (Google) and the absence of a PV panel (IGN) are located over the PV panel.

Visually, the IGN image is more "blurred" than the Google image, meaning it misses details at the finest scales. We can see that on the Google image, these finest scales contribute the prediction (components **(a)** on the WCAM of the Google image, on the bottom row of Figure 3.16) but are no longer present on the IGN image, explaining why the model no longer recognizes the PV panel. Indeed, the disappearance of the finest scales, i.e., lines of the grid that are less visible, fooled the model, as these lines were necessary for the prediction.

On the IGN image, the model relies on coarser details to make its prediction (mostly the details in **(b)**). We also see that the critical component is no longer visible on the IGN image. This critical component (located in **(c)** corresponds in our case to the frame of the PV system. In appendix C, section 2.1, we provide additional illustrations of the disappearance of important components.

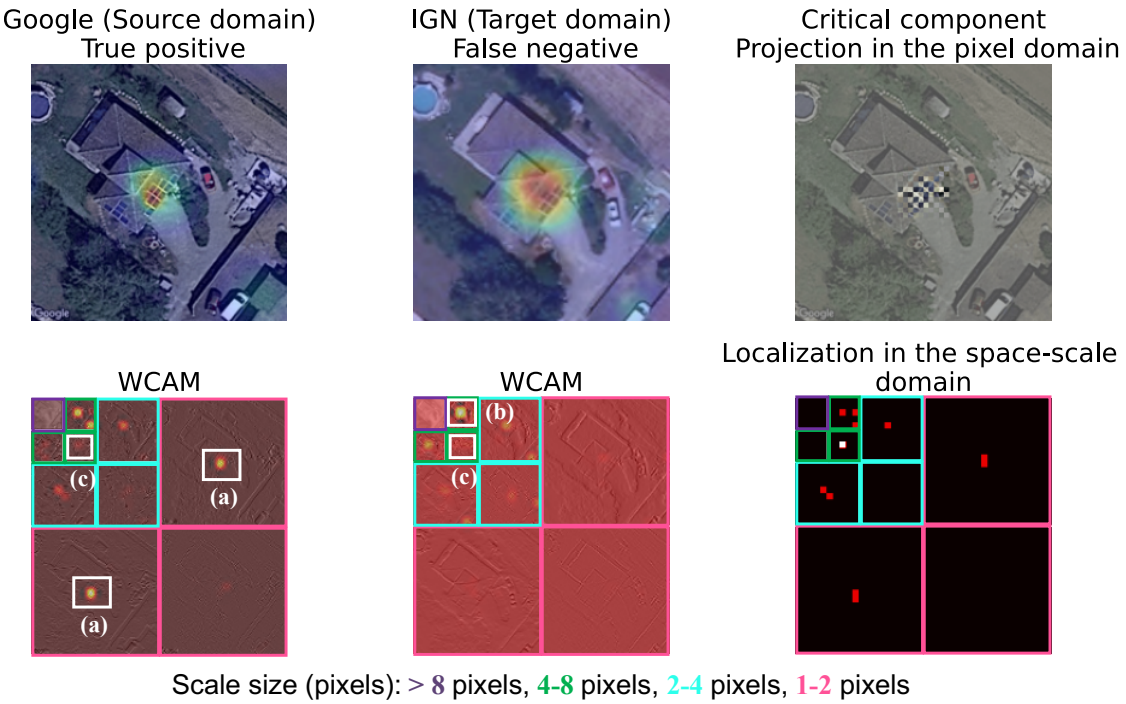


Figure 3.16 – Predictions on Google image (left, upper row) and IGN image (middle, upper row) and associated WCAMs (bottom row). The brighter the highlighted region for the prediction, the more important it is. The rightmost column plots the most important components of Google Images and the critical components. Adapted from Kasmi et al. (2023b).

Finding the missing components These results raise the question of understanding why this component disappeared. To address this question, we will introduce in the next section a model for acquisition conditions to show that these acquisition conditions are an instance of image corruptions (Hendrycks and Dietterich, 2019) that affect specific frequencies of the input image. If the model relies on components that are likely to be disrupted (such as the components **(a)** on Figure 3.16), then the it no longer recognizes the PV panel.

2.3 Acquisition conditions as image corruptions

2.3.1 A model for acquisition conditions

Where do varying acquisition conditions come from? Acquisition condition refers to the technical pipeline (imaging sensor, plane, postprocessing) to take the image but also to the meteorological conditions and the time of the day during which the picture was taken. Many factors could explain the variation in acquisition conditions. The following section looks for a simplified model to model acquisition conditions. This model will connect acquisition conditions and spatial frequencies or

scales. Therefore, using the WCAM, we can comprehensively assess the robustness of acquisition conditions using the scale-frequency characterization brought by the WCAM.

Acquisition conditions as a combination of Gaussian noise and blur The conversion of an observed scene into a digital image can be summarized in three main steps, as depicted in figure 3.17.

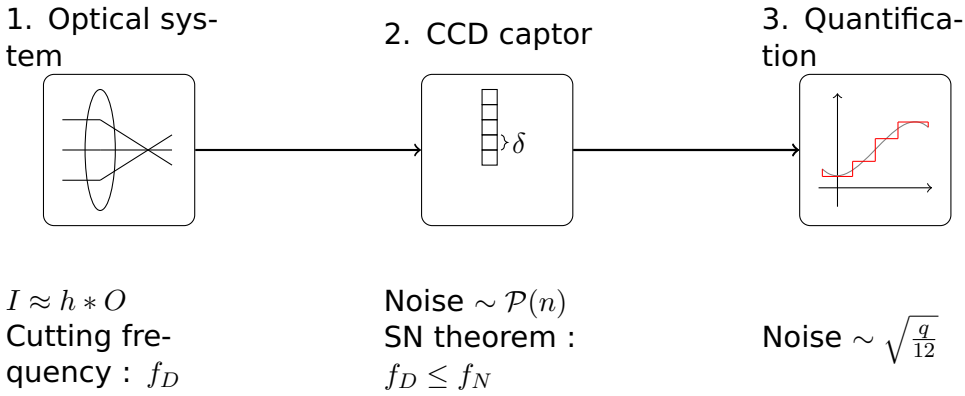


Figure 3.17 – Image acquisition process. SN: Shannon-Nyquist.

These steps are:

1. The acquisition of the signal by the optical lens can be approximated by a convolution between the object O and a Gaussian filter g. This Gaussian filter approximates the real Point Spread Function (PSF) accounting for the optical system and its defaults, the integration in time, and the space of the detectors in the focal plane. This PSF is the inverse Fourier transform of the autocorrelation of the entrance pupils of the optical system convoluted with the sinc of the detector.
2. The conversion of the analogical signal to a numerical signal by the complementary metal oxide semiconductor (CMOS) or charge-coupled device (CCD) sensor.
3. The subsampling of the signal depending on the sampling resolution of the CMOS or CCD image sensor.

Equation 3.8 summarizes the acquisition process as a convolution between the optical system h and the original image O,

$$I = h * O + \varepsilon, \tag{3.8}$$

where $*$ denotes the convolution operation and ε captures random noise while capturing the photons on the CCD captor. As the number n of photons is large, we can approximate this noise by a Gaussian noise. A key property of the optical lens

is the range of frequencies it is sensitive to. We denote f_D as the highest frequency the analogical sensor is sensible to.

A fundamental property of our model for acquisition conditions, described by Equation 3.8, is that the Gaussian filter is a low-pass filter and the Gaussian noise is a white noise with the same level of "energy" over the Fourier spectral range between 0 and $1/f_D$ ($f_D = 2f_N$ is the highest frequency in the analogical signal and f_N denotes the Nyquist frequency). Therefore, the contrast from the observed scene compared to the noise along this spectral range decreases exponentially, affecting the highest frequencies first.

2.3.2 Highlighting the sensitivity of the model to corrupted images

The noise and blur experiment We now set up an experiment to reproduce the model's sensitivity to acquisition conditions by mimicking the acquisition process of an image using a combination of Gaussian noise and blur. We denote h_{σ_b} a Gaussian filter parameterized by its standard deviation σ_b and ε_{σ_b} a Gaussian noise parameterized by its standard deviation σ_n. In order to simulate the acquisition process defined in Equation 3.9, we apply a Gaussian filter and add Gaussian noise to simulate the acquisition process of an object O into an image I:

$$I = h_{\sigma_b} * O + \varepsilon_{\sigma_n} \tag{3.9}$$

The parameters σ_b and σ_n capture the variability in the quality of the sensors. We consider the test images of BDAPPV as the ground truth images O. We generate altered datasets by adjusting the values of σ_b and σ_n. The values chosen are the following:

— $\sigma_b \in \{0\} \cup [1, 5]$. The size r of the convolution kernel is set as $r = \lceil 4\sigma_b \rceil + 1$. We denote $\Sigma_b \equiv \{0\} \cup [1, 5]$

— $\sigma_n \in [0, 0.017] \equiv \Sigma_n$

We chose these values to balance the loss of information and have enough granularity to visualize the gradual effect of the acquisition conditions. Figure 3.18 presents examples of the varying acquisition conditions modeled that way.

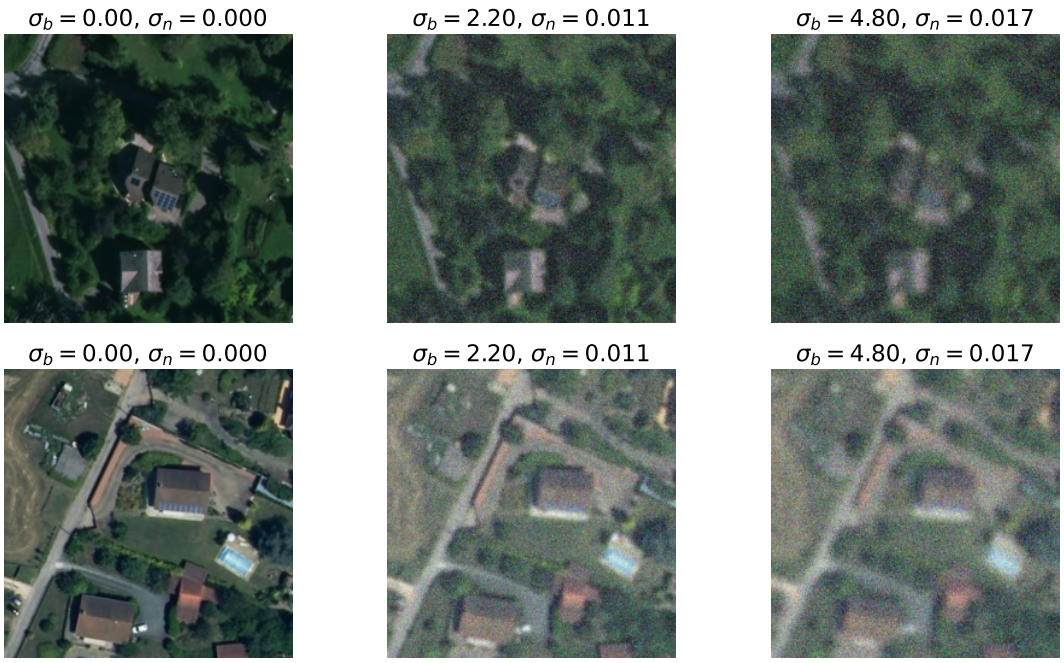


Figure 3.18 – Examples of varying acquisition conditions modeled after the Gaussian blur and noise model. We assume that the object O corresponds to the image without alteration (leftmost column).

We evaluate the model's[4] sensitivity at the global scale by computing the F1 score over the complete dataset. To do so, we generate altered test datasets for all combinations in $\Sigma_b \times \Sigma_n$. Figure 3.19 displays the results. We can see that the F1 score decreases until it eventually reaches 0. Interestingly, we can see regions in the (σ_b, σ_n) plane where the F1 score is constant. It means that the accuracy is equal despite different parameterizations of noise and blur. We interpret these results as a response to iso-quality levels. Different combinations of noise and blur produce a similar image quality, and the model responds to this image quality; this is in line with sizing rules for image quality that state that the quality follows iso-curves that depend on the noise and the value of the cutting frequency. In other words, a blurred but clean image will have the same quality as a neat but noisy image (for given levels of blur and noise).

4. By "model", I refer in this section to a standard classification model. In these experiments, I used a ResNet-50 (He et al., 2016), trained on BDAPPV.

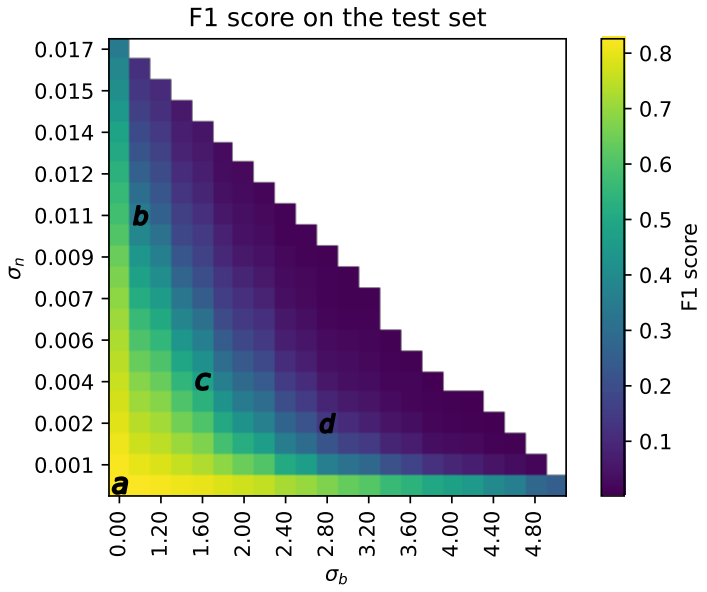


Figure 3.19 – Evolution of the F1 score on the test set depending on the noise and blur levels of the test set. Each letter marks a combination whose confusion matrix is unwrapped in Table 3.2.

Finally, on Table 3.2, we decompose the F1 score as done in section 2.2.1 to verify that a surge in the number of false negatives causes the drop in the F1 score. We can see that this is indeed the case. Whether perturbing the noise or the blur level leads to an increase in the number of false negatives, similar to the experiment of section 2.2.1. The rise in the number of false negatives drives the drop in accuracy.

Table 3.2 – **F1 score** and decomposition in terms of true positives, true negatives, false negatives, and false positives. Each line corresponds to a given level of corruption of the dataset, parameterized by the noise and blur level, σ_n and σ_b.

(σ_b, σ_n)	F1 score (↑)	TP	TN	FP	FN
a: (0.00, 0.000)	0.83	1302	1615	392	157
b: (1.00, 0.011)	0.30	270	1937	70	1189
c: (1.60, 0.004)	0.47	491	1882	125	968
d: (2.80, 0.002)	0.22	183	1963	44	1276

Results on a single image Now that we have shown the model's sensitivity to varying levels of noise and blur globally, we focus on single images. We consider one image and corrupt it for varying noise and blur strengths, following equation (3.9) and the associated values for σ_b and σ_n. For each combination $\sigma_b \times \sigma_n$ of blur and noise, we generate N corrupted images and pass these images to the model. We measure the accuracy as the ratio between correct predictions N_c and the total

number of corrupted variants of the image N. Initially, the image is correctly predicted. We report this accuracy in the lower left quadrant of Figure 3.20. We can see that for low values of σ_b and σ_n, the prediction is not sensitive to the addition of noise and blur (yellow area).

On the other hand, past a given threshold of noise and blur, the model can no longer recognize the PV panel, as the corruption of the image removed the necessary content. Interestingly, we can see a transition region where the model sensitivity to corruptions is the largest. In these cases, the model is correct for between 20% and 80% of the samples.

We analyze the model's decision process in these three cases (correct prediction, **a**, uncertain prediction, **b**, and wrong prediction, **c**. We can see how the corruption destroys the information in the finer scales and how the model changes the important factors to make the prediction. In the case **c**, the perturbation is so strong that the critical information can no longer be recognized and the model makes a wrong prediction. Case **b** is more puzzling: the model model makes a correct prediction in this case, but its accuracy is lower. Thanks to the WCAM, we can understand why the predictions are more uncertain. The model relies on high-frequency components (rather than low frequencies, as in **a**) because of the noise that has been added to the image. As the model relies on these frequencies, its prediction is more uncertain: we can see that the information at that scales (1-2 pixel scale) is not readable for a human interpreter.

Conclusion: acquisition condition as frequency perturbations Our small model illustrated that we can reproduce the effect of acquisition conditions by altering the frequency content of the image. The blur level affects the highest frequencies, while the noise level affects all the spectrum. Like in the experiment of section 2.2.1, altering the noise and blur levels leads to a decrease in the F1 score driven by a rise in false negatives.

In the next section, we will discuss how we can limit the sensitivity to acquisition conditions by lowering the reliance on the high-frequency components of the image and by incentivizing the model to rely on several scales instead of one.

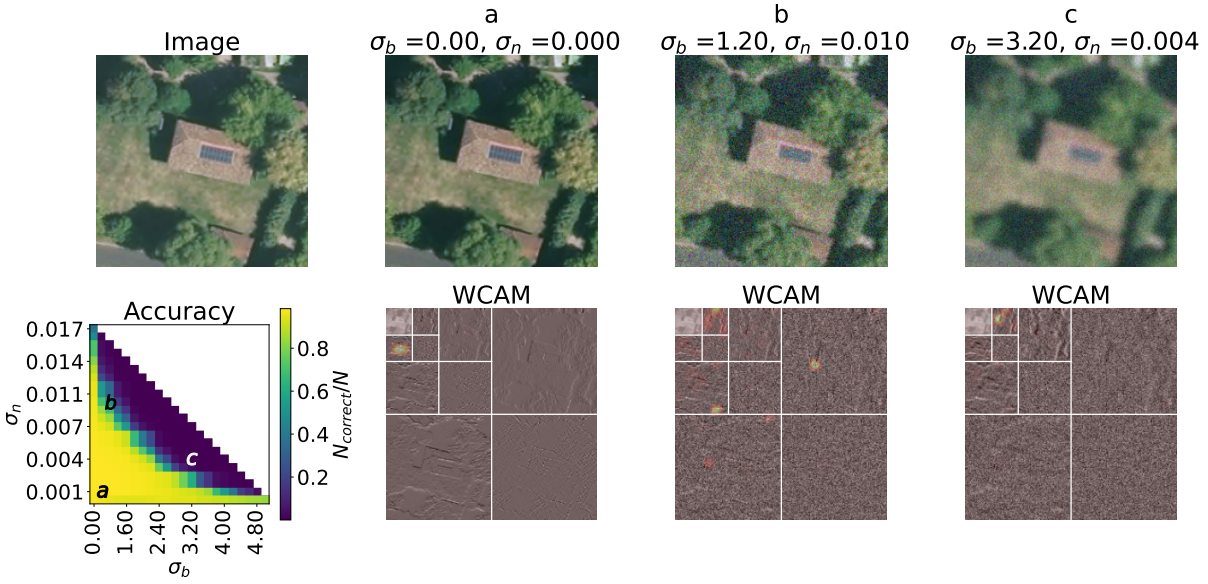


Figure 3.20 – Accuracy of a model's predictions under different levels of blur and noise and plot of some corrupted images and their associated WCAMs. We can see that for the same result at the macroscopic scale (a lower F1-score caused by a rise in the false negatives), the model behaves in two different ways at the microscopic level. If blurring increases, it tends to look for new components. If the noise increases, it tends to be disrupted by this noise and to focus on higher frequencies than if there were no noise.

3 Reliably improving the robustness of the model's decision process to acquisition conditions

In order to improve the robustness of our model to acquisition conditions, we review the literature and analyze whether it reliably improves the robustness to acquisition conditions. Finally, we introduce our data augmentation strategy to reliably improve the robustness of acquisition conditions, i.e., having a model that sees relevant and robust features on the input image.

3.1 Robustness to image corruptions: review of existing works

3.1.1 Acquisition conditions as a natural image corruption

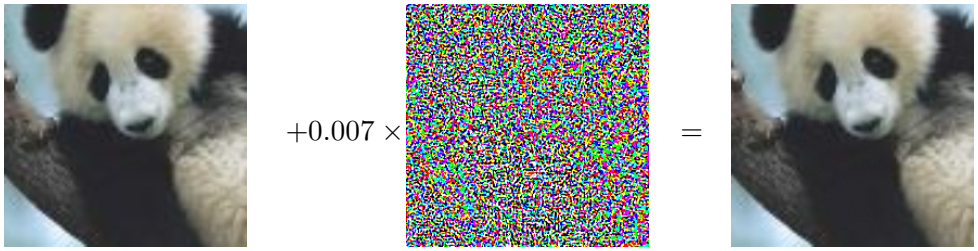


Figure 3.21 – A demonstration of fast adversarial example generation applied to GoogLeNet (Szegedy et al., 2016) on ImageNet. By adding an imperceptibly small vector whose elements are equal to the sign of the elements of the cost function gradient concerning the input, we can change GoogLeNet's classification of the image. Here our ϵ of 0.007 corresponds to the magnitude of the smallest bit of an 8-bit image encoding after GoogLeNet's conversion to real numbers. Taken from Goodfellow et al. (2015).

Traditionally, researchers in machine learning have been worried about the effects of *adversarial* perturbations (Goodfellow et al., 2015), i.e., input perturbations specifically targeted at fooling the model. These perturbations are generally invisible to the human observer. Figure 3.21 presents the well-known example of a panda predicted as a gibbon after a small perturbation is applied to the image.

However, adversarial perturbations are not the only kind of perturbations to which models can be sensitive. Hendrycks and Dietterich (2019) refer to this class of perturbations as *natural image corruptions* as they emerge naturally. These natural image corruptions cover a wide spectrum of perturbations, ranging from blur, which can occur when the observer takes the image, to noise or compression artifacts such as `.jpeg` compression. Hendrycks and Dietterich (2019) introduced a framework to benchmark the robustness of deep learning models to such corruptions. Figure 3.22 presents examples of natural image corruptions covered by Hendrycks and Dietterich's ImageNet-C(orruptions) dataset.

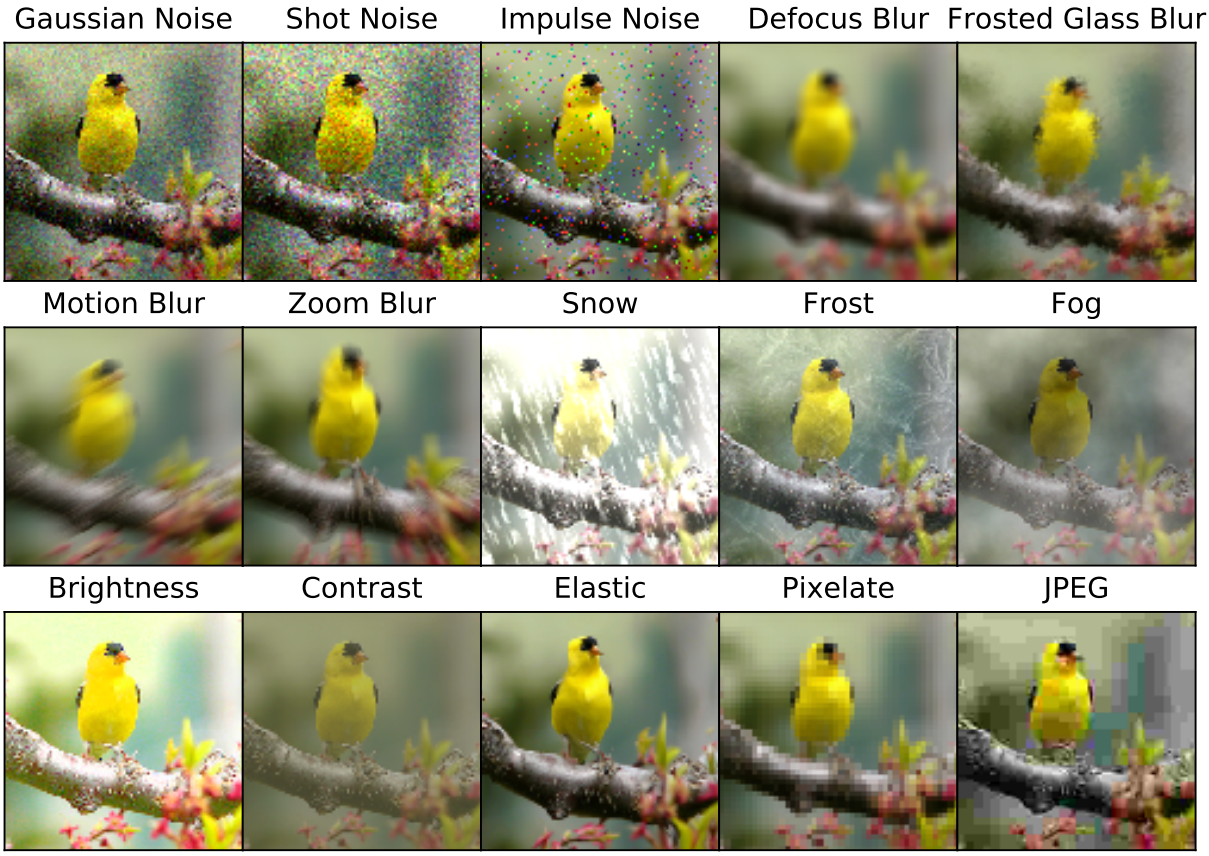


Figure 3.22 – Examples of images coming from the ImageNet-C dataset of Hendrycks and Dietterich (2019). Each corruption has different levels of severity. Source: Hendrycks and Dietterich (2019).

From Figure 3.22, we can view the acquisition conditions modeled after the model of section 2.3 as a combination of blur and Gaussian noise and, therefore, as a specific instance of natural image corruptions.

3.1.2 Robustness to natural image corruptions

Improving the robustness through data augmentations Various methods have been proposed to improve the robustness of CNNs to natural image corruption. Hendrycks et al. (2020) introduced AugMix, a data augmentation technique aiming at generating a high diversity of augmented images from an input sample. A set of operations (perturbations) op to be applied to the images and sampling weights ω are sampled. The resulting image x_{aug} is obtained through the composition $x_{aug} = \omega_1 op_1 \circ \cdots \circ \omega_n op_n(x)$ where x is the original image. Then, the augmented image is interpolated with the original image with a weight $m \in [0, 1]$ that is also randomly sampled. We have $x_{augmix} = mx + (1 - m)x_{aug}$. Similarly, Hendrycks et al. (2022) augment an input image with fractal patterns, and Sun et al. (2022a) perturb the Fourier spectrum of the input image. Alternatively, some methods focused on finding the best data augmentation strategy (or policy) for a given dataset. Cubuk

et al. (2019) determined the best augmentations strategy S as the outcome of a reinforcement learning problem: a controller predicts an augmentation policy from a search space. Then, the authors train a model, and the controller updates its sampling strategy S based on the train loss. The goal is that the controller generates better policies over time. The authors derive optimal augmentation strategies for various datasets, including ImageNet (Russakovsky et al., 2015), and show that the optimal policy for ImageNet generalizes well to other datasets.

Enforcing inductive biases Data augmentation strategies aim to learn an *inductive bias* (Mitchell, 1980) during training. Inductive biases correspond to assumptions that a model uses to make a prediction, and therefore generalize better, on unseen data. The translational invariance of CNNs is an example of inbuilt inductive bias. Several works have shown that another bias is the texture bias (Geirhos et al., 2019): models rely on texture rather than shape to make a prediction. As this bias can fool models, Geirhos et al. (2019) proposed to train models on Synthetised-ImageNet (SIN) to force models to rely on shapes rather than textures. We refer the reader to appendix C, section 2.3, for more details on these data augmentation techniques.

3.2 A benchmark of existing approaches

Overview We consider several methods registered in the ImageNet-C leaderboard. These methods include AugMix, AutoAugment, and RandAugment. Figure 3.23 presents examples of augmented images with these data augmentation techniques. We train a ResNet-50 model with these augmentations on Google images and evaluate it on the IGN test images. We report the F1 score and the number of true positives, false positives, true negatives, and false negatives.

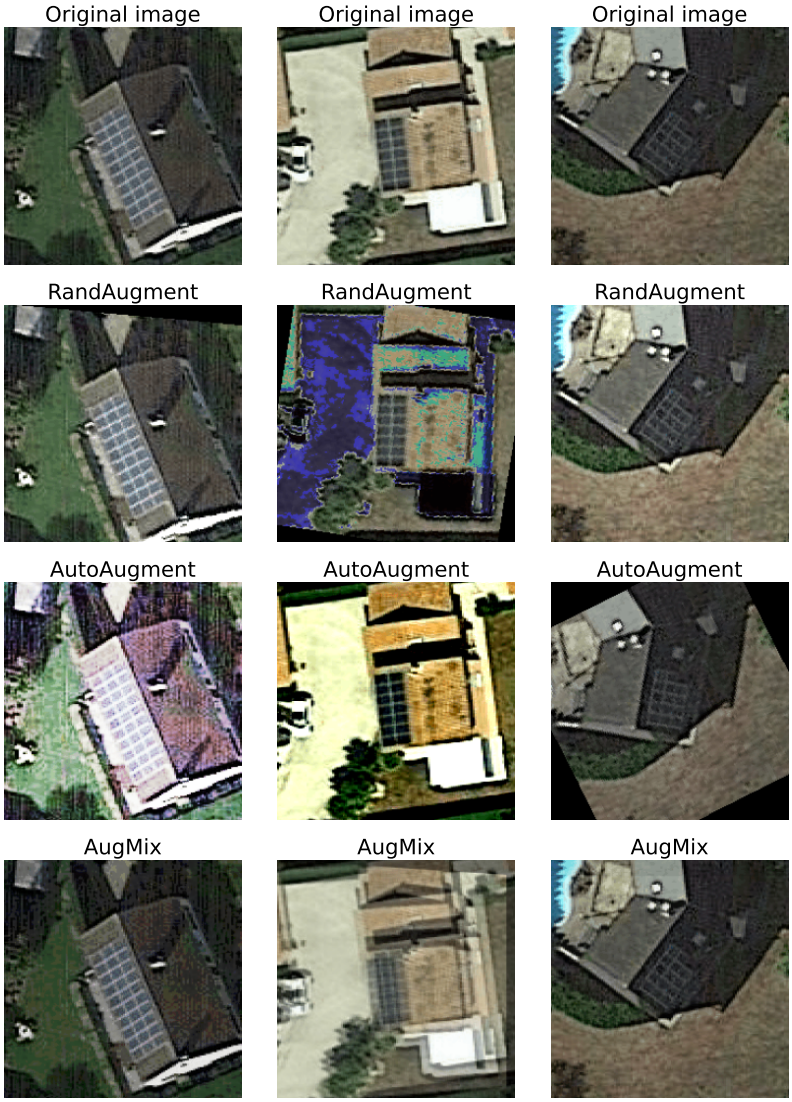


Figure 3.23 – Illustration of augmented images with the selected data augmentation techniques.

Results Table 3.3 presents the results. To evaluate the maximum attainable performance, we compare these results with the "Oracle," i.e., a model trained without augmentations on IGN images. We can see that these strategies yield modest improvement over the standard empirical risk minimization (ERM, Vapnik, 1999) method and barely reduce the number of false negatives. In Table C.3 in appendix C, section 2.4, we present the accuracy of these models on their source domain.

Table 3.3 – **F1 Score** and decomposition in true positives, true negatives, false positives, and false negatives for models trained on Google with different mitigation strategies. Evaluation of IGN images. The Oracle corresponds to a model trained on IGN images with standard augmentations. The best results are **bolded** and second best underlined.

	F1 Score (↑)	TP	TN	FP	FN
Oracle	0.88	1818	1992	428	83
ERM (Vapnik, 1999)	0.44	566	2321	99	1335
AutoAugment (Cubuk et al., 2019)	0.46	598	2318	102	1303
AugMix (Hendrycks et al., 2020)	<u>0.48</u>	624	2318	102	1277
RandAugment (Cubuk et al., 2020)	**0.51**	707	2280	140	1194

3.3 A novel data augmentation technique for improving the robustness to acquisition conditions

3.3.1 Motivation and approach

Motivation Implementing existing data augmentation techniques resulted in very modest improvements in the model's robustness to acquisition conditions. Recalling the key results from our study of section 2.3.2, we would like our model (1) not to rely on components that are likely to be disrupted by the change of image provider and (2) incentivize the model to rely on different scales rather than a single scale to make a prediction. To this end, we blur the image so that high-frequency components are discarded for addressing (1). To address (2), we randomly perturb the wavelet transform of the image so that the model learns that scales can be altered. Indeed, the components most likely to be perturbed by a shift in provider are located in the high-frequencies of the image, as highlighted in section 2.3.2.

Overview of our data augmentation method We call our data augmentation method the Blurring and Wavelet Perturbation (WP). This method consists in (1) blurring the image to remove the high-frequency components. Then, we randomly perturb the wavelet coefficients of the image to force the model to rely on different scales rather than a single scale. To perturb the wavelet transform, we compute the wavelet coefficients of the image and randomly set some of them to 0. We compute the wavelet transform with five levels of details and perturb the coefficients across all scales (including the approximation coefficients). After empirical investigation, we set the share of coefficients to 0 at 20% (independently in each channel) to balance the perturbation across scales while keeping the image's semantic content. As for the blur, we choose a blur value so that the shape of the PV panel remains visible to a human observer. It corresponds to a blurring value $\sigma = 2.$ in the `ImageFilter.GaussianBlur` method of the Python Imaging Library (PIL). We apply the same blur in both x and y directions. Figure 3.24 presents augmented

images with our Blurring and Wavelet Perturbation (WP) technique.

Baselines: Blur and Noise and Blur In addition to our strategy, we test two baselines: the effect of a simple Gaussian blur on the image to discard the high-frequency components and a random composition of noise and blur to mimic the perturbation as in the model introduced in section 2.3.2. These augmentations are intended as baselines to see whether the perturbation of the wavelet transform improves over more straightforward approaches. Figure 3.24 presents augmented images with these data augmentation techniques. In appendix C, section 2.4, we present additional training results for these data augmentation techniques.

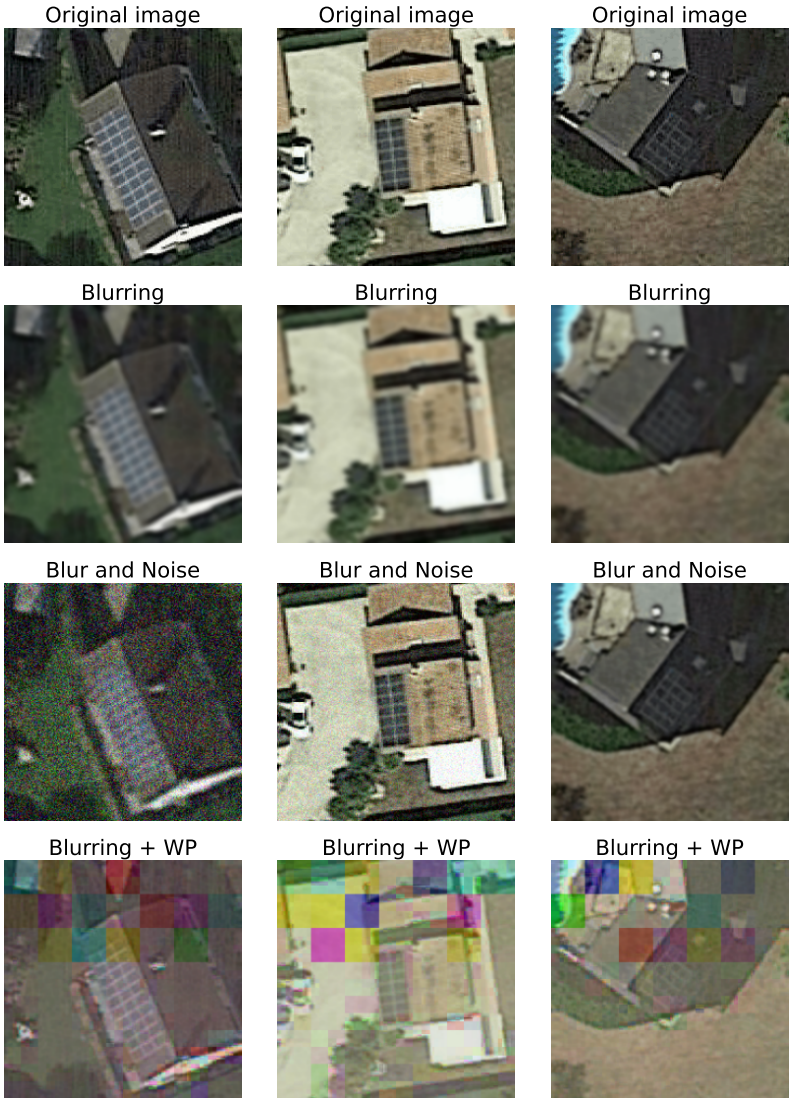


Figure 3.24 – Illustration of augmented images with our data augmentation techniques. The colored pixels that appear with the Blurring + WP augmentation are a consequence of the fact that we hide some information in channels and not others.

3.3.2 Results: increasing the robustness to acquisition conditions

Blurring and wavelet perturbation improve accuracy Table 3.4 reports the results of our data augmentation techniques and compares them with existing methods. We can see that augmentations that explicitly discard small scales (high frequencies) information perform the best[5]. However, the blurring method sacrifices the recall (which drops to 0.6) to improve the F1 score. On Table 3.4, this can be seen by the increase in false positives. Therefore, this method is unreliable for improving the robustness to acquisition conditions.

On the other hand, adding wavelet perturbation (WP) improves the accuracy of the classification model without sacrificing the precision or the recall. While the drop in accuracy is still sizeable compared to the Oracle, the gain is consistent compared to other data augmentation techniques. Compared to RandAugment, the best-benchmarked method, our Blurring + WP is closer to the targets regarding true positives and true negatives and makes lower false negatives. Work is still needed to close the gap with the Oracle. However, this experiment shows that it is possible to consistently and reliably improve the robustness of acquisition conditions using a data augmentation technique, which does not leverage any information on the IGN dataset.

Table 3.4 – **F1 Score** and decomposition in true positives, true negatives, false positives, and false negatives for models trained on Google with different mitigation strategies. Evaluation on IGN images. The Oracle corresponds to a model trained on IGN images with standard augmentations. The best results are **bolded** and second best underlined.

	F1 Score (↑)	TP	TN	FP	FN
Oracle	0.88	1818	1992	428	83
ERM (Vapnik, 1999)	0.44	566	2321	99	1335
AutoAugment (Cubuk et al., 2019)	0.46	598	2318	102	1303
AugMix (Hendrycks et al., 2020)	0.48	624	2318	102	1277
RandAugment (Cubuk et al., 2020)	0.51	707	2280	140	1194
Noise and blur	0.48	636	2287	133	1265
Blurring	**0.74**	1855	1196	1224	46
Blurring + WP	0.58	896	2114	306	1005

Relying on consistent scales Figure 3.25 compares the scales on which the best-performing methods rely. We want our models to rely on the largest scales (i.e.,

5. This result is further underlined by the fact that the Oracle – trained on IGN – evaluated on Google performs better than the model trained on Google images and tested on IGN. As IGN images have less information in the highest frequencies, lowering the reliance on the highest frequencies is essential to guarantee a good generalization to new acquisition conditions. We refer the reader to Table C.3 in appendix C, section 2.4 for more information.

lowest frequencies) to entail robustness (Zhang et al., 2023), in our case against image quality alterations.

We can see that the Blurring + WP method relies on more reliable factors than the other methods: it mainly relies on low-frequency components centered on the PV panel. The blurring in this case is very dispersed, and the RandAugment has a behavior that is qualitatively similar to the ERM. More generally, the WCAM lets us compare methods that perform quantitatively similarly.

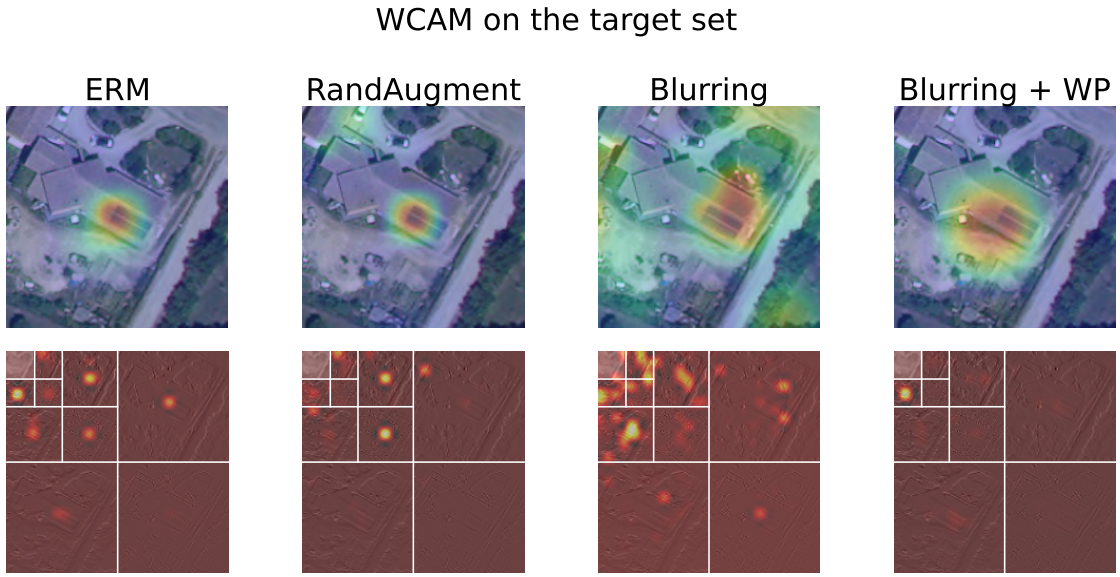


Figure 3.25 – WCAMs on IGN of models trained on Google with different augmentation techniques.

Conclusion of the chapter

In this chapter, we introduced a new feature attribution method, the WCAM, to identify what regions of the space-scale domain contribute the most to the model's decision. The regions of the space-scale domain correspond to the image's structural components, which can be simultaneously interpreted semantically and can be related to frequency ranges.

The WCAM enables us to assess the *relevance* of the model's decision process: we witnessed the fact that it relied on critical components related to a visual feature of the PV panel (e.g., the meshed patterns of the PV panel). However, when we remove this critical component, we can flip the model's prediction.

What causes the disappearance of the critical component (and thus the sensitivity to distribution shifts highlighted by the literature)? We set up an experiment where we disentangle the familiar sources of distribution shifts: geographical

variability, varying ground sampling distance and varying acquisition conditions. Results showed that most of the variability comes from the varying acquisition conditions, which lead to a surge in the false negatives. We interpreted this result as the model missing its critical component due to the change in the image provider.

How can we effectively improve the robustness of the decision process to varying acquisition conditions? We modeled acquisition conditions as a combination of Gaussian noise and blur, showing that they can be related to image corruption. We reviewed works that proposed methods to improve the robustness to acquisition conditions and introduced a new method based on the perturbation of the wavelet spectrum of the image and the removal of the high-frequency components through blurring, effectively improving the robustness to acquisition conditions.

The WCAM thus enables us to assess the relevance and robustness of the model's decision process. Together with the DTA, it closes the definition of the reliability introduced in chapter 1. In the next chapter, we will introduce DeepPVMapper. This algorithm aims for reliable and scalable mapping of rooftop PV installations. We will use the DTA of chapter 2 and the WCAM and the data augmentation that we introduced in this chapter to enhance the robustness to varying acquisition conditions of current mapping algorithms.

Chapter 4

Constructing a reliable and scalable algorithm for mapping rooftop PV installations in France

Summary

This chapter focuses on the third and final pillar of reliability: enhancing the model's robustness to acquisition conditions. We introduce DeepPVMapper, advancing the state-of-the-art in three main aspects: reviewing classification and segmentation models, optimizing the pipeline to minimize false detections, and standardizing the extraction of the characteristics of rooftop PV systems using our new library, `PyPVRoof`. Evaluation involves metrics reflecting real-life conditions for a comprehensive performance assessment. DeepPVMapper is 16% more accurate and 31% faster than a DeepSolar-based architecture. We deploy DeepPVMapper on the problematic cases from chapter 2, section 3 to demonstrate its effectiveness in addressing issues encountered with our replication of DeepSolar. We also discuss the broader applicability of our method. Implementing the DTA elsewhere in Europe is possible with databases equivalent to the RNI. Besides, the WCAM facilitates the audit of the model's decision process, guiding training methods for improved accuracy on unseen images.

In the two last chapters, we introduced two methods to evaluate the reliability of any model's data and behavior. In this chapter, we will leverage our framework to improve current algorithms to make them more reliable and scalable for mapping rooftop PV installations in France.

1 Identifying the limitations of the current approaches for PV systems mapping

We can distinguish two components from our definition of reliability: extrinsic and intrinsic. Extrinsic reliability is external to the model, while intrinsic reliability is a property of the model. This section reviews the current state-of-the-art and sees where we can improve the intrinsic reliability.

1.1 Where should we focus on ?

1.1.1 Extrinsic and intrinsic reliability

Definition Our definition of reliability rests on three pillars: the ability to monitor the model's outputs, the ability to inspect and understand the decision process of the model, and ensuring that the decision process is robust regarding a given perturbation (in our case, the heterogeneous acquisition conditions). In chapter 2, we introduced the DTA to enable the user to monitor the model's output. In chapter 3, we introduced the WCAM to assess the reliability of the decision process and improve the robustness to varying acquisition conditions.

The ability to monitor the model's outputs or evaluate the relevance of its decision process is model agnostic or *extrinsic* to the model. On the other hand, a model designed properly can be more robust than another. Therefore, the robustness is *intrinsic* to the model.

This chapter is devoted to improving the intrinsic reliability of current mapping algorithms. To this end, we identify in section 1.1.2 of this chapter the current limitations of the mapping algorithm. Our starting point is the widespread DeepSolar-based architecture, which we introduced in chapter 2, section 3.1.1.

A review of DeepSolar The original DeepSolar architecture (Yu et al., 2018) was a semi-supervised approach. The main model was a classification model trained on approximately 473,000 labeled images, including 20,300 (about 4.3%) positives. The motivation for this choice is that classification models require image-level labels, which are less labor-intensive to gather than pixel-level labels, necessary to train segmentation models. The classification model, an Inception-v3 (Szegedy et al., 2016) model, was fine-tuned on this dataset.

The goal of DeepSolar is to estimate the size of the PV systems. The classification model only returns an image-level label (1 if the image contains a PV panel and 0 otherwise). Therefore, it is necessary to convert this prediction into a segmentation mask (i.e., a binary image where pixels equal to 1 indicate that the pixel corresponds to a PV panel and pixels equal to 0 indicate that there is no PV panel). Traditionally, a segmentation model would be required for such a task. Instead, the authors introduced a semi-supervised approach, i.e., a learning method that predicts segmentation masks using only image-level labels. Figure 4.1 summarizes the flowchart of DeepSolar for detecting and segmenting PV panels on orthoimagery.

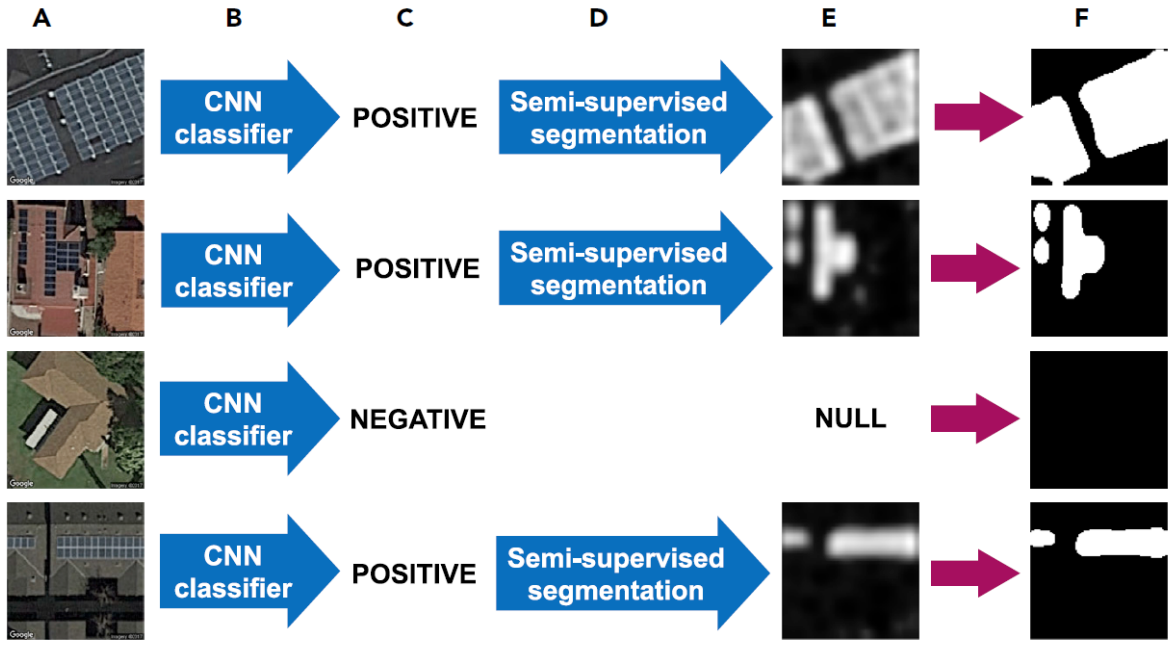


Figure 1. Schematic of DeepSolar Image Classification and Segmentation Framework
(A) Input satellite images are obtained from Google Static Maps.
(B) Convolutional neural network (CNN) classifier is applied.
(C) Classification results are used to identify images containing systems.
(D) Segmentation layers are executed on positive images and are trained with image-level labels rather than actual outlines of the solar panel, so it is "semi-supervised."
(E) Activation maps generated by segmentation layers where whiter pixels indicate higher likelihood of solar panel visual patterns.
(F) Segmentation is obtained applying a threshold to the activation map and finally both panel size and system counts can be obtained.

Figure 4.1 – Original DeepSolar pipeline. Source: Yu et al. (2018).

The semi-supervised approach consists in training a small model to predict the localization and shape of the PV panels using the classification model's feature maps and the label. The feature maps (i.e., the intermediate layer in the convolutional part of the model) contain features learned by the model, from the most elementary (edges, shape) to the most general (overall shape of the object). Gen-

eral features capture the overall location of the object at the expense of the accuracy. On the other hand, elementary features are much more detailed but lack global information. Combining them enables the construction of well-defined, sharp segmentation masks encompassing all the PV panels, as depicted on the rightmost column of Figure 4.2. The authors call this combination of high-level and low-level features a "greedy" feature extraction.

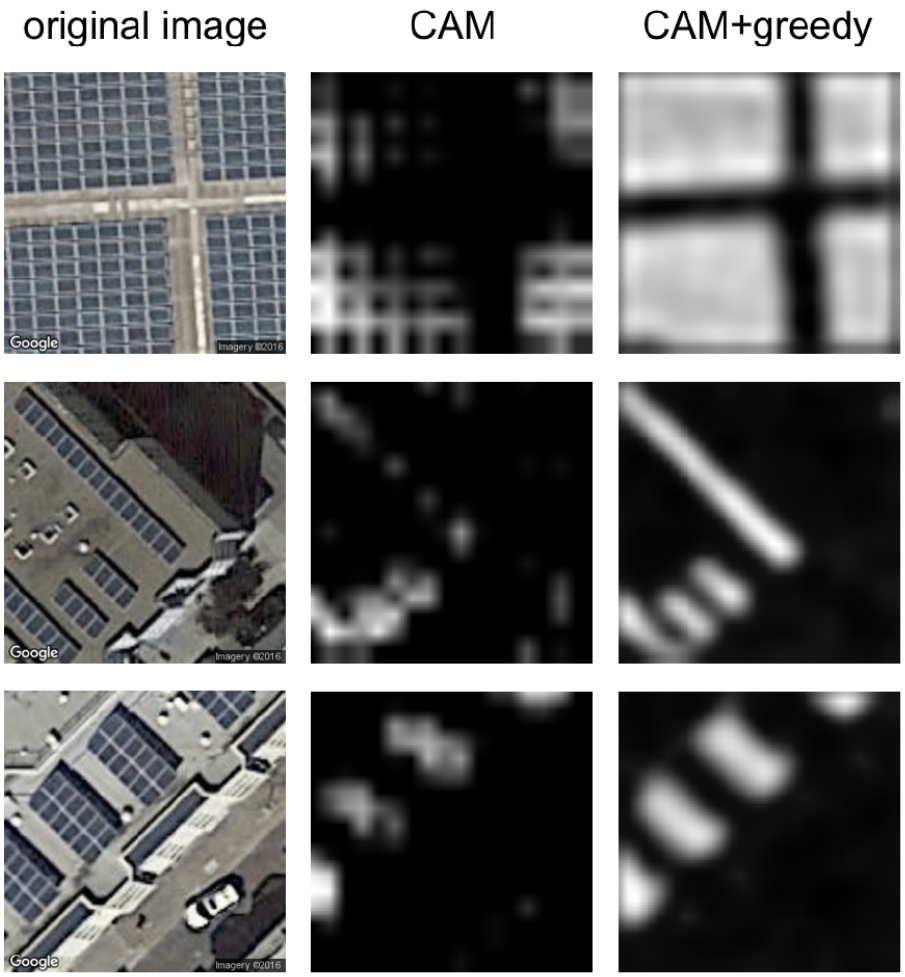


Figure 4.2 – Semi-supervised approach of DeepSolar. The left column contains original images. The middle column contains the original images' Class Activation Maps (CAMs) without greedy layer-wise training. The right column is the CAMs of the original images with greedy layer-wise training. Taken from Yu et al. (2018).

As the training database of DeepSolar (Awala, 2020) featured ground truth segmentation masks, subsequent works, notably Rausch et al. (2020) and Mayer et al. (2022), abandoned the self-supervised approach to replace the second step with an actual segmentation model, Deeplab-v3 (Chen et al., 2018) model. Therefore, we refer to these approaches (as well as our implementation in chapter 2, section 3.1.1) as "DeepSolar-based" or "DeepSolar" approaches.

1.1.2 Improving the intrinsic reliability of DeepSolar

Reviewing the two-step approach Even if current mapping algorithms no longer rely on a semi-supervised approach, the two-step approach is widely spread in the field (Parhar et al., 2021). The main reason is that one expects to see PV panels on a tiny subset of the images, so classifying first enables one to filter the database and run image segmentation, which is computationally more expensive, only when necessary.

Even if the computational cost is higher, a line of work (Frimane et al., 2023; Zhang et al., 2020, 2021a; Huang et al., 2018; Malof et al., 2019) directly performs image segmentation. To the best of our knowledge, the benefits of the two-step approach (i.e., classification and segmentation) have not been evaluated against the one-step (i.e., only segmentation). The main reason is that segmentation accuracy is evaluated with the Intersection-over-Union and the classification accuracy with the F1 score[1]. Therefore, we miss a more general evaluation framework.

Reviewing the recent advances in computer vision Existing works (Mayer et al., 2022, 2020; Rausch et al., 2020) used an Inception-V3 model (Szegedy et al., 2016). However in the last couple of years, several breakthroughs were made in computer vision with the advent of vision transformers (ViT, (Dosovitskiy et al., 2021)) and hybrid convolutional-vision transformer models such as ConvNext (Liu et al., 2022), DeiT (data efficient transformer, Touvron et al., 2021) or ConvMixer (Trockman and Kolter, 2023). So far, two works have shown modest gains of the all-transformer architecture (Luzi et al., 2023) for classification and gains in terms of generalizability for transformer-based image segmentation (Guo et al., 2024).

Characteristics extraction Over the years, the scope of remote PV mapping in general, and DeepSolar-based approaches in particular, gradually expanded. The goal was to estimate the surface area of PV panels, the installed capacity, and other statistics, such as the tilt and azimuth angles. To this end, various approaches have been proposed, depending on the additional GIS data available in the different use cases. This led to a great diversity of statistics and methods to extract these characteristics. This diversity makes the comparisons between the approaches difficult. Also, it limits the reproducibility and transferability of existing approaches to new settings, as the architectures and data requirements vary from one study to the other. To overcome these difficulties, namely the lack of standardization and the heterogeneous data requirements, we present a standardized approach for extracting PV characteristics in section 1.2 of this chapter.

1. We define these accuracy metrics in section 2.1.1 of the present chapter.

1.2 Standardized characteristics extraction: the `PyPVRoof` library

1.2.1 Motivation and overview

Extracting the PV system's characteristics is necessary to derive spatial statistics on the rooftop PV fleet. Depending on the study, the targeted characteristics range from the surface of the installations to a more comprehensive characterization that includes the installed capacity and the tilt and azimuth angles. The approaches to extracting the characteristics vary from one study to the other as the needs and the available data differ. This heterogeneity makes it complex to compare the approaches, derive consistent and standardized characteristics, and know what best practices should be favored. To address this limitation, we introduced `PyPVRoof` (Trémenbert et al., 2023), a Python library designed to extract rooftop PV system characteristics from geolocalized PV polygons.

`PyPVRoof` accommodates additional data sources, such as preexisting registries (i.e., auxiliary data) or digital surface models (DSM), depending on their availability for the user. The list of characteristics that we extract is the following:

— Localization (latitude and longitude)

— Tilt angle (in degrees)

— Azimuth angle (in degrees, relative to North)

— Surface (in m^2). Estimating the surface requires knowing the tilt, as only the *projected* surface is derived from the input polygon.

— Installed capacity (in kW_p). The surface is needed to estimate the installed capacity as its first-order approximation is the surface multiplied by an efficiency factor (So et al., 2017).

Figure 4.3 summarizes the workflow of `PyPVRoof`. `PyPVRoof` combines methods for characteristics extraction based on a review of existing works in the field. These methods were chosen based on accuracy, simplicity, and efficiency. In particular, we retained the most simple between two equally performing methods (e.g., the lookup table over the random forest). Finally, we restricted ourselves to methods that require as few additional inputs as possible. These methods reflect the current state-of-the-art of PV characteristics extraction.

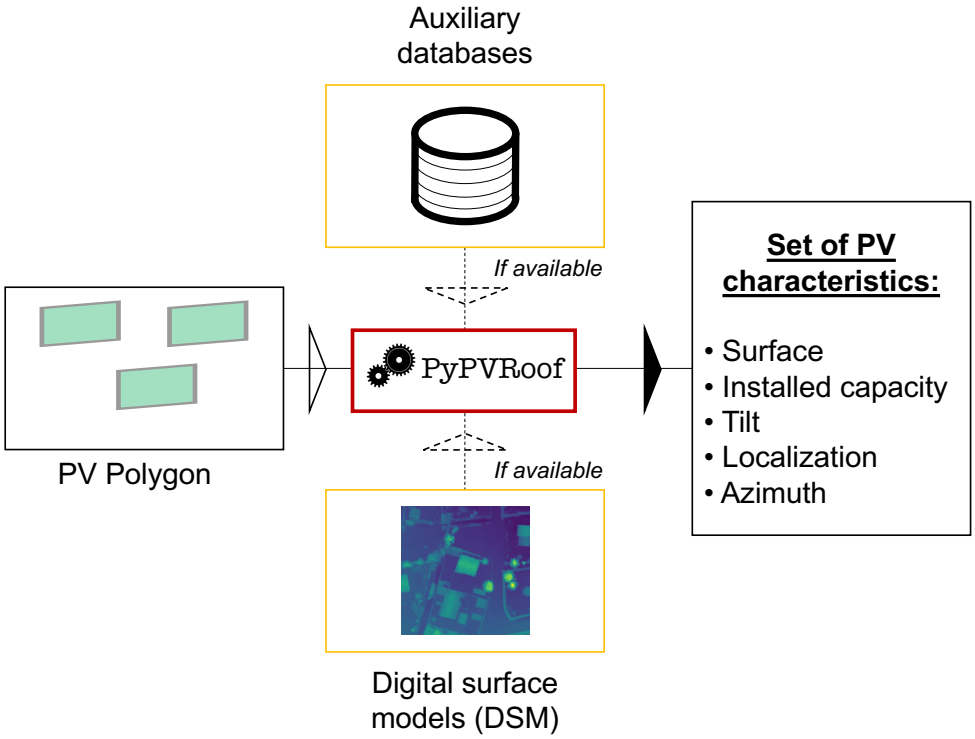


Figure 4.3 – Flowchart of the proposed method to extract installations' characteristics. The installed capacity depends on the surface of the PV system, as the installed capacity is equal to the surface area multiplied by the efficiency of the PV modules. Source: Trémenbert et al. (2023).

`PyPVRoof` accommodates the most common use cases when mapping PV installations. We refer the reader to Trémenbert et al. (2023) for a thorough description of these use cases. For DeepPVMapper, we are in a case where we want to extract many PV system characteristics (installed capacity, tilt, and azimuth angles) but do not have access to 3D data over the whole of France. However, as we have an external registry with PV characteristics (the BDPV database), we can use it to calibrate methods that enable us to derive the targeted characteristics. Figure 4.4 presents the flowchart and the associated methods to extract PV characteristics if the user has only access to auxiliary data. For our case, `PyPVRoof` leverages BDPV data to calibrate the panel efficiency module coefficient to correlate the installation's surface with an installed capacity. A lookup table (LUT) is also computed from this input data for the tilt angle estimation (see chapter 2, section 3.1.1 for more details on the computation of the lookup table). Finally, we apply a bounding box algorithm to estimate the azimuth angle.

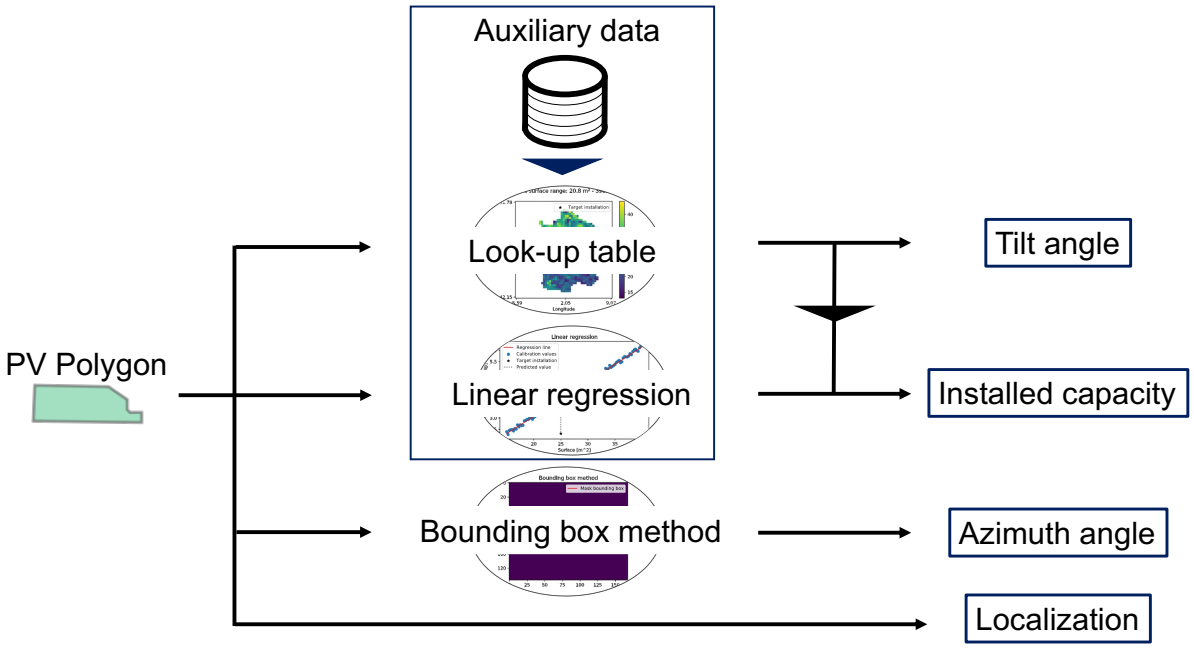


Figure 4.4 – `PyPVRoof` flowchart when only auxiliary data is available. Adapted from Trémenbert et al. (2023).

1.2.2 Identifying the best methods for extracting PV systems' characteristics

Table 4.1 – Overview of the methods considered for building `PyPVRoof`. The column "data requirements" indicates the additional requirements besides the geolocalized polygon.

Method name	Data requirements	Target characteristic	Original work
Direct computation	PV polygon	Surface Tilt Installed capacity	Yu et al. (2018) Walch et al. (2020) Rausch et al. (2020)
Hough algorithm	RGB Image	Azimuth	Edun et al. (2021)
Linear regression	LiDAR data Labelled data	Tilt, azimuth Installed capacity	Rausch et al. (2020) So et al. (2017) Malof et al. (2019)
Theil-Sen regression	LiDAR data	Tilt, azimuth	
Random forest	Labelled data	Tilt, installed capacity	

Evaluated methods To design `PyPVRoof`, we reviewed the existing works' methods and use cases. We then sorted these methods depending on their data requirements and the characteristics that they extracted. We replicated these methods

and evaluated them on a benchmark on our dataset BDAPPV. Table 4.1 summarizes the methods evaluated for designing `PyPVRoof`. For a more detailed presentation of these methods, we refer the reader to the appendix C, section 3.1 or to Trémenbert et al. (2023).

Benchmarking approach We evaluate each method based on the execution time and on the following metrics:

— Mean error (bias) (ME): $\frac{1}{n}\sum_{i=1}^{n}\hat{x}_i - x_i$

— Mean absolute error (MAE): $\frac{1}{n}\sum_{i=1}^{n}|\hat{x}_i - x_i|$

— Root Mean Square Error (RMSE): $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{x}_i - x_i)^2}$

— Mean absolute percentile error (MAPE): $\frac{1}{n}\sum_{i=1}^{n}\frac{|\hat{x}_i - x_i|}{x_i}$

1.2.3 Results of the benchmark of the methods

Surface estimation The direct estimation method achieves a mean error of 3.62m², a mean absolute error of 5.01m², and a RMSE of 6.86%. The runtime is instantaneous. We observe minor differences between annotated and predicted masks, which assess the overall quality of the predicted masks. However, we give particular attention to the positive bias between masks and the surface reported in the characteristics file of BDAPPV, highlighting a tendency to overestimate the referenced surface area. Such bias is irrelevant since BDAPPV's surface area values cannot be assessed; for instance, an overrepresentation of installations of 20m² could result from a systematic roundup in the database.

Tilt angle estimation Table 4.2 presents the results. For tilt estimation, it turned out that the LUT was a surprisingly strong baseline over the other methods: the random forest yielded only minor improvements, but the runtime was an order of magnitude larger. Although not significant for a single installation, such a difference in runtime is significant when scaling the method to thousands of PV polygons. As for the methods that require surface models, we can see that their accuracy relies on the quality of the input data. We tested the Theil-Sen method on photogrammetry-based surface models and LiDAR surface models. We can see a noticeable improvement when shifting from photogrammetry to LiDAR. However, the LUT method's superiority over the Theil-Sen method with LiDAR data is questionable. Indeed, we would expect the LiDAR method to be the best-performing method. A possible reason for that might be that the LUT, being trained on BDPV,

reproduces the biases of the testing dataset. In appendix C, section 4.2, we discuss how we can control the quality of the tilt angles reported in BDPV using LiDAR data and ground truth measurements.

Table 4.2 – Performance metrics for the estimation of the tilt angle. The two lines for the Theil-Sen method report the accuracy results whether photogrammetry DSM or LiDAR DSM are passed as inputs. The best results are **bolded** and second best underlined.

Method	ME [°]	MAE [°]	RMSE [°]	Runtime [sec]
Random Forest	**6e-4**	**5.34**	**7.03**	0.28
Lookup table (LUT)	-2.40	7.68	10.29	**6e-6**
Theil-Sen (Photogrammetry)	3.99	14.10	17.50	0.09
Theil-Sen (LiDAR)	2.06	11.08	14.69	0.09
Hough with DSM	2.90	13.45	16.62	2.47

Azimuth angle estimation For azimuth estimation, we replicated the method of Edun et al. (2021) using the Hough algorithm. They report MAEs ranging from 15.62 to 30.53 degrees depending on the type of panel considered, the most significant errors associated with rooftop panels, and the smallest with ground panels. Our replication is, therefore, in line with theirs, as we report an MAE of 22.70 degrees. Surprisingly, we see very few improvements brought by the Hough method with surface models. On the other end, the bounding-box method, which solely relies on the PV polygon, is a very accurate approach, even outperforming the Theil-Sen algorithm (in the case of photogrammetry DSM). The Theil-Sen method with LiDAR data is the most accurate, as shown in Table 4.3.

Table 4.3 – Performance metrics for the estimation of the azimuth angle. The two lines for the Theil-Sen method report the accuracy results whether photogrammetry DSM or LiDAR DSM are passed as inputs. The best results are **bolded** and second best underlined.

Method	ME [°]	MAE [°]	RMSE [°]	Runtime [sec]
Hough (Edun et al., 2021)	2.10	22.70	40.26	0.04
Hough with DSM	-0.73	23.78	43.66	2.50
Theil-Sen (Photogrammetry)	-6.65	15.54	35.64	0.09
Theil-Sen (LiDAR)	**-0.08**	**3.10**	**4.38**	0.09
Bounding-box	-1.39	12.90	32.76	**0.02**

Installed capacity estimation Estimating the installed capacity requires the tilt angle and the module efficiency. Indeed, we use the real surface rather than the

projected surface as input to estimate the installed capacity. Rausch et al. (2020) reported a nine percentage point increase in the median absolute percentage error (MedAPE)[2] for estimating the installed capacity when considering the tilt angle. We compared variants of the random forest estimator, with θ coming from different methods, to see how potential errors propagated. As it can be seen from Table 4.4, all random forests perform equally. We can also see that these methods are only slightly better than the clustered linear regression, which improves over So et al. (2017). On a different dataset, the authors reported mean squared errors ranging from 1.64 to 1.69, corresponding to a RMSE of 1.28 to 1.30 kW_p.

Table 4.4 – Performance metrics for the estimation of the installed capacity. Column θ indicates the method used to derive the tilt necessary to compute the estimated surface S_{est}, taken as input to estimate the installed capacity. "RF" indicates random forest, and "TS" Theil-Sen. The best results are **bolded** and second best underlined.

Method	θ	**ME** [kW_p]	**MAE** [kW_p]	**RMSE** [kW_p]	**MAPE** [%]	**Runtime** [sec]
Random forest (with S_{est})	RF	0.022	**0.328**	0.750	**9.37**	1.1e-1
Random forest (with S_{proj} and θ)	TS	0.061	0.393	0.848	11.48	4.4e-4
Random forest (with S_{proj} and θ)	RF	0.079	0.379	0.921	10.66	4.3e-2
Clustered linear regression	RF	**-0.015**	0.376	**0.687**	11.57	**7.2e-7**

1.2.4 Choice of the methods for DeepPVMapper

Summarized results Table 4.5 summarizes the accuracy results of these methods. We can see a significant improvement gain when using LiDAR data, especially on the azimuth angle. We focused on rasters for a fair comparison with the photogrammetry DSM, so the improvement could be more significant if we implemented Theil-Sen directly on the raw LiDAR data (i.e., the point cloud). Besides, a surprising result is that LiDAR data is better for azimuth angle than tilt estimation. An explanation for this is that azimuth estimation is less sensitive to noisy data points in the (z) elevation direction than tilt estimation. We highly recommend using the Theil-Sen method if the DSM is precise enough (e.g., from LiDAR data). Otherwise, the bounding-box method is competitive at a much lower computational cost.

2. Rausch et al. (2020) define the MedAPE as

$$MedAPE = Median\left(\frac{|y_1 - \hat{y_1}|}{y_1}, \ldots, \frac{|y_n - \hat{y_1}|}{y_n}\right),$$

where y_i denotes the true value and $\hat{y_i}$ the estimated value.

Table 4.5 – Accuracy results of `PyPVRoof`'s methods for PV panels characteristics extraction.

Characteristic	Method	Accuracy (RMSE) [unit]
Surface	Direct computation	6.9 [m²]
Tilt	LUT	10.29 [°]
	Theil-Sen (LiDAR)	14.7 [°]
Azimuth	Bounding-box	32.8 [°]
	Theil-Sen (LiDAR)	4.4 [°]
Installed capacity	Linear regression	0.69 [kW_p]

Selected methods Based on the results of Table 4.5, we chose the following methods for each characteristic. The method in italics will be used for DeepPVMapper.

— Surface (horizontal projection): *direct computation*.

— Tilt angle: constant imputation, a *lookup table*, and Theil-Sen estimation. The constant imputation works in all cases; the LUT turned out to be very competitive compared to the random forests, and Theil-Sen is competitive when surface models are available.

— Azimuth angle: We keep the *lookup table* method and the Theil-Sen estimation to be used when surface models are available.

— Installed capacity: We keep the constant imputation of the module efficiency and the *linear models*, as it turned out to be very competitive with the random forests.

We integrate `PyPVRoof` after the classification and segmentation step. This step returns a geolocalized PV polygon. Using the BDPV database, we calibrate the regression coefficients for estimating the installed capacity from the surface of the installation and the lookup table for deriving the tilt angle. The result is a data frame where each line is a PV installation, and the columns record the localization (latitude, longitude, and city), surface, installed capacity, tilt, and azimuth angles.

2 From DeepSolar to DeepPVMapper: how to make state-of-the-art more reliable?

This section presents our approach for improving the reliability of current state-of-the-art PV mapping algorithms. First, we introduce new evaluation metrics based on the DTA defined in chapter 2. These metrics are more representative of the accuracy in operating conditions. These accuracy metrics enable a more comprehensive evaluation of the algorithm's configurations. Then, we discuss how we can enhance

the architecture of the algorithm to reduce the number of false detections. To this end, we introduce a new data preprocessing method whose aim is to reduce the occurrence of false detections. This method involves inducing an overlap between the thumbnails and focusing only on relevant areas.

2.1 Evaluation on metrics that are more representative of the operational conditions

2.1.1 Overview of the usual accuracy metrics

F1 score (classification) The F1-Score is most widely used for evaluating the accuracy of a classifier. In addition, some works (e.g., Mayer et al., 2020) use Cohen's κ (Cohen, 1960).

The F1 score is the harmonic mean between the precision and the recall. The precision corresponds to the ratio of correct detections among detections $P = \frac{TP}{TP+FP}$ and the recall to the ratio of correct detections among the population, $R = \frac{TP}{TP+FN}$. The F1 score is then computed as

$$F1 = 2 * \frac{P * R}{P + R}. \tag{4.1}$$

The F1 score is a particular instance of the more general F_β score,

$$F_\beta = \left(1 + \beta^2\right) \frac{\left(1 + \beta^2\right) P \times R}{\beta^2 P + R},$$

where β is chosen such that the recall is β times as important as precision.

Cohen's κ (classification) This index was introduced by Cohen (1960) and measures how the classifiers' performance (measured by p_0) differs from an expected classifier (measured by p_e). Following Mayer et al. (2020), we have

$$\kappa = \frac{p_0 - p_e}{1 - p_e}, \tag{4.2}$$

where

$$p_0 = \frac{TP + TN}{TP + FP + TN + FN},$$

and

$$p_e = \frac{((TP + FP)(TP + FN) + (FN + TN)(FP + TN))}{(TP + FP + TN + FN)^2},$$

and TP, TN, FP, and FN stand for true positives, true negatives, false positives, and false negatives, respectively. Cohen's κ is lower or equal to 1. If $\kappa \leq 0$, then the classifiers disagree. The closer κ to 1, the higher the agreement between the classifiers.

Matthews correlation coefficient (classification) This score was introduced by Matthews (1975) and is also known as the ϕ coefficient or the mean square contingency coefficient. It is computed as

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}. \tag{4.3}$$

Li et al. (2020) discussed this score in the context of PV panels detection. They argued that Matthews correlation coefficient (MCC), which is more robust in reporting accuracy in imbalanced settings, should be preferred. The MCC takes values between -1 and 1, -1 indicating that the classifier is always wrong, 0 indicating that the classifier is random, and 1 indicating that the classifier is perfect.

Intersection-over-Union (segmentation) The Jaccard Index or Intersection-over-Union (IoU) score is a ratio that measures how well the predicted polygon overlaps with a reference polygon. If both polygons overlap perfectly, their intersection equals their union; thus, the ratio equals 1. On the other hand, if both sets are completely disjoint, the intersection is null, so the ratio is equal to 0. The IoU between two sets A and B is written as

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}. \tag{4.4}$$

2.1.2 Deriving a representative testbench using the DTA

Limitation of existing metrics First, these metrics require ground truth annotations and thus cannot be computed outside the test dataset or on a subset of the mapping area that has been manually labeled. We have shown in chapter 2, section 2.2, that the evaluation on the test set is not necessarily representative of the performance on the mapping area, mainly due to distribution shifts (see also Wang and Deng, 2018). To the best of our knowledge, evaluation for remote sensing of PV installations only focuses on reporting the methods' precision and recall. Malof et al. (2019) evaluated the generalization of their method SolarMapper to Connecticut for an algorithm trained over California using a small annotated sample in Connecticut.

More importantly, what ultimately matters to the user is not the classification nor the segmentation accuracy but rather metrics related with rooftop PV deployment, e.g., the installed capacity or the number of systems.

Towards more comprehensive model evaluation Mitchell et al. (2019) propose to use *model cards*, i.e., reporting documents that describe the performance of machine learning models in a variety of settings as well as the cases in which they are intended to be used.

In our case, to report more faithfully the accuracy of the model and to report it in a more relevant way regarding the task at hand, we propose to use the downstream task accuracy, introduced in chapter 2 section 2.2. We focus on the accuracy of the estimation of the installed capacity as our analyses in chapter 2, section 3.2.1 showed that the tilt and azimuth angle estimation are already satisfying.

We recall the three metrics for evaluating the accuracy of the estimation of the installed capacity: the MAPE, which compares the overall estimation of the installed capacity with the reference; the detection ratio, which compares the number of detections with the actual number of installations and the AIPE, which compares the estimated average size of the installation with the actual average size of the installation. The MAPE measures the mismatch between the registered and the estimated installed capacity at the city level. The detection ratio ensures that the algorithm detects the correct number of installations. The AIPE indicates whether we under or overestimate the size of the installations. By construction, a negative (resp. positive) AIPE indicates that, on average, we underestimate (resp. overestimate) the size of the installations.

We evaluate the model with the DTA metrics on an area of 120km² near Lyon, France. This area is sufficient for evaluating the benefits of our approach and sufficiently small to enable multiple evaluations of variants of the mapping algorithm in a limited time. We chose this area among several others in France as the geographical conditions vary with a densely populated urban area and a countryside surrounding. The density of PV installations is also rather inhomogeneous, making the area quite challenging for the algorithm. This benchmark is fully reproducible by following the instructions on our public repository Kasmi et al. (2023c).

The evaluation using the DTA rather than the F1 Score or the IoU enables a comprehensive assessment of the mapping algorithm. Therefore, we will be able to evaluate the impact of the different models (classification and segmentation) but also the impact of architectural choices such as the one-step or two-step approach and the effect of various modules such as the filtering module introduced in section 2.2 on the accuracy.

Proposed approach: reporting accuracy with standard metrics and DTA metrics Our approach for evaluating the models is twofold: We evaluate the classification and segmentation models with the standard metrics and then the whole pipeline with the DTA metrics. We designed the DTA to be more representative of the fitness for the use of the algorithm for rooftop PV mapping. As the characteristics extraction module of the mapping algorithm remains the same, differences in the estimation are imputable to differences in the classification and segmentation models.

2.1.3 Comprehensive evaluation of the classification and segmentation branches

Figure 4.5 sketches the relationships between the different evaluations carried out in this study. Headings in the black boxes correspond to steps of the mapping algorithm. Green boxes correspond to evaluation metrics. The size of the square represents the perimeter of the evaluation. For instance, the perimeter of the evaluation with the F1-Score and Cohen's κ is the classification module of the algorithm.

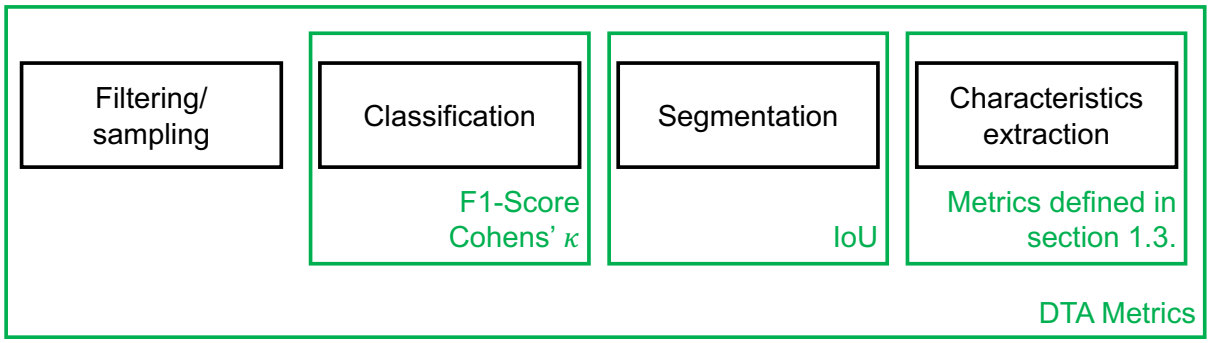


Figure 4.5 – Perimeter of the evaluations carried out in this study. Headers in the black boxes correspond to the different modules of the mapping algorithm.

Evaluation of the classification models We consider several recent and popular classification models based on the Vision transformer architecture (Dosovitskiy et al., 2021). We benchmark this model, alongside the ConvNext (Liu et al., 2022), DeiT (data efficient transformer, Touvron et al.,2021) and ConvMixer (Trockman and Kolter, 2023).

In addition to these models, we evaluate the gains from the data augmentation strategies introduced in the previous chapter. We consider the Blurring + Wavelet perturbation (WP) method. We do not consider alternative methods, as our evaluation in chapter 3, section 3.3.2 revealed that our data augmentation method outperforms existing methods. We implement these data augmentation techniques on a ResNet-50 backbone.

Segmentation branch Recent works in the field leveraged either DeepLab-V3 or variants of the U-net (Ronneberger et al., 2015) architecture. We benchmarked the U-net architecture on BDAPPV to see whether it brought significant accuracy gains.

Evaluation of the architecture and sampling methods Finally, we compare the benefits of the two-step architecture with the one-step architecture. To do so, we set the classification threshold to 0 so that all images are passed to the segmentation model. We also evaluate the benefits of the sampling approach introduced in sections 2.2 and 2.3 using the DTA.

2.2 Reducing the occurrence of false negatives through overlapping the thumbnails

2.2.1 Motivation: an empirical observation

Do false negatives lie at the edge of the image? While analyzing the model's results, we observed that false negatives appeared more often when the PV panel lies at the edge of the input image. We set up a small experiment on the BDAPPV dataset to verify this assumption.

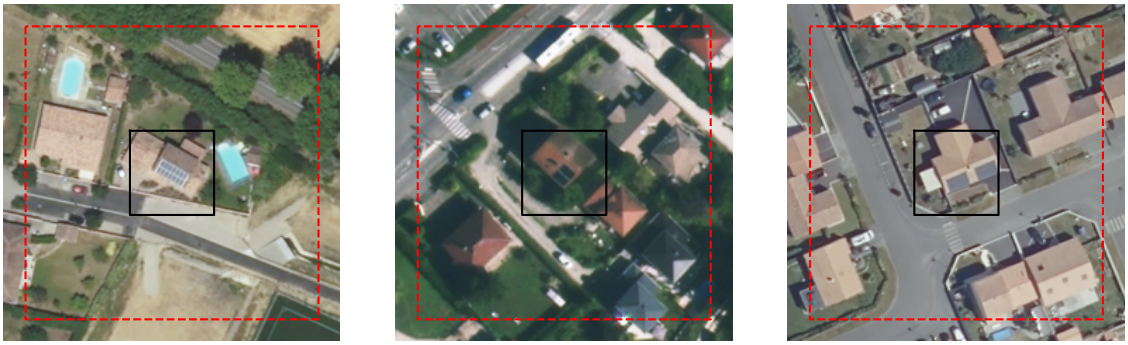


Figure 4.6 – Illustration of the cropping of thumbnails of a size of 224×224 pixels from the raw BDAPPV image with a size of 400×400 pixels to simulate various locations of the PV panel on the image. The black square is always included in the smaller thumbnails, the red dashed lines indicate the boundaries of the thumbnails, such that this black square is contained in the thumbnail.

Proposed approach We consider a model trained on BDAPPV using the standard training procedure. As the raw images have a size of 400×400 and the input size of the model is lower (typically 224×224), we can simulate the effect of the location of the panel on the image by cropping the image using different cropping centers. We consider a subset of 768 images for which the PV panel is entirely contained in the center of the image (black squares on Figure 4.6). By definition, the image center's coordinate (in pixels) is (0,0). We consider cropping centers varying from -124 to 124, resulting in thumbnail boundaries within the red dashed lines in Figure 4.6. This way, we ensure that the panel always lies in the image, but its location varies. In appendix C, section 3.2, we display an example of thumbnails generated from the raw image so that the position of the PV panel varies.

Results With our procedure, we obtain $J^{(i)}$ variants of the i^{th} image. We index each variant by the cropping center's coordinates (x, y). By construction, all images contain a PV panel. We compute the average predicted probability over the number of images for each cropping center (x, y). Figure 4.7 (a) depicts the results. We can see that the average predicted probability decreases as the cropping center moves

away from the panel's location. Applying the classification threshold, we can define a region where the prediction remains a true positive (green area on figure 4.7 (b)) or becomes a false negative (red area).

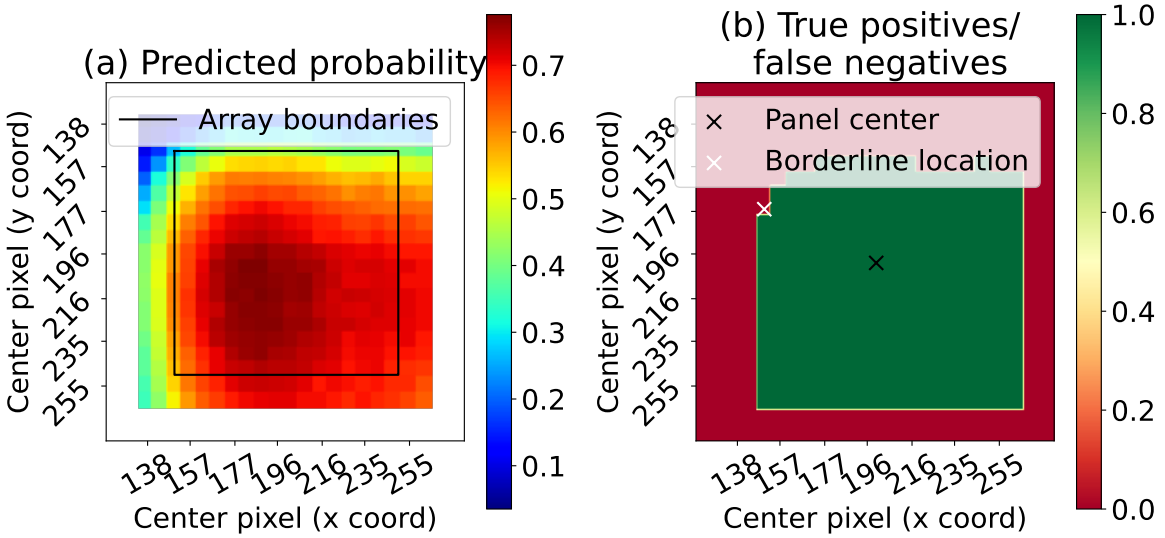


Figure 4.7 – Average predicted probability (a) and true positives and false positives domains (b) as the center of the thumbnail moves away from the location of the panel.

2.2.2 Determining the maximal admissible distance

Definition From Figure 4.7 (b), we can define the *borderline location* b as the closest cropping center (x_b, y_b) such that the model predicts a false negative. This point corresponds to the white cross in Figure 4.7 (b).

Then, we can convert the distance between the borderline location and the image center in meters. This distance corresponds to the *maximal admissible distance* d^*: if one wants to avoid false negatives, then one should place thumbnails center no farther than this distance from each other. This way, we can reduce the probability of having false negatives. Equation (4.5) gives the formula for the maximal admissible distance.

$$d^* \text{ [m]} = \sqrt{x_b^2 + y_b^2} \text{ [px]} \times GSD \text{ [m / px]}, \tag{4.5}$$

where (x_b, y_b) corresponds to the coordinates of the borderline location and GSD corresponds to the ground sampling distance of the image. We assume that the panel's coordinates are $(0, 0)$.

Calibration By convention, in the standard (non-overlapping) case, the distance d_0 between two centers of neighboring thumbnails equals the thumbnail's width. As

the thumbnail size for ResNets is usually 224 pixels, d_0 =224 [px]×0.2 [m / px] = 44 [m].

According to Figure 4.7, the borderline location is (58,97), so

$$d^* = \sqrt{(58)^2 + (97)^2} \text{ [px]} \times GSD \text{ [m / px]} \approx 113.02 \times 0.2 \text{ [m]} \approx 22.60 \text{ [m]}. \tag{4.6}$$

Moreover, the distance between two thumbnail centers decreases by about 50 % compared to the non-overlapping baseline. The computations are made taking as reference $(0, 0)$ the center of the thumbnail.

2.3 Reducing the occurrence of false positives by focusing on relevant areas

2.3.1 Motivation: identifying the adequate region-of-interest boosts accuracy

As underlined by Krapf et al. (2021), existing rooftop PV mapping pipelines struggle to scale at the size of power systems or countries. One reason for that is the fact that current methods do not target specific areas when carrying out mapping. By definition, rooftop PV panels are located in anthropized areas, corresponding only to a tiny fraction of a territory's overall area.

The fact that the so-called region-of-interest (ROI) is orders of magnitude smaller than the target area is a common feature of numerous remote sensing tasks (Uzkent and Ermon, 2020) such as development mapping (Sheehan et al., 2019), poverty mapping (Ayush et al., 2021) or object counting (Gao et al., 2020). However, knowing *a priori* which areas to target can be challenging (Meng et al., 2022). To overcome this issue, Meng et al. (2022) introduced IS-Count, a strategy based on importance sampling (IS), which consists to pick only representative areas for object counting. Their approach achieves good accuracy while only mapping a small fraction of the target area. Other works (Uzkent and Ermon, 2020; Ayush et al., 2021) used reinforcement learning to choose which areas to scan. In addition to lowering the computational burden, prior identification of the ROI also increases the accuracy of deep learning-based remote sensing (Kong and Henao, 2022).

Narrowing the focus to rooftops and urbanized areas The works mentioned above introduced methods that help determine the ROI when the latter is unknown. However, in our case, the ROI can be deduced from the localization of the buildings.

By definition, rooftop PV panels are located on rooftops. The location of buildings is generally easily accessible (e.g., on OpenStreetMap or, in our case, using the BD TOPO of the IGN).

2.3.2 Finding the optimal sampling strategy

Preprocessing as spatial sampling In section 2.2.2, we showed that we need to induce an overlap between the thumbnails passed to the classification model to reduce the occurrence of false negatives.

We extract these thumbnails from a larger image tile. Extracting the thumbnails can be viewed as the definition of a set of locations (i.e., the center of each thumbnail). The set of locations can be referred to as a mesh, and we may wonder whether there is an efficient way of sampling this set of locations.

We state our problem as follows: We wish to find a mesh $M = (m_{x_0}, \ldots, m_{x_n})$, i.e., a set of n thumbnail centers such that the distance between two thumbnails centers m_{x_i} and m_{x_j}, $i \neq j$ should be lower than a *maximal admissible distance* d^* that depends on the ground sampling distance of the input overhead imagery.

For this, we determine a sampling strategy such that the distance between two points is at most d^* and the number of sampled locations is the lowest. We then have a generic set of locations from which we pick only the points near a building.

Sampling strategies We consider several methods for generating the thumbnail centers: the random strategy and the deterministic strategy. For the random strategy, we consider two sampling schemes: the vanilla Monte-Carlo approach and the Sobol approach, a quasi-Monte-Carlo (QMC) method known to increase the sampling efficiency over the vanilla Monte-Carlo method. The vanilla Monte-Carlo approach rests on the `np.random.uniform()` method, based on the permuted congruential generator (PCG) family of algorithms for generating pseudo-random sequences of numbers (O'Neill and College, 2014). The QMC or Sobol sampling method roughly consists of forming successively finer interval partitions and reordering the coordinates in each dimension. It spreads points more evenly in space than with the vanilla random sampling method (Sobol, 1967). Our deterministic strategy defines the points' location according to a precomputed grid on the thumbnail.

Figure 4.8 illustrates the three strategies on a dummy grid. Each dot indicates the center of a thumbnail. We can see that the deterministic strategy paves the space better than the random strategies, leaving no empty locations.

Once we determine our best strategy, we evaluate the gains brought by our method, measured by the number of target locations to investigate and the resulting computational cost it brings: the higher the number of locations, the higher the computational cost.

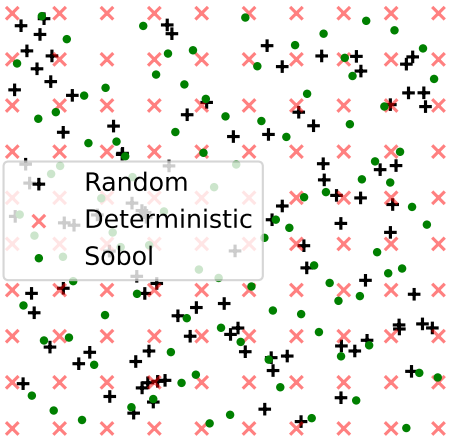


Figure 4.8 – Illustration of how sampling strategies would cover a tile. All strategies represent $n = 100$ points.

2.3.3 Picking the most efficient sampling strategy

The deterministic sampling strategy is the most efficient We compared our three sampling strategies and evaluated which one enabled us to reach the distance d^* between centers while generating as few thumbnails as possible. Figure 4.9 depicts the results: the best strategy is the deterministic strategy.

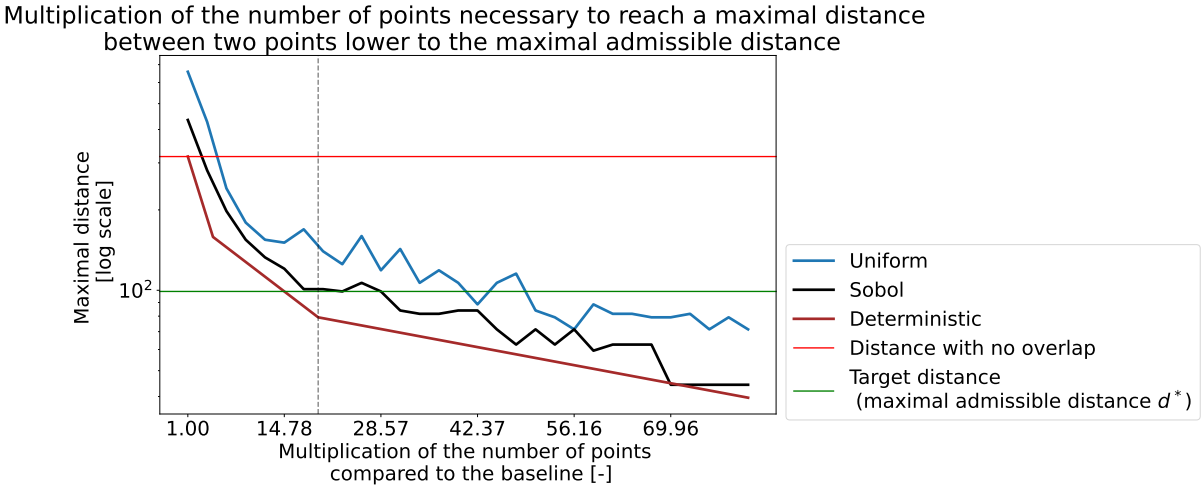


Figure 4.9 – Increase in the number of points to reach a distance of at most d^* between two thumbnail centers according to the deterministic and random (Sobol and Uniform) sampling strategies.

Overall, our approach leads to a decrease in the absolute number of thumbnails Inducing an overlap increases the number of thumbnails. However, as we combine this method with filtering using the BD TOPO, the overall number of thumbnails decreases as we focus on a smaller area. To quantify the gain yielded by the building filtering, we generated a mesh of thumbnail centers and filtered it to keep only the centers intersecting with a building. Results showed that, on average, over 300 tiles of 25km² each the sampling with BD TOPO leads to removing 83.1 % of the points. This amounts to a decrease of 32.4 % in the number the points (hence thumbnails) compared to the case with no sampling and overlapping strategy. We discuss in section 3.1.3 and in appendix A, the broader implications of this result.

3 Results

In this section, we present the results of the benchmarks of the classification and segmentation branches and discuss the evaluation of the model with the DTA metrics. The evaluation with the DTA metrics highlights differences in the accuracy of models whose F1-Score is similar, thus highlighting the relevance of our proposed evaluation method. We show that our sampling with overlapping and filtering algorithm increases the accuracy. We discuss the implications regarding the computational cost of this filtering process. Finally, to illustrate the practical improvements brought by DeepPVMapper, we reexamine the cases introduced in chapter 2, section 3 in the light of our new mapping algorithm.

3.1 The DTA reveals accuracy differences among models and the better performance brought by the sampling

3.1.1 Accuracy on the test set

Classification We report the accuracy results on the training dataset BDAPPV (Kasmi et al., 2023d) for various classification models in table 4.6. According to the F1 score and Cohen's κ, the performance of the Inception-v3 is high (the F1 score is 0.83 and Cohen's κ is 0.69), and alternative models only match its performance. CNNs architectures (ResNet and Inception) show similar performance. In line with Luzi et al. (2023), the benefits of transformer components only appear in hybrid architectures. Overall, the performance improvements are modest. A reason for this modest improvement could be that transformers have less inductive biases than CNNs and, therefore, require more training data than accessible in popular remote sensing datasets (e.g., the Inria dataset (Maggiori et al., 2017) or DeepGlobe (Demir et al., 2018) used by Luzi et al., 2023) to achieve good accuracy. Overall, this benchmark shows that the accuracy measured by the F1 score saturates and that recent and heavy architectures do not bring significant accuracy gains.

Table 4.6 – Benchmark of various model architectures on BDAPPV (Kasmi et al., 2023d). The best results are **bolded** and the second best results underlined.

	Model	F1 score (↑)	κ (↑)
	Inception v3 (Szegedy et al., 2016) (baseline)	0.83	0.69
CNN	ResNet-101 (He et al., 2016)	0.82	0.65
CNN	ResNet-50 (He et al., 2016)	**0.84**	0.68
ViT	ViT-B16 (Dosovitskiy et al., 2021)	0.81	0.64
ViT	ViT-B8 (Dosovitskiy et al., 2021)	0.81	0.68
Hybrid	ConvNext (Liu et al., 2022)	**0.84**	**0.72**
Hybrid	DeiT (Touvron et al., 2021)	**0.84**	0.67
Hybrid	ConvMixer (Trockman and Kolter, 2023)	0.83	0.67
Robust	ResNet-50 + WP (Kasmi et al., 2023b)	0.82	0.62

Segmentation Table 4.7 presents the results of the benchmark of the segmentation models. We evaluated the U-Net model and saw that the performance lags behind the baseline. de Luis et al. (2023) report similar results for the segmentation branch on the IGN images of BDAPPV, with IoUs ranging from 0.45 to 0.56. The main reason is that there are not enough IGN samples (17,000) in BDAPPV to achieve satisfactory performance. de Luis et al. (2023) achieve a good accuracy on BDAPPV's Google images (28,000 samples), in line with the accuracy reached on Bradbury et al. (2016) (37,000 samples).

Table 4.7 – Benchmark of various model architectures for the segmentation branch. The best results are **bolded** and second best underlined.

Model	IoU (↑)
Baseline (Kasmi et al., 2023a)	**0.85**
U-Net (Ronneberger et al., 2015)	0.48

3.1.2 Downstream task accuracy

Overall results Table 4.8 presents the results according to the downstream task accuracy. We can see that alternative models and the sampling process yield significant accuracy improvements compared to the baseline. Using a ResNet or a ConvNext model in the algorithm significantly improves the DTA (by all metrics) compared to the baseline. The effect of the sampling also yields significant accu-

racy improvements. These results demonstrate that the accuracy measured by the F1 score on the test dataset (Table 4.6) is unreliable and that the DTA is more representative of the performance in an operational setting. The baseline, the ResNet, and the ConvNext models performed similarly on the test set. However, we can see that they significantly differ when deployed over a larger area and in an operational setting.

Table 4.8 – Accuracy results (DTA) on a 120km² area around Lyon, representative of the operational conditions. The best results are **bolded** and the second best results underlined.

Pipeline	DTA		
	MAPE [%]	Ratio [-]	AIPE [%]
Baseline (Kasmi et al., 2022a)	55.7	1.29	15.4
ResNet-50	46.9	1.09	15.5
ConvNext	45.5	1.11	15.3
ResNet + Sampling/filtering	39.5	**0.91**	**12.6**
ConvNext + Sampling/filtering	**38.8**	0.84	18.2
Segmentation only	89.52	1.66	24.32
ResNet + WP	48.22	0.57	16.97
ResNet + Sampling + WP	40.62	0.82	21.48

Discussion Results from Table 4.8 show that the performance measured by the F1 score (Table 4.6) is not representative of the true performance. Indeed, on our dataset, models saturated at around 0.80, but when evaluating these models using the DTA, we can see sizeable differences in the accuracy. These results show the relevance of going beyond standard accuracy metrics and benchmarks on test datasets to evaluate model accuracy, as Vishniakov et al. (2023) also underlined.

Besides, the DTA metrics benchmark shows that the segmentation-only approach will likely generate many false positives by overestimating the installations' number and size. On the other hand, adding the sampling process is very efficient in increasing the accuracy of the mapping algorithm.

3.1.3 The filtering also yields efficiency gains

Overall, the filtering increases the efficiency To demonstrate these gains, we evaluate the computational cost of the pipeline over our test area of 120km². Table 4.9 shows the results. We can see that the filtering increases by 31% the speed of the process. Scaling up (linearly) to France could results in a spare of 39 days of computations compared to the baseline. It highlights the benefit of mapping at the grid scale by considering the relevant ROI, in our case, the areas that contain a rooftop. In appendix A we discuss a simplified framework to estimate the gains in energy consumption brought by this sampling method.

Table 4.9 – Computational gains brought by the sampling. The best results are **bolded**.

Variant	Runtime [sec/km²]	Scale-up [days]
Baseline (Kasmi et al., 2022a)	19.39	122.08
Sampling	**13.19**	**83.04**

3.2 Building and deploying DeepPVMapper

3.2.1 The pipeline of DeepPVMapper

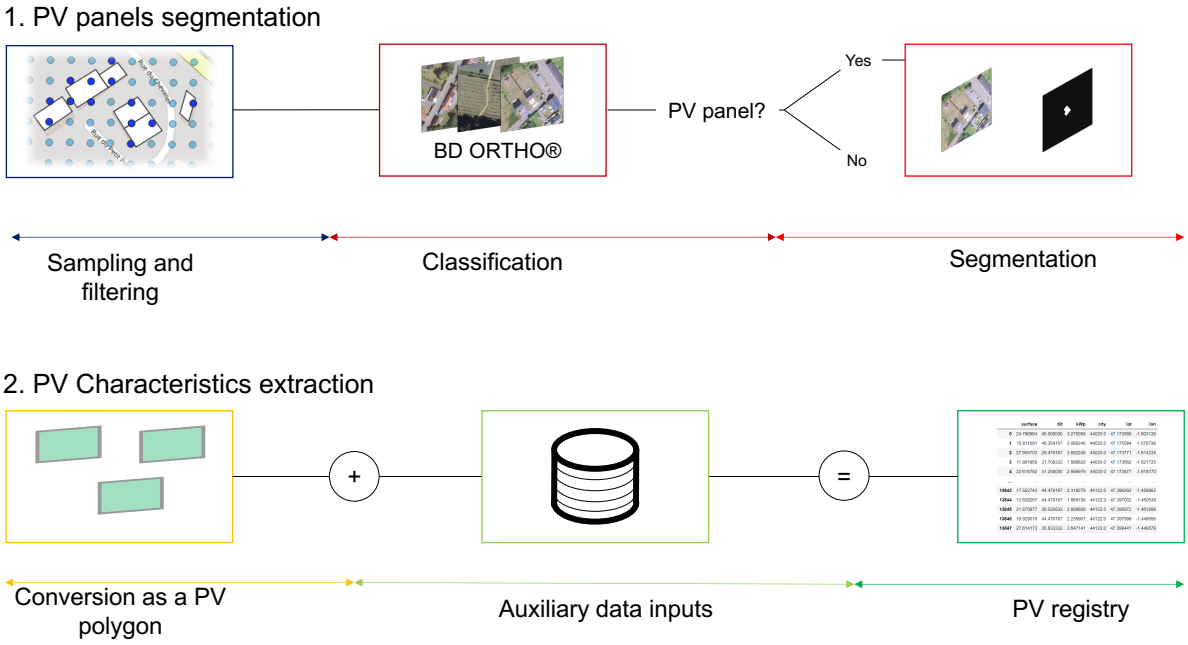


Figure 4.10 – Flowchart of DeepPVMapper.

Overview Figure 4.10 shows our resulting mapping algorithm. The main difference with DeepSolar-based models is that we include a filtering module at the beginning of the algorithm to select the relevant areas to inspect. The second main difference is that our postprocessing module is based on `PyPVRoof` and does not necessarily require having access to 3D data as it accommodates various use cases (Trémenbert et al., 2023).

Preprocessing module Before passing images to the classification model, we restrict ourselves to the relevant areas, which correspond to areas where there are buildings. Besides, we also implement an overlap between the thumbnails passed to the classification model to minimize the probability of false negatives.

Classification and segmentation models We use the ResNet-50 model for classification, as the ConNext model does not outperform the ResNet-50 whlie being significantly larger. Using the ConvNext would not yield accuracy gains compared to the ResNet but would require more computing power for inference.

Regarding the segmentation model, we stick with the DeepLab-v3 model as BDAPPV appears insufficient to train a segmentation model from scratch, and we do not have larger datasets to pre-train a segmentation model on. Finally, the performance of the segmentation branch is already satisfying.

Characteristics extraction Our characteristics extraction module is generic and accommodates the absence of 3D data for estimating a tilt angle. As such data is increasingly available in France, it will be straightforward to integrate this data into the pipeline.

For our version of DeepPVMapper, we extract the characteristics of the PV installations using the bounding box for the azimuth angle and the lookup table for the tilt angle. We use a clustered linear regression to estimate the installed capacity of the installation. This parameterization is the same as our baseline model introduced in chapter 2, section 3.1.1.

3.2.2 Returning to Cobrieux

Table 4.10 – Extract of the registry generated by DeepSolar and DeepPVMapper for the city of Cobrieux (Nord).

	DeepSolar			**DeepPVMapper**		
	ID	Surface	Installed capacity	ID	Surface	Installed capacity
	29925	12.35	1.96	37677	42.59	5.61
	29897	275.51	27.40	37682	20.73	2.77
	29904	24.18	3.84	37685	21.73	2.90
	29924	36.01	3.94	37692	12.26	1.81
	29926	17.69	2.81	37695	21.98	2.93
	29927	24.34	3.87	37703	27.91	3.68
	29921	39.61	3.94			
	29913	12.59	2.00			
	29908	14.94	2.37			
	29910	19.78	3.14			
	29912	46.26	4.60			
	29901	38.16	4.17			
Average size			5.34			**3.28**
(curated)			3.33			
Reference			*2.72*			*2.72*

False positives are no longer present In chapter 2, section 3.3.2, we discussed a problematic case where the number of installations was relatively accurate (12

estimations for ten installations). However, the overall estimation of the installed capacity was too high (64 kW_p, while the target was 27 kW_p). After running DeepPVMapper over this departement instead of our replication of DeepSolar, we focused on the city of Cobrieux. This time, the algorithm detects only six installations out of 10, but the model no longer detects any outliers, such as the barn's roof depicted on Figure 2.17. This translates into the fact that the average installation size is more in line with what is expected from the RNI, as shown in Table 4.10. The accuracy at the installation scale is even higher than the "curated" registry obtained from DeepSolar. As the actual installed capacity is known, it is less damaging not to estimate all the installations than to estimate a wrong distribution of the installed capacities.

Analysis with the WCAM The WCAM further enables us to understand why DeepSolar identified the barn's roof as a PV panel and why DeepPVMapper did not. Figure 4.11 shows the results. We can see that DeepSolar relies on small-scale components (bottom of the WCAM), whereas DeepPVMapper does not. Therefore, DeepSolar has been confused by these factors.

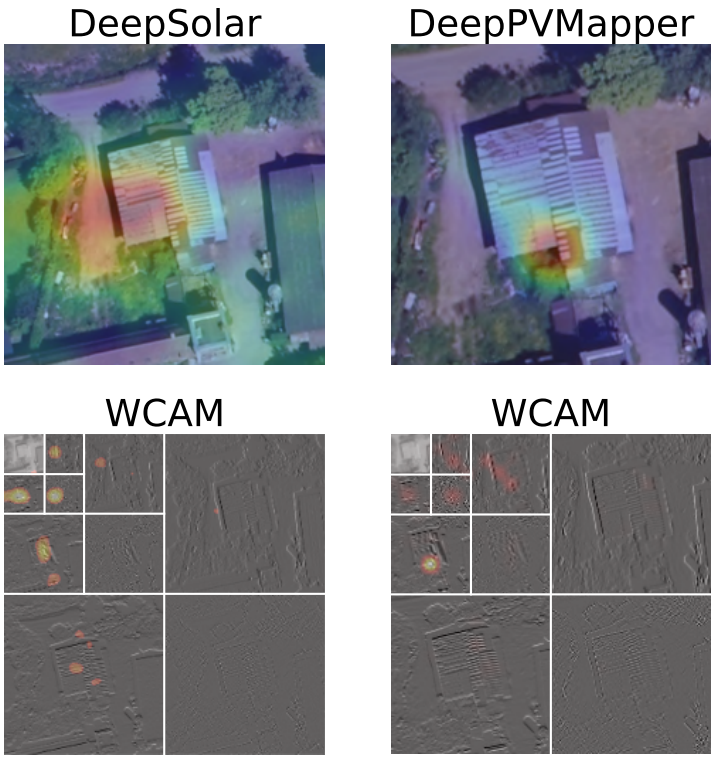


Figure 4.11 – Comparison of the behavior of DeepSolar (left) and DeepPVMapper (right) on the barn (false positive) that was detected by DeepSolar and avoided by DeepPVMapper.

3.3 Broader impact

Applying DeepPVMapper elsewhere The principle of the DTA can be expanded to other European countries. Indeed, the RNI has counterparts in many other European countries. It can be substituted by the *Marktstammdatenregister (MaStR)* for Germany (Bundesnetzagentur, 2022) , the *Stamdataregister* for Denmark (Energistyrelsen, 2022), the *Datenregister* in the Netherlands (CBS, 2024) or the *Anlagenregister* in Austria (E-Control, 2023).

Reliable mapping The triplet DeepPVMapper, WCAM, and the DTA enable a reliable mapping of PV installations in France. As counterparts of the RNI exist elsewhere in Europe, the same framework can be applied to other countries, using these local counterparts as reference data for monitoring the model's data. The model will eventually see its performance decrease. However, the DTA enables the quantification of this drop, and the WCAM enables the visualization of how the model makes its prediction and can guide data processing or future fine-tuning. Besides, it is unnecessary to acquire a lot of fine-tuning data to improve the model's performance (Freitas et al., 2023), at least if the region of interest lies in Europe.

A generic takeaway for applying the model elsewhere will be to rescale the input images to have a ground sampling distance of 20 cm/pixel, aligned with the resolution of the IGN, to keep the performance drop as low as possible. Indeed, such downscaling will lower the noise in the image, which is the main driver for the performance drop (see chapter 3, section 2.2.1 for more details).

Conclusion of the chapter

This chapter discussed the third pillar: improving the model's robustness to acquisition conditions. To this end, we introduce DeepPVMapper, which improves upon the current state-of-the-art following three main directions: improvement of the classification and segmentation models and the pipeline to reduce the probability of false detections, standardized extraction of rooftop PV characteristics using our newly introduced Python wlibrary `PyPVRoof` and evaluation of the accuracy using metrics that are more representative of the real-life conditions and enable a comprehensive evaluation of the performance of the model. Our evaluation showed that DeepPVMapper is 16% more accurate and requires 31% less computing time than an architecture based on DeepSolar. We also deployed DeepPVMapper over the problematic cases identified in chapter 2, section 3 to show that it effectively addresses the issues that we encountered with our replication of DeepSolar in these cases.

Finally, we discussed the broader applicability of our method in new areas. The

DTA can be estimated using equivalent databases available in some European countries. The WCAM enables the audit of the model's decision process and can guide training methods or image filtering to improve the accuracy of unseen images. These results show that DeepPVMapper enables reliable mapping of PV installations as it is built to generate as few errors as possible, can be monitored with the DTA, and is an auditable model thanks to the WCAM. In the next chapter, we will discuss how we use the data of this registry for estimating rooftop PV power production and improving rooftop PV observability.

Chapter 5

Assessing the gains of the registry for estimating the rooftop PV power production

Summary

This chapter studies whether the registry can improve rooftop PV observability, defined as the transmission system operator's ability to estimate real-time and future power production accurately. We propose an approach that estimates the individual PV rooftop power production with the basic conversion model from Dobos (2014) and weather data. For this method to improve PV observability, it should enable accurate estimation, achieve better performance compared to other approaches that do not require conversion models, and scale up to the size of power systems. Using ground truth measurements of 900 rooftop PV systems, we demonstrate that our approach meets these three requirements and thus has the potential to improve rooftop PV observability. The main limitations of our approach lie in the fact that we were limited by the size of our ground truth measurements to study how our approach scales up. Further work is needed to see how this approach performs compared to current methods employed to estimate rooftop PV power production.

1 Additional requirements for improving rooftop PV observability

The PV registry that we've built following the methodology introduced in the previous chapters is not sufficient for improving rooftop PV observability. This registry contains the characteristics of the PV systems (tilt and azimuth angle, installed capacity and localization). In the introduction, we defined PV observability as the ability of the TSO to estimate a power unit's real-time and future production accurately. In addition to knowing the true localization and characteristics of the installations, we need weather data, and ground truth measurements to measure the accuracy of the estimation of the PV power curves from the PV characteristics and the weather data. This section introduces the ground truth measurements and the weather data that we will use in addition to our PV registry.

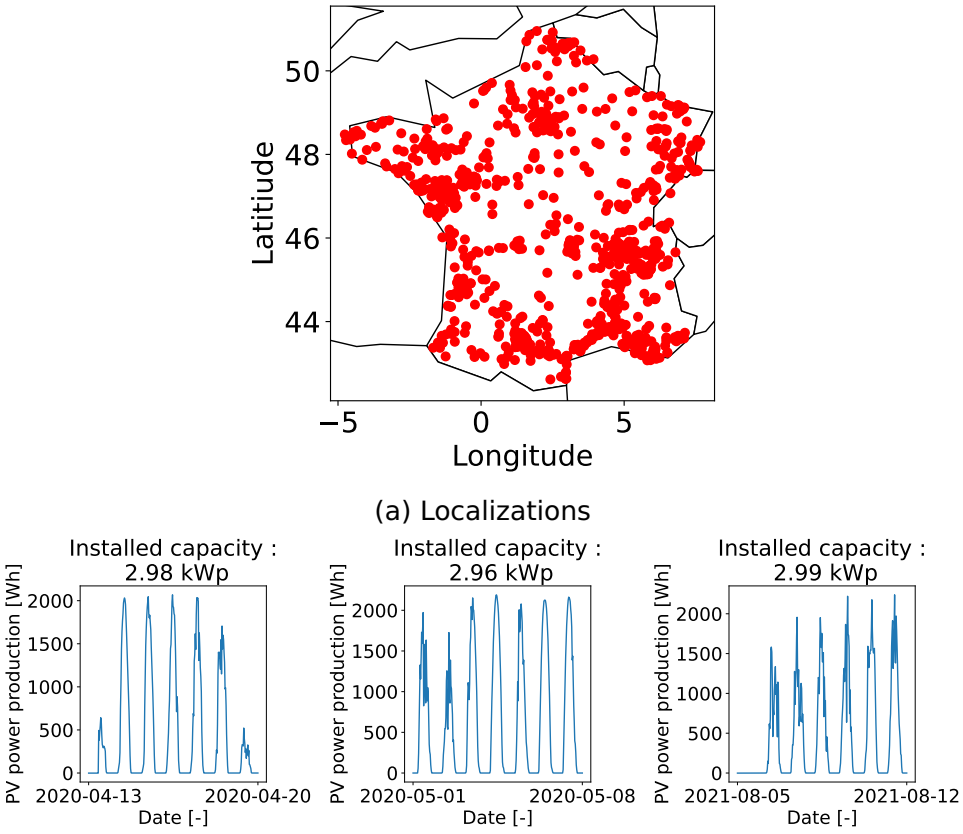
1.1 BDPV ground measurement data

1.1.1 Overview

Description Thanks to the non-profit association *Asso BDPV*, we had access to ground truth PV power production measurements of 1,793 individual PV systems. These measurements span over France and have a granularity of 30 minutes. Existing datasets contain less installations, for instance de Hoog et al. (2021) had access to PV systems measurements for 740 systems (at a time resolution of 5 minutes) of businesses and homes in Western Australia, and Perera et al. (2022) used the Pecan Street's Dataport (Pecan Street, 2024), which contains measurements at a 1-minute resolution for 73 households across the United States. The closest database from ours is the data provided by IBW, the local utility of Wohlen in Switzerland. This dataset contains PV power generation profiles of 15 homes and the PV systems' technical characteristics, used by Walch et al. (2021).

This data complements our existing data, which contains the localization and technical characteristics of the PV installation. This means that with these measurements, we have all the necessary information to carry out rooftop PV power estimation, from detecting the installation on aerial images to estimating the accuracy of the power production estimation. We also know the precise location of the installations, contrary to de Hoog et al. (2021), who relied on postcodes.

Visualizations Figure 5.1 plots the localization (Figure 5.1a) of the PV measurements and examples (Figure 5.1b) of power output time series coming from our dataset. In the following chapter, "power output time series" and "PV yield time series", "PV power production" refer to the same data.



(a) Localizations

(b) Power curves

Figure 5.1 – Localizations and example of power curves contained in our PV measurements dataset.

1.1.2 Quality checks

To control the quality of the data, we controlled the alignment between the PV measurements and a PV power estimation generated from solar irradiance and temperature data. We also manually inspected the data to remove faulty and biased measurements. We remove installations for which we have too few measurements or too many periods with no measurements. Over the 1,793 raw installations, 906 passed our quality checks. We refer the reader to appendix C, section 4.1 for examples of production reports leveraged to filter the faulty measurements.

1.1.3 Descriptive statistics

On our curated dataset, we have, on average, 27,346 measures per installation, corresponding to an average duration of 569,7 days of observations. The time window spans from December 2, 2020, to February 15, 2023. Table 5.1 summarizes the key descriptive statistics regarding our data.

Table 5.1 – Summary statistics on the PV power measurements. The load factor is the ratio between the PV power production at time t and the installed capacity of the installation.

Variable	Unit	Min	Max	Mean	Median
StartDate	[-]	2020-02-12 23:00:00	2023-01-14 23:00:00	-	-
EndDate	[-]	2020-05-31 22:00:00	2023-02-15 22:30:00	-	-
Max capacity factor	[W/W$_p$]	0.00	1.00	0.82	0.83
Mean capacity factor	[W/W$_p$]	0.00	0.21	0.13	0.13
Median capacity factor	[W/W$_p$]	0.00	0.04	0.00	0.00
Number of measures	[-]	289	51716	27346	28242
Installed capacity	[W$_p$]	0.60	62.04	3.65	2.96

1.2 Solar radiation and temperature data

Our conversion model will take two main kinds of weather variables as inputs: solar radiation and temperature. The solar radiation data comes from The Copernicus Atmospheric Monitoring Service (CAMS), and the temperature and weather data from the ECMWF Reanalysis v5 (ERA5).

1.2.1 Solar radiation: Copernicus Atmospheric Monitoring Service (CAMS)

The CAMS solar radiation services (Qu et al., 2017) provide historical values (2004 to present) of global (GHI), direct (BHI), and diffuse (DHI) solar irradiation, as well as direct normal irradiation (BNI). We also have clear-sky values (i.e., irradiation values with no clouds). These clear-sky values are obtained using aerosol, ozone, and water vapor information from the CAMS global forecasting system. Other properties, such as ground albedo and ground elevation, are also considered. Similar

time series are available for cloudy (or all-sky) conditions. However, since the high-resolution cloud information is directly inferred from satellite observations, these are currently only available inside the field-of-view of the Meteosat Second Generation (MSG) satellite, which is roughly Europe, Africa, the Atlantic Ocean, and the Middle East with a nadir (directly below the satellite) spatial resolution of 3 km and a temporal resolution of 15 minutes.

Table 5.2 – Description of the main variables included in CAMS.

Name	Unit	Description
BHI	$Wh.m^{-2}$	Direct horizontal all sky irradiation
BHIc	$Wh.m^{-2}$	Direct horizontal clear sky irradiation
BNI	$Wh.m^{-2}$	Direct normal all sky irradiation
BNIc	$Wh.m^{-2}$	Direct normal clear sky irradiation
DHI	$Wh.m^{-2}$	Diffuse horizontal all sky irradiation
DHIc	$Wh.m^{-2}$	Diffuse horizontal clear sky irradiation
GHI	$Wh.m^{-2}$	Global horizontal all sky irradiation
GHIc	$Wh.m^{-2}$	Global horizontal clear sky irradiation

Table 5.2 summarizes the main variables. We consider the 15-minute values for a time interval covering the period between 2020 and 2023. The data has various spatial resolutions, but we can interpolate the values at the point of interest. In our case, we consider the latitude and longitude of the PV installations as points of interest. Figure 5.2 presents a sample of the time series provided by CAMS for two consecutive days.

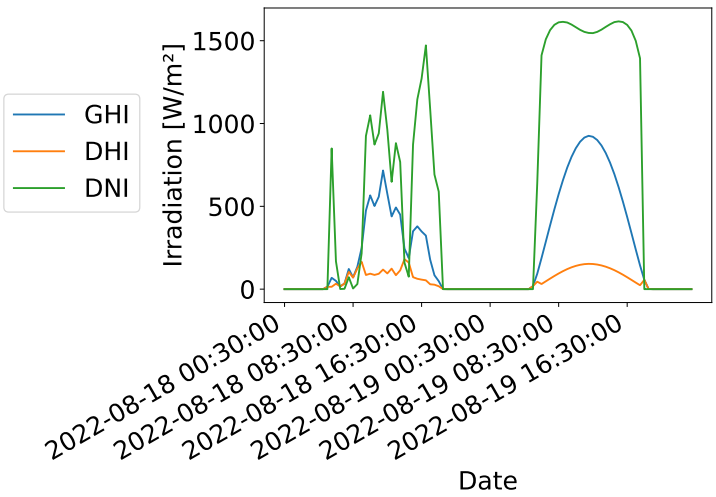


Figure 5.2 – Example of solar irradiation time series provided by CAMS for an installation located near Toulouse, France. The first day is cloudy, and the second day is sunny.

1.2.2 Temperature data: Copernicus Climate Change Service (C3S)

The Copernicus Climate Change Service (C3S), operated by the European Centre for Medium-Range Weather Forecasts (ECMWF), provides the temperature data. Our data source is the 5th reanalysis of the ECMWF data (ERA5, Hersbach et al., 2020)

A reanalysis combines climate model data with observations from across the world into a globally complete and consistent dataset using the laws of physics. This principle, called data assimilation, is based on the method used by numerical weather prediction centers, where every so many hours (12 hours at ECMWF), a previous forecast is combined with newly available observations in an optimal way to produce a new best estimate of the state of the atmosphere, called analysis, from which an updated, improved forecast is issued. Reanalysis works similarly but at a reduced resolution to provide a dataset spanning several decades. Reanalysis does not have the constraint of issuing timely forecasts, so there is more time to collect observations and, when going further back in time, to allow for the ingestion of improved versions of the original observations, which all benefit the quality of the reanalysis product.

We only use the 2m temperature for a single latitude and longitude point and at a temporal resolution of 1 hour (linearly interpolated to get a temperature value every 30 minutes). Finally, we always consider the same reanalysis data for our study to make our results homogeneous. Figure 5.3 displays an example of a temperature time series we use.

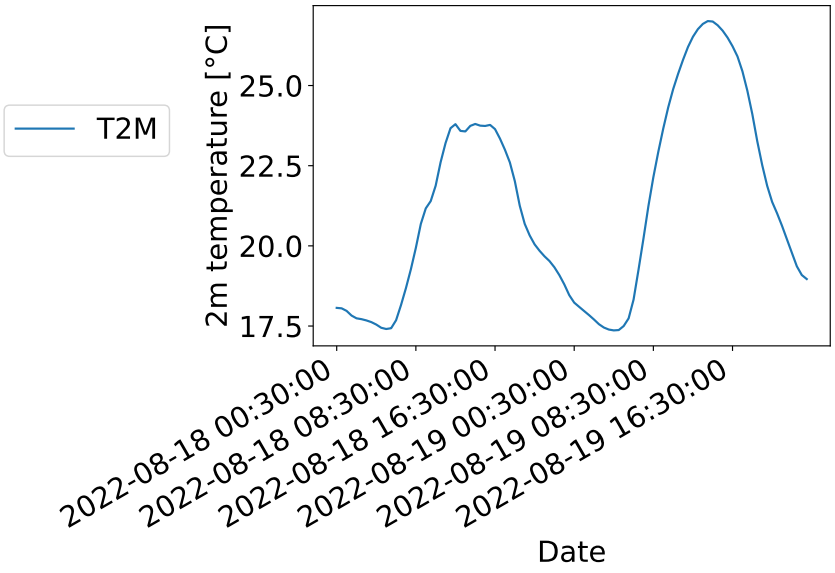


Figure 5.3 – Sample of 2m air temperature from ERA5.

1.3 PV registry

We consider the PV registry generated with DeepPVMapper. At the time of writing, this registry contains more than 100,000 installations. However, in this study, we aim to evaluate the data quality to estimate rooftop PV power production. Therefore, we intersect this registry with the installations from the training database BDAPPV for which we have ground truth characteristics and PV power measurements. We have 1,485 installations in the test set of BDAPPV, which we intersect with the 904 clean power measurements. We end up with 294 unique installations. Table 5.3 presents some descriptive statistics of the PV registry used in this chapter.

Table 5.3 – Descriptive statistics of the PV systems' characteristics extracted from the PV registry for the systems used in this study.

Variable	Unit	Min	Max	Mean	Median	n
Installed capacity	kW_p	1.29	38.84	3.12	2.68	276
Tilt angle	Degrees	11.88	51.63	26.83	26.12	276
Azimuth angle	Degrees	-90.00	90.00	4.23	0.00	276

2 Assessing the relevance of our approach

The previous section introduced the additional requirements besides the PV registry for improving rooftop PV observability. In this section, we present our method for improving rooftop PV observability and assessing the relevance of our approach towards this end. Our approach will improve PV observability if it enables an accurate measurement of the production of the PV systems and if it can scale up to thousands of installations. Finally, our approach is relevant if it provides an improvement compared to competing approaches.

2.1 Evaluation metrics

Throughout our study, we will evaluate the rooftop PV power estimation accuracy using the root mean squared error (RMSE) and the percentage RMSE (pRMSE). We define the RMSE as

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y)}, \tag{5.1}$$

where $\hat{y}_i$ is the estimated target value (e.g., PV power production in W), y_i is the true value, and n is the number of samples. Besides, we define the pRMSE as

$$pRMSE_j = \frac{RMSE_j}{p_{PV,j}} \times 100, \tag{5.2}$$

where j is the index of the PV system that we consider. In other words, the pRMSE is the RMSE normalized by the installed capacity. The pRMSE also corresponds to the normalized RMSE (nRMSE).

2.2 Proposed approach for improving rooftop PV observability

2.2.1 An installation-based approach

Following Saint-Drenan (2016), we restricted our search for characteristics to the set of minimal characteristics that impact PV power estimation the most. These parameters are the system size, tilt, and azimuth angles. Therefore, we input these parameters leveraging the data from the registry. The other parameters are set to their default values. As we rely only on limited information, we used the simplest possible model, the PVWatts model (Dobos, 2014).

2.2.2 Conversion model

Overview We use the conversion model PVWatts (Dobos, 2014). This model takes as input the effective plane-of-array (POA) irradiance and the module temperature and returns the DC power of the module. The effective POA irradiance corresponds to the POA irradiance, taking into account the optical losses of the module. These losses are accounted for following the method of Martin and Ruiz (2001). The computation of the POA irradiance requires knowing the module's tilt and azimuth angle. Dobos (2014) takes into account several other parameters, as displayed on Table 5.4.

Table 5.4 – Set of minimal PV system characteristics for the conversion model Parameters that we input are **bolded**, advanced parameters are in *italics*. Source: Dobos (2014).

Field	Unit	Default value
System size	kW	4
Module type	{Standard, Premium, Thin film}	Standard
System losses	%	14
Array type	{Fixed open rack, Fixed roof mount, 1-Axis, Backtracked 1-Axis, 2-Axis}	Fixed open rack Fixed open rack
Tilt angle	degrees	Site latitude
Azimuth angle	degrees	180° (Northern hemisphere), 0° (Southern hemisphere)
DC/AC ratio	ratio	1.1
Inverter efficiency	%	96
Ground coverage ratio (1 axis only)	fraction	0.4

Computation of the POA irradiance The POA irradiance corresponds to the solar irradiance incident on a surface that is adjusted to the tilt and azimuth angle of the array. It represents the solar energy reaching a surface considering its

orientation towards the sun. The POA irradiance can be decomposed into three components:

— A direct component (POA direct or beam irradiance): This is the solar radiation that reaches the surface in a direct line from the sun. It is the sunlight that travels directly through the atmosphere without being scattered or reflected,

— A diffuse component (POA diffuse irradiance): This is the solar radiation that reaches the surface after being scattered by molecules and particles in the atmosphere. It includes the sunlight that comes from all directions other than the direct path from the sun,

— A reflected component (reflected irradiance): The portion of sunlight that is reflected off nearby surfaces, such as the ground or surrounding structures, and reaches the surface of the PV module

The sum of these components gives the total POA irradiance. We leverage the Python library `pvlib` (Holmgren et al., 2018) to compute the POA irradiance. The function takes as input the solar zenith angle (SZA), the solar azimuthal angle, the top-of-atmosphere (TOA) sun position, and the three components of solar radiation (GHI, DHI, and DNI).

Computation of the module temperature The performance of a PV module depends on its temperature and decreases when the temperature increases. We estimate the module temperature following **?**, given by (5.3):

$$T_{module,t} = T_{2m,t} + \frac{k_{therm} G_{POA,t}}{G_{stc}} \tag{5.3}$$

In other words, the module temperature at time t corresponds to the sum of the temperature at 2 meters, and the temperature increases due to the exposition of the module to the solar radiation. The temperature increase is weighted by the factor k_{therm}, meaning that we assume a linear relationship between the increase in temperature and the global POA (GPOA) irradiance at time t. G_{stc} denotes the irradiance under standard test conditions (STC) and is equal to 1000 W/m^2

Computation of the effective POA irradiance The effective POA irradiance corresponds to the POA irradiance after accounting for the optical losses of the module. To account for these losses, we implement Martin and Ruiz's IAM (incident angle modifier) model (Martin and Ruiz, 2001; Martín and Ruiz, 2002, 2005). This model returns incident angle modifiers (IAMs) applied to the POA irradiance to obtain the effective POA irradiance. Intuitively, this model accounts for the fact that the glass on the PV module reflects that the angular losses (AL) of PV modules are a function of the solar incident angle θ_{AOI} (Martin and Ruiz, 2001). We considered reference values for a monocrystalline module.

$$AL\left(\theta_{AOI}\right) = 1 - \frac{\bar{T}\left(\theta_{AOI}\right)}{\bar{T}\left(0\right)} = 1 - \frac{1 - \exp\left(-\cos\left(\theta_{AOI}\right)/a_r\right)}{1 - \exp(-1/a_r)} \approx 1 - \frac{1 - \bar{R}\left(\theta_{AOI}\right)}{1 - \bar{R}\left(0\right)} \tag{5.4}$$

where $\bar{T}(x)$ is the weighted transmittance at incident angle x, $\bar{R}(x)$ the weighted reflectance at incident angle x and a_r the angular losses.

Then an angular factor f_{I_α}, corresponding to the IAM, is defined as the ratio between the module's short circuit current I_{sc} at indecent angle θ_{AOI} to the I_{sc} at normal incidence.

$$f_{I_\alpha} = \frac{I_{sc}\left(\theta_{AOI}\right)}{I_{sc}(0)\cos\left(\theta_{AOI}\right)} \approx \frac{1 - \bar{R}\left(\theta_{AOI}\right)}{1 - \bar{R}\left(0\right)} \tag{5.5}$$

We compute the IAM for the three components of POA irradiance, and the effective POA is given by

$$POA_{eff} = f_{beam} \times G_{POA} + f_{diff,sky} \times D_{POA} + f_{diff,ground} \times R_{POA} \tag{5.6}$$

Where $f_\bullet$ corresponds to the IAM, G_{POA} the direct POA, D_{POA} the diffuse POA component, and R_{POA} the reflected POA component. Figure 5.4 illustrates the different components of solar radiation considered to compute the effective plane-of-array incidence. "Direct," "Diffuse," and "Reflected" correspond to the components of solar radiation.

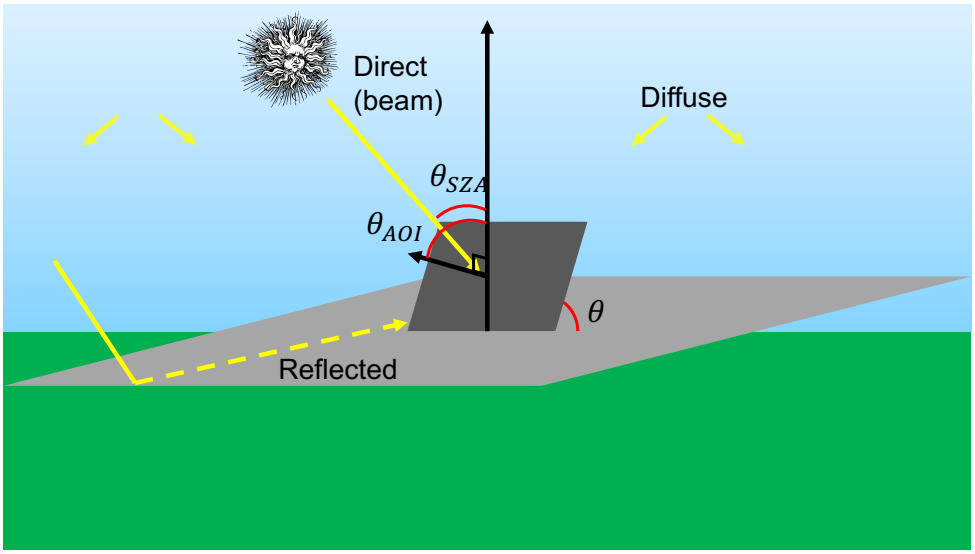


Figure 5.4 – Illustration of the POA irradiation modeled with our approach. θ indicates angles, "AOI": "angle of incidence" and "SZA": solar zenith angle. The light gray surface is flat, and the dark grey surface is tilted with tilt angle θ.

2.2.3 Wrap-up and visualization

Model summary The PVWatts PV model (Dobos, 2014) estimates the DC PV power $p_{PV,t}$ production at time t according to Equation 5.7,

$$p_{PV,t} = \frac{POA_{eff}(\theta,\phi)}{G_{stc}} \times P_{PV} \times (1 + \gamma_{pdc}(T_{module,t} - T_{stc})), \tag{5.7}$$

where $POA_{eff}(\theta,\phi)$ is a function of the tilt angle θ and the azimuth angle ϕ of the installation, P_{PV} is the installation's installed capacity and γ_{pdc} is an efficiency factor that reflects the decrease in the module's performance with the temperature. The temperature corresponding to the standard test conditions is 25°C and γ_{pdc} to -0.002 K^{-1}.

Visual check Figure 5.5 presents the estimation of some generation curves using this model. This small example illustrates that our estimation is well calibrated: the PV power production is well estimated, and there are no lags in the temporal variables.

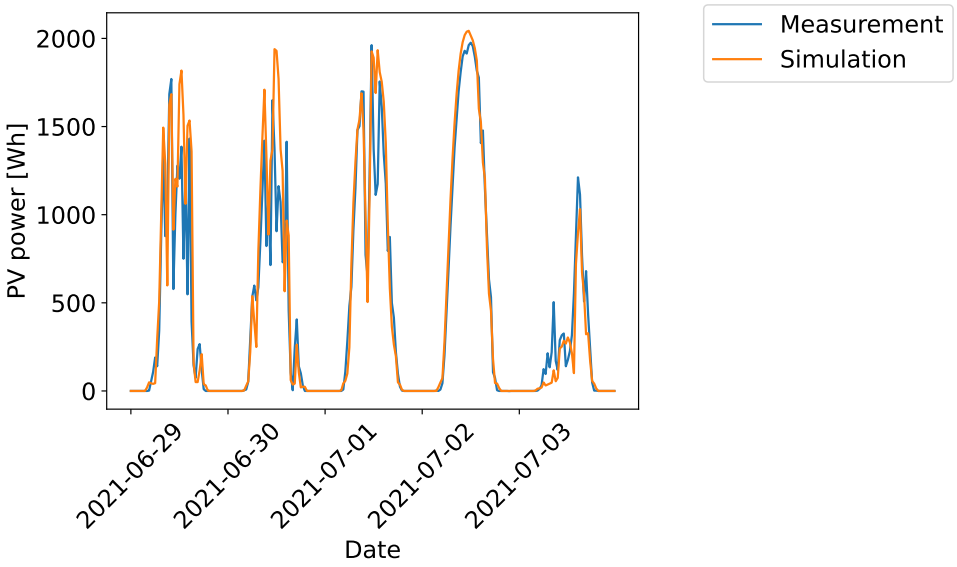


Figure 5.5 – Power curves generated with our conversion model plotted against the corresponding ground truth measurement for one installation.

2.3 Assessing the relevance for improving PV observability

Assessing the relevance of our approach for improving rooftop PV observability requires addressing three questions: First, how accurately does our approach estimate the individual rooftop PV power production? Second, does this approach improve over existing approaches? Finally, how does this approach scale up? Indeed, in practice, we aggregate individual estimations for thousands of PV systems.

2.3.1 Assessing the accuracy and relevance of our approach

To assess whether our approach is accurate and relevant for improving PV observability, we compare our approach with ground truth measurements and define a set of baselines against which we compare our model. We define this set of baselines because approaches typically used to infer PV production and presented in the introduction rely on the assumption that there are many PV systems in a given grid point, which does not hold with our ground truth measurements.

The information-free baseline We consider an information-free or naive baseline consisting in parameterizing the conversion model with default values of 30° for the tilt angle and 180° for the azimuth angle. We always consider that the installed capacity is known as it is the only information accessible at the disaggregated scale for RTE. This approach is inspired by the method of Walch et al. (2020) who derive PV power measurements from solar irradiance data with a workflow that assumes a tilt angle of 15° or 0° depending on the surface of the installation and an azimuth angle of 180° We refer the reader to Walch et al. (2020) for more details on their modeling approach.

Statistical or implicit baselines We consider another instance of baselines, which we call *statistical*. These approaches aim to evaluate the relevance of our proposed modeling approach in an operational setting and obtain comparisons in terms of accuracy with simpler approaches. For this statistical baseline, we make the strong assumption that we have access to PV power measurement records and can estimate the PV power production using a model trained on these metered installations.

Our statistical models are trained on a held-out dataset of PV power measurements coming from BDPV. We take as input the solar irradiance and temperature data, the sun position, and the installations' installed capacity. We also refer to this approach as implicit because the parameters of the conversion model are inferred during training. We consider two instances of models: a linear regression and a neural network. Our neural network model is a simple one-hidden layer network with 128 hidden neurons. We trained it using a learning rate of 0.001 over 15 epochs (iterations on the whole training dataset) and a mean squared error (MSE) loss.

Table 5.5 presents the RMSE and pRMSE reached by our statistical models after training on the test set. We looked for the best model hyperparameters for the neural network through grid search.

Comparison with the Oracle We evaluate all our approaches against an Oracle, that is, our conversion model parameterized with the actual tilt and azimuth values taken from BDPV. We include this Oracle as an indication of the upper bound on

Table 5.5 – Accuracy (RMSE in [W] and pRMSE in [%] in parenthesis) of the statistical models considered in this study on their test dataset. n indicates the number of samples (here: number of power measurements in the test set).

Model	RMSE [W] (pRMSE) [%]	n
Linear regression	666.24 (8.86)	5705320
Neural network	986.28 (11.41)	5705320

accuracy that it is possible to achieve with our approach, combining a conversion model and weather data. The error of the Oracle encompasses the uncertainties inherent to CAMS solar radiation, the ERA5 air temperature, and the conversion model, neglected factors in our modeling (shadows), and errors in the tilt and azimuth angles reported in BDPV.

2.3.2 Scaling-up: evaluating the impact of aggregation on the accuracy of the PV power estimation

Definitions We define the **characterization error** as the effect of an error in the estimation of the PV system's parameters for estimating the PV power production. We denote ϕ^* and θ^* as the true parameters, assumed to be those imputed in BDPV. Given an installation, we denote $[\underline{x}, \bar{x}]$ the range of perturbations around a variable x and $p^*_{PV}(t)$ and $p^x_{PV}(t)$ the PV power estimation at time t with the true parameters and with the parameters x, respectively. We focus on the estimation error of the tilt and azimuth angles, as we assume that the true installed capacity is known.

Estimation of the individual characterization error We study the behavior of the estimation error as the number of plants included in the estimation of the PV power production increases. Saint-Drenan et al. (2016), decomposes the RMSE of the aggregated PV power estimation as

$$RMSE = \sqrt{\frac{1}{N_t}\sum_{t=1}^{N_t}\left(\frac{\sigma_0^2(t)}{N} + \mu_0^2(t)\right)}, \tag{5.8}$$

where N denotes the number of plants, N_t the number of time steps, and $\sigma_0(t)$ and $\mu_0(t)$ the standard deviation and the mean of the original power plants at time t. It follows that from Equation 5.8,

$$RMSE \overset{n\to\infty}{\longrightarrow} \sqrt{\frac{1}{N_t}\sum_{t=1}^{N_t}\mu_0^2(t)}, \tag{5.9}$$

if the individual estimation errors are independent (central limit theorem). The sources of error in the framework of Saint-Drenan et al. (2016) are the irradiation variability and the variability in plant characteristics. Our framework adds an additional source of variability, the variability in the characterization error of the PV plants. This source of error should also be independent if one wants the RMSE to converge towards the population mean.

We introduce a framework to empirically study how the aggregated error behaves when the characterization error is not independent from one system to another, i.e., that it has systematic biases. We then identify which regime DeepPVMapper belongs to empirically assess how the PV power estimation error will behave with our methodology.

Evaluation framework We consider the simulation p^*_{PV} computed with the true parameters, as the reference. Then, we define ranges of perturbations around these true values. To avoid irrelevant values, the range of perturbations is bounded by 0 and 90 for the tilt angle and remains in the [-180, 180] interval for the azimuth angle. For each combination $(\phi, \theta) \in [\underline{\phi}, \bar{\phi}] \times [\underline{\theta}, \bar{\theta}]$, we simulate the PV power production and compute the pRMSE between p^*_{PV} and $p^{\phi,\theta}_{PV}$. We iterate through all installations contained in our database.

Figure 5.6 presents examples of the error matrices that we obtain for individual installations. We can see that the error is at its lowest when the parameters correspond to the true parameters. By definition when $(\phi, \theta) = (\phi^*, \theta^*)$ the error is null. We can also see that, as expected, when $\theta = 0$, the error does not depend on the azimuth angle.

Aggregation to a fleet of systems We aggregate power curves generated with wrong configurations to study the effect of the individual characterization error on the aggregated PV power production. We assume that all other sources of error remain identical so that the only variation in the error is caused by a variation in the PV system's parameters. We list all possible biases and label them from "No bias" to "VIII." Table 5.6 summarizes these cases.

We first consider a given number of installations n. Then, we generate a bias during the aggregation by picking tilt and azimuth values for each installation $i \in \{1, \dots, n\}$. We constrain the way of picking tilt and azimuth values depending on the bias case we are interested in. For instance, if we are interested in case **(II)**, we pick tilt values in $[\underline{\theta}, \bar{\theta}]$ (no bias in the estimation of the tilt) and azimuth values in $[\phi^*, \bar{\phi}]$ (upward bias in the estimation of the azimuth angle).

Then, we compute and aggregate the PV power production using the sampled values and compare the error of the aggregated production with the true production. The number of installations n we consider for our aggregation varies from 6

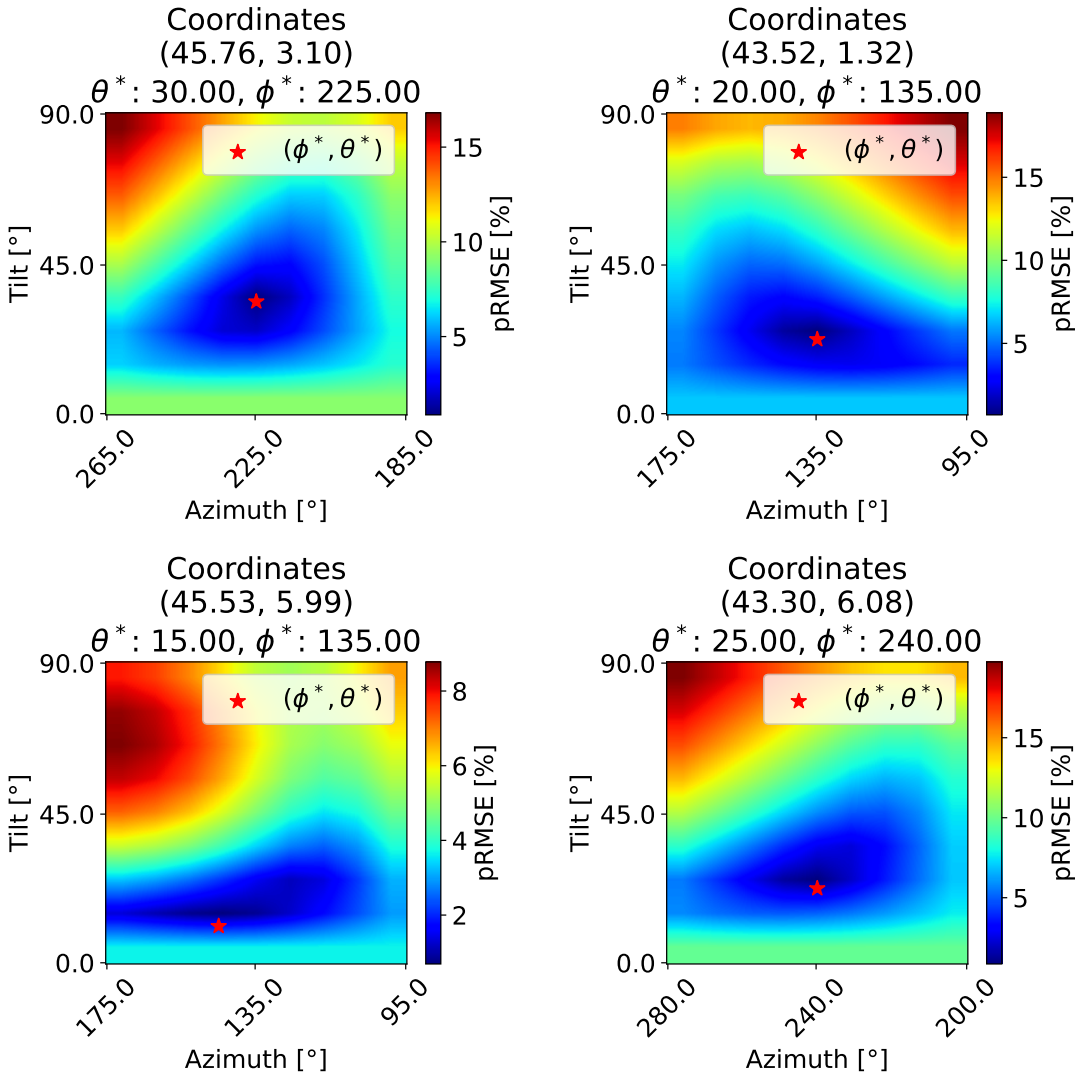


Figure 5.6 – Examples of individual characterization error. The red star shows the localization of the true parameter values, where the characterization error is equal to 0 by construction.

to 24. We defined these clusters of installations from our database by considering nearby installations.

Table 5.6 – Summary of the cases for aggregating the characterization errors from the installation level to the representative cell level.

Case	Direction of the bias on ϕ	Direction of the bias on θ
No bias	None	None
(I)	Negative	None
(II)	Positive	None
(III)	None	Negative
(IV)	None	Positive
(V)	Positive	Positive
(VI)	Positive	Negative
(VII)	Negative	Positive
(VIII)	Negative	Negative

3 Our approach paves the way towards better PV observability

In this section, we present our results. We first stress out the added value of estimating rooftop PV characteristics for estimating the PV power production, at the individual level. We then underline some conditions that need to be satisfied for the estimation to remain accurate when we aggregate the installations. Finally, we lay out directions towards comparing our method with the TSO's current methods for estimating the PV power production.

3.1 Improving rooftop PV observability

3.1.1 Rooftop PV is observable

Table 5.7 shows the results of the comparison of our method with alternative baselines. The main result is that it is possible to improve rooftop PV observability. We can derive accurate rooftop PV power measurements at the installation scale using a simple conversion model and limited information on the PV system. The estimation error, measured by the pRMSE, is about 10%. This approach does not require access to ground truth PV power measurements and can be used as a first approach to estimate or reconstruct rooftop PV power measurements.

The error between our approach, the information-free approach, and the Oracle are not statistically significant: this shows that the parameterization has little effect on the estimation error compared to other factors, which are not accounted for in this study (e.g., the shadings).

We can see that the linear regression approach performs surprisingly well. The difference between our approach and the linear regression is not significant. This result is two-sided: on the one hand, it tells us that our method achieves about the

Table 5.7 – Comparison of the RMSE [W] and pRMSE [%] (in parenthesis) of the estimation at the individual installation scale with parameters from DeepPVMapper. Best results are **bolded** and second best underlined. n indicates the number of installations used in this study.

	Case	Min [W]	Max [W]	Mean [W]	Median [W]	n
	Oracle	114.61 (3.90)	2137.82 (26.49)	281.53 (8.36)	223.06 (7.66)	255 255
Explicit	DeepPVMapper	**119.56** **(4.15)**	3001.42 (43.39)	**332.57** **(10.10)**	**245.33** **(8.18)**	255 255
Explicit	Information-free	147.43 (6.15)	**2972.53** **(26.15)**	353.63 (10.37)	283.37 (9.90)	255 255
Statistical	Linear regression	134.42 (4.67)	7663.42 (33.27)	392.97 (10.18)	257.21 (8.86)	255 255
Statistical	Neural Network	408.24 (8.67)	9182.32 (32.78)	749.90 (21.11)	609.91 (20.94)	255 255

same accuracy as a linear regression[1]. On the other hand, it underlines that *if* PV power measurements are available, our conversion model does not bring significant improvements with a statistical approach calibrated on these measurements. The central assumption for this second result is that the training dataset is representative of the test dataset.

3.1.2 Temporal and spatial patterns

Seasonal patterns Figure 5.7 decomposes the pRMSE of the estimation of the PV power production with DeepPVMapper and the conversion model for each time of the day. A time-of-the-day (TOD) timestep corresponds to a 30-minute interval. For each TOD timestep and each installation, we compute the pRMSE of the PV power estimation. We obtain 255 estimations for each time interval, enabling us to derive a global mean and median (blue and orange curves, respectively) and interquartile ranges showing the dispersion of the errors across installations as a function of the TOD timestep.

On the "all year" chart, we only filter dates according to their hour of the day. We can see that the error is null during the night and increases to reach a peak during the day.

On the bottom charts, we filter the dates to keep only the Winter and Summer months. Summer months correspond to the months from June to September, and

1. It is not entirely unexpected as the PVWatts conversion model is fundamentally a linear transformation of the temperature and the G_{POA}.

Winter months from December to March (4 months each).

We can see that the error is largest during winter and is high throughout the day. On the other hand, during Summer, the error is high at dawn and dusk but decreases around noon.

The poorer performance of our model during winter, dawn, and dusk may be attributable to the fact that we do not take shadings into account. Indeed, Walch et al. (2020) reported a similar behavior with their model, which did not consider shadings. Another possible explanation is that our model neglects the self-consumption of power inverters and the behavior of the modules with low luminance.

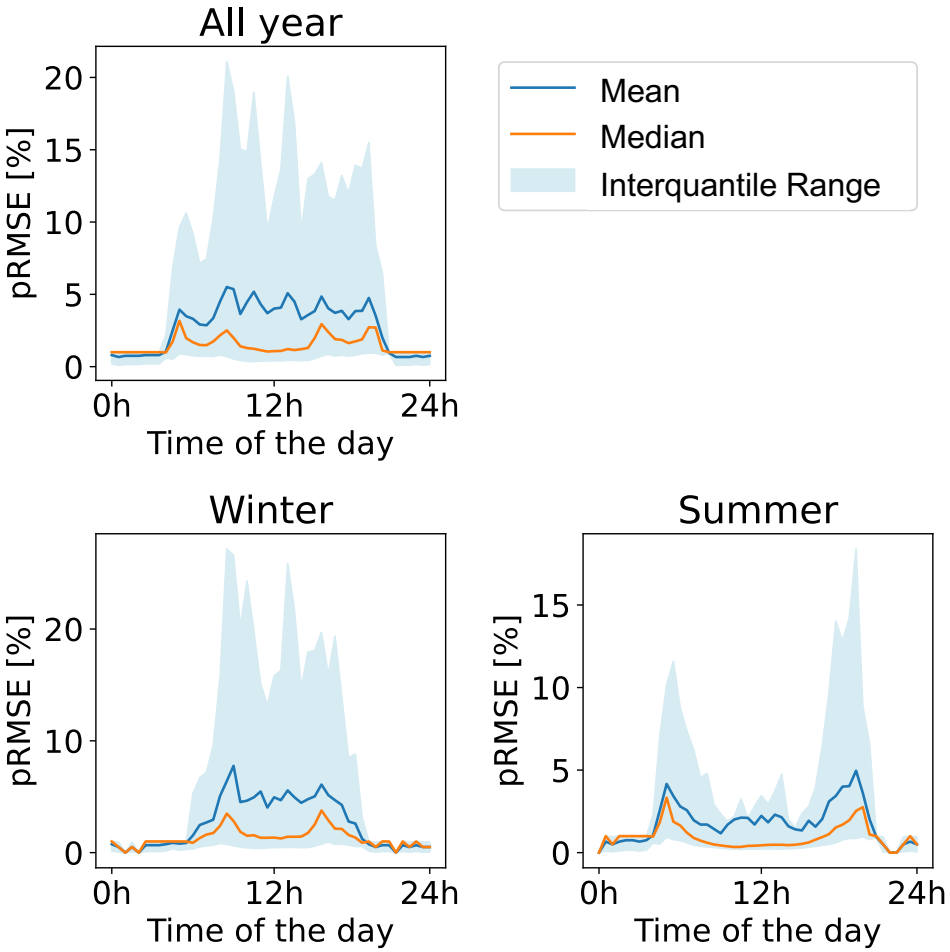


Figure 5.7 – 30-minute pRMSE of the estimation of the PV power production using DeepPVMapper and the conversion model. The interquantile range plots the range between the 5th and 95th percentile.

Geographical variability Figure 5.8 plots the pRMSE for each installation depending on its recorded localization. The error in the East of France is higher than in the West of France. It could again be a consequence of the absence of shadings in the model. Indeed, Eastern France is more hilly than Western France, so shadings may be more critical. It could also be caused by differences in solar irradiation be-

tween the West and East due to differences in the types of climates between these regions. The geographical variability of the Oracle (see appendix C, section 4) is similar.

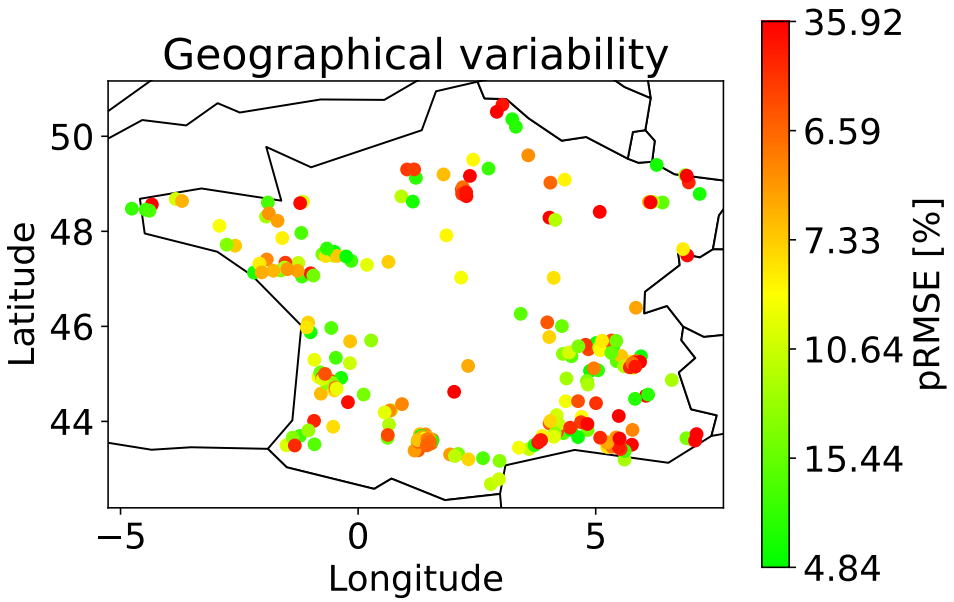


Figure 5.8 – Geographical variability of the pRMSE [%] of the PV power estimation depending on the localization of the installation.

3.2 The impact of characteristics estimation biases on accuracy

Table 5.8 – pRMSE [%] of the aggregated rooftop PV power production estimation under different estimation biases of the PV systems' characteristics.

Case	Min	Max	Mean	Median	n
No bias	2.68	9.59	5.59	5.35	461
(I) Negative bias on the azimuth	4.03	15.18	9.88	9.38	461
(II) Positive bias on the azimuth	1.76	5.81	2.57	2.49	461
(III) Negative bias on the tilt	3.10	9.55	6.30	6.10	461
(IV) Positive bias on the tilt	3.01	10.03	6.01	5.72	461
(V) Positive bias on both	1.85	5.70	3.02	2.97	461
(VII) Negative on azimuth, positive on tilt	1.93	7.24	3.36	3.28	461
(VIII) Negative on both	4.82	16.07	10.66	10.16	461
Oracle for an individual installation	*3.90*	*26.49*	*8.36*	*7.66*	255

Overall results To estimate the effect of the aggregation on the characterization error, consider the 9 cases described in section 2.3.2. Table 5.8 presents the results. We can see that the average characterization error is the largest when there is a negative sampling bias on the tilt or the azimuth. A "negative" bias means that

the azimuth angle is estimated slightly westwards compared to the true value for the azimuth. On the other hand, a positive bias on either the tilt or the azimuth is not a problem. Most of the sensitivity comes from misspecification of the azimuth angle rather than the tilt angle, in line with Saint-Drenan et al. (2015). Compared to the estimation of the PV power production for a single installation, aggregation nonetheless brings improvements, except in the worst cases. Our results suggest some errors in the tilt and azimuth angles reported in BDPV. For instance, a positive bias (i.e., the azimuth pointing more eastwards than in reality) entails a lower pRMSE than without bias.

Aggregation dynamics Table 5.8 discusses static results. We then considered the effect of increasing the number of installations in the sample on the characterization error, depending on the latent estimation bias. Figure 5.9 plots the results for the unbiased case, the worst cases **(I)** and **(VIII)** and the best cases **(II)** and **(V)**. On the charts of Figure 5.9, we plot the pRMSE as a function of the number of installations. We generated increasingly larger clusters of installations based on their geographical proximity. A limitation of this approach is that we have more small clusters with six installations than large clusters with 24 installations. Therefore, estimating the standard deviation of the pRMSE in large clusters is not necessarily possible. Nevertheless, we compute the mean and median pRMSE and, if relevant, the standard deviations. Surprisingly, in the unbiased case, the increase in the number of installations does not decrease the overall PV power estimation error. This decrease only happens in the case **(II)**.

Our results show no decrease in the RMSE when we aggregate the power curves as we would have expected, except in one case. However, we should note that we deal with very small sample sizes. Our study indicates that the error does not necessarily decrease as the sample size increases *for small sample sizes and* that the aggregated PV power production error depends on the latent sampling bias. The magnitude of the effect of the latent sampling bias can be sizeable.

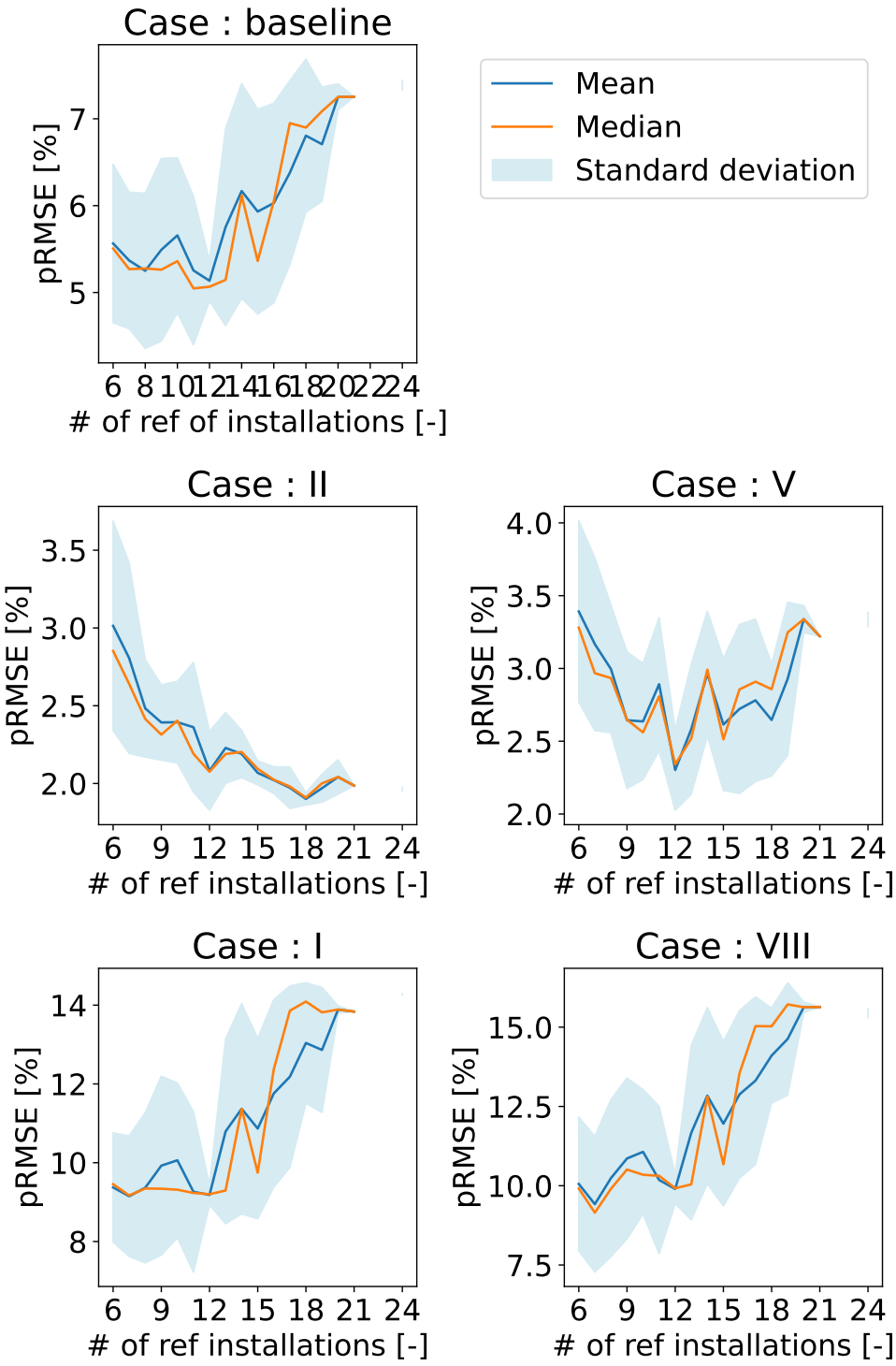


Figure 5.9 – Characterization errors as a function of the number of installations for the unbiased case and a set of worst (**(I)** and **(VIII)**) and best cases (**(II)** and **(V)**).

3.2.1 On the relevance of the DTA

This study of the effect of systematic biases on the aggregation of individual PV power curve estimations stresses the relevance of monitoring the model's outputs with tools such as the DTA. In chapter 2 section 3.2.1, we compared the distribution of the tilt and azimuth angles estimated with our mapping algorithm with the reference values recorded in BDPV.

We found no evidence of a systematic bias in estimating the azimuth angle (the estimation is slightly more concentrated than the true values). However, the model slightly overestimates the tilt angle (see the left panel of Figure 2.12, the mean tilt angle is slightly above the mean value of BDPV, and we estimate less small tilt angles than in the reference).

This leads us to consider that with DeepPVMapper; we will aggregate tilt angles as in case **IV** of Table 5.8: we have no bias in the azimuth angle, but a positive bias on the tilt. Figure 5.10 provides us insights into the behavior of the error in the aggregation of the estimations of the PV power curves.

We can see that under the estimation bias of DeepPVMapper, our results do not indicate that the error will decrease when aggregating the PV power curves. As this should be interpreted carefully, the error should remain within a five-percentage margin around the individual installation error. This means that the error at the aggregated scale of the estimation of the PV power production is expected to be around 10%. Compared to the other bias cases, in the case of DeepPVMapper, we are in a favorable situation where the error in the PV power production estimation remains moderate.

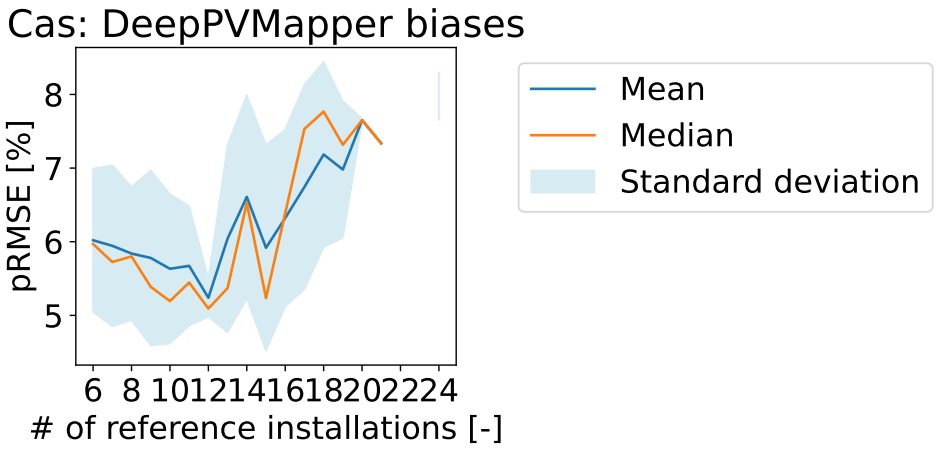


Figure 5.10 – Behavior of the error of the aggregation of the estimations of the PV power curves in the case of tilt and azimuth angles estimated with DeepPVMapper.

3.3 Broader impact: closing the gap with the TSO's aggregated approaches

Differences of scales In this study, we analyzed the behavior of a conversion model-based estimation of the PV power production at the scale of the individual power system. At most, we consider aggregation of PV systems comprising up to 20 systems. On the other hand, current methods for improving PV observability, such as the probabilistic approach (Saint-Drenan, 2016), assume that the density of installations is much higher. The aggregation results hold, provided we deal with hundreds of PV systems.

Avoiding breaking assumptions We chose not to directly implement the probabilistic approach as its assumptions regarding the number of installations would not hold. Instead, we evaluated our approach with individual-level PV power estimation and approaches that do not require assumptions on the number of installations to work.

Towards bridging the gap If one were to compare our approach with current methods used by the TSO, it is necessary to scale our approach by computing the PV power production for all registered systems in our registry. It would then be possible to implement a method such as the probabilistic method. However, there will not be ground truth measurements to assess which approach is the best at this scale; only intercomparisons are possible. Most importantly, we demonstrated in this study that it is possible to accurately estimate the PV power production of rooftop PV systems with a scalable approach whose error remains stable when scaling up.

Time horizons Finally, we focused on real-time rooftop PV power production estimation. We used reanalysis and real-time data to minimize the share of the PV power estimation that can be attributed to the weather data. However, to fully address observability, we need to be able to compute forecasts of the PV power production. To do this, we can extend our approach using forecasting weather data. The difference in accuracy observed between our results and the forecasting results will be attributable to the weather data.

Conclusion of the chapter

In this chapter, we evaluated the relevance of the registry for improving rooftop PV observability. We defined observability as the ability of the TSO to estimate a power unit's real-time and future production accurately. To evaluate the relevance

of the registry, we need to address the following questions: First, how accurately can we estimate the rooftop PV power production with information derived from the PV registry? Second, how does this estimation compare with competing approaches? Third, can it scale up to thousands of PV systems?

We leveraged ground truth measurements of individual rooftop PV systems to address these questions. We had curated information for about 900 individual PV systems, along with their PV characteristics. We also used solar irradiation data from the Copernicus Atmospheric Monitoring Services (CAMS) and the European Centre for Medium-Range Weather Forecasts (ECMWF) temperature data. As we only have limited information - tilt and azimuth angles, installed capacity, solar irradiance, and temperature data - at the scale of the country and potentially have thousands of systems to consider, we implemented the model from Dobos (2014). We compared our approach with information-free modeling that considers constant values for the tilt and azimuth angles and statistical approaches that do not require knowing the tilt and azimuth angles.

Our results showed that our system-based modeling approach allows an accurate estimation of the individual PV power production. Our results did not show significant differences in accuracy between different parameterizations of the conversion model (whether using the true parameters, parameters from DeepPVMapper, or constant parameters).

Studying how our approach scales up meant studying how estimation biases in the PV systems' parameters compensate or accumulate when considering the sum of individual PV power curves. We proposed a framework to study the effect of these biases on the PV power estimation and empirically showed that, in our case, errors do not accumulate too much when the number of installations increases. However, due to our ground truth data limitations, we could only scale up to tens of PV systems, not thousands. We laid out directions to continue scaling up our approach and adequately compare it with the TSO's current methods.

Chapter 6

Conclusion and discussion

1 Conclusion

1.1 Answer to the scientific question

The fact that the transmission system operator (TSO) in France lacks precise power production measurements for about 22% of the photovoltaic (PV) installed capacity motivated this thesis work. Improving the so-called observability of rooftop PV power production is crucial in the context of the sharp increase in the PV installed capacity. To this end, constructing a technical registry of rooftop PV installations (i.e., installations with an installed capacity lower than 36 kW_p) is necessary. This registry aims to provide the necessary technical characteristics of the PV systems for accurately estimating their power production: localization, tilt and azimuth angles, and installed capacity. Given the high number of installations to map –about 600,000 in France– remote sensing on aerial orthoimagery using deep learning methods appears to be the best methodology for this task.

However, current state-of-the-art methods are not reliable enough for our industrial problem, which led me to formulate the following scientific question: **is deep learning-based remote sensing on orthoimagery a suitable method for constructing a nationwide registry of rooftop photovoltaic (PV) installations intended to improve the observability of PV power production in France?**

Methodology To address this question, I first defined reliability as the combination of three components: the ability of a user to monitor the model's output, audit its decision process, and enforce a robust decision process. Based on this definition, I defined what requirements the registry should satisfy and how I could check that it met these requirements (sub-question (SQ) 1). Second, how could we ensure that a deep learning model reliably mapped PV panels (SQ2)? Finally, I built a reliable

algorithm capable of producing a registry meeting the reliability requirements, and I introduced a methodology to assess its relevance for improving PV observability (SQ3).

(SQ1) What requirements should the data have, and how can we check whether the registry meets these requirements? In chapter 2, I established that the registry's main requirement is to correctly reflect the spatial distribution of rooftop PV installed capacity. Second, the tilt angle distribution must vary with the PV system's latitude, and the azimuth angle distribution should match the true distribution of the azimuth angle. As I had access to the distribution of the installed capacity, aggregated at the city level and a self-reported dataset that features the tilt and azimuth angles of the installations, I leveraged these datasets, namely the *registre national d'installation* (RNI) and BDPV to define the *downstream task accuracy* (DTA). The DTA evaluates the accuracy of the registry by comparing aggregations from the registry with available data. The DTA does not require labeled data and is available for the whole territory. Therefore, it enables monitoring the model outputs and pointing the user toward the failure cases directly.

I evaluated the existing state-of-the-art PV mapping algorithms. I highlighted a 30 percentage points accuracy drop (regarding the estimation of the installed capacity at the city level) between the training dataset and the mapping area (i.e., the territory where we deploy the model to map PV installations). This drop is very uneven as the accuracy can drop dramatically in some cities, while in others, the registry is very accurate. A question arising is why such a drop occurs and why there is such heterogeneity. I found no evidence of a geographical pattern explaining the loss of accuracy. Using the GradCAM, a well-established feature attribution method, I then visually compared the important areas on the input images depending on the models' decisions. This analysis led me to formulate a hypothesis regarding the model's decision process: to predict a PV panel, a model relies on a limited set of characteristics correlated with the PV panels of the training dataset. Given an image, the model will predict the PV panel if the *right* feature appears on the image.

(SQ2) How can we ensure that deep learning models reliably map rooftop PV installations? The hypothesis formulated in chapter 2 required the introduction of a new feature attribution method to understand better *what* models see on input images (through the lenses of scales) and not only to assess *where* they look. In chapter 3, I expanded existing feature attribution methods from the pixel domain to the space-scale domain. Rather than perturbing input images and evaluating the model's response to these perturbations in the pixel domain, I perturbed the wavelet transform of the images and evaluated the model's response. The proposed method is the wavelet scale attribution method (WCAM). Using the WCAM, I could assess the reliability of the decision process as it highlights which *scales*

are important in the model's decision. Besides, scales simultaneously correspond to structural elements of the image, enabling the user to interpret the model's decision and dyadic frequency ranges, characterizing the potential "brittleness" of these features to perturbations of the input images.

I then set up an experiment to understand what mainly caused the drop in accuracy between the change in the acquisition conditions, the varying ground sampling distance, and the geographical backgrounds. This experiment showed that the acquisition conditions were the main cause for the drop in accuracy, driven by a rise in false negatives. Using the WCAM, I could show that panels were no longer recognized when the acquisition conditions changed because these acquisition conditions disrupted important high-frequency patterns such as the gridded common among many PV systems. This result comforts the working hypothesis established in chapter 2. I then introduced a model based on Gaussian noise and blur to reproduce the effects of varying acquisition conditions and proposed a data augmentation technique to reduce the sensitivity of deep learning models to varying acquisition conditions. This method outperformed existing techniques on our benchmark dataset.

The new feature attribution (the WCAM) paves the way for a finer understanding of the decision process of a model by disentangling it into different scales while maintaining the understandnig on the localization. The prediction between two images will be similar not based on their similarity in the image space but on their similarity in the space-scale space. To avoid such behavior leading to too many false detections, I introduced a data augmentation that lowers the reliance on the most fragile components of the image (i.e., the finest scales or highest frequencies).

(SQ3) How to build and integrate the registry for rooftop PV power production estimation and evaluate its relevance for improving PV observability?
First, in chapter 4, I reviewed the current state-of-the-art algorithms for mapping PV installations and identified where I could improve the algorithm's reliability. Improvements could be made to the characteristics extraction part of the algorithm, the image preprocessing process, and the evaluation metrics. To improve the characteristics extraction part of the algorithm, I introduced a standardized approach for extracting PV characteristics. I then improved the image preprocessing method so that it minimizes the occurrence of false negatives by inducing an overlap between the generated thumbnails and minimizes the occurrence of false negatives by extracting images only in relevant (i.e., anthropized) areas. Finally, I argued that an evaluation of the model's performance using the DTA rather than the F1-Score or the Intersection-over-Union should be preferred, given our goals with the mapping algorithm. I evaluated a wide range of classification models, data augmentation techniques, and algorithm architectures on my new benchmark.

The results also indicated that data augmentation strategies based on the blur-

ring and the perturbation of the wavelet transform of the image do not significantly improve the accuracy in operational conditions. Nevertheless, our proposed approach outperforms the existing state-of-the-art algorithms for mapping rooftop PV installations, and the DTA and the WCAM provide monitoring and auditing tools to the user that increase the trust in the model and the reliability of the registry.

Our industrial goal was to derive accurate PV power estimations of the rooftop PV fleet; I had access to PV power production measurements of 900 curated rooftop PV systems. In chapter 5, I introduced a simplified conversion model that uses the characteristics obtained from our registry and solar irradiance and temperature data to derive PV power estimations at the installation level. I stated that the observability of rooftop PV installations would be increased if this approach could accurately estimate the PV power production and scale up to thousands of installations. On the one hand, it turns out that estimating the rooftop PV power production using the characteristics provided by our technical registry and weather data accurately estimates the individual power curves and seems to scale well, at least at the scale of a couple of dozen systems. This result shows that rooftop PV observability can be improved by acquiring few information on rooftop PV systems, namely the tilt and azimuth angles, the localization, and the installed capacity. The performance of this model-based approach is about the same as a linear regression, calibrated on rooftop PV plants. Our study shows that the model-based approach used with the technical registry is a good starting point if one has no information nor measurements on the rooftop PV systems, which is generally the case.

Answer to the main question The scientific question was whether deep learning-based remote sensing on orthoimagery could be a suitable method for constructing a nationwide registry of rooftop photovoltaic (PV) installations intended to improve the observability of PV power production in France. More broadly, it raises the question of whether deep learning models are mature enough to be safely used in industrial pipelines. This work's central contribution is identifying quality and dependability standards and proposing a methodology to verify that the deep learning model and the generated data meet these standards. The necessary conditions are the ability to monitor the model's data and to audit its decision process. Therefore, it is necessary to have complementary data and to define relevant KPIs against which the data produced by the deep learning model will be monitored. Standard feature attribution techniques are insufficient for auditing the model's decision as they do not assess what models see. Our WCAM provides a first step towards addressing this issue. Finally, having a robust and accurate model is desirable but insufficient to achieve the required level of trust in the data and the decision process, as user's trust comes from his or her ability to monitor the data and audit the model. Therefore, deep learning and Earth observation data are suitable because one has enough additional data to monitor the model during its deployment.

Answer to the industrial question This work contributes to improving PV observability to the extent that it provides valuable additional information regarding the geographical distribution of small-scale PV systems. It also enables an accurate estimation of PV power production, but its relevance in an operational setting is questionable, as the model-based method does not significantly outperform alternative methods. Nonetheless, our assessment of the gains for improving rooftop PV observability can be complemented by collecting more reference data to empirically demonstrate our approach's greater accuracy compared to the TSO's current practices.

1.2 Discussion

Finding where models fail and stepping out of the "black room" When starting the project, I was left with numerous approaches, some very specific for addressing one issue. At the same time, nothing had been published regarding rooftop PV panels mapping in France, and no training data was available. This contrast encouraged me first to gather training data and deploy a prototype in France, which I could improve afterward by identifying its most critical issues. Discussions with potential users about the first results of this prototype showed how the standard accuracy metrics failed to address these users' concerns regarding the model. At least they wanted to evaluate it against existing data sources (which led to the standardization of this process with the DTA), and the most skeptical users wanted to assess whether the model recognized PV panels or not some spurious factors. The GradCAM was enough to rule out obvious spurious correlations such as swimming pools but not enough to explain why the model took a track and field for a PV panel.

I believe that for applied works, off-the-shelf models are sufficient for most of the use cases or at least for building prototypes. Besides, the posterity of seeking state-of-the-art performance seems very limited: deep learning is a quick-evolving field, the rate of progress is sometimes absurdly fast, and it is only a matter of months (or even weeks) before someone surpasses the proposed methodology.

On the other hand, addressing the simple question "Does the model work properly?" and deriving a protocol for monitoring and auditing deep models when they fail turned out to be a dense journey. I am convinced there is still much to do in what some would call AI auditing, especially in the context of the everyday use of deep learning models by non-specialists in numerous workflows. Non-specialist humans are increasingly interacting with deep learning models, and I think that providing them with the right tools to enable them to understand better what models really do, how they work, and what are their limitations can improve the overall trust and critical thinking towards these systems, indispensable for a sound integration into many workflows.

Towards a reasoned use of deep learning I think that another key parameter to foster deep learning models beyond fancy APIs and computer science benchmarks is to show that deep learning researchers are aware of the general public's concerns. In the context of improving rooftop PV observability, I could not ignore the environmental impact of my methods. I present in appendix A a study of the environmental impact of our approach. Our framework aims to encourage researchers to report the environmental impact of their models more systematically and consider this impact as a choice variable. Our simplified framework quantified DeepPVMapper's energy cost represents about half of the yearly expected production of an average individual 3 kW_p rooftop PV system in France [1] and showed that most of the energy was consumed during the inference.

I was also struck by the impact of the grid intensity on the resulting environmental impact. The impact of the algorithm ranges between 75 to more than 1000 kg CO_2e, i.e., depending on the localization of the server's localization, mapping PV installations in France can be equivalent to the production of 2 kg of beef or one round trip travel between Paris and New York. Therefore, the decarbonization of the grid contributes to reducing the environmental impact of deep learning.

I was delighted to see that even if environmental reports are not yet standard practice, the machine learning community is well aware of the environmental issues with deep learning. Many research efforts have been put into assessing the environmental impact of deep learning and deep learning-based systems, and much work remains to be done to provide a comprehensive assessment of the environmental impact of deep learning.

1.3 Contributions

Fields of contributions This thesis work contributes to two main domains: deep learning and power systems. I leveraged the case study of mapping small-scale PV installations to evaluate under which conditions we could reliably use deep learning models. To this end, I introduced a methodology for monitoring a model's output using indirect measurements, enabling the evaluation of the quality of the data produced by the model when no labels are available (Kasmi et al., 2022a). I also introduced a novel attribution method that decomposes a model's prediction into scales using the wavelet transform and show that this method can effectively improve the reliability of deep learning models by providing a finer interpretation of their decision process (Kasmi et al., 2023a). This attribution method contributes to closing the gap in assessing *what* deep learning algorithms see on input images. Finally, I introduced a novel training dataset, BDAPPV (Kasmi et al., 2023d), containing nearly 50,000 annotated images and coming from two image providers. This enables to train models for mapping PV installations in France but can also serve as

1. Assuming that the yearly expected yield for a 3 kW_p installation in France lies between 2.5 (North) and 3.5 (South) MWh/year (BDPV, 2023).

a benchmark for evaluating the robustness to varying acquisition conditions of *any* model, as done by Guo et al. (2024).

On the power system's side, I contributed to improving the knowledge of the rooftop PV fleet by mapping rooftop PV installations over 38 French departements (at the time of writing of this thesis, and ultimately to the whole metropolitain France) using deep learning and orthoimagery. The registry records the localization, tilt, and azimuth angles and the nameplate capacity of the installation. The proposed registry is currently the world's second-largest for small-scale PV and the world's largest with this level of detail. I also discussed how we could improve rooftop PV observability using the data from this registry (Kasmi et al., 2024). Overall, this work improved the observability of the French rooftop PV installations. The tools introduced in this thesis can be applied in other countries where the same issue arises. The only requirement is access to the true installed capacity at the desired (e.g., the city) level.

Additional contributions of this thesis work include an open-source Python library for extracting rooftop PV characteristics from geolocalized polygons (Trémenbert et al., 2023), and DeepPVMapper, an open-source mapping algorithm that can be reused and improved by anyone (Kasmi et al., 2023c).

Applications for RTE and beyond For the TSO, the proposed modeling approach of the PV power production shows that it is possible to derive accurate measurements of the rooftop PV power production using limited information regarding the PV systems. This approach paves the way for improving the rooftop PV power estimation and, thus, the overall PV power estimation at different spatial and temporal scales, from individual estimations to nationwide aggregates and from reanalysis to forecasts.

The registry provides a current view of the PV fleet. Therefore, the TSO can use it to calibrate the PV potential models used in prospective studies. It can also be used to analyze the geographical, social, and economic drivers behind PV adoption, as done in works such as Wang et al. (2022) or Freitas et al. (2023). Such registries can also be relevant for public authorities seeking a straightforward way of assessing the current state of PV deployment over their territory.

2 Limitations

2.1 On the power system's side...

Fair comparisons with regional PV power modeling Despite leveraging a large ground truth rooftop PV measurements dataset, I could not compare our approach with standard regional PV power estimation methods such as the probabilistic approach. The probabilistic approach estimates the PV power production,

aggregated at the scale of a small area (typically a grid point of a couple tens of km^2), using the distribution of tilt and azimuth angles of the installations located in this small area. For the empirical distribution of tilt and azimuths to be statistically representative, one needs at least a thousand installations in this gridpoint (Saint-Drenan et al., 2016). In our case study, I had ground truth PV power measurements and installations' characteristics for at most twenty installations in such constrained areas, too few to properly implement the probabilistic approach.

Coarse mapping Our goal was to design a method enabling the construction of a registry over France. The resulting registry roughly characterizes the rooftop PV fleet and makes several assumptions regarding the PV fleet. First, I did not distinguish PV technologies (e.g., monocrystalline or polycrystalline) nor distinguish PV panels from solar thermal panels. I also assumed that the efficiency of the PV panels was constant for all PV panels. Finally, I did not rely on LiDAR data to compute the tilt and azimuth angle of the installations, as this data was not available for the whole coverage of the French territory. I also did not rely on multispectral or hyperspectral images due to a lack of training data and appropriate models for dealing with these images. I think that the accuracy of the characterization can be significantly improved by taking into account the technologies explicitly dealing with thermal solar systems (e.g., using multispectral data and labeled data from Garioud et al., 2023) and using surface models to compute the tilt and azimuth angles.

On the relevance of model-based modeling for improving rooftop PV observability Our results showed that the accuracy of the estimation of the PV power production of a single installation is not significantly different between linear regression and our conversion model (no matter how parameterized). This shows the limitation of such model-based approaches, as it is hard to get more information on the PV panels at the scale of a country, as discussed in the previous point.

On the other hand, taking shadings into account for the PV power estimation could increase the computational burden of the model and thus limit its scaling-up ability. I think conversion models are better suited for constructing or reconstructing accurate power measurements of individual installations destined to be used as synthetic reference data since gathering ground truth individual PV power curves is challenging.

2.2 ... and on the deep learning side

Geographical variability I did not find evidence of a significant impact of sensitivity to the geographical localization once the acquisition conditions and the effects of the ground sampling distance were accounted for. This result is limited to our

use case and may be explained by our training dataset capturing the geographical variability encountered in France and neighboring regions.

Beyond data augmentation for improving robustness I restricted myself to data augmentation techniques when looking for methods to improve the robustness to acquisition conditions. The reason for this was that data augmentation techniques are more accessible and more straightforward to apply in operating conditions than other methods that I would call explicit regularization techniques, i.e., methods that consist in evaluating the model against a custom loss function or training a model that differs from the standard model architectures (e.g., Geirhos et al. (2019), Arjovsky et al., 2019). Another method for improving the model's robustness could be to use so-called foundation models (Bommasani et al., 2022), i.e., large models trained on vast amounts of data, so vast that the notion of distribution shifts eventually disappears. It could be interesting to see how such models would compare against "traditional" methods (data augmentation and explicit regularization) for improving the robustness to varying acquisition conditions and geographical variability. Some large-scale models have already been introduced for remote sensing data (Liu et al., 2024; Cha et al., 2023).

Improving the understandability of the WCAM Our attribution method, the WCAM, decomposes a model's decision in the space-scale domain by showing a heatmap over the wavelet decomposition of the image. As a result, the interpretation is tricky as it requires familiarity with the wavelet transform. One of the promises of the WCAM is to disentangle the various structural components of the image that contribute to the prediction (e.g., the gridded pattern of the PV panel and its overall shape). Providing an explicit visualization of these structural components could significantly improve the practical usability of the WCAM.

3 Perspectives

3.1 Power system perspectives

Registries for accurate estimation of the installed capacity If physical model-based methods may not necessarily be the best method for improving rooftop PV observability, the relevance of detailed technical registries remains. Indeed, throughout our study, I assumed that I knew the exact installed capacity. However, reference data may be inaccurate, as shown by Rausch et al. (2020). Therefore, constructing a PV registry for controlling the quality of other data sources is relevant, even if, afterward, PV power production estimations are carried out using statistical models.

Self-consumption and netload estimation Individual rooftop PV installations are increasingly used for self-consumption. At the end of September 2020, when I started the Ph.D. thesis, less than 70,000 PV installations of less than 36 kW_p (amounting to 16.6% of the connected installations) were self-consuming at least partially their production. At the end of 2023, more than 364,000 (50.1%) connected installations self-consume their production (Enedis, 2024).

The French current legislative framework (République française, 2017, 2023) is such that individual and collective self-consumption is increasingly more interesting than total selling to the grid. New and future connected PV installations will self-consume a part of their production. Therefore, rooftop PV observability will shift from observing the PV power production to observing the netload injections.

Good knowledge of rooftop PV power production will remain important, especially since we can expect that ground truth PV power measurements will remain scarce: the only information available will be the netload (i.e., the difference between the PV power production and household consumption). However, these power production estimations will have to be integrated into larger models that also take into account electric consumption, either at the scale of individual households or at the scale of neighborhoods.

Citizen science and quality checks Following the crowdsourcing campaigns that led to BDAPPV, a set of collaborative tools have been developed to enable system owners to delineate their installation on an image and verify the accuracy of the azimuth angles recorded by the users. These tools contributed to improving the quality of BDPV's data. However, even for the system owners, reporting the tilt angle is difficult. Using a conversion model, the ground truth measurements and the LiDAR data could be an efficient way to improve the reporting of the tilt angles and thus contribute to a large, curated, and high-quality dataset of rooftop PV characteristics and power measurements.

3.2 Deep learning perspectives

Multilabel and specialized classification Another possible extension of this work would be to focus on the classification branch. One could do multilabel classification to identify the PV system type. These labels are already featured in the BDPV database. Besides, one could gather more training data of larger installations (e.g., PV on shading roofs) to enhance the detector's accuracy for these installations. The relevance of combining specialized detectors for the different instances of PV installations could also be discussed, as well as the effect of potential uneven performance on the aggregated estimation of the rooftop PV installed capacity.

The present work also showed that in the binary classification of PV panels, the gridded pattern of old systems is an essential feature for the model, which tends to

look for such patterns in images. It could be interesting to see how the importance of this pattern evolves if we distinguish different classes of PV panels.

Predictive features and inductive biases Our work showed that false detections, arising mainly because the model confuses gridded patterns with PV panels, contributed to overestimating the PV installed capacity. The WCAM enabled us to highlight this phenomenon and draw directions to mitigate it. However, it could be interesting to understand why the model favored this pattern. This question is more theoretical and falls beyond the scope of the present thesis. In my current understanding of the question, figuring out why the gridded pattern ends up being the "favorite" feature of the model requires understanding how the model constructs predictive features from the input data during training.

Understanding the representations of convolutional neural networks The WCAM is a *post-hoc* explainability method, i.e., it explains the decision of a black-box model. The explanation provided by the WCAM highlights important regions in the scale-space domain. Another approach for explainable machine learning consists of using intrinsically interpretable models. The Scattering transform (Bruna and Mallat, 2013) can be considered an intrinsically interpretable model. It could be interesting to compare the representations (i.e., how the model compresses the information from the input image) of the Scattering transform with those of convolutional neural networks derived with the WCAM. Checking the representations' alignment and accuracy of both models could guide us toward using computationally less demanding and more interpretable models. Cheng and Ménard (2021) already showed the relevance of the Scattering transform in astrophysics but lacked a tool to analyze the representations of CNNs from a space-scale perspective.

References

R. Achtibat, M. Dreyer, I. Eisenbraun, S. Bosse, T. Wiegand, W. Samek, and S. Lapuschkin. From "Where" to "What": Towards Human-Understandable Explanations through Concept Relevance Propagation, June 2022. arXiv:2206.03208 [cs]. xliii, 50

J. Alireza. area: Calculate the area inside of any GeoJSON geometry. This is a port of Mapbox'sgeojson-area for Python, 2018. URL https://github.com/scisco/area. 197

M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. 147

E. Arnaudo, G. Blanco, A. Monti, G. Bianco, C. Monaco, P. Pasquali, and F. Dominici. A Comparative Evaluation of Deep Learning Techniques for Photovoltaic Panel Detection from Aerial Images. *IEEE Access*, pages 1–1, 2023. ISSN 2169-3536. doi: 10.1109/ACCESS.2023.3275435. xxxi, xxxii, 7, 9

H. Awala. DeepSolar dataset, 2020. URL https://www.kaggle.com/datasets/husseinawala/deepsolar. 88

K. Ayush, B. Uzkent, K. Tanmay, M. Burke, D. Lobell, and S. Ermon. Efficient Poverty Mapping from High Resolution Remote Sensing Images. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(1):12–20, May 2021. ISSN 2374-3468. doi: 10.1609/aaai.v35i1.16072. Number: 1. 103

BDPV. Carte des installations - BDPV, 2023. URL https://www.bdpv.fr/fr/carteInstallation.php. 144, 177

U. Bhatt, A. Weller, and J. M. F. Moura. Evaluating and Aggregating Feature-based Model Explanations. In C. Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3016–3022. International Joint Conferences on Artificial Intelligence Organization, July 2020. doi: 10.24963/ijcai.2020/417. xliii, 50

R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, E. Brynjolfsson, S. Buch, D. Card, R. Castellon, N. Chatterji, A. Chen, K. Creel, J. Q. Davis, D. Demszky, C. Donahue, M. Doumbouya, E. Durmus, S. Ermon, J. Etchemendy, K. Ethayarajh, L. Fei-Fei, C. Finn, T. Gale, L. Gillespie, K. Goel, N. Goodman, S. Grossman, N. Guha, T. Hashimoto, P. Henderson, J. Hewitt, D. E. Ho, J. Hong, K. Hsu, J. Huang, T. Icard, S. Jain, D. Jurafsky, P. Kalluri, S. Karamcheti, G. Keeling, F. Khani, O. Khattab, P. W. Koh, M. Krass,

R. Krishna, R. Kuditipudi, A. Kumar, F. Ladhak, M. Lee, T. Lee, J. Leskovec, I. Levent, X. L. Li, X. Li, T. Ma, A. Malik, C. D. Manning, S. Mirchandani, E. Mitchell, Z. Munyikwa, S. Nair, A. Narayan, D. Narayanan, B. Newman, A. Nie, J. C. Niebles, H. Nilforoshan, J. Nyarko, G. Ogut, L. Orr, I. Papadimitriou, J. S. Park, C. Piech, E. Portelance, C. Potts, A. Raghunathan, R. Reich, H. Ren, F. Rong, Y. Roohani, C. Ruiz, J. Ryan, C. Ré, D. Sadigh, S. Sagawa, K. Santhanam, A. Shih, K. Srinivasan, A. Tamkin, R. Taori, A. W. Thomas, F. Tramèr, R. E. Wang, W. Wang, B. Wu, J. Wu, Y. Wu, S. M. Xie, M. Yasunaga, J. You, M. Zaharia, M. Zhang, T. Zhang, X. Zhang, Y. Zhang, L. Zheng, K. Zhou, and P. Liang. On the Opportunities and Risks of Foundation Models, July 2022. arXiv:2108.07258 [cs]. 147

K. Bradbury, R. Saboo, T. L Johnson, J. M. Malof, A. Devarajan, W. Zhang, L. M Collins, and R. G Newell. Distributed solar photovoltaic array location and extent dataset for remote sensing object identification. *Scientific data*, 3(1):1–9, 2016. Publisher: Nature Publishing Group. xxxii, 8, 107, 180, 185

J. Bruna and S. Mallat. Invariant Scattering Convolution Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1872–1886, Aug. 2013. ISSN 1939-3539. doi: 10.1109/TPAMI.2012.230. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence. 149

Bundesnetzagentur. Marktstammdatenregister, 2022. URL https://www.marktstammdatenregister.de/. 112

J. Camilo, R. Wang, L. M. Collins, K. Bradbury, and J. M. Malof. Application of a semantic segmentation convolutional neural network for accurate automatic detection and mapping of solar photovoltaic arrays in aerial imagery, Jan. 2018. arXiv:1801.04018 [cs]. xxxii

A. Casanova, M. Careil, J. Verbeek, M. Drozdzal, and A. Romero Soriano. Instance-Conditioned GAN. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 27517–27529. Curran Associates, Inc., 2021. 9

CBS. CBS Open data portal, 2024. URL https://opendata.cbs.nl/statline/portal.html?_la=en&_catalog=CBS&tableId=80030eng&_theme=1080. 112

K. Cha, J. Seo, and T. Lee. A Billion-scale Foundation Model for Remote Sensing Images, Apr. 2023. arXiv:2304.05215 [cs]. 147

S. Chabrier, B. Emile, C. Rosenberger, and H. Laurent. Unsupervised Performance Evaluation of Image Segmentation. *EURASIP Journal on Advances in Signal Processing*, 2006(1):096306, Dec. 2006. ISSN 1687-6180. doi: 10.1155/ASP/2006/96306. 28

B. Chen, Y. Li, and N. Zeng. Centralized Wavelet Multiresolution for Exact Translation Invariant Processing of ECG Signals. *IEEE Access*, 7:42322–42330, 2019. ISSN 2169-3536. doi: 10.1109/ACCESS.2019.2907249. xv, 52

G. Chen, P. Peng, L. Ma, J. Li, L. Du, and Y. Tian. Amplitude-Phase Recombination: Rethinking Robustness of Convolutional Neural Networks in Frequency Domain. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 448–457, Montreal, QC, Canada, Oct. 2021a. IEEE. ISBN 978-1-66542-812-5. doi: 10.1109/ICCV48922.2021.00051. xxxix

J. Chen, F. Liu, B. Avci, X. Wu, Y. Liang, and S. Jha. Detecting Errors and Estimating Accuracy on Unlabeled Data with Self-training Ensembles. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 14980–14992. Curran Associates, Inc., 2021b. 27

L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, Apr. 2018. ISSN 01628828. doi: 10.1109/TPAMI.2017.2699184. Publisher: IEEE Computer Society _eprint: 1606.00915. 31, 33, 88

Y. Chen, Q. Ren, and J. Yan. Rethinking and Improving Robustness of Convolutional Neural Networks: a Shapley Value-based Approach in Frequency Domain. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 324–337. Curran Associates, Inc., 2022. xliv, 51, 57, 58

S. Cheng and B. Ménard. How to quantify fields or textures? A guide to the scattering transform, Nov. 2021. arXiv:2112.01288 [astro-ph, physics:physics]. 149

D. Chicco. Siamese neural networks: An overview. *Artificial neural networks*, pages 73–94, 2021. Publisher: Springer. 10

CodeCarbon. CodeCarbon documentation, 2023. URL https://mlco2.github.io/codecarbon/. 176

J. Cohen. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46, Apr. 1960. ISSN 0013-1644, 1552-3888. doi: 10.1177/001316446002000104. 97

Commission de Régulation de l'Énergie. Présentation des réseaux d'éléctricité, 2024. URL https://www.cre.fr/Electricite/Reseaux-d-electricite/presentation-des-reseaux-d-electricite. 20

J. Crabbé and M. van der Schaar. Evaluating the Robustness of Interpretability Methods through Explanation Invariance and Equivariance. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 71393–71429. Curran Associates, Inc., 2023. xliii, 50

G. Csurka. A Comprehensive Survey on Domain Adaptation for Visual Applications. In G. Csurka, editor, *Domain Adaptation in Computer Vision Applications*, pages 1–35. Springer International Publishing, Cham, 2017. ISBN 978-3-319-58347-1. doi: 10.1007/978-3-319-58347-1_1. xxxiii, 10

G. Csurka, R. Volpi, and B. Chidlovskii. Unsupervised Domain Adaptation for Semantic Image Segmentation: a Comprehensive Survey, Dec. 2021. arXiv:2112.03241 [cs]. xxxiii, 10

E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le. AutoAugment: Learning Augmentation Strategies From Data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. li, lii, 76, 79, 81, 196, 197

E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le. RandAugment: Practical automated data augmentation with a reduced search space. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3008–3017, June 2020. doi: 10.1109/CVPRW50498.2020.00359. ISSN: 2160-7516. li, lii, 79, 81, 196, 197

D. Czirjak. Detecting photovoltaic solar panels using hyperspectral imagery and estimating solar power production. *Journal of Applied Remote Sensing*, 11(2): 026007, Apr. 2017. ISSN 1931-3195. doi: 10.1117/1.JRS.11.026007. 8

P. Dardouillet, A. Benoit, E. Amri, P. Bolon, D. Dubucq, and A. Credoz. Explainability of Image Semantic Segmentation Through SHAP Values. In J.-J. Rousseau and B. Kapralos, editors, *Pattern Recognition, Computer Vision, and Image Processing. ICPR 2022 International Workshops and Challenges*, pages 188–202, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-37731-0. doi: 10.1007/978-3-031-37731-0_19. 41

J. de Hoog, S. Maetschke, P. Ilfrich, and R. R. Kolluri. Using Satellite and Aerial Imagery for Identification of Solar PV: State of the Art and Research Opportunities. In *Proceedings of the Eleventh ACM International Conference on Future Energy Systems*, pages 308–313, Virtual Event Australia, June 2020. ACM. ISBN 978-1-4503-8009-6. doi: 10.1145/3396851.3397681. xxxi, 7

J. de Hoog, M. Perera, K. Bandara, D. Senanayake, and S. Halgamuge. Solar PV Maps for Estimation and Forecasting of Distributed Solar Generation. In *ICML 2021 Workshop on Tackling Climate Change with Machine Learning*, 2021. 116

T. De Jong, S. Bromuri, X. Chang, M. Debusschere, N. Rosenski, C. Schartner, K. Strauch, M. Boehmer, and L. Curier. Monitoring Spatial Sustainable Development: semi-automated analysis of Satellite and Aerial Images for Energy Transition and Sustainability Indicators, 2020. xxxii, 9, 10

A. de Luis, M. Tran, T. Hanyu, A. Tran, L. Haitao, R. McCann, A. Mantooth, Y. Huang, and N. Le. SolarFormer: Multi-scale Transformer for Solar PV Profiling, Oct. 2023. arXiv:2310.20057 [cs]. 107

I. Demir, K. Koperski, D. Lindenbaum, G. Pang, J. Huang, S. Basu, F. Hughes, D. Tuia, and R. Raskar. DeepGlobe 2018: A Challenge to Parse the Earth through Satellite Images. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 172–17209, June 2018. doi: 10.1109/CVPRW.2018.00031. arXiv:1805.06561 [cs]. 106

A. Devarajan, B. Kellish, C. Kido, A. Newman, and K. Bradbury. Automated Rooftop Solar PV Detection and Power Estimation through Remote Sensing, 2016. xxxii

A. Dobos. PVWatts Version 5 Manual. Technical Report NREL/TP-6A20-62641, 1158421, NREL, Sept. 2014. xxiii, lvi, 115, 122, 125, 138

A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 89, 100, 107

E-Control. Anlagenregister, 2023. URL https://anlagenregister.at/. 112

A. S. Edun, K. Perry, J. B. Harley, and C. Deline. Unsupervised azimuth estimation of solar arrays in low-resolution satellite imagery through semantic segmentation and Hough transform. *Applied Energy*, 298:117273, Sept. 2021. ISSN 03062619. doi: 10.1016/j.apenergy.2021.117273. 92, 94

Enedis. Enedis Open Data, 2024. URL https://data.enedis.fr/pages/accueil/. 148

Energistyrelsen. Stamdataregister, 2022. URL https://ens.dk/service/statistik-data-noegletal-og-kort/data-oversigt-over-energisektoren. 112

ESA. Pléiades Neo full archive and tasking - Earth Online, 2024. URL https://earth.esa.int/eogateway/catalog/pleiades-neo-full-archive-and-tasking. xiv, 16

European Union. Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions 'Fit for 55': delivering the EU's 2030 Climate Target on the way to climate neutrality, 2021. 2

European Union. Overview - NUTS - Nomenclature of territorial units for statistics - Eurostat, 2024. URL https://ec.europa.eu/eurostat/web/nuts/overview. 18

T. Fel, R. Cadene, M. Chalvidal, M. Cord, D. Vigouroux, and T. Serre. Look at the Variance! Efficient Black-box Explanations with Sobol-based Sensitivity Analysis. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 26005–26014. Curran Associates, Inc., 2021. xliii, xlvi, 50, 51, 54, 55, 56

T. Fel, M. Ducoffe, D. Vigouroux, R. Cadène, M. Capelle, C. Nicodème, and T. Serre. Don't Lie to Me! Robust and Efficient Explainability with Verified Perturbation Analysis. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16153–16163, June 2023a. doi: 10.1109/CVPR52729.2023.01550. ISSN: 2575-7075. 50

T. Fel, A. Picard, L. Béthune, T. Boissin, D. Vigouroux, J. Colin, R. Cadène, and T. Serre. CRAFT: Concept Recursive Activation FacTorization for Explainability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2711–2721, June 2023b. xliii, 50

S. Freitas, M. Silva, E. Silva, A. Marceddu, M. Miccoli, P. Gnatyuk, L. Marangoni, and A. Amicone. An Artificial Intelligence-Based Framework to Accelerate Data-Driven Policies to Promote Solar Photovoltaics in Lisbon. *Solar RRL*, n/a(n/a): 2300597, 2023. ISSN 2367-198X. doi: 10.1002/solr.202300597. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/solr.202300597. lx, lxi, 112, 145

Frimane, R. Johansson, J. Munkhammar, D. Lingfors, and J. Lindahl. Identifying small decentralized solar systems in aerial images using deep learning. *Solar Energy*, 262:111822, Sept. 2023. ISSN 0038-092X. doi: 10.1016/j.solener.2023.111822. xxxii, 9, 89

G. Gao, Q. Liu, and Y. Wang. Counting Dense Objects in Remote Sensing Images. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4137–4141, May 2020. doi: 10.1109/ICASSP40776.2020.9053690. ISSN: 2379-190X. 103

H. Gao, Y. Tang, L. Jing, H. Li, and H. Ding. A Novel Unsupervised Segmentation Quality Evaluation Method for Remote Sensing Images. *Sensors*, 17(10):2427, Oct. 2017. ISSN 1424-8220. doi: 10.3390/s17102427. 28

A. Garioud, N. Gonthier, L. Landrieu, A. De Wit, M. Valette, M. Poupée, S. Giordano, and b. Wattrelos. FLAIR : a Country-Scale Land Cover Semantic Segmentation Dataset From Multi-Source Optical Imagery. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 16456–16482. Curran Associates, Inc., 2023. 146

R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019. 77, 147

R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann. Shortcut Learning in Deep Neural Networks. *Nature Machine Intelligence*, 2(11):665–673, Nov. 2020. ISSN 2522-5839. doi: 10.1038/s42256-020-00257-z. arXiv:2004.07780 [cs, q-bio]. 51, 61

V. Golovko, S. Bezobrazov, A. Kroshchanka, A. Sachenko, M. Komar, and A. Karachka. Convolutional neural network based solar photovoltaic panel detection in satellite photos. In *2017 9th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*, pages 14–19, Bucharest, Sept. 2017. IEEE. ISBN 978-1-5386-0697-1. doi: 10.1109/IDAACS.2017.8094501. xxxii, 8

V. Golovko, A. Kroshchanka, S. Bezobrazov, A. Sachenko, M. Komar, and O. Novosad. Development of Solar Panels Detector. In *2018 International Scientific-Practical Conference Problems of Infocommunications. Science and Technology (PIC S&T)*, pages 761–764, Kharkiv, Ukraine, Oct. 2018. IEEE. ISBN 978-1-5386-6609-8 978-1-5386-6611-1. doi: 10.1109/INFOCOMMST.2018.8632132. xxxii

I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and Harnessing Adversarial Examples. In Y. Bengio and Y. LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 75

N. Gorelick, M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore. Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote sensing of Environment*, 202:18–27, 2017. Publisher: Elsevier. xlix, 25, 179

H. Guan and M. Liu. Domain Adaptation for Medical Image Analysis: A Survey. *IEEE Transactions on Biomedical Engineering*, 69(3):1173–1185, Mar. 2022. ISSN 1558-2531. doi: 10.1109/TBME.2021.3117407. Conference Name: IEEE Transactions on Biomedical Engineering. xxxiii, 10

I. Gulrajani and D. Lopez-Paz. In Search of Lost Domain Generalization. In *International Conference on Learning Representations*, 2021. xxxiii, 10

Z. Guo, J. Lu, Q. Chen, Z. Liu, C. Song, H. Tan, H. Zhang, and J. Yan. TransPV: Refining photovoltaic panel detection accuracy through a vision transformer-based deep

learning model. *Applied Energy*, 355:122282, Feb. 2024. ISSN 0306-2619. doi: 10.1016/j.apenergy.2023.122282. lx, 89, 145

N. M. Haegel, R. Margolis, T. Buonassisi, D. Feldman, A. Froitzheim, R. Garabedian, M. Green, S. Glunz, H.-M. Henning, B. Holder, and others. Terawatt-scale photovoltaics: Trajectories and challenges. *Science*, 356(6334):141–143, 2017. Publisher: American Association for the Advancement of Science. 9

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. liv, 65, 71, 107

D. Hendrycks and T. Dietterich. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. In *International Conference on Learning Representations*, 2019. xvii, li, 68, 75, 76

D. Hendrycks, N. Mu, E. D. Cubuk, B. Zoph, J. Gilmer, and B. Lakshminarayanan. AugMix: A Simple Data Processing Method to Improve Robustness and Uncertainty. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. li, lii, 76, 79, 81, 196, 197

D. Hendrycks, A. Zou, M. Mazeika, L. Tang, B. Li, D. Song, and J. Steinhardt. PixMix: Dreamlike Pictures Comprehensively Improve Safety Measures. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Mar. 2022. doi: 10.48550/arXiv.2112.05135. URL http://arxiv.org/abs/2112.05135. arXiv:2112.05135 [cs]. 76

H. Hersbach, B. Bell, P. Berrisford, S. Hirahara, A. Horányi, J. Muñoz-Sabater, J. Nicolas, C. Peubey, R. Radu, D. Schepers, A. Simmons, C. Soci, S. Abdalla, X. Abellan, G. Balsamo, P. Bechtold, G. Biavati, J. Bidlot, M. Bonavita, G. De Chiara, P. Dahlgren, D. Dee, M. Diamantakis, R. Dragani, J. Flemming, R. Forbes, M. Fuentes, A. Geer, L. Haimberger, S. Healy, R. J. Hogan, E. Hólm, M. Janisková, S. Keeley, P. Laloyaux, P. Lopez, C. Lupu, G. Radnoti, P. de Rosnay, I. Rozum, F. Vamborg, S. Villaume, and J.-N. Thépaut. The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146 (730):1999–2049, 2020. ISSN 1477-870X. doi: 10.1002/qj.3803. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/qj.3803. lv, 120

W. Hoeffding. A Class of Statistics with Asymptotically Normal Distribution. In S. Kotz and N. L. Johnson, editors, *Breakthroughs in Statistics: Foundations and Basic Theory*, Springer Series in Statistics, pages 308–334. Springer, New York, NY, 1992. ISBN 978-1-4612-0919-5. doi: 10.1007/978-1-4612-0919-5_20. 53

W. Holmgren, C. Hansen, and M. Mikofski. pvlib python: a python package for modeling solar energy systems. *Journal of Open Source Software*, 3(29):884, Sept. 2018. ISSN 2475-9066. doi: 10.21105/joss.00884. 123

T. Homma and A. Saltelli. Importance measures in global sensitivity analysis of nonlinear models. *Reliability Engineering & System Safety*, 52(1):1–17, Apr. 1996. ISSN 0951-8320. doi: 10.1016/0951-8320(96)00002-6. 54

W. Hu, K. Bradbury, J. M. Malof, B. Li, B. Huang, A. Streltsov, K. Sydny Fujita, and B. Hoen. What you get is not always what you see—pitfalls in solar array assessment using overhead imagery. *Applied Energy*, 327:120143, Dec. 2022. ISSN 0306-2619. doi: 10.1016/j.apenergy.2022.120143. 9

B. Huang, L. M. Collins, K. Bradbury, and J. M. Malof. Deep Convolutional Segmentation of Remote Sensing Imagery: A Simple and Efficient Alternative to Stitching Output Labels. In *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, pages 6899–6902, Valencia, July 2018. IEEE. ISBN 978-1-5386-7150-4. doi: 10.1109/IGARSS.2018.8518701. xxxii, 89

Hugging Face. CO2 Emissions and the HF Hub: Leading the Charge, 2023. URL https://huggingface.co/blog/carbon-emissions-on-the-hub. 12, 175

IEA. Tracking Clean Energy Progress 2023. Technical report, IEA, Paris, 2023. xxix, 1

IGN. BD TOPO (R) | Géoservices, 2023. URL https://geoservices.ign.fr/bdtopo. xxxv, 18

IGN. BD ORTHO® | Géoservices, 2024a. URL https://geoservices.ign.fr/bdortho. xiv, xxxv, xlix, 16, 17, 25

IGN. LiDAR HD | Géoservices, 2024b. URL https://geoservices.ign.fr/lidarhd. xxxv, 20

IGN. Rapports de contrôle qualité | Géoservices, 2024c. URL https://geoservices.ign.fr/documentation/donnees/ortho/bdortho/rapport-controle-qualite. 18

IPCC. *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, UK and New York, NY, USA, 2021a. doi: 10.1017/9781009157896. Type: Book. xxix, 1

IPCC. Summary for Policymakers. In V. Masson-Delmotte, P. Zhai, A. Pirani, S. L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M. I. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J. B. R. Matthews, T. K. Maycock, T. Waterfield, O. Yelekçi, R. Yu, and B. Zhou, editors, *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, UK and New York, NY, USA, 2021b. doi: 10.1017/9781009157896.001. Type: Book Section. xiv, xxix, 2

M. J. W. Jansen. Analysis of variance designs for model output. *Computer Physics Communications*, 117(1):35–43, Mar. 1999. ISSN 0010-4655. doi: 10.1016/S0010-4655(98)00154-4. 54, 56

C. Ji, M. Bachmann, T. Esch, H. Feilhauer, U. Heiden, W. Heldens, A. Hueni, T. Lakes, A. Metz-Marconcini, M. Schroedter-Homscheidt, S. Weyand, and J. Zeidler. Solar photovoltaic module detection using laboratory and airborne imaging spectroscopy data. *Remote Sensing of Environment*, 266:112692, Dec. 2021. ISSN 00344257. doi: 10.1016/j.rse.2021.112692. 8

J. Jo and Y. Bengio. Measuring the tendency of CNNs to Learn Surface Statistical Regularities, Nov. 2017. arXiv:1711.11561 [cs, stat]. 51

Z.-Q. John Xu, Y. Zhang, T. Luo, Y. Xiao, and Z. Ma. Frequency Principle: Fourier Analysis Sheds Light on Deep Neural Networks. *Communications in Computational Physics*, 28(5):1746–1767, June 2020. ISSN 1815-2406, 1991-7120. doi: 10.4208/cicp.OA-2020-0085. 51

JRC. PVGIS Online Tool, 2023. URL https://joint-research-centre.ec.europa.eu/pvgis-online-tool_en. 198

M. Karoui, F. Benhalouche, Y. Deville, K. Djerriri, X. Briottet, T. Houet, A. Le Bris, and C. Weber. Partial Linear NMF-Based Unmixing Methods for Detection and Area Estimation of Photovoltaic Panels in Urban Hyperspectral Remote Sensing Data. *Remote Sensing*, 11(18):2164, Sept. 2019. ISSN 2072-4292. doi: 10.3390/rs11182164. 8

M. S. Karoui, F. z. Benhalouche, Y. Deville, K. Djerriri, X. Briottet, and A. L. Bris. Detection And Area Estimation For Photovoltaic Panels In Urban Hyperspectral Remote Sensing Data By An Original Nmf-Based Unmixing Method. In *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, pages 1640–1643, Valencia, July 2018. IEEE. ISBN 978-1-5386-7150-4. doi: 10.1109/IGARSS.2018.8518204. 8

G. Kasmi, L. Dubus, Y.-M. Saint-Drenan, and P. Blanc. Towards Unsupervised Assessment with Open-Source Data of the Accuracy of Deep Learning-Based Distributed PV Mapping. In T. Corpetti, D. Ienco, R. Interdonato, M.-T. Pham, and S. Lefèvre, editors, *Proceedings of MACLEAN: MAChine Learning for EArth ObservatioN Workshop co-located with the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD 2022), Grenoble, France, September 18-22, 2022*, volume 3343 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2022a. xiii, xiv, xix, xxi, xxii, xxxiv, xxxix, xli, lv, lx, 14, 30, 32, 36, 108, 109, 144, 176, 177, 200, 201

G. Kasmi, Y.-M. Saint-Drenan, D. Trebosc, R. Jolivet, J. Leloux, B. Sarr, and L. Dubus. A crowdsourced dataset of aerial images with annotated solar photovoltaic arrays and installation metadata, July 2022b. URL https://zenodo.org/records/7358126. 24

G. Kasmi, L. Dubus, Y.-M. Saint-Drenan, and P. Blanc. Assessment of the Reliablity of a Model's Decision by Generalizing Attribution to the Wavelet Domain. In *XAI in Action: Past, Present, and Future Applications workshop at NeurIPS 2023*. arXiv, Sept. 2023a. doi: 10.48550/arXiv.2305.14979. arXiv:2305.14979 [cs, stat]. xiii, xv, xvi, xliii, xlvi, lx, 14, 55, 57, 58, 64, 107, 144

G. Kasmi, L. Dubus, Y.-M. Saint-Drenan, and P. Blanc. Can We Reliably Improve the Robustness to Image Acquisition of Remote Sensing of PV Systems? In *NeurIPS 2023 Workshop on Tackling Climate Change with Machine Learning*, 2023b. xiii, xv, xvi, xix, xxi, xxii, xliii, xliv, xlvii, l, lx, 14, 49, 59, 60, 66, 67, 68, 107, 194

G. Kasmi, D. Laurent, B. Philippe, and Y.-M. Saint-Drenan. DeepPVMapper, Sept. 2023c. URL https://doi.org/10.5281/zenodo.8380321. lii, liii, lxi, 99, 145

G. Kasmi, Y.-M. Saint-Drenan, D. Trebosc, R. Jolivet, J. Leloux, B. Sarr, and L. Dubus. A crowdsourced dataset of aerial images with annotated solar photovoltaic arrays and installation metadata. *Scientific Data*, 10(1):59, Jan. 2023d. ISSN 2052-4463. doi: 10.1038/s41597-023-01951-4. xiii, xiv, xviii, xxiii, xxxiv, xl, xlix, lx, 14, 24, 25, 32, 65, 106, 107, 144, 180, 183, 186, 187, 188

G. Kasmi, A. Touron, P. Blanc, Y.-M. Saint-Drenan, M. Fortin, and L. Dubus. Remote-Sensing-Based Estimation of Rooftop Photovoltaic Power Production Using Physical Conversion Models and Weather Data. *Energies*, 17(17):4353, Aug. 2024. ISSN 1996-1073. doi: 10.3390/en17174353. lii, lxi, 14, 145

B. B. Kausika. *GIS4PV: A technological impact assessment of the application of GIS for Photovoltaic Solar Energy.* PhD thesis, Utrecht University, May 2022. 9

B. B. Kausika, D. Nijmeijer, I. Reimerink, P. Brouwer, and V. Liem. GeoAI for detection of solar photovoltaic installations in the Netherlands. *Energy and AI*, 6:100111, Dec. 2021. ISSN 26665468. doi: 10.1016/j.egyai.2021.100111. xxxii, 9

S. Killinger, D. Lingfors, Y.-M. Saint-Drenan, P. Moraitis, W. van Sark, J. Taylor, N. A. Engerer, and J. M. Bright. On the search for representative characteristics of PV systems: Data collection and analysis of PV system azimuth, tilt, capacity, yield and shading. *Solar Energy*, 173:1087–1106, Oct. 2018. ISSN 0038092X. doi: 10.1016/j.solener.2018.08.051. xxxviii, 22, 27, 29

D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. In Y. Bengio and Y. LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 33

P. W. Koh and P. Liang. Understanding Black-box Predictions via Influence Functions. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1885–1894. PMLR, July 2017. ISSN: 2640-3498. xxxiii, 51

P. W. Koh, S. Sagawa, H. Marklund, S. M. Xie, M. Zhang, A. Balsubramani, W. Hu, M. Yasunaga, R. L. Phillips, I. Gao, T. Lee, E. David, I. Stavness, W. Guo, B. Earnshaw, I. Haque, S. M. Beery, J. Leskovec, A. Kundaje, E. Pierson, S. Levine, C. Finn, and P. Liang. WILDS: A Benchmark of in-the-Wild Distribution Shifts. In *Proceedings of the 38th International Conference on Machine Learning*, pages 5637–5664. PMLR, July 2021. ISSN: 2640-3498. 10

F. Kong and R. Henao. Efficient Classification of Very Large Images With Tiny Objects. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2384–2394, 2022. 103

S. Krapf, N. Kemmerzell, S. Khawaja Haseeb Uddin, M. Hack Vázquez, F. Netzler, and M. Lienkamp. Towards Scalable Economic Photovoltaic Potential Analysis Using Aerial Images and Deep Learning. *Energies*, 14(13):3800, June 2021. ISSN 1996-1073. doi: 10.3390/en14133800. 103

S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K.-R. Müller. Unmasking clever hans predictors and assessing what machines really learn. *Nature communications*, 10(1):1–8, 2019. Publisher: Nature Publishing Group. xli, 41

M. Levandowsky and D. Winter. Distance between Sets. *Nature*, 234(5323):34–35, Nov. 1971. ISSN 1476-4687. doi: 10.1038/234034a0. Number: 5323 Publisher: Nature Publishing Group. 185

P. Li, H. Zhang, Z. Guo, S. Lyu, J. Chen, W. Li, X. Song, R. Shibasaki, and J. Yan. Understanding rooftop PV panel semantic segmentation of satellite and aerial images for better using machine learning. *Advances in Applied Energy*, 4:100057, Nov. 2021. ISSN 26667924. doi: 10.1016/j.adapen.2021.100057. 18

Q. Li, Y. Feng, Y. Leng, and D. Chen. SolarFinder: Automatic Detection of Solar Photovoltaic Arrays. In *2020 19th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*, pages 193–204, Sydney, NSW, Australia, Apr. 2020. IEEE. ISBN 978-1-72815-497-8. doi: 10.1109/IPSN48710.2020.00024. xxxii, 98

J. Lindahl, R. Johansson, and D. Lingfors. Mapping of decentralised photovoltaic and solar thermal systems by remote sensing aerial imagery and deep machine learning for statistic generation. *Energy and AI*, page 100300, Sept. 2023. ISSN 2666-5468. doi: 10.1016/j.egyai.2023.100300. xxxii, 9

C.-W. Liu, R. Lowe, I. Serban, M. Noseworthy, L. Charlin, and J. Pineau. How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas, Nov. 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1230. 28

F. Liu, D. Chen, Z. Guan, X. Zhou, J. Zhu, Q. Ye, L. Fu, and J. Zhou. RemoteCLIP: A Vision Language Foundation Model for Remote Sensing. *IEEE Transactions on Geoscience and Remote Sensing*, pages 1–1, 2024. ISSN 1558-0644. doi: 10.1109/TGRS.2024.3390838. Conference Name: IEEE Transactions on Geoscience and Remote Sensing. 147

Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie. A ConvNet for the 2020s. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Mar. 2022. doi: 10.48550/arXiv.2201.03545. arXiv:2201.03545 [cs]. liv, 89, 100, 107

E. Lorenz, T. Scheidsteger, J. Hurka, D. Heinemann, and C. Kurz. Regional PV power prediction for improved grid integration. *Progress in Photovoltaics: Research and Applications*, 19(7):757–771, Nov. 2011. ISSN 1062-7995, 1099-159X. doi: 10.1002/pip.1033. 6

A. S. Luccioni, Y. Jernite, and E. Strubell. Power Hungry Processing: Watts Driving the Cost of AI Deployment?, Nov. 2023. arXiv:2311.16863 [cs]. 175

A. S. Luccioni, S. Viguier, and A.-L. Ligozat. Estimating the carbon footprint of BLOOM, a 176B parameter language model. *The Journal of Machine Learning Research*, 24(1):253:11990–253:12004, Mar. 2024. ISSN 1532-4435. 175, 176

S. M. Lundberg and S.-I. Lee. A Unified Approach to Interpreting Model Predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 50

F. Luzi, A. Gupta, L. Collins, K. Bradbury, and J. Malof. Transformers for Recognition in Overhead Imagery: A Reality Check. pages 3778–3787, 2023. 89, 106

E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez. Can semantic labeling methods generalize to any city? the INRIA aerial image labeling benchmark. In *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 3226–3229. IEEE, 2017. 106

S. Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11 (7):674–693, July 1989. ISSN 1939-3539. doi: 10.1109/34.192463. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence. xlv, 52

S. Mallat. *A wavelet tour of signal processing*. Elsevier, 1999. 52

J. M. Malof, Rui Hou, L. M. Collins, K. Bradbury, and R. Newell. Automatic solar photovoltaic panel detection in satellite imagery. In *2015 International Conference on Renewable Energy Research and Applications (ICRERA)*, pages 1428–1431, Palermo, Italy, Nov. 2015. IEEE. ISBN 978-1-4799-9982-8. doi: 10.1109/ICRERA.2015.7418643. xxxi, 7

J. M. Malof, K. Bradbury, L. M. Collins, and R. G. Newell. Automatic detection of solar photovoltaic arrays in high resolution aerial imagery. *Applied Energy*, 183:229–240, Dec. 2016a. ISSN 03062619. doi: 10.1016/j.apenergy.2016.08.191. xxxii, 8

J. M. Malof, K. Bradbury, L. M. Collins, R. G. Newell, A. Serrano, H. Wu, and S. Keene. Image features for pixel-wise detection of solar photovoltaic arrays in aerial imagery using a random forest classifier. In *2016 IEEE International Conference on Renewable Energy Research and Applications (ICRERA)*, pages 799–803. IEEE, 2016b. 8

J. M. Malof, L. M. Collins, K. Bradbury, and R. G. Newell. A deep convolutional neural network and a random forest classifier for solar photovoltaic array detection in aerial imagery. In *2016 IEEE International Conference on Renewable Energy Research and Applications (ICRERA)*, pages 650–654. IEEE, 2016c. xxxii

J. M. Malof, L. M. Collins, and K. Bradbury. A deep convolutional neural network, with pre-training, for solar photovoltaic array detection in aerial imagery. In *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 874–877, Fort Worth, TX, July 2017. IEEE. ISBN 978-1-5090-4951-6. doi: 10.1109/IGARSS.2017.8127092. 8

J. M. Malof, B. Li, B. Huang, K. Bradbury, and A. Stretslov. Mapping solar array location, size, and capacity using deep learning and overhead imagery, 2019. xxxi, xxxii, xl, xli, 9, 34, 37, 89, 92, 98

N. Martin and J. M. Ruiz. Calculation of the PV modules angular losses under field conditions by means of an analytical model. *Solar Energy Materials and Solar Cells*, 70(1):25–38, Dec. 2001. ISSN 0927-0248. doi: 10.1016/S0927-0248(00)00408-6. 122, 123

N. Martín and J. M. Ruiz. A new model for PV modules angular losses under field conditions. *International Journal of Solar Energy*, 22(1):19–31, Jan. 2002. ISSN 0142-5919. doi: 10.1080/01425910212852. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01425910212852. 123

N. Martín and J. M. Ruiz. Annual angular reflection losses in PV modules. *Progress in Photovoltaics: Research and Applications*, 13(1): 75–84, 2005. ISSN 1099-159X. doi: 10.1002/pip.585. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/pip.585. 123

B. W. Matthews. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica Et Biophysica Acta*, 405(2):442–451, Oct. 1975. ISSN 0006-3002. doi: 10.1016/0005-2795(75)90109-9. 98

K. Mayer, Z. Wang, M.-L. Arlt, D. Neumann, and R. Rajagopal. DeepSolar for Germany: A deep learning framework for PV system mapping from aerial imagery. In *2020 International Conference on Smart Energy Systems and Technologies (SEST)*, pages 1–6, Istanbul, Turkey, Sept. 2020. IEEE. ISBN 978-1-72814-701-7. doi: 10.1109/SEST48500.2020.9203258. xxxii, xxxix, 9, 89, 97

K. Mayer, B. Rausch, M.-L. Arlt, G. Gust, Z. Wang, D. Neumann, and R. Rajagopal. 3D-PV-Locator: Large-scale detection of rooftop-mounted photovoltaic systems in 3D. *Applied Energy*, 310:118469, Mar. 2022. ISSN 03062619. doi: 10.1016/j.apenergy.2021.118469. xxxii, xxxix, xl, liii, 9, 27, 28, 29, 30, 33, 34, 88, 89, 198

C. Meng, E. Liu, W. Neiswanger, J. Song, M. Burke, D. B. Lobell, and S. Ermon. IS-Count: Large-Scale Object Counting from Satellite Images with Covariate-Based Importance Sampling. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 12034–12042. AAAI Press, 2022. doi: 10.1609/AAAI.V36I11.21462. 103

Ministère de la Transition Ecologique et Solidaire. Stratégie Nationale Bas Carbone. Technical report, Ministère de la Transition Ecologique et Solidaire, 2020. 3

M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 220–229, Jan. 2019. doi: 10.1145/3287560.3287596. arXiv:1810.03993 [cs]. 98

T. M. Mitchell. The Need for Biases in Learning Generalizations. *Rutgers University*, 1980. 77

W. J. Morokoff and R. E. Caflisch. Quasi-Monte Carlo Integration. *Journal of Computational Physics*, 122(2):218–230, Dec. 1995. ISSN 0021-9991. doi: 10.1006/jcph.1995.1209. 54

J. Murray, D. Marcos, and D. Tuia. Zoom In, Zoom Out: Injecting Scale Invariance into Landuse Classification CNNs. In *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, pages 5240–5243. IEEE, 2019. 65

V. Nagarajan. *Explaining generalization in deep learning: progress and fundamental limits*. PhD thesis, Carnegie Mellon University, 2021. 209

M. Nielsen. *True Orthophoto generation*. PhD thesis, DTU, 2004. xiv, 17

P. Novello, T. Fel, and D. Vigouroux. Making Sense of Dependence: Efficient Black-box Explanations Using Dependence Measure. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 4344–4357. Curran Associates, Inc., 2022. 50

NREL. Best Research-Cell Efficiency Chart, 2023. URL https://www.nrel.gov/pv/cell-efficiency.html. 198, 201

ODRÉ. Open Data Réseaux Énergies (ODRÉ), 2024. URL https://opendata.reseaux-energies.fr/. 20

Our World in Data. Carbon intensity of electricity, 2024. URL https://ourworldindata.org/grapher/carbon-intensity-electricity. xxiii, 177

M. E. O'Neill and H. M. College. PCG: A Family of Simple Fast Space-Efficient Statistically Good Algorithms for Random Number Generation. *ACM Transactions on Mathematical Software*, 2014. 104

P. Parhar, R. Sawasaki, A. Todeschini, C. Reed, H. Vahabi, N. Nusaputra, and F. Vergara. HyperionSolarNet: Solar Panel Detection from Aerial Images. In *NeurIPS 2021 Workshop on Tackling Climate Change with Machine Learning*, 2021. xl, 9, 34, 89, 197

J. K. Parrish, T. Jones, H. K. Burgess, Y. He, L. Fortson, and D. Cavalier. Hoping for optimality or designing for inclusion: Persistence, learning, and the social network of citizen science. *Proceedings of the National Academy of Sciences*, 116(6):1894–1901, 2019. Publisher: National Acad Sciences. 188

D. Patterson, J. Gonzalez, Q. Le, C. Liang, L.-M. Munguia, D. Rothchild, D. So, M. Texier, and J. Dean. Carbon Emissions and Large Neural Network Training, Apr. 2021. arXiv:2104.10350 [cs]. 175

Pecan Street. Dataport – Pecan Street Inc., 2024. URL https://www.pecanstreet.org/dataport/. 116

X. Peng, B. Usman, N. Kaushik, J. Hoffman, D. Wang, and K. Saenko. VisDA: The Visual Domain Adaptation Challenge, Nov. 2017. arXiv:1710.06924 [cs]. xxxiii, 11

M. Perera, J. De Hoog, K. Bandara, and S. Halgamuge. Multi-resolution, multi-horizon distributed solar PV power forecasting with forecast combinations. *Expert Systems with Applications*, 205:117690, Nov. 2022. ISSN 09574174. doi: 10.1016/j.eswa.2022.117690. 116

V. Petsiuk, A. Das, and K. Saenko. RISE: Randomized Input Sampling for Explanation of Black-box Models. Sept. 2018. doi: 10.48550/arXiv.1806.07421. arXiv:1806.07421 [cs]. xliii, 50

M. Pierro, F. R. Liolli, D. Gentili, M. Petitta, R. Perez, D. Moser, and C. Cornaro. Impact of PV/Wind Forecast Accuracy and National Transmission Grid Reinforcement on the Italian Electric System. *Energies*, 15(23):9086, Nov. 2022. ISSN 1996-1073. doi: 10.3390/en15239086. xxx

E. H. P. Pooch, P. Ballester, and R. C. Barros. Can We Trust Deep Learning Based Diagnosis? The Impact of Domain Shift in Chest Radiograph Classification. In J. Petersen, R. San José Estépar, A. Schmidt-Richberg, S. Gerard, B. Lassen-Schmidt, C. Jacobs, R. Beichel, and K. Mori, editors, *Thoracic Image Analysis*, Lecture Notes in Computer Science, pages 74–83, Cham, 2020. Springer International Publishing. ISBN 978-3-030-62469-9. doi: 10.1007/978-3-030-62469-9_7. xxxiii, 10

S. Puttemans, W. V. Ranst, and T. Goedemé. Detecting of photovoltaic installations in RGB aerial imaging: a comparative study. In *Proceedings of GEOBIA 2016 : Solutions and synergies, 14-16 September 2016, Enschede, Netherlands*. University of Twente, Faculty of Geo-Information Science and Earth Observation (ITC), Sept. 2016. doi: 10.3990/2.429. xxxi, 7

Z. Qu, A. Oumbe, P. Blanc, B. Espinar, G. Gesell, B. Gschwind, L. Klüser, M. Lefèvre, L. Saboret, M. Schroedter-Homscheidt, and L. Wald. Fast radiative transfer parameterisation for assessing the surface solar irradiance: The Heliosat-4 method. *Meteorologische Zeitschrift*, 26(1):33–57, Feb. 2017. ISSN 0941-2948. doi: 10.1127/metz/2016/0781. lv, 118

N. Rahaman, A. Baratin, D. Arpit, F. Draxler, M. Lin, F. Hamprecht, Y. Bengio, and A. Courville. On the Spectral Bias of Neural Networks. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5301–5310. PMLR, June 2019. 51

B. Rausch, K. Mayer, M.-L. Arlt, G. Gust, P. Staudt, C. Weinhardt, D. Neumann, and R. Rajagopal. An Enriched Automated PV Registry: Combining Image Recognition and 3D Building Data. In *NeurIPS 2020 Workshop on Tackling Climate Change with Machine Learning*, 2020. 9, 88, 89, 92, 95, 147, 198

B. Recht, R. Roelofs, L. Schmidt, and V. Shankar. Do ImageNet Classifiers Generalize to ImageNet? In *Proceedings of the 36th International Conference on Machine Learning*, pages 5389–5400. PMLR, May 2019. ISSN: 2640-3498. 10

M. T. Ribeiro, S. Singh, and C. Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2939778. xliii, 50

O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing. ISBN 978-3-319-24574-4. doi: 10.1007/978-3-319-24574-4_28. 100, 107

A. S. Ross, M. C. Hughes, and F. Doshi-Velez. Right for the right reasons: training differentiable models by constraining their explanations. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, pages 2662–2670, Melbourne, Australia, 2017. AAAI Press. ISBN 978-0-9992411-0-3. xxxiii, 11

D. Rotman, J. Hammock, J. Preece, D. Hansen, C. Boston, A. Bowser, and Y. He. Motivations Affecting Initial and Long-Term Participation in Citizen Science Projects in Three Countries. *iConference 2014 Proceedings*, Mar. 2014. doi: 10.9776/14054. Publisher: iSchools. 188

RTE France. Energy Pathways to 2050. Technical report, RTE France, 2022. xiii, xxx, xxxi, 5

RTE France. Comprendre et piloter l'électrification d'ici 2035. Technical report, RTE France, Paris, 2023. xxix, 1

RTE France and IEA. Conditions and Requirements for the Technical Feasibility of a Power System with a High Share of Renewables in France Towards 2050. Technical report, Paris, 2021. xxx, 4

O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, Dec. 2015. ISSN 1573-1405. doi: 10.1007/s11263-015-0816-y. 77, 196

République française. Décret n° 2017-676 du 28 avril 2017 relatif à l'autoconsommation d'électricité et modifiant les articles D. 314-15 et D. 314-23 à D. 314-25 du code de l'énergie, 2017. 148

République française. Décret n° 2020-456 du 21 avril 2020 relatif à la programmation pluriannuelle de l'énergie, 2020. xiii, xxxi, 3

République française. Arrêté du 22 décembre 2023 modifiant l'arrêté du 6 octobre 2021 fixant les conditions d'achat de l'électricité produite par les installations implantées sur bâtiment, hangar ou ombrière utilisant l'énergie solaire photovoltaïque, d'une puissance crête installée inférieure ou égale à 500 kilowatts telles que visées au 3° de l'article D. 314-15 du code de l'énergie et situées en métropole continentale, 2023. 148

K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting Visual Category Models to New Domains. In K. Daniilidis, P. Maragos, and N. Paragios, editors, *Computer Vision – ECCV 2010*, Lecture Notes in Computer Science, pages 213–226, Berlin, Heidelberg, 2010. Springer. ISBN 978-3-642-15561-1. doi: 10.1007/978-3-642-15561-1_16. xxxiii, 10

Y.-M. Saint-Drenan. *A probabilistic approach to the estimation of regional photovoltaic power generation using meteorological data*. PhD thesis, Sept. 2016. 7, 122, 137

Y.-M. Saint-Drenan, S. Bofinger, R. Fritz, S. Vogt, G. Good, and J. Dobschinski. An empirical approach to parameterizing photovoltaic plants for power forecasting and simulation. *Solar Energy*, 120:479–493, Oct. 2015. ISSN 0038092X. doi: 10.1016/j.solener.2015.07.024. xiv, xxxvii, 4, 6, 134

Y.-M. Saint-Drenan, G. Good, M. Braun, and T. Freisinger. Analysis of the uncertainty in the estimates of regional PV power generation evaluated with the upscaling method. *Solar Energy*, 135:536–550, Oct. 2016. ISSN 0038092X. doi: 10.1016/j.solener.2016.05.052. lvi, 6, 127, 128, 146

Y.-M. Saint-Drenan, S. Vogt, S. Killinger, J. M. Bright, R. Fritz, and R. Potthast. Bayesian parameterisation of a regional photovoltaic model–Application to forecasting. *Solar Energy*, 188:760–774, 2019. Publisher: Elsevier. 198

H. Sauermann and C. Franzoni. Crowd science user contribution patterns and their implications. *Proceedings of the National Academy of Sciences*, 112(3):679–684, Jan. 2015. doi: 10.1073/pnas.1408907112. Publisher: Proceedings of the National Academy of Sciences. 188

P. Schulam and S. Saria. Can You Trust This Prediction? Auditing Pointwise Reliability After Learning. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, pages 1022–1031. PMLR, Apr. 2019. ISSN: 2640-3498. xxxiii, 10, 11

A. Segal, Y. K. Gal, R. J. Simpson, V. Victoria Homsy, M. Hartswood, K. R. Page, and M. Jirotka. Improving Productivity in Citizen Science through Controlled Intervention. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, pages 331–337, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 978-1-4503-3473-0. doi: 10.1145/2740908.2743051. 188

R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision*, 128(2):336–359, Feb. 2020. ISSN 1573-1405. doi: 10.1007/s11263-019-01228-7. xiii, xv, xvi, xli, xlii, 41, 42, 48, 50, 64

P. K. Sen. Estimates of the Regression Coefficient Based on Kendall's Tau. *Journal of the American Statistical Association*, 63 (324):1379–1389, Dec. 1968. ISSN 0162-1459. doi: 10.1080/01621459.1968.10480934. Publisher: Taylor & Francis _eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.1968.10480934. 198

L. S. Shapley. *A Value for N-Person Games*. RAND Corporation, Santa Monica, CA, 1952. doi: 10.7249/P0295. 41

E. Sheehan, C. Meng, M. Tan, B. Uzkent, N. Jean, M. Burke, D. Lobell, and S. Ermon. Predicting Economic Development using Geolocated Wikipedia Articles. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, pages 2698–2706, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 978-1-4503-6201-6. doi: 10.1145/3292500.3330784. 103

A. Shrikumar, P. Greenside, A. Shcherbina, and A. Kundaje. Not Just a Black Box: Learning Important Features Through Propagating Activation Differences, Apr. 2017. arXiv:1605.01713 [cs]. 49

K. Simonyan, A. Vedaldi, and A. Zisserman. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. In Y. Bengio and Y. LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Workshop Track Proceedings*, 2014. xv, 49, 50

B. So, C. Nezin, V. Kaimal, S. Keene, L. Collins, K. Bradbury, and J. M. Malof. Estimating the electricity generation capacity of solar photovoltaic arrays using only color aerial imagery. In *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 1603–1606, Fort Worth, TX, July 2017. IEEE. ISBN 978-1-5090-4951-6. doi: 10.1109/IGARSS.2017.8127279. xxxix, 31, 90, 92, 95, 201

I. M. Sobol. On the distribution of points in a cube and the approximate evaluation of integrals. *USSR Computational Mathematics and Mathematical Physics*, 7(4): 86–112, Jan. 1967. ISSN 0041-5553. doi: 10.1016/0041-5553(67)90144-9. 54, 104

I. M. Sobol. On sensitivity estimation for nonlinear mathematical models. *Matematicheskoe modelirovanie*, 2(1):112–118, 1990. Publisher: Russian Academy of Sciences, Branch of Mathematical Sciences. 54

E. Strubell, A. Ganesh, and A. McCallum. Energy and Policy Considerations for Deep Learning in NLP, June 2019. arXiv:1906.02243 [cs]. 175

E. Strubell, A. Ganesh, and A. McCallum. Energy and Policy Considerations for Modern Deep Learning Research. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(09):13693–13696, Apr. 2020. ISSN 2374-3468. doi: 10.1609/aaai.v34i09.7123. Number: 09. 175

J. Sun, A. Mehra, B. Kailkhura, P.-Y. Chen, D. Hendrycks, J. Hamm, and Z. M. Mao. A Spectral View of Randomized Smoothing Under Common Corruptions: Benchmarking and Improving Certified Robustness. In S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, editors, *Computer Vision – ECCV 2022*, volume 13664, pages 654–671. Springer Nature Switzerland, Cham, 2022a. ISBN 978-3-031-19771-0 978-3-031-19772-7. doi: 10.1007/978-3-031-19772-7_38. Series Title: Lecture Notes in Computer Science. 76

T. Sun, M. Segu, J. Postels, Y. Wang, L. Van Gool, B. Schiele, F. Tombari, and F. Yu. SHIFT: A Synthetic Driving Dataset for Continuous Multi-Task Domain Adaptation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21371–21382, 2022b. xxxiii, 10

M. Sundararajan, A. Taly, and Q. Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pages 3319–3328, Sydney, NSW, Australia, 2017. JMLR.org. 49

C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 30, 33, 75, 86, 89, 107

M. R. Taesiri, G. Nguyen, S. Habchi, C.-P. Bezemer, and A. Nguyen. ImageNet-Hard: The Hardest Images Remaining from a Study of the Power of Zoom and Spatial Biases in Image Classification. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 35878–35953. Curran Associates, Inc., 2023. 65

H. Theil. A Rank-Invariant Method of Linear and Polynomial Regression Analysis. In B. Raj and J. Koerts, editors, *Henri Theil's Contributions to Economics and Econometrics: Econometric Theory and Methodology*, Advanced Studies in Theoretical and Applied Econometrics, pages 345–381. Springer Netherlands, Dordrecht, 1992. ISBN 978-94-011-2546-8. doi: 10.1007/978-94-011-2546-8_20. 198

N. Thompson, K. Greenewald, K. Lee, and G. F. Manso. The Computational Limits of Deep Learning. In *Computing within Limits*. LIMITS, June 2023. doi: 10.21428/bf6fb269.1f033948. 175

A. Torralba and A. A. Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528, June 2011. doi: 10.1109/CVPR.2011.5995347. ISSN: 1063-6919. 10

H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jegou. Training data-efficient image transformers & distillation through attention. In *Proceedings of the 38th International Conference on Machine Learning*, pages 10347–10357. PMLR, July 2021. ISSN: 2640-3498. 89, 100, 107

A. Trockman and J. Z. Kolter. Patches Are All You Need? *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. 89, 100, 107

Y. Trémenbert, G. Kasmi, L. Dubus, Y.-M. Saint-Drenan, and P. Blanc. PyPVRoof: a Python package for extracting the characteristics of rooftop PV installations using remote sensing data, Sept. 2023. arXiv:2309.07143 [eess]. xvii, xix, liv, lxi, 14, 32, 90, 91, 92, 93, 109, 145, 199, 205

D. Tuia, C. Persello, and L. Bruzzone. Domain Adaptation for the Classification of Remote Sensing Data: An Overview of Recent Advances. *IEEE Geoscience and Remote Sensing Magazine*, 4(2):41–57, June 2016. ISSN 2168-6831. doi: 10.1109/MGRS.2016.2548504. Conference Name: IEEE Geoscience and Remote Sensing Magazine. xxxiii, 9, 10, 64

USGS. The National Map | U.S. Geological Survey, 2024. URL https://www.usgs.gov/programs/national-geospatial-program/national-map. xiv, 16

B. Uzkent and S. Ermon. Learning When and Where to Zoom With Deep Reinforcement Learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12345–12354, 2020. 103

V. Vapnik. *The nature of statistical learning theory*. Springer science & business media, 1999. xxxiii, lii, 10, 78, 79, 81, 197

K. Vishniakov, Z. Shen, and Z. Liu. ConvNet vs Transformer, Supervised vs CLIP: Beyond ImageNet Accuracy, Nov. 2023. arXiv:2311.09215 [cs]. 108

A. Walch, R. Castello, N. Mohajeri, and J.-L. Scartezzini. Big data mining for the estimation of hourly rooftop photovoltaic potential and its uncertainty. *Applied Energy*, 262:114404, 2020. Publisher: Elsevier. 92, 126, 132

A. Walch, M. Rüdisüli, R. Castello, and J.-L. Scartezzini. Quantification of existing rooftop PV hourly generation capacity and validation against measurement data. *Journal of Physics: Conference Series*, 2042(1):012011, Nov. 2021. ISSN 1742-6588, 1742-6596. doi: 10.1088/1742-6596/2042/1/012011. lv, 116

H. Wang, X. Wu, Z. Huang, and E. P. Xing. High-Frequency Component Helps Explain the Generalization of Convolutional Neural Networks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8681–8691, June 2020. doi: 10.1109/CVPR42600.2020.00871. 51, 58

M. Wang and W. Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153, 2018. Publisher: Elsevier. 98

M. Wang, Q. Cui, Y. Sun, and Q. Wang. Photovoltaic panel extraction from very high-resolution aerial imagery using region–line primitive association analysis and template matching. *ISPRS Journal of Photogrammetry and Remote Sensing*, 141: 100–111, July 2018. ISSN 09242716. doi: 10.1016/j.isprsjprs.2018.04.010. xxxii

R. Wang, J. Camilo, L. M. Collins, K. Bradbury, and J. M. Malof. The poor generalization of deep convolutional networks to aerial imagery from new geographic locations: an empirical study with solar array detection. In *2017 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, pages 1–8, Washington, DC, Oct. 2017. IEEE. ISBN 978-1-5386-1235-4. doi: 10.1109/AIPR.2017.8457965. xli, 9, 37, 65

S. Wang, R. Veldhuis, C. Brune, and N. Strisciuglio. What do neural networks learn in image classification? A frequency shortcut perspective. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1433–1442, Oct. 2023. 51

Z. Wang, Z. Wang, A. Majumdar, and R. Rajagopal. Identify Solar Panels in Low Resolution Satellite Imagery with Siamese Architecture and Cross-Correlation. In *NeurIPS 2019 Workshop on Tackling Climate Change with Machine Learning*, 2019. 10

Z. Wang, M.-L. Arlt, C. Zanocco, A. Majumdar, and R. Rajagopal. DeepSolar++: Understanding residential solar adoption trajectories with computer vision and technology diffusion models. *Joule*, 6(11):2611–2625, Nov. 2022. ISSN 25424351. doi: 10.1016/j.joule.2022.09.011. lxi, 145

D. Yin, R. Gontijo Lopes, J. Shlens, E. D. Cubuk, and J. Gilmer. A Fourier Perspective on Model Robustness in Computer Vision. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d. Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. xliv, 51, 58

J. Yu, Z. Wang, A. Majumdar, and R. Rajagopal. DeepSolar: A Machine Learning Framework to Efficiently Construct a Solar Deployment Database in the United States. *Joule*, 2(12):2605–2617, Dec. 2018. ISSN 2542-4351. doi: 10.1016/j.joule.2018.11.021. xvii, xxxii, xxxix, 8, 9, 86, 87, 88, 92

J. Yuan, H.-H. L. Yang, O. A. Omitaomu, and B. L. Bhaduri. Large-scale solar panel mapping from aerial images using deep convolutional networks. In *2016 IEEE International Conference on Big Data (Big Data)*, pages 2703–2708, Washington DC,USA, Dec. 2016. IEEE. ISBN 978-1-4673-9005-7. doi: 10.1109/BigData.2016.7840915. xxxii, 8

M. Zech and J. Ranalli. Predicting PV Areas in Aerial Images with Deep Learning. In *2020 47th IEEE Photovoltaic Specialists Conference (PVSC)*, pages 0767–0774, Calgary, AB, Canada, June 2020. IEEE. ISBN 978-1-72816-115-0. doi: 10.1109/PVSC45281.2020.9300636. xl, 9, 34

M. D. Zeiler and R. Fergus. Visualizing and Understanding Convolutional Networks. In D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 818–833, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10590-1. doi: 10.1007/978-3-319-10590-1_53. xliii, 50

H. Zhang, J. E. Fritts, and S. A. Goldman. Image segmentation evaluation: A survey of unsupervised methods. *Computer Vision and Image Understanding*, 110(2): 260–280, May 2008. ISSN 1077-3142. doi: 10.1016/j.cviu.2007.08.003. xxxix, 27

J. Zhang, S. A. Bargal, Z. Lin, J. Brandt, X. Shen, and S. Sclaroff. Top-Down Neural Attention by Excitation Backprop. *International Journal of Computer Vision*, 126 (10):1084–1102, Oct. 2018. ISSN 1573-1405. doi: 10.1007/s11263-017-1059-x. 41

J. Zhang, X. Jia, and J. Hu. Pseudo Supervised Solar Panel Mapping based on Deep Convolutional Networks with Label Correction Strategy in Aerial Images. In *2020 Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–8, Melbourne, Australia, Nov. 2020. IEEE. ISBN 978-1-72819-108-9. doi: 10.1109/DICTA51227.2020.9363379. 89

J. Zhang, X. Jia, and J. Hu. SP-RAN: Self-paced Residual Aggregated Network for Solar Panel Mapping in Weakly Labelled Aerial Images. *IEEE Transactions on Geoscience and Remote Sensing*, 2021a. Publisher: IEEE. 89

J. Zhang, H. Chao, A. Dhurandhar, P.-Y. Chen, A. Tajer, Y. Xu, and P. Yan. When Neural Networks Fail to Generalize? A Model Sensitivity Perspective. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(9):11219–11227, June 2023. ISSN 2374-3468. doi: 10.1609/aaai.v37i9.26328. Number: 9. 82

Y. Zhang, D. Hong, D. McClement, O. Oladosu, G. Pridham, and G. Slaney. Grad-CAM helps interpret the deep learning models trained to classify multiple sclerosis types using clinical brain magnetic resonance imaging. *Journal of Neuroscience Methods*, 353:109098, Apr. 2021b. ISSN 01650270. doi: 10.1016/j.jneumeth.2021.109098. xli, 41

Z. Zhang, D. Meng, L. Zhang, W. Xiao, and W. Tian. The range of harmful frequency for DNN corruption robustness. *Neurocomputing*, 481:294–309, Apr. 2022. ISSN 09252312. doi: 10.1016/j.neucom.2022.01.087. 51, 58

B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning Deep Features for Discriminative Localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 49

K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy. Domain Generalization: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4396–4415, Apr. 2023. ISSN 1939-3539. doi: 10.1109/TPAMI.2022.3195549. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence. xxxiii, 10

Appendices

Appendix A

Discussion of the environmental impact

1 Literature and proposed approach

Environmental impact of deep learning I would also like to broaden the scope of the results obtained in this PhD thesis. Recent studies (Strubell et al., 2019, 2020; Luccioni et al., 2024; Patterson et al., 2021; Thompson et al., 2023; Luccioni et al., 2023) raised concerns regarding the environmental impact of deep learning. For instance, Strubell et al. (2019) showed that the environmental impact of training a large language model (LLM) is equivalent to the impact of 5 cars over their whole life cycle. However, I was surprised that few studies report the environmental impact of the models that they benchmark. A proper assessment of the environmental impact of deep learning models is still an open problem (Luccioni et al., 2024), and data is often lacking for a comprehensive assessment of the environmental impact using the lifecycle assessment (LCA) methodology.

Proposed approach and metrics In chapter 4, sections 2.3.3 and 3.1.3, I showed that our sampling method reduced the number of thumbnails to generate, which led to an overall gain in efficiency which amounts to days when scaling to the size of France. In this section, I go a step further and introduce a simplified framework for expressing these efficiency gains in terms of the environmental impact of the mapping algorithm.

The main goal of this framework is to trigger further research and encourage researchers to report the environmental impact of their models. Reporting the environmental impact of models gradually becomes standard practice, as the platform HuggingFace now reports the carbon intensity of its models (Hugging Face, 2023), and I wish to subscribe to this trend. I propose a simplified framework to enable fast computations with low overhead. This framework aims to estimate the model's

energy consumption in Wh for its training and deployment. I assume that the energy consumption scales linearly with the number of model parameters for training and the inference time for deployment.

Our framework requires only two inputs: the number of parameters of the model at hand and the total time for inference. To evaluate the energy consumption of training, I took the number reported by Luccioni et al. (2024) while training the LLM BLOOM. This model has 176B (billion) parameters and requires 433,195,000 Wh of electricity for training. This gives us an energy consumption of $E_{train} = 0.0025$ Wh of electricity per parameter. To evaluate the energy consumption during inference, I used the Python library CodeCarbon (CodeCarbon, 2023) and measured the electricity consumption while running the inference scripts. On average, I measured an electricity consumption $E_{inference} = 805$ Wh with my hardware. The total energy consumption is given as

$$\begin{aligned} C_{total}[\text{Wh}] = \#\text{ parameters} \times E_{train} \\ + T_{inference} \times E_{inference} \end{aligned} \tag{A.1}$$

where $T_{inference}$ denotes the time to run the pipeline in hours, I denote $C_{inference} \equiv T_{inference} \times E_{inference}$ and $C_{train} \equiv \#\text{ parameters} \times E_{train}$ Our framework is simplified and depends on our hardware to estimate $E_{inference}$. Nonetheless, it will serve as a baseline for further discussions and encourage systematically reporting the energy consumption of deep learning models in benchmarks.

2 Results

2.1 Energy consumption

In Table A.1, I first report C_{train} and $C_{inference}$ for a representative set of our models. I can see how the sampling module dramatically reduces the inference time and, thus, the associated energy cost.

Table A.1 – Computation of C_{train} and $C_{inference}$ for some selected variants of the mapping algorithm.

Variant	Training # Parameters [-]	C_{train} [Wh]	Runtime [sec/km²]	Inference Scale-up [days]	$C_{inference}$ [Wh]
Baseline (Kasmi et al., 2022a)	25M	62,500	19.39	122.08	2 372,496
ResNet-50	25M	62,500	16.78	105.50	2,038,260
ConvNext + Sampling	87M	217,500	12.77	80.37	1,552,748
ResNet-50 + Sampling	25M	62,500	13.19	83.04	1,604,333

Then, in Table A.2, I sum C_{train} and $C_{inference}$ to report an estimation of the total energy consumption of different variants of the pipeline, according to equation

(A.1). I consider the baseline and variants that yielded close results regarding the DTA. The impact of training is small (10 % of the total energy consumption) compared to the impact of inference. It is even more true as training is done only once, whereas one performs inference regularly. My measure of the environmental impact suggests that the most significant room for improvement lies in the efficiency of the mapping algorithm (i.e., non-deep learning factors). As the yearly expected yield for a 3kW$_p$ installation in France lies between 2.5 (North) and 3.5 (South) MWh/year BDPV (2023), the consumption of the algorithm represents about half of the yearly expected production of an average individual 3kWp rooftop PV system in France.

Table A.2 – Total energy consumption in Wh of deploying variants of our pipeline. Best results are **bolded**.

Pipeline	**Energy consumption [Wh]**		
	C_{train}	$C_{inference}$	C_{total}
Baseline (Kasmi et al., 2022a)	62,500	2,372,496	2,434,996
ConvNext + Sampling	217,500	1,552,748	1,770,248
ResNet+ Sampling	62,500	1,604,333	**1,666,833**

2.2 Environmental impact

Using this framework, it is easy to convert the electric consumption into a carbon intensity using the carbon intensity of the electric grid. As my model was trained and deployed on servers located in France, taking France's average carbon intensity for 2022, I obtained an overall impact in CO_2e. Table A.3 shows the carbon intensity of our algorithm depending on its location. These results indicate that the decarbonization of the grid reduces the environmental impact of deep learning.

Table A.3 – Carbon intensity of DeepPVMapper (ResNet + Sampling). Source of the carbon intensities: Our World in Data (2024).

Country	Grid carbon intensity (2022) [gCO_2e/kWh]	Environmental impact [kg CO_2e]
France	85	141.68
Sweden	45	75.00
Germany	385	641.73
Poland	635	1058.44
US	367	611.73

Appendix B

The training dataset BDAPPV

1 Additional details on the data extraction and the raw data records

1.1 Extraction of the raw data

Raw data extraction Our annotation campaign leverages the database of PV systems operated by the non-profit association *Asso BDPV* (*Base de données Photovoltaïque* - Photovoltaic database). *Asso BDPV* (BDPV) gathers metadata (geolocation and metadata of the PV systems) and the energy production data of PV installations provided by individual system owners, mainly in France and Western Europe. The primary purpose of the BDPV database is to monitor system owners' energy production. BDPV also promotes PV energy by disseminating information and data to the general public and authorities.

The BDPV data contains the localization of more than 28,000 installations. We used this localization to extract the panels' thumbnails. During the first annotation campaign, we extracted 28,807 thumbnails using Google Earth Engine (GEE, Gorelick et al. (2017) application programming interface (API). For the second campaign, we extracted 17,325 thumbnails from the IGN Geoservices portal (https://geoservices.ign.fr/bdortho).

Our thumbnails all have a resolution of 400×400 pixels. Thumbnails extracted from GEE API correspond to a ground sampling distance (GSD) of 0.1 m/pixel. The API directly generates this thumbnail by setting the zoom level to 20, the localization to the ground truth localization contained in BDPV, and the output size to be 400×400 pixels. For IGN images, the resolution of the thumbnails corresponds to a GSD of 0.2 m/pixel. The procedure for generating IGN thumbnails differs from Google. First, we downloaded geo-localized tiles from IGN's Geoservices portal. These tiles have a resolution of 25,000×25,000 pixels, covering an area of 25km². Then, we extracted the thumbnail by generating a 400×400 pixels raster centered

around the location of the PV panel. Finally, we export this raster as a `.png` file. We do not publish the exact location of the panels for confidentiality reasons. We illustrate our training dataset generation workflow in Figure B.1. It comprises three main steps: raw data extraction, thumbnail annotation, and metadata matching.

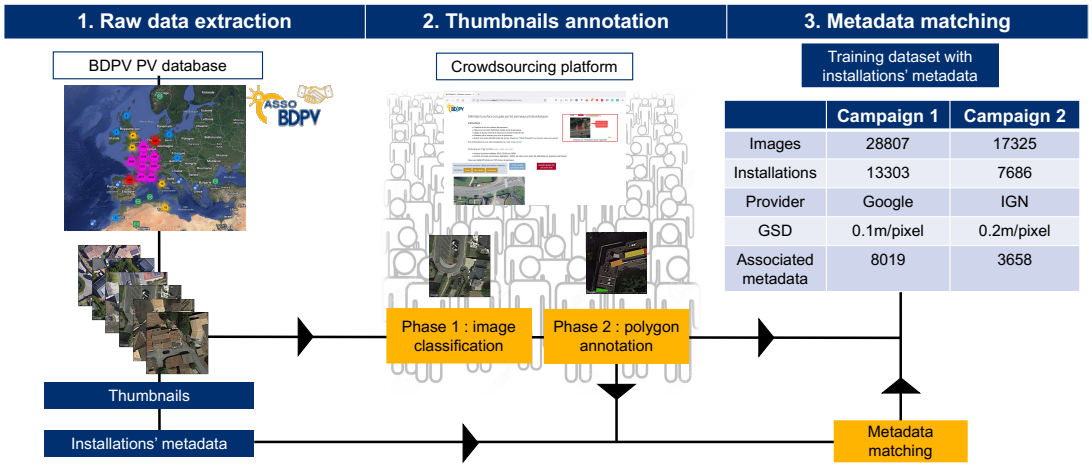


Figure B.1 – Flowchart of the training dataset generation based on the BDPV PV data and crowdsourcing. "GSD" stands for the ground sampling distance, i.e., the distance between the centers of two adjacent pixels measured on the ground. Taken from Kasmi et al. (2023d).

Image classification and polygon annotation This paragraph presents the main steps we followed to construct our database. In section 2, we detail the crowdsourcing campaign itself.

We extracted thumbnails based on the geolocation of the installations recorded in the BDPV dataset. However, this geolocation can be inaccurate, so before asking users to draw polygons of PV installations, we asked them to classify the images. This corresponds to the first phase of the annotation campaign. Once users classified images, we asked them to draw the PV polygons on the remaining images. This corresponds to the second phase of the crowdsourcing campaign.

We designed our campaign to get at least five annotations per image. It enabled us to derive so-called *consensus metrics*, targeted at measuring the quality of our labels. This way, we go further than the consensus between two annotators reported in previous work (Bradbury et al., 2016) to measure annotation quality. The analysis of the users' annotations during phases 1 and 2 are reproducible using the notebook `annotations` available on the public repository [1].

During the first phase, the user clicks on an image if it depicts a PV panel. We recorded the localization of the user's click and instructed them to click *on* the PV panel if there was one. We collected an average of 10 actions (click with localization or no click) per image. We apply the kernel density estimate (KDE) algorithm to the

1. The repository is accessible at this URL: https://github.com/gabrielkasmi/bdappv

annotations to estimate a confidence level for the annotations and the approximate localization of the PV panel on the image. The likelihood $f_\sigma(x_i)$ of presence of a panel for each pixel x_i is given by

$$f_\sigma(x_i) = \frac{1}{N} \sum_{k=1}^{N} K_\sigma(x_k - x_i), \tag{B.1}$$

where K_σ is a Gaussian kernel with a standard deviation σ, x_k is the coordinate of the k^{th} annotation, and N is the total number of annotations.

After an empirical investigation, we calibrated the standard deviation of the kernel to reflect the approximate spatial extent of an array on the image. We set its value to 25 pixels for Google images and 12 for IGN images. It corresponds to a distance of 2.5 m. The maximum value of the KDE quantifies the confidence level of the annotation. We refer to it as the *pixel annotation consensus* (PAC). This metric is proportional to the number of annotations. We use the PAC to determine whether an image contains an array.

During the second phase, annotators delineate the PV panels on the images validated during phase 1. Users can draw as many polygons as they want. On average, we collected five polygons per image. We collect the coordinates of the polygons drawn by the annotators. However, these false positives have fewer annotations than true positives. To select only the true positives, we compute the PAC through the following steps:

1. We convert each user's polygon into a binary raster;
2. We compute the normalized PAC by summing all rasters and dividing by the number of annotators,
3. We apply a relative threshold and keep only the pixels whose PAC is greater than the threshold;
4. We compute the coordinates of the resulting mask using OpenCV's polygon detection algorithm (https://docs.opencv.org/3.4/d4/d73/tutorial_py_contours_begin.html).

In step 2., the unnormalized PAC takes values between 0 and the number N_i of annotators for the i^{th} image. 0 means no user included the pixel into his polygon, and N_i means that *all* annotators encapsulated the corresponding pixel in their polygons.

Metadata matching Once we generate our PV panel polygons (i.e., segmentation masks), we match them with the installations' metadata reported in the BDPV dataset. Our matching procedure follows three steps: internal consistency, unique matching, and external consistency. Note that we only apply these filters when matching the metadata and the masks.

Internal consistency ensures that the entries in the BDPV dataset are coherent before any matching. It is simply a cleaning of the raw dataset. To do this cleaning, we verify whether the information in one column of the BDPV dataset is coherent with the records from the other columns. For instance, if a PV system's record says it has ten modules and a surface of $3m^2$, this would mean that each PV module has a surface of $0.3m^2$, which is impossible (the smallest size being $1.7m^2$).

Our segmentation masks may depict more than one array. It occurs if, for instance, more than one panel is on the image shown to the annotators. In this case, we adopt a conservative view: if the segmentation mask depicts more than one panel, we cannot know which corresponds to the installation reported in the BDPV dataset. In this case, we do not match the segmentation mask with an installation.

After internal consistency filtering and unique matching, we are left with segmentation masks depicting single panels with coherent metadata. A final filtering step consists in making sure that the characteristics reported in the database match those that can be deduced from the segmentation mask. We assess the adequacy between the surface of the installation's mask and its true surface, which is reported in the BDPV dataset, by computing the ratio between them. We keep only installations whose ratio equals 1 (with a tolerance bandwidth of $\pm$ 25%). We apply this bandwidth to accommodate the possible approximations in the segmentation mask. The reported surface excludes the inter-panel space and the distortions induced by the panel's projection on the image, as images are not perfectly orthorectified.

1.2 Data and quality checks

A training dataset The training dataset containing RGB images, ready-to-use segmentation masks of the two campaigns, and the file containing PV installations' metadata is accessible on our Zenodo repository at this URL: https://zenodo.org/record/7358126. It is organized as follows:

- `bdappv/` Root data folder
  - `google` / `ign` One folder for each campaign
    - `img` Folder containing all the images presented to the annotators. This folder contains 28,807 images for Google and 17,325 for IGN. We provide all images as `.png` files.
    - `mask` Folder containing all segmentation masks generated from the polygon annotations of the annotators. This folder contains 13,303 masks for Google and 7,686 for IGN. We provide all masks as `.png` files.
  - `metadata.csv` The `.csv` file with the metadata of the installations. Table B.1 describes the attributes of this table.

Table B.1 – Data attributes and description of the `metadata.csv` data file. Taken from Kasmi et al. (2023d).

Field	Attribute name	Description	Format	Unit
Installation ID	idInstallation	The ID of the installation	Integer	-
Identifier	identifiant	The name of the image of the installation	String	-
Inverter ID	idInverter	The ID of the inverter of the installation	Integer	-
Inverter name	nameInverter	The name of the inverter of the installation	String	-
Number of inverters	countInverters	The number of inverters of the installation	Integer	-
Arrays ID	idArrays	The ID of the solar arrays used by the installation	Integer	-
Arrays' name	nameArrays	The name of the solar arrays used by the installation	Float	-
Number of arrays	countArrays	The number of PV arrays (modules) of the installation	Integer	-
Surface	surface	The surface (in m²) of the installation	Float	m^2
Azimuth	azimuth	The azimuth angle in degrees relative to the north (south = 180) of the installation.	Float	Degrees
Installation type	typeInstallation	Indicates on which infrastructure the installation is mounted: - 0: rooftop - 1: unknown - 2: rooftop of a non-livable building - 3: ground - 4: other - 5: shade house - 6: sunshade - 7: solar tracker with one axis - 8: solar tracker with two axes	Integer	-
Tilt	tilt	The tilt angle of the installation	Integer	Degrees
Installed capacity	kWp	The installed capacity of the installation in kWp	Float	kWp
Date of installation	dateInstalled	The date (month, year) the installation has been installed	String	Date
Is integrated	isIntegrated	Indicates if the installation is integrated (on the rooftop)	Boolean	-
Self-consumption	selfConsumption	Indicates if the installation is used for self-consumption (alternative is that PV power is reinjected into the grid)	Boolean	-
Département	departement	The *département* (county) in which the installation is located	Integer	-
City	city	The city where the installation is located	String (UTF-8)	-
Controlled	Controlled	Indicates whether the installations' metadata are clean	Boolean	-
Matched with IGN image	IGNControlled	Indicates whether the installation corresponds to a unique segmentation mask corresponding to an IGN image	Boolean	-
Matched with Google image	GoogleControlled	Indicates whether the installation corresponds to a unique segmentation mask corresponding to a Google image	Boolean	-

Figure B.2 presents some examples of images from the BDAPPV dataset.

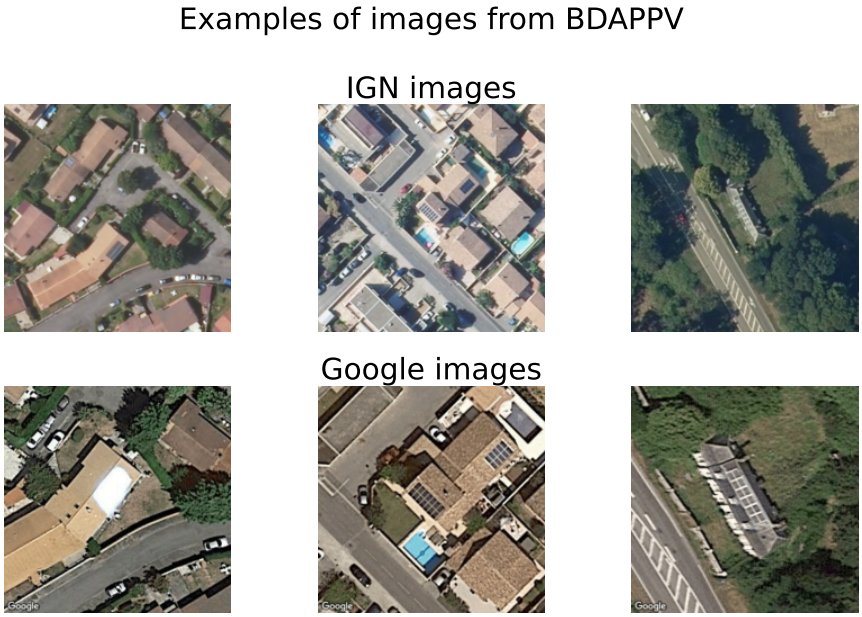


Figure B.2 – Examples of images from the BDPV training database.

The raw crowdsourcing data In addition to the training dataset, we also released the raw crowdsourcing data. It is structured as follows: the `raw` subfolder contains the raw annotation data from the two annotation campaigns and the raw PV installations' metadata. The `replication` subfolder contains the compiled data for generating our segmentation masks. The `validation` subfolder contains the compiled data necessary to replicate the analyses presented in the technical validation section.

- `data/` Root data folder
  - `raw/` Folder containing the raw crowdsourcing data and raw metadata;
    - `input-google.json`: Input data containing all information on images and raw annotators' contributions for both phases (clicks and polygons) during the first annotation campaign;
    - `input-ign.json`: Input data containing all information on images and raw annotators' contributions for both phases (clicks and polygons) during the second annotation campaign;
    - `raw-metadata.csv`: The file containing the PV systems' metadata extracted from the BDPV database before filtering. It can be used to replicate the association between the installations and the segmentation masks, as done in the notebook `metadata`.
  - `replication/` Folder containing the compiled data used to generate the segmentation masks;

- `campaign-google` / `campaign-ign`. One folder for each campaign
  - `click-analysis.json`: Output on the click analysis, compiling raw input into a few best-guess locations for the PV arrays. This dataset enables the replication of our annotations;
  - `polygon-analysis.json`: Output of polygon analysis, compiling raw input into a best-guess polygon for the PV arrays.
- `validation/` Folder containing the compiled data used for technical validation.
  - `campaign-google` / `campaign-ign`. One folder for each campaign
    - `click-analysis-thres=1.0.json`: Output of the click analysis with a lowered threshold to analyze the effect of the threshold on image classification, as done in the notebook `annotations`;
    - `polygon-analysis-thres=1.0.json`: Output of polygon analysis, with a lowered threshold to analyze the effect of the threshold on polygon annotation, as done in the notebook `annotations`.
  - `metadata.csv` the filtered installations' metadata.

Quality checks Throughout the generation of the training dataset, we tested whether the threshold values chosen to classify the images, construct the polygon, and associate the polygons with the installations' metadata yielded as few errors as possible. We base our approach on a consensus metric to classify images and construct the polygons, namely the pixel annotation consensus (PAC). Thus, we improve on Bradbury et al. (2016), who proposed a confidence value based on the Jaccard Similarity Index (Levandowsky and Winter, 1971) between the two annotations. As for the association between the polygons and installations' metadata, we balance between accuracy and keeping as many installations as possible.

As mentioned in the methods section, the choice criterion for image classification during phase 1 is the consensus among users. We empirically investigated a range of thresholds and determined that a value of 2.0 yielded the most accurate classification results. In other words, we require that at least three annotators click around the same point to validate the classification.

We use an *absolute* (unnormalized by the number of annotators for this image) threshold to decide whether the image contains a panel. The threshold is absolute because users could only click once on the image during the annotation campaign, even if the latter contained more than one array. As such, an absolute threshold does not dilute the consensus among users when there is more than one panel on the image.

The leftmost plot of Figure B.3 plots the histogram of the absolute PAC. Visual inspection revealed that the peak for values below 2.0 corresponded to false positives. We enable replication of the threshold analysis in the notebook `annotation`.

We used a consensus metric to merge the users' annotations like the click annotation. After empirical investigations, we found that a *relative* threshold (expressed as a share of the total number of annotators) was the most effective for yielding the most accurate masks and that its value should be 0.45. In other words, we consider that a pixel depicts an installation if at least 45% of the annotators included it in their polygons.

The center plot of Figure B.3 depicts the histogram of the relative PAC. Visual inspection revealed that the few values below 0.45 corresponded to remaining false positives (e.g., roof windows). The use of a relative threshold is motivated by the fact that the users can annotate as many polygons as they want. We enable replication of the threshold analysis in the notebook `annotation`.

We link segmentation masks and installation metadata according to the steps described in the section "Metadata matching ." To measure the quality of this linkage, we measure the Pearson correlation coefficient (PCC) between the surface reported in the installation metadata dataset (referred to as the "target" surface) and the surface estimated from the segmentation masks (referred to as the "estimated" surface). The higher the PCC, the better our matching procedure.

Figure B.3 plots estimated and target surfaces. After filtering, we obtain a PCC coefficient of 0.99 between the target and estimated surfaces. Without filtering, the PCC coefficient equals 0.68 for Google images and 0.61 for IGN images. It shows that our metadata-matching procedure enabled us to pick the installations with the best fit between the reported surface and the surface estimated from the segmentation masks.

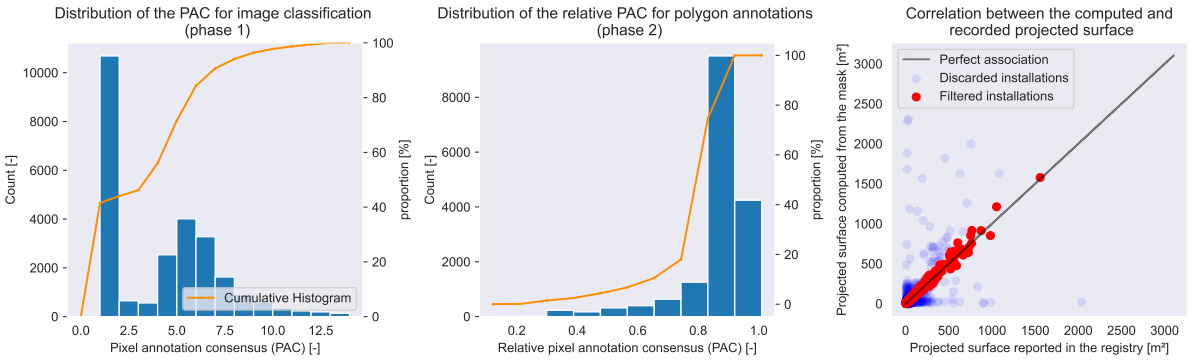


Figure B.3 – Validation by comparison of the surface estimated from the masks and the surface reported in the PV installations' metadata. Taken from Kasmi et al. (2023d).

Our matching procedure comprises three steps: internal consistency, uniqueness, and external consistency. Each of these steps discards installations from the BDPV database. Figure B.4 summarizes the number of installations filtered at each process step. We can see that most filtering happens when we discard segmentation masks with multiple installations.

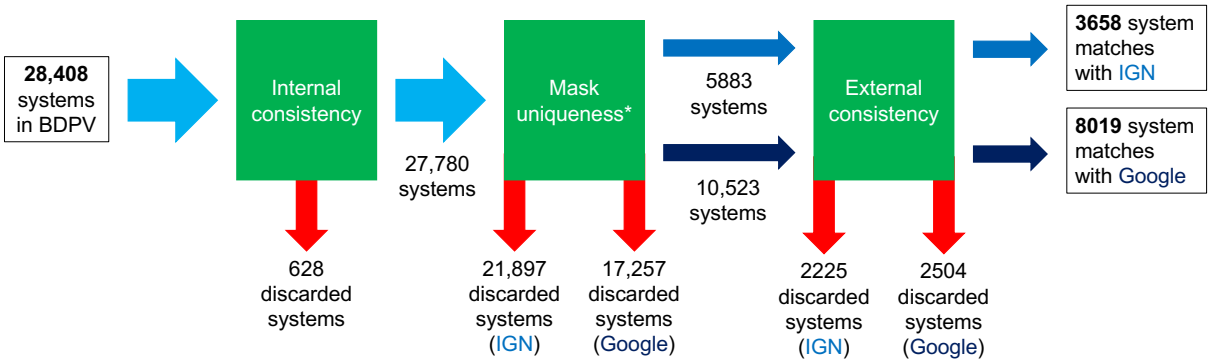


Figure B.4 – Number of installations filtered through the different filtering steps during the association between the masks and the installations' metadata. *During the mask uniqueness step, we account for the fact that (a) not all BDPV installations were identified on images (13,303 were identified on Google images and 7,686 on IGN images) and (b) among these identified installations, some of the masks contained more than one polygon. Adapted from Kasmi et al. (2023d).

2 Crowdsourcing campaign analysis

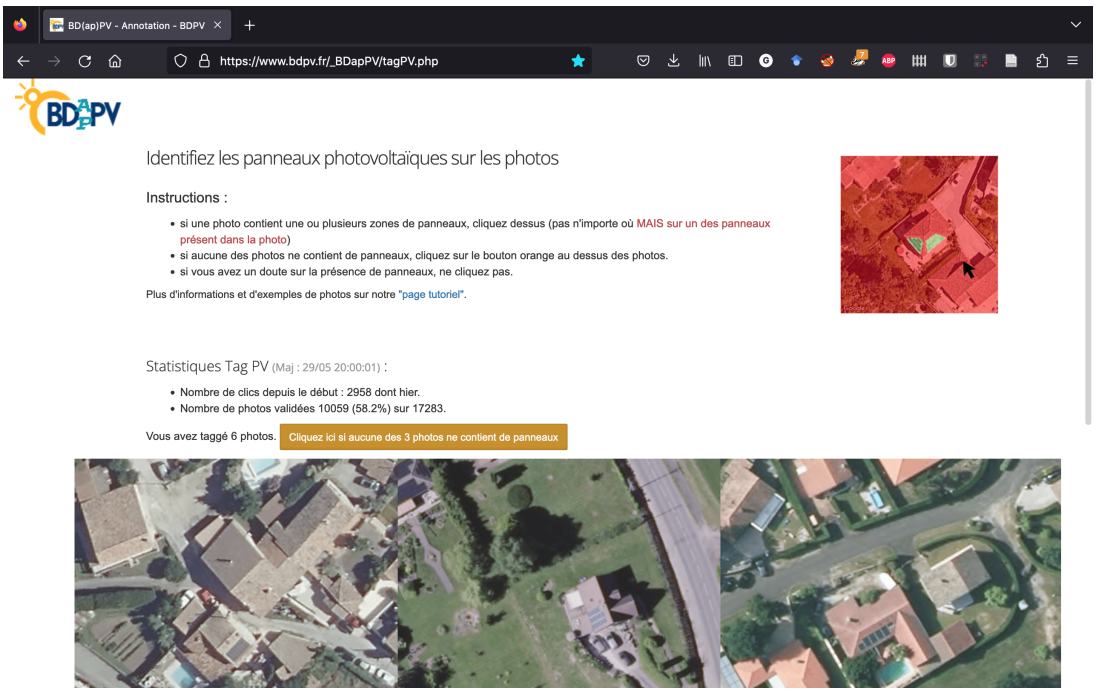


Figure B.5 – Screenshot of the BDAPPV crowdsourcing platform (first phase).

The crowdsourcing platform Figure B.5 depicts screenshots of the crowdsourcing platform. This platform is accessible at the URL `https://www.bdpv.fr/_BDapPV/` although it does not receive contributions anymore. Each campaign phase has its dedicated webpage where users can annotate the PV panels.

Overall statistics Table B.2 summarizes the annotators contributions during the crowdsourcing campaigns. We can see more annotators for the first campaign than for the second. This may be because we advertised more for this campaign.

Table B.2 – Summary statistics of the contributions during the crowdsourcing campaigns. Source: Kasmi et al. (2023d).

	Google	**IGN**
Total number of actions	349,394	119,528
Total number of annotators	1,901	1,021
Actions during phase 1	291,597	90,084
Actions during phase 2	68,162	29,444
Active annotators during phase 1	1,043	51
Active annotators during phase 2	960	980
Active annotators during both phases	102	10

The contributions are very concentrated, as seen from Figure B.6. Thirty users make almost two-thirds of the contributions during the crowdsourcing campaign. We can see that for both phases, the pattern is the same. This result is consistent with prior studies (Parrish et al., 2019; Rotman et al., 2014; Segal et al., 2015; Sauermann and Franzoni, 2015) on crowdsourcing efforts, which documented such a Pareto law for the contributions.

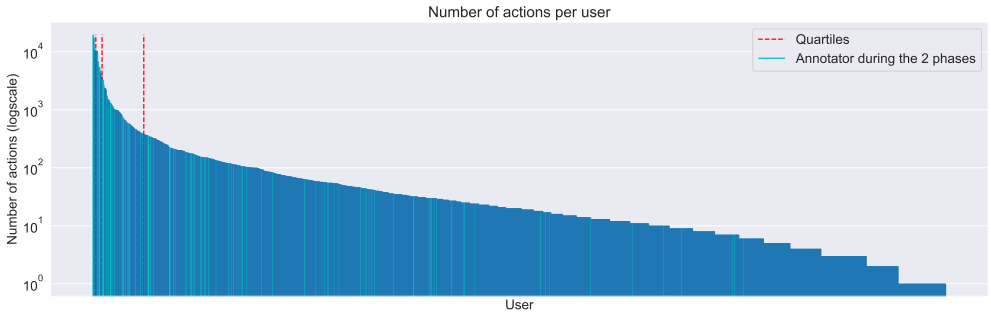


Figure B.6 – Number of annotations per user.

Temporal dynamics Figure B.7 presents the daily number of contributions during the first campaign. We can see about ten times more daily contributions during the first phase than during the second. Also, we can see that annotators' engagement decreases over time. On Figure B.7, "communication" indicates a public post on social media. We can see a response to the first post, less to the second one.

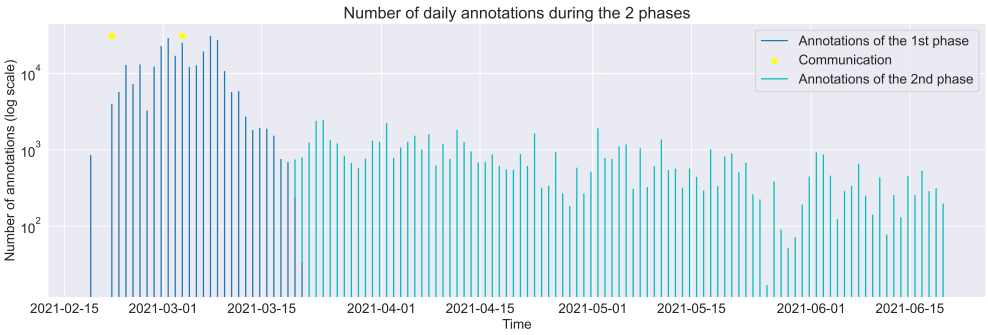


Figure B.7 – Number of daily annotations (log scale) during the first crowdsourcing campaign. Light blue indicates the second phase, and dark blue the first phase.

Appendix C

Supplementary materials

1 Supplementary material to chapter 2

1.1 Additional plots of the distribution of the tilt angle

Figures C.1 to C.3 present additional plots based on the model of Figure 2.12. We can see that the mean tilt angle values depend on the latitude, with angles steeper in the North than in the South. Localizations are taken in the North (Figure C.1), South (Figure C.2) and East (Figure C.3).

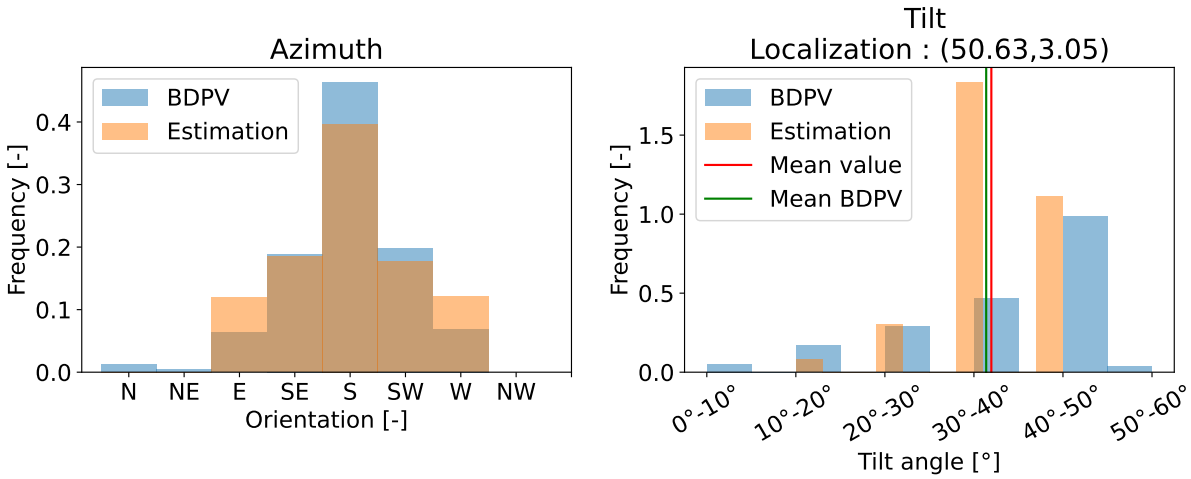


Figure C.1 – Comparison of the distribution of the azimuth (left) and tilt (right) angles obtained from our mapping algorithm and BDPV.

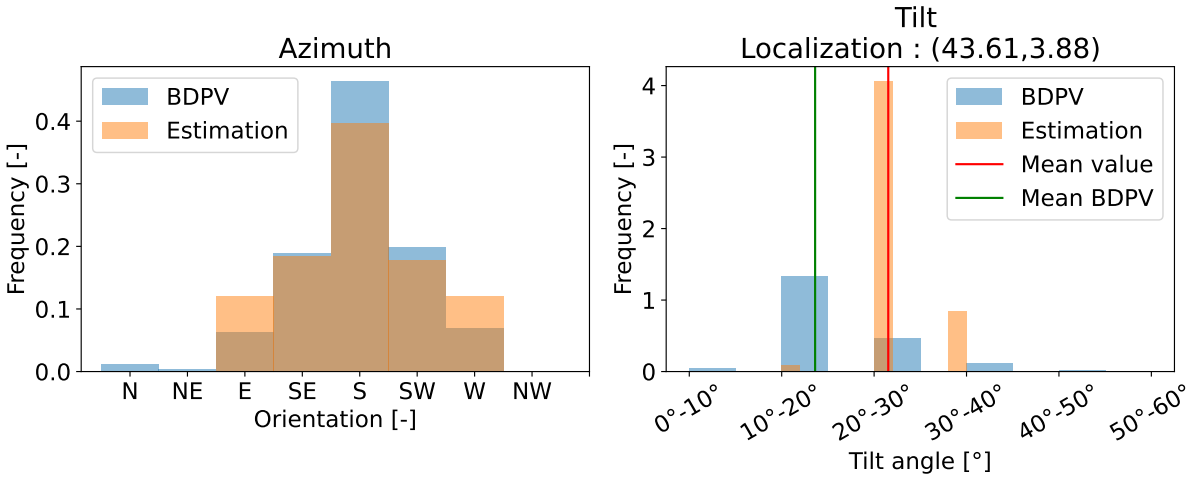


Figure C.2 – Comparison of the distribution of the azimuth (left) and tilt (right) angles obtained from our mapping algorithm and BDPV.

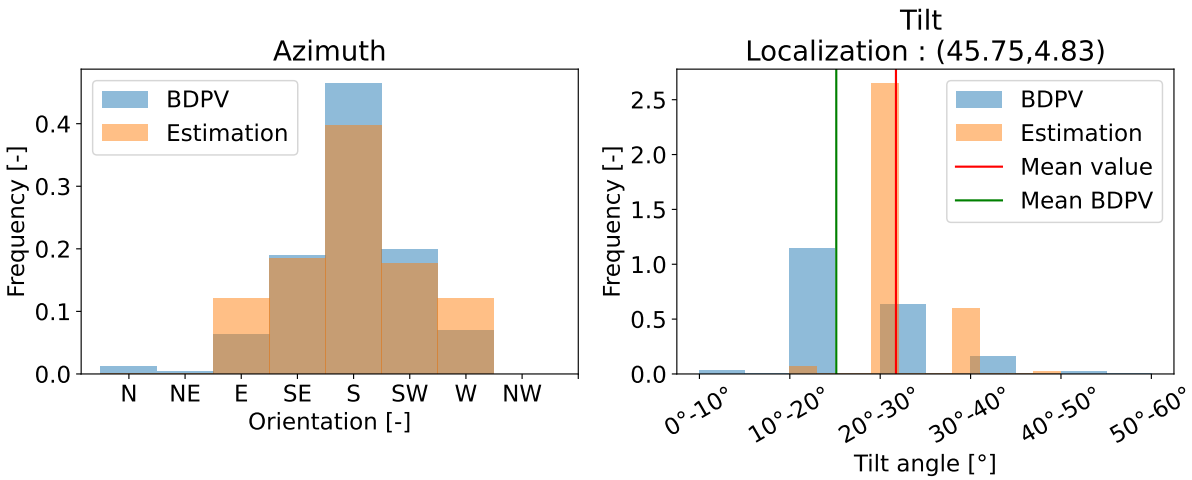


Figure C.3 – Comparison of the distribution of the azimuth (left) and tilt (right) angles obtained from our mapping algorithm and BDPV.

1.2 Complementary regressions of the city's coordinates on its error

In this section, we report different parameterizations that model relationships other than linear between the city's geographical coordinates and its error measured with the DTA. The transformation of the independent variable does not change our interpretation of the results. We can only notice that the latitude is significant at the 5% level for the APE if we consider the square of the latitude.

Logarithmic transformation Table C.1 plots the estimation results with a logarithmic transformation of the coordinates.

Table C.1 – Results of estimating the linear model defined in Equation 2.1 for the dependent variables APE, AIPE, and ratio.

	APE	**ratio**	**AIPE**
β_1 (Log(Latitude))	361.9940	-1.2030	83.0664
	(365.550)	*(3.837)*	*(276.282)*
β_2 (Log(Longitude))	-0.7179	0.0351*	2.6796
	(4.943)	*0.019)*	*(2.719)*
α	-9.337e-10	3.473e-12	-2.383e-10
	(9.1e-10)	*(9.51e-12)*	*(6.92e-10)*
N	1620	1839	1620
Adjusted-R^2	0.414	0.572	0.412

Clustered standard errors in parenthesis.
$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Quadratic transformation Table C.2 plots the estimation results with a quadratic transformation of the coordinates.

Table C.2 – Results of estimating the linear model defined in Equation 2.1 for the dependent variables APE, AIPE, and ratio.

	APE	**ratio**	**AIPE**
β_1 (Latitude^2)	0.0841	-2.769e-06	0.0369
	(0.063)	*(0.001)*	*(0.062)*
β_2 (Longitude^2)	-0.9757**	0.0085*	0.7663
	(4.943)	*(0.411)*	*(0.477)*
α	-8.36e-11	3.25e-13	-7.603e-11
	(6.13e-11)	*(6.93e-13)*	*(6.22e-11)*
N	1620	1839	1620
Adjusted-R^2	0.415	0.585	0.413

Clustered standard errors in parenthesis.
$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

2 Supplementary material to chapter 3

2.1 Additional visualization of the dissappearance of important components

Figure C.4 presents another example of the effect of the change of image provider on the ability of the model to detect PV panels. We can see that in this case, on

Google, the important factor was the factor **(b)** (on the leftmost image), which is no longer important on the IGN image (on the right). On the IGN image, due to the absence of information in **(b)**, the model relied on **(a)** to predict that there is no PV panel. the model instead relied on the factor **(a)** . Finally, we can see two regions, highlighted by **(c)** on the Google image, that also contribute to the prediction, but only in the case of the Google image. We may wonder whether one factor is not the critical component. Without this factor, the model no longer recognizes the panel and instead switches to a random factor to make a prediction. The fact that the shading evolved on this image does not seem to be the key driver for the model not to recognize the image, as the example on Figure 3.16 does not depict shadings.

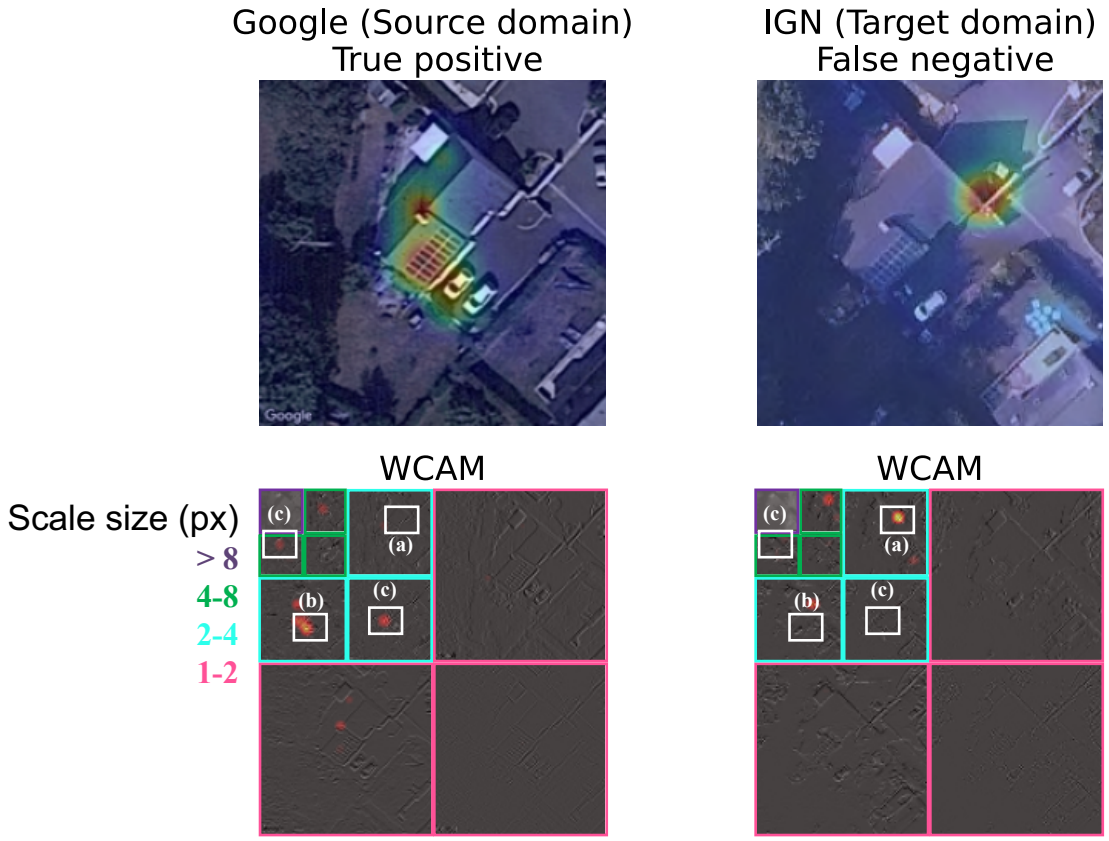


Figure C.4 – Predictions on Google image (left, upper row) and IGN image (right, upper row) and associated WCAMs (bottom row, displayed in logscale and with the same color scale). The brighter the highlighted region for the prediction, the more important it is. Taken from: Kasmi et al. (2023b).

2.2 Additional examples of the extraction of the critical component

Figure C.5 and Figure C.6 depict additional examples of identification of the critical components using the WCAM.

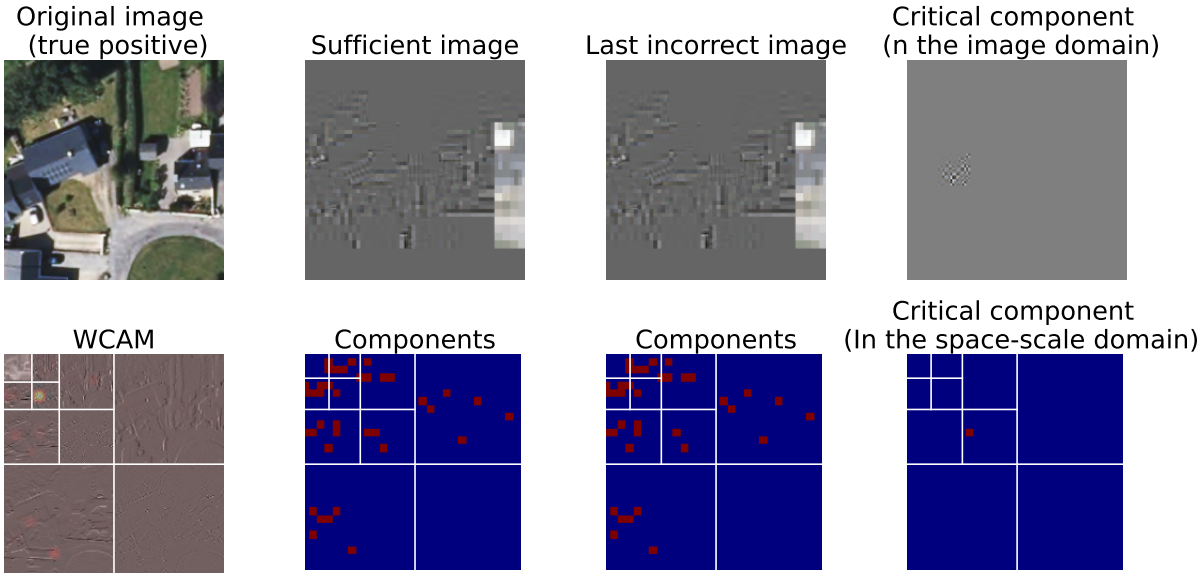


Figure C.5 – Identification of the critical component (highlighted in white on the "Critical component" plot on the bottom right of the image. Without this component, the model does not predict the PV panel. The sufficient image is the image reconstructed with the minimal set of components.

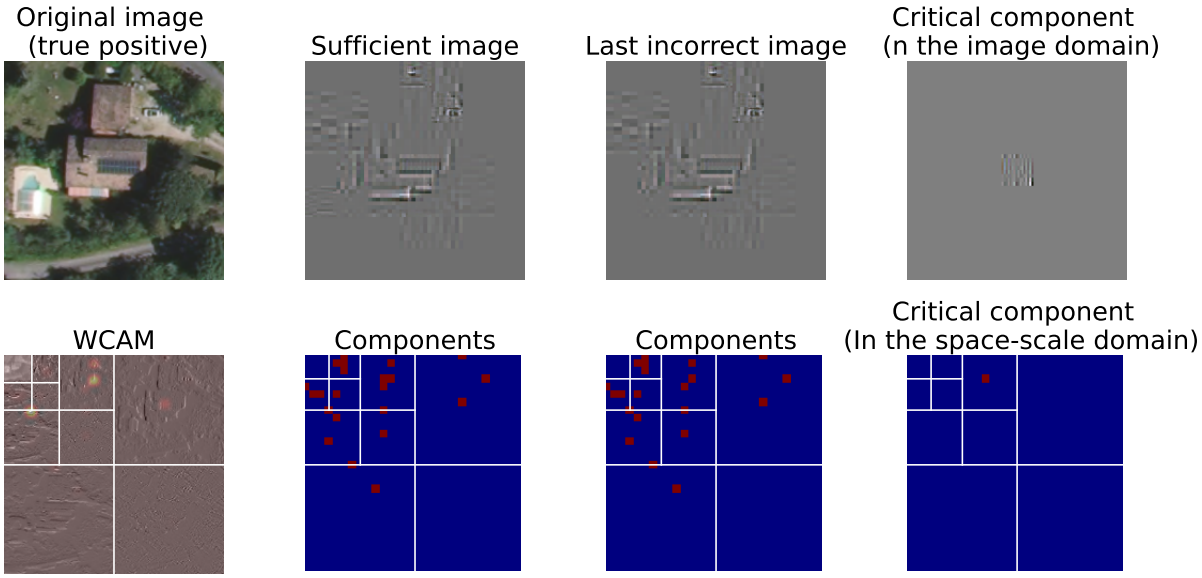


Figure C.6 – Identification of the critical component (highlighted in white on the "Critical component" plot on the bottom right of the image. Without this component, the model does not predict the PV panel. The sufficient image is the image reconstructed with the minimal set of components.

2.3 Details on the data augmentation techniques

AugMix (Hendrycks et al., 2020) The data augmentation strategy "Augment-and-Mix" (AugMix) consists in producing a high diversity of augmented images from an input sample. A set of operations (perturbations) to be applied to the images are sampled, along with sampling weights. The image resulting x_{aug} is obtained through the composition $x_{aug} = \omega_1 op_1 \circ \dots \omega_n op_n(x)$ where x is the original image. Then, the augmented image is interpolated with the original image with a weight m that is also randomly sampled. We have $x_{augmix} = mx + (1-m)x_{aug}$.

AutoAugment (Cubuk et al., 2019) This strategy aims at finding the best data augmentation for a given dataset. The authors determined the best augmentations strategy S as the outcome of a reinforcement learning problem: a controller predicts an augmentation policy from a search space. Then, the authors train a model, and the controller updates its sampling strategy S based on the train loss. The goal is for the controller to generate better policies over time. The authors derive optimal augmentation strategies for various datasets, including ImageNet Russakovsky et al. (2015), and show that the optimal policy for ImageNet generalizes well to other datasets.

RandAugment (Cubuk et al., 2020) This strategy's main goal is to remove the need for a computationally expansive policy search before model training. Instead of searching for transformations, random probabilities are assigned to the transformations. Then, each resulting policy (a weighted sequence of K transformations) is graded depending on its strength. The number of transformations and the strength are passed as input when calling the transformation.

2.4 Training results

Table C.3 reports the training results of our methods on the source (Google) test set. We can see that a small margin outperforms our spectral method compared to other methods on the source dataset. We can see that all models achieve nearly perfect accuracy on the source dataset.

Table C.3 – **F1 Score** and decomposition in true positives, true negatives, false positives, and false negatives for models trained on Google images with different strategies to mitigate the sensitivity to acquisition conditions. Evaluation computed on the Google (source) dataset.

	F1 Score (↑)	TP	TN	FP	FN
ERM (Vapnik, 1999)	0.98	1891	2355	36	39
AutoAugment (Cubuk et al., 2019)	0.98	1906	2340	51	24
AugMix (Hendrycks et al., 2020)	0.98	1894	2354	37	36
RandAugment (Cubuk et al., 2020)	0.98	1907	2342	49	23
Noise and blur	0.98	1897	2339	52	33
Blurring	0.82	1636	1958	433	294
Blurring + WP	0.90	1798	2135	256	132
Oracle	0.91	1815	2127	264	115

Interestingly, when evaluating the Oracle (the model trained on IGN) on Google, it remains competitive, even outperforming the accuracy of our methods on the source dataset. This further underlines that removing high-frequency content is very important for guaranteeing a good ability to generalize to new acquisition conditions.

3 Supplementary material to chapter 4

3.1 Methods evaluated for constructing `PyPVRoof`

Direct computation Overhead imagery is usually orthorectified (i.e., with a uniform scale). One can only compute the *projected* surface from the polygon. The computation is straightforward, and the only requirement is considering the projection. Parhar et al. (2021) describe the Mercator case. In practice, packages such as `area` (Alireza, 2018) estimate the surface of `geojson` polygons, taking into account the deformation induced by the projection system.

Once the projected surface is known, one needs the tilt angle to compute the real surface. Denoting S_{proj} the projected surface and θ the tilt angle of the installation (in degrees), the real surface is given by Equation C.1:

$$S = S_{proj} / \cos\left(\theta \times \frac{\pi}{180}\right). \tag{C.1}$$

3.1.1 Constant parameters

Constant tilt Tilt is necessary to compute the real surface of the installation. When neither registries nor surface models are available, it is still possible to infer

a tilt angle from the remaining data (i.e., the PV polygon). However, in practice, the optimal tilt angle of an installation is known. Typically, a tilt angle of around 30 degrees is optimal in most European countries. Regional models estimating the PV yield of solar plants consider this value by default (JRC, 2023; Saint-Drenan et al., 2019). In our case, we allow the user to input a default coefficient if necessary. This case can be seen as a worst-case situation if no surface models or auxiliary data is available.

Constant efficiency An efficiency factor relates the surface of a PV installation and its installed capacity. The PV panel efficiency increased due to the cell efficiency increase over the last couple of decades (NREL, 2023). This efficiency is usually measured in kW_p/m^2. The efficiency depends on many criteria (e.g., module technology of the panel, aging, manufacturer), which are not necessarily publicly available. However, average efficiencies can be used. For instance, Rausch et al. (2020); Mayer et al. (2022) used a value of 6 kW_p/m^2 as a reference value to estimate the installed capacity from the surface. As for the tilt angle, we allow the user to input this efficiency value.

3.1.2 Theil-Sen estimation

The Theil-Sen estimator (TSE) initially proposed by Theil (1992) and Sen (1968) is a robust regression method. It consists in considering the median of the slopes of all lines (or planes in higher dimensions) through pairs of points. This method is more robust to outliers than ordinary least squares.

We use this method to fit a plane $z(x, y) = ax + by + c$ parameterized by only three parameters, a, b and c, to a set of points corresponding to altitudes. These altitudes come from the digital surface model (DSM) passed as input. Figure C.7 depicts an example of LiDAR DSM provided by the IGN. Lighter areas correspond to higher altitudes.

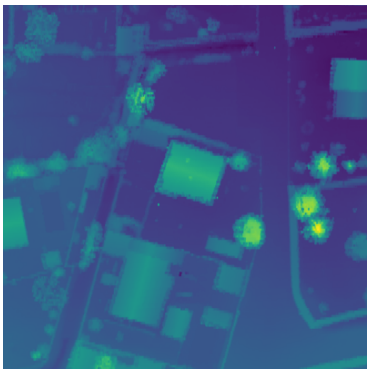


Figure C.7 – Example of DSM: the rasterization of the LiDAR from the IGN.

The direction of the gradient of the plane gives the azimuth angle φ. The slope

value along this gradient corresponds to the tilt angle θ. The gradient of the plane $\nabla z(x, y)$ is given in Equation C.2:

$$\nabla z(x,y) = \left(\frac{\partial z}{\partial x}(x,y), \frac{\partial z}{\partial y}(x,y)\right) = (a,b), \tag{C.2}$$

and

$$\varphi = \arctan\left(\frac{a}{b}\right), \ \theta = \arctan\left(\frac{h}{d}\right). \tag{C.3}$$

where $h = a^2+b^2$ and $d = \sqrt{a^2+b^2}$. Figure C.8 depicts the principle of the Theil-Sen method to compute the tilt and azimuth angles.

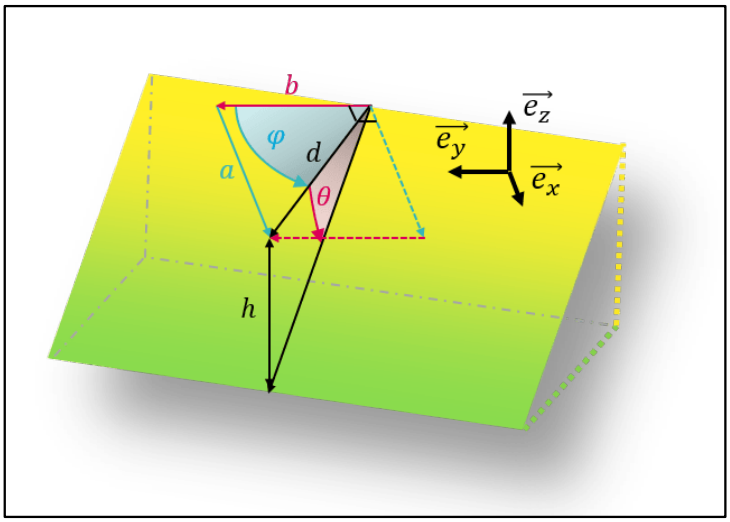


Figure C.8 – Theil-Sen method principle. The plane is deduced from the raster and is parameterized as $z(x, y) = ax + by + c$. φ corresponds to the azimuth angle and θ to the tilt angle. $\overrightarrow{e_x}$, $\overrightarrow{e_y}$, $\overrightarrow{e_z}$ correspond to the canonical basis of $\mathbb{R}^3$. Taken from Trémenbert et al. (2023).

3.1.3 Lookup table

If surface models are unavailable, we can still recover a tilt angle more accurately than with direct computation. To achieve this, we only need a sample of tilt angles for the desired area, e.g., a smaller PV or building database. We can then reflect the spatial variability of the tilt angle by computing an average tilt angle per grid point. The reference value associated with the installation corresponds to the average of the existing installations located in this grid point. The lookup table requires that the auxiliary data frame span the complete area or interest (e.g., a region or a country).

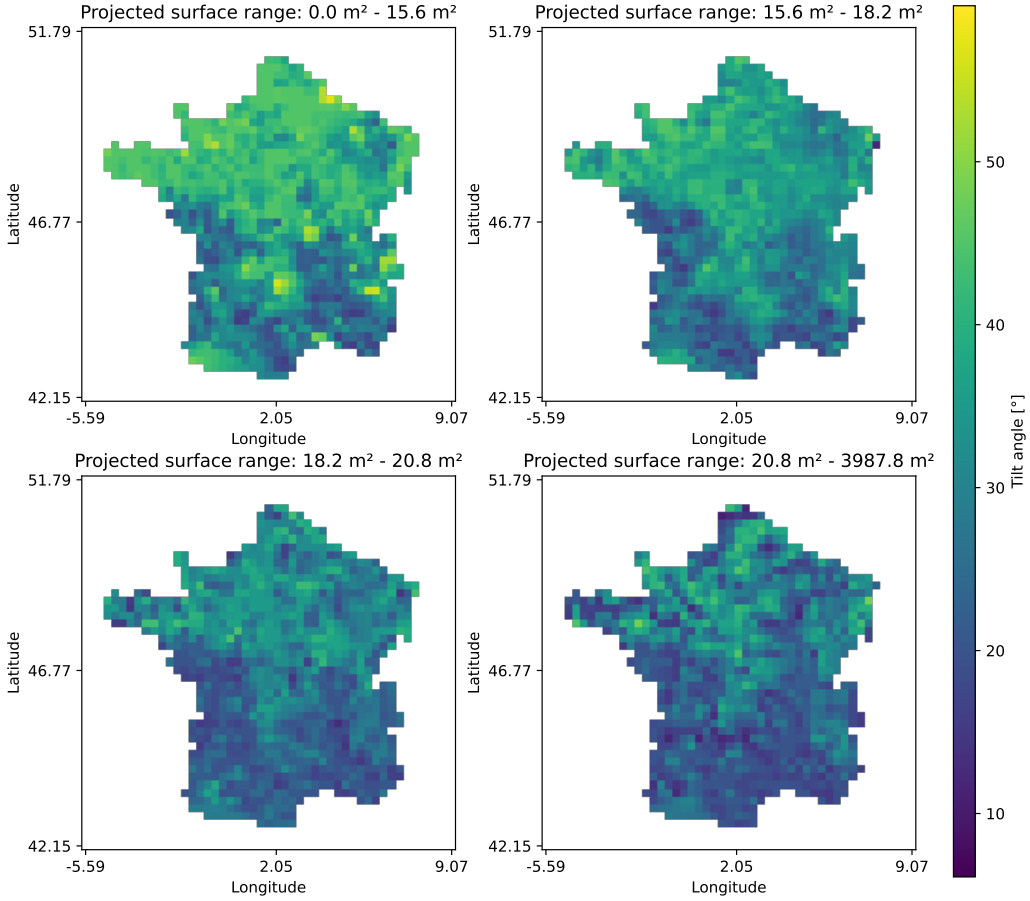


Figure C.9 – Lookup table for 50×50 grid-points and four surface categories computed for the PV mapping algorithm of Kasmi et al. (2022a). Surface categories correspond to quartiles of the distribution of the surface in the auxiliary data.

We compute this so-called lookup table (LUT) only once, and the user can pass a precomputed LUT as input. Computation is done as follows: we first define the spatial extent by setting easternmost E, northernmost N, westernmost W, and southernmost S boundaries. These boundaries are expressed in geographical coordinates. We then define a grid by dividing the numerical intervals defined by E and W and S and N respectively. We end up with K longitude intervals and L latitude intervals. Besides, we cluster the auxiliary data frame by (projected) surface category to define T surface categories. After empirical investigations, defining intervals as quantiles yielded the best results.

We then aggregate all sample points $x_1^{k,l,t}, \dots x_n^{k,l,t}$ whose coordinates belong to the $k \times l - th$ grid point and projected surface that belong the $t - th$ category. We then compute the reference tilt angle for this $(k,l,t) - th$ box, denoting $\theta^{k,l,t}$ by averaging

the tilt values of the n sample points falling into this bin. If no sample is available, we do not input a value.

Once this step is finished, we end up with a subset of grid points for which no reference value is available. We estimate a value by interpolating a $\theta^{k,l,t}$ by interpolating the neighboring values. We do not interpolate across surface categories. Figure C.9 displays the LUT obtained for the PV mapping algorithm of Kasmi et al. (2022a) using this method.

3.1.4 Bounding-box

The bounding box method only requires the polygon to compute the azimuth angle φ. The bounding-box method is an alternative when no surface models are accessible. We simplify the polygon's geometry by computing its bounding box. Then, we compute the azimuth angles associated with the "long" and "short" sides of the rectangle. We input as azimuth angle the angle corresponding to the longest side. We implicitly assume that the PV panel tends to be wider than high. The main limitation of this method is that it cannot distinguish between a panel facing eastwards or westwards, northwards or southwards. In the latter case, however, we can assume that the PV panel should not point northwards (at least in the Northern Hemisphere). If our bounding-box heuristic estimates that the polygon points between -45 and 45 degrees (0 being the reference for the North), we correct the estimation by applying a horizontal symmetry.

3.1.5 Linear regressions

So et al. (2017) showed that it is possible to accurately estimate the installed capacity by fitting a linear regression between the surface and the installed capacity. We build on this method. The linear model is given by Equation C.4,

$$c = \gamma_0 + \gamma S, \tag{C.4}$$

where S is the surface in m² and c is the capacity in kW_p of the installation. As pointed out by So et al. (2017), γ_0 is a bias coefficient; in the true model, γ_0 should equal zero. In our case, we consider $\gamma_0 = 0$ and estimate γ from BDAPPV.

Efficiencies can differ depending on the PV installation's surface NREL (2023). To accommodate this, we introduce another estimation for the installed capacity, namely the clustered linear regression. Clusters are defined depending on the surface of the installation. The goal is to reflect the different efficiencies while keeping the number of parameters as low as possible. This approach is inspired by the second model of So et al. (2017), which estimated a panel-wise coefficient γ. Their approach, however, required additional unobservable information, such as the manufacturer's design.

Figure C.10 represents the linear regression of the installed capacity on the surface and shows the relatively low dispersion of points around this mean. The leftmost plot shows the different coefficients depending on the surface cluster. We focus on surfaces lower than 200 m², where the density of installations is the highest. We can see that the efficiencies recorded in our reference registry are higher for smaller installations.

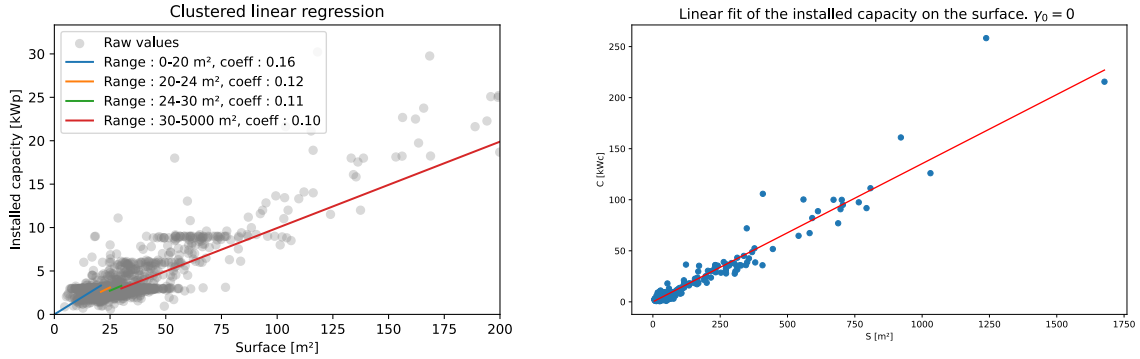


Figure C.10 – Left: clustered linear regression. Right: linear regression with a single coefficient.

3.2 Example of shifted thumbnails generated from a larger image

Figure C.11 presents an example of thumbnails extracted from the 400×400 image.

Figure C.11 – Example of thumbnails generated from a 400×400 image with the panel's position on the image being shifted. The coordinates indicate the position of the center of the thumbnail (in pixels relative to the upper-left corner of the image).

4 Supplementary material to chapter 5

Geographical variability of the estimation error by the Oracle Figure C.12 depicts the geographical variability of the error of the estimation of the PV power production using the conversion model, with parameters for the PV installations coming from BDPV. We can see that the geographical variability is about the same

as that of DeepPVMapper. Overall, the error is lower.

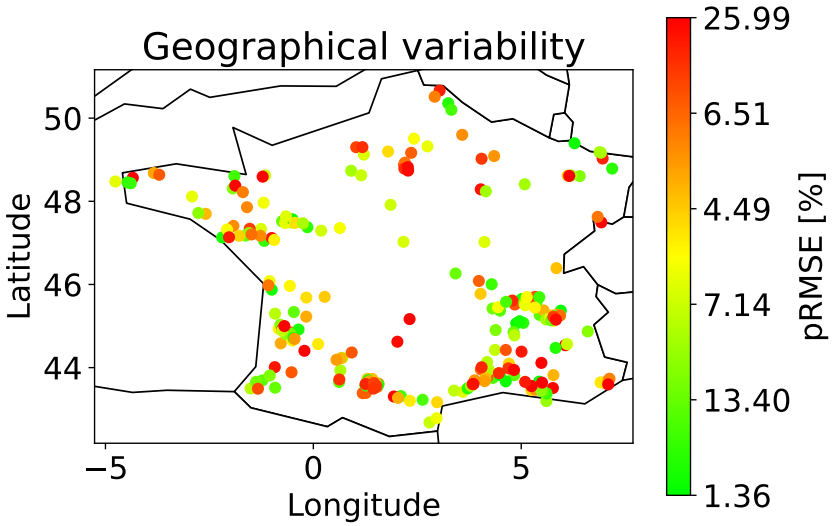


Figure C.12 – Geographical variability of the pRMSE [%] of the PV power estimation depending on the localization of the installation (Oracle).

4.1 Examples of reports generated to inspect the quality of the PV power measurements from BDPV

In addition to the quantitative inspections described in chapter 5, our quality check also relied on the qualitative examination of production reports such as those shown on Figure C.13 and Figure C.14. These reports consist in fitting a simulation model (different from our simple simulation model as the latter accommodates for all parameters of the PV installation) and comparing the production with the estimation. Figure C.13 and Figure C.14 present two examples of images that passed and failed our quality checks (QC), respectively. We can see that in the case of installation # 2248, which failed the quality check, the estimation between the production and the estimation does not fit at all, indicating that either the localization of the installation is wrong (so the weather data is inaccurate) and that the installation does not work correctly as it produces much less than expected.

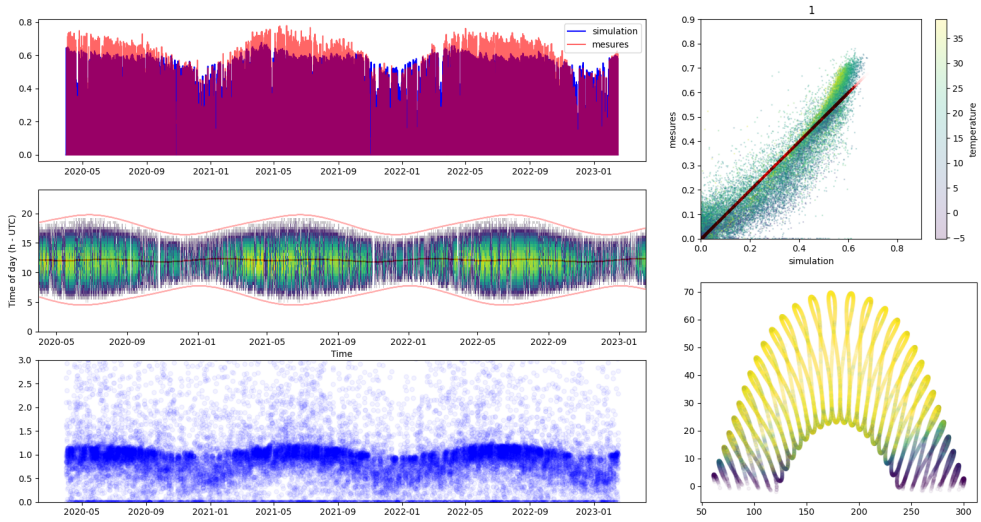


Figure C.13 – Example of an installation that passed the QC.

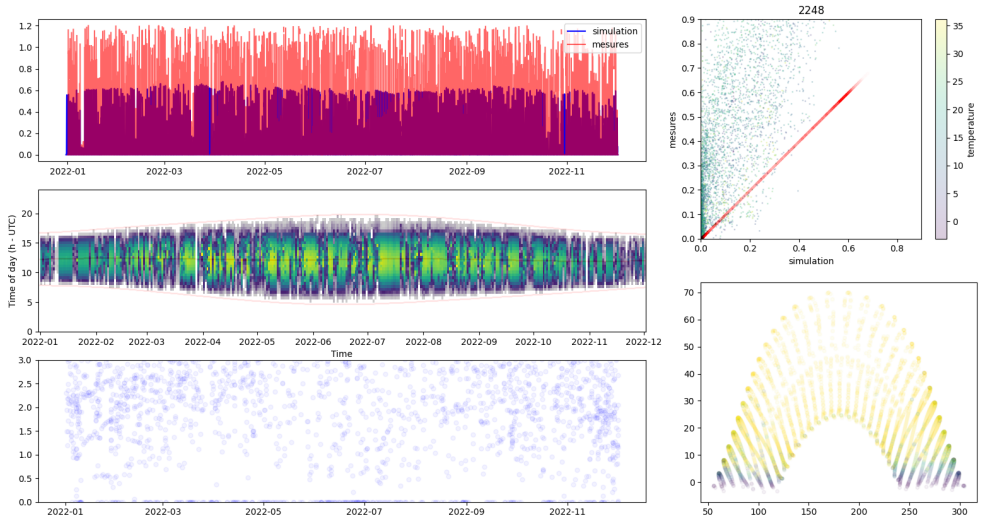


Figure C.14 – Example of an installation that failed the QC.

4.2 Using the LiDAR and the ground truth power measurements to evaluate the accuracy of the tilt angles reported in BDPV

Proposed approach and illustrative example Assuming that the LiDAR data provides an accurate estimation of the inclination of the rooftops, we estimate the tilt and azimuth angles using a Theil-Sen regression, available in Trémenbert et al. (2023). We then estimate the PV power production for one installation with

our model, using three different parameterizations: the tilt and azimuth angles of BDPV, the tilt and azimuth estimated from the LiDAR data, and the tilt angle estimated from the LiDAR data, and the azimuth angle from BDPV. We compare the different variants with the ground truth data. In all cases, the installed capacity is the installed capacity recorded in BDPV.

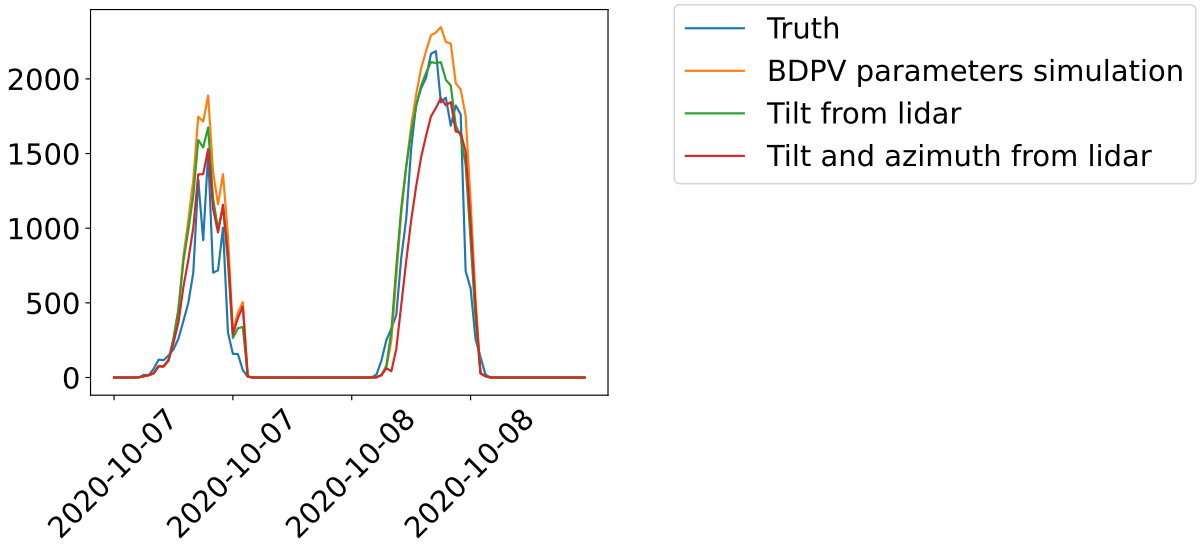


Figure C.15 – Fit of the PV power estimation with different model parameterization.

Figure C.15 illustrates the load curves obtained with our different model parameterization, compared to the ground truth. We can see that the fit seems to be better when we use the tilt and azimuth angles estimated with the LiDAR data rather than the values of BDPV. Table C.4 presents the results of the PV power estimation for one installation using the three different parameterizations of the conversion model. The column "case" indicates the name of the configuration, the columns θ and ϕ indicate the values of the tilt and azimuth angles in each configuration, and the column "pRMSE" indicates the corresponding pRMSE. We can see that the better fit observed on Figure C.15 for two days of data is quantitatively backed: for this installation, using the tilt angle derived from BDPV leads to a better estimation of the PV power production of the installation.

Table C.4 – Accuracy of the PV power production estimation using parameterizations from the LiDAR data. Best results are **bolded**.

Case	θ [°]	ϕ [°]	pRMSE [%]
θ and ϕ from BDPV	40.00	205.00	4.18
θ and ϕ from LiDAR	21.37	251.57	4.21
θ from LiDAR, ϕ from BDPV	21.37	205.00	**3.38**

Towards an independent quality check method We investigated whether we could assess the data quality using this approach. We scaled the approach to 16 installations and computed the pRMSE of the PV power production estimation. Table C.5 presents the results. On average, the pRMSE of the modeling error using the parameters from BDPV is the best, but it is not unbeatable. This shows that there are cases in which the BDPV parameterization is inaccurate.

Table C.5 – Average pRMSE for different configurations of parameterizations of the conversion model.

	pRMSE [%]				
Case	Min	Max	Mean	Median	n
θ and ϕ from BDPV	1.24	**5.44**	**3.01**	**3.00**	16
θ and ϕ from LiDAR	**0.95**	27.87	5.19	3.22	16
θ from LiDAR, ϕ from BDPV	0.99	34.05	5.52	2.68	16

We can then inspect these cases to see which of the LiDAR parameterizations beat the BDPV parameterization. Figure C.16 presents an example of an installation for which the estimation yielded by the LiDAR parameterization achieves better results than the estimation with the parameters from BDPV.

We can see that the tilt in BDPV seems to be overestimated while the azimuth angle of BDPV is visually correct. The overall profile of the model using the LiDAR parameterization is similar to the BDPV parameterization. The sensitivity of the predictions to the temperature also seems to be similar. However, the correlation coefficient (R^2) between the simulation and the prediction is higher using LiDAR parameters than BDPV parameters.

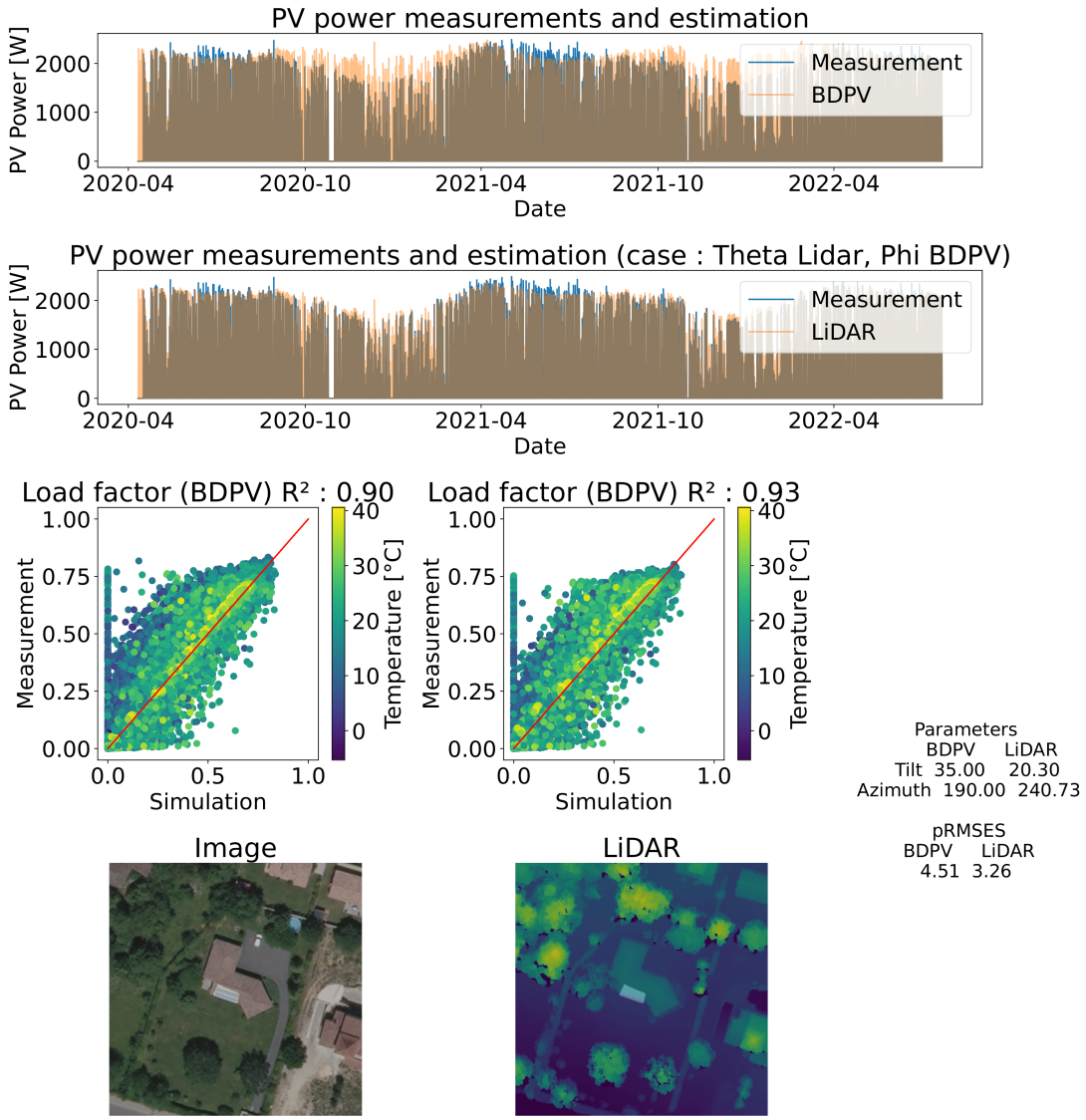


Figure C.16 – Report comparing the estimation using BDPV parameters and LiDAR parameters. The pRMSEs are reported in %, and the tilt and azimuth angles are reported in degrees.

Appendix D

Introduction to machine learning

Remark *This appendix is based on the lecture notes of the lecture on machine learning given by Arnak Dalalyan during the final year at ENSAE during the academic year 2017-2018 and the PhD thesis of V. Nagarajan.*

1 Notations and definitions

Overview Statistical learning or machine learning aims to design automatic procedures to uncover general rules based on examples. The starting point is a sample (x_i, y_i) used to infer the general rules. Depending on the assumptions on this sample, several learning paradigms can be distinguished :

- **Offline (batch)** or **online** learning, depending on whether the samples (x_i, y_i) are available all at once or sequentially to the learner,
- **Supervised** learning when one has access to samples x_i (features) and their associated labels y_i,
- **Unsupervised** learning, when one has only access to the x_i's,
- **Weakly** or **semi** supervised learning if only a part of the x_i's have labels,
- **Self-supervised** learning (SSL) if labels are generated from the features. As in unsupervised learning, in SSL, we only assume access to the x_i's (features) and generate labels or pseudo-labels from the features.

In the following, we focus on the most widespread machine learning framework, (batch) supervised learning.

Dataset We assume that we have access to a dataset $\mathcal{D}_n = \{(x_1, y_2), \dots, (x_n, y_n)\}$ comprised of input features $x_i \in \mathcal{X}$ and their associated labels $y_i \in \mathcal{Y}$, where $\mathcal{X}$ and $\mathcal{Y}$ are measurable sets. For instance, features x_i can be images. In this case, $\mathcal{X} = \mathbb{R}^{W \times L \times C}$ where $W \times L$ is the resolution of the image and C is the number of

color channels. On the other hand, labels are integers -1 or 1 encoding whether there is a solar array on the image. Then, $\mathcal{Y} = \{-1, 1\}$. We assume further assume that $\mathcal{D}_n \overset{i.i.d.}{\sim} P_{XY}$ where P_{XY} (or P in short) is unknown.

Predictor and hypothesis class Given the dataset $\mathcal{D}_n$, our goal is to find a predictor f belonging to a hypothesis class $\mathcal{F}$. Ideally, we want this predictor to infer the underlying classification or regression rule from the dataset. A predictor is, therefore, a function $f : \mathcal{X} \to \mathcal{Y}$. The hypothesis class corresponds to the set of functions from which we can pick our predictor. For instance, if we do linear classification, our hypothesis class would be the set of all linear predictors. The cardinality of $\mathcal{F}$, denoted $|\mathcal{F}|$, corresponds to the complexity of the hypothesis class.

Loss function A loss function $\mathcal{L}$ penalizes a predictor for making a wrong prediction.

Risk of a predictor The quality of a predictor f is assessed through its risk. The risk measures the average loss on a given sample $(x, y) \sim P_{XY}$ when using the predictor f. Formally,

$$R_{P_{X,Y}}(f) \coloneqq \mathbb{E}_{P_{XY}}\left[\mathcal{L}(Y, f(X))\right] \tag{D.1}$$

and the best prediction function is the predictor such that:

$$f^* \in \underset{f \in \mathcal{F}}{\arg\min}\, R_{P_{X,Y}}(f) \tag{D.2}$$

This function is also called the *Bayes predictor* or *oracle*.

Consistency of an algorithm In practice, we compute a surrogate for f based the data $\mathcal{D}_n$. Since $\mathcal{D}_n$ is a random variable, so is the surrogate, denoted $\hat{f}_n$. We say that an algorithm is :

— **Consistant** with respect to $P_{X,Y}$ if and only if (iif) :

$$\lim_{n\to\infty} \mathbb{E}_{P_{X,Y}}\left[R_{P_{X,Y}}(\hat{f}_n)\right] = \mathbb{E}_{P_{X,Y}}\left[R_{P_{X,Y}}(f^*_{P_{X,Y}})\right]$$

— **Consistent with respect to** $\mathcal{P}$ if it is consistent for all $P_{X,Y} \in \mathcal{P}$

— **Univsersally consistant** iif it is consistant for all probability $P_{X,Y}$ defined on $\mathcal{X} \times \mathcal{Y}$.

Based on consistency, we can consider a learning algorithm as a *good* algorithm if the latter is universally consistent *and* uniformly convergent over a given distribution family $\mathcal{P}$. Indeed, obtaining a uniform convergence property for *all* distribution families is impossible. In practice, the choice of $\mathcal{P}$ is crucial since finding a good $\mathcal{P}$,

i.e., a good way to model the problem, enables a fast convergence of the learning algorithm.

2 Empirical risk minimization

Overview We want to find a function f whose risk is as close as possible to the risk of the oracle. Unfortunately, neither the distribution of the data P_{XY} nor f^* are known. However, it is possible to approximate the risk by its empirical counterpart:

$$\hat{R}_n(f) \coloneqq \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}\left(Y_i, f(X_i)\right) \tag{D.3}$$

Furthermore, if we assume that $\mathbb{E}\left[\mathcal{L}(Y, f(X))^2\right] < \infty$, then, by the (strong) law of large numbers and the central limit theorem, we get that :

$$\hat{R}_n(f) \xrightarrow[n\to\infty]{a.s.} R_{P_{X,Y}}(f) \tag{D.4}$$

and

$$\sqrt{n}\left(\hat{R}_n(f) - R_{P_{X,Y}}(f)\right) \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}\left(0, Var\left[\mathcal{L}(Y, f(X))\right]\right) \tag{D.5}$$

Empirical risk minimization Based on these results, we can derive a simple learning algorithm based on the minimization of (D.3). This algorithm is the empirical risk minimization (ERM).

$$\hat{f}_n = \underset{f\in\tilde{\mathcal{F}}}{\arg\min}\, \hat{R}_n(f) \tag{D.6}$$

where $\tilde{\mathcal{F}} \subseteq \mathcal{F}$ is a subset of the hypothesis class. We consider a subset of the hypothesis in order to avoid overfitting. On the other hand, we need $\tilde{\mathcal{F}}$ to be large enough to find a good approximation of f^*.

Exess risk, stochastic and systematic errors Denoting $f^*_{\tilde{\mathcal{F}}}$ the oracle on $\tilde{\mathcal{F}}$, we can decompose the **excess risk** of $\hat{f}_n$ as :

$$R_{P_{X,Y}}(\hat{f}_n) - R_{P_{X,Y}}(f^*) = \underbrace{R_{P_{X,Y}}(\hat{f}_n) - R_{P_{X,Y}}(f^*_{\tilde{\mathcal{F}}})}_{\text{stochastic error}} + \underbrace{R_{P_{X,Y}}(f^*_{\tilde{\mathcal{F}}}) - R_{P_{X,Y}}(f^*)}_{\text{systematic error}} \tag{D.7}$$

The excess risk corresponds to the difference between the risk of a given function and the minimum risk possible for a hypothesis class. Rewriting it as the sum between the stochastic and the systematic error amounts to rephrasing the excess risk in terms of the bias-variance trade-off. The stochastic error can be reduced by finding a good subset $\tilde{\mathcal{F}}$, which introduces a bias in the modeling.

3 Excess risk bounds

Supremum bound The supremum bound tells us that the stochastic risk can be bounded as follows :

$$R_{P_{X,Y}}(\hat{f}_n) - R_{P_{X,Y}}(f^*_{\tilde{\mathcal{F}}}) \leq 2 \sup_{f \in \mathcal{F}} \left| R_{P_{X,Y}}(f) - \hat{R}_n(f) \right| \tag{D.8}$$

However, it is possible to obtain tighter bounds for the stochastic error. These bounds are probabilistic, meaning that they are statements such as :

$$\mathbb{P}\left[R_{P_{X,Y}}(\hat{f}_n) - R_{P_{X,Y}}(f^*_{\tilde{\mathcal{F}}}) \leq \delta_n(\varepsilon) \right] > 1 - \varepsilon, \ \forall \varepsilon \in (0,1) \tag{D.9}$$

Hoeffding's bound This bound states that

$$R_{P_{X,Y}}(\hat{f}_n) - R_{P_{X,Y}}(f^*_{\tilde{\mathcal{F}}}) \leq (b-a)\sqrt{\frac{2\log(2|\tilde{\mathcal{F}}|/\varepsilon)}{n}} \tag{D.10}$$

The main problem with this bound is that the bound becomes vacuous as soon as $|\tilde{\mathcal{F}}| = \infty$.

Vapnik-Chervonenkis (VC) dimension The key idea behind the VC bound is that rather than considering the cardinality of $\mathcal{F}$, one should instead consider the number of functions making a different prediction that it is possible to construction from $\mathcal{F}$. Indeed, two different functions making the same predictions on $\mathcal{D}_n$ will have the same ERM, and it is therefore not necessary to "count them twice." The VC bound relies on the VC dimension, corresponding to the largest shattering coefficient of $\mathcal{F}$. In the context of binary classification, for a given set $\mathcal{X}$ of dimension d, the shattering coefficient of $\mathcal{F}$ is the cardinality of the largest subset S that can be shattered by $\mathcal{F}$. Since the shattering coefficient depends on $\mathcal{D}_n$, we refer to the VC-dimension of f on $\mathcal{D}_\setminus$, denoted $V_{\mathcal{F}}(\mathcal{D}_n)$. We can check that $V_{\mathcal{F}}(\mathcal{D}_n) \leq n$.

VC-bound for binary classifiers We can derive an excess risk bound based on the VC dimension. This bound is valid in the context of binary classification since it is restricted to the set of loss functions $\mathcal{L} : \mathcal{Y} \times \mathcal{Y} \to \{-1, 1\}$. With probability at least $1 - \varepsilon$, we have :

$$R_{P_{X,Y}}(\hat{f}_n) - R_{P_{X,Y}}(f^*_{\tilde{\mathcal{F}}}) \leq 2\sqrt{\frac{2V_{\mathcal{F}}((2n)\log(4(2n+1)/\varepsilon)}{n}} \tag{D.11}$$

4 From the excess risk to generalization bounds

Overview So far, we have bounded the excess risk, which compares the *true* risk of a learning algorithm to the *true* oracle risk. In practice, however, we only have

access to the *empirical* risk of the learning algorithm, so we are also interested in knowing how far the empirical risk is far from the true risk. For simplicity, denoting f a given learning algorithm, $\hat{f}$ a learning rule obtained from the data, R the true risk and $\hat{R}$ the empirical risk, we can distinguish between :

— The **excess risk** defined as $R(\hat{f}) - \inf_{f \in \mathcal{F}} R(f)$, and for which we provide some bounds in section 3,

— The **generalization gap** (or generalization bound) which is defined as $R(\hat{f}) - \hat{R}(\hat{f})$

Consistency of ERM So far, we have bounded the stochastic error. However, we showed that the excess risk can be decomposed as the sum of the stochastic and systematic errors. If we want to achieve consistency, we need the systematic error to be as close as possible, and ideally, we would like the Bayes predictor f^* to be comprised in $\tilde{\mathcal{F}}$. A minimal condition for the oracle predictor to be in $\tilde{\mathcal{F}}$ is to allow $\tilde{\mathcal{F}}$ to grow as a function of n. Therefore, when $n \to \infty$, $\tilde{\mathcal{F}} \to \mathcal{F}$ and the ERM can achieve consistency. In the following, we consider that $\tilde{\mathcal{F}} = \mathcal{F}_n$.

Oracle bounds Let $\tilde{\mathcal{F}} = \mathcal{F}_n$. Denote $C(\mathcal{F})$ the complexity of the hypothesis class $\mathcal{F}$. Excess risk bounds obtained in section 3 can be rewritten as :

$$R_{P_{X,Y}}(\hat{f}_n) \leq C_1 R_{P_{X,Y}}(f^*_{\mathcal{F}_n}) + C_2 \sqrt{\frac{\log(C(\mathcal{F}_n)/\varepsilon)}{n}} \tag{D.12}$$

with probability at least $1 - \varepsilon$ and where C_1 and C_2 are universal constants.

Generalization gap The last inequality bounds the *true* risk of $\hat{f}_n$, but the only quantity we have access to is $\hat{R}_n(\hat{f}_n)$. However, it is possible to deduce from (D.11) a bound on the true error based on the empirical error, which is the generalization gap we are interested in. Using the same notations, we have, with probability at least $1 - \varepsilon$,

$$R_{P_{X,Y}}(\hat{f}_n) \leq \hat{R}_n(\hat{f}_n) + K \sqrt{\frac{\log(C(\mathcal{F}_n))/\varepsilon}{n}} \tag{D.13}$$

Conclusion We can see that the generalization gap can be closed using the VC theory. We can compute the empirical risk using (D.13) to compute the true risk and Equation D.11 to evaluate how far is the true risk from the oracle risk.

Appendix E

Publications associated with this thesis

1 Peer-reviewed journal papers

Kasmi, G., Saint-Drenan, Y.-M., Trebosc, D., Jolivet, R., Leloux, J., Sarr, B. & Dubus, L.. "*A crowdsourced dataset of aerial images with annotated solar photovoltaic arrays and installation metadata*". Scientific Data, 10(1), 59.
DOI: https://doi.org/10.1038/s41597-023-01951-4

Abstract Photovoltaic (PV) energy generation plays a crucial role in the energy transition. Small-scale PV installations are deployed at an unprecedented pace, and their integration into the grid can be challenging since public authorities often lack quality data about them. Overhead imagery is increasingly used to improve the knowledge of residential PV installations with machine learning models capable of automatically mapping these installations. However, these models cannot be easily transferred from one region or data source to another due to differences in image acquisition. To address this issue known as domain shift and foster the development of PV array mapping pipelines, we propose a dataset containing aerial images, annotations, and segmentation masks. We provide installation metadata for more than 28,000 installations. We provide ground truth segmentation masks for 13,000 installations, including 7,000 with annotations for two different image providers. Finally, we provide installation metadata that matches the annotation for more than 8,000 installations. Dataset applications include end-to-end PV registry construction, robust PV installations mapping, and analysis of crowdsourced datasets.

Kasmi, G., Touron, A., Blanc, P. & Saint-Drenan, Y.-M., Fortin, M., Dubus, L.. "*Remote Sensing-Based Estimation of Rooftop Photovoltaic Power Production Using Physical Conversion Models and Weather Data*". Energies 17(17), 4353.
DOI: https://doi.org/10.3390/en17174353

Abstract The global photovoltaic (PV) installed capacity, vital for the electric sector decarbonation, has reached 1,552.3 GW_p in 2023. In France, the capacity stood in April 2024 at 19.9 GW_p. The growth of the PV installed capacity over a year was nearly 32% worldwide and 15.7% in France. However, integrating PV electricity into grids is hindered by poor knowledge of rooftop PV systems, constituting 20% of France's installed capacity, and the lack of measurements of the production stemming from these systems. This problem of lack of measurements of the rooftop PV power production is referred to as the lack of observability. Using ground truth measurements of individual PV systems, available at an unprecedented temporal and spatial scale, we show that estimating the PV power production of an individual rooftop system by combining solar irradiance and temperature data, the characteristics of the PV system inferred from remote sensing methods and an irradiation-to-electric power conversion model provides accurate estimations of the PV power production. Our study shows that we can improve rooftop PV observability, and thus its integration into the electric grid, using little information on these systems, a simple model of the PV system and weather data.

2 Conference and workshop papers (peer-reviewed)

Kasmi, G., Dubus, L., Blanc, P. & Saint-Drenan, Y.-M.. "*Assessment of the Reliablity of a Model's Decision by Generalizing Attribution to the Wavelet Domain*". In *XAI In Action: Past, Future, and Present Applications workshop at NeurIPS 2023*.
DOI: https://doi.org/10.48550/arXiv.2305.14979

Abstract Neural networks have shown remarkable performance in computer vision, but their deployment in numerous scientific and technical fields is challenging due to their black-box nature. Scientists and practitioners need to evaluate the reliability of a decision, i.e., to know simultaneously if a model relies on the relevant features and whether these features are robust to image corruptions. Existing attribution methods aim to provide human-understandable explanations by highlighting important regions in the image domain, but fail to fully characterize a decision process's reliability. To bridge this gap, we introduce the Wavelet sCale Attribution Method (WCAM), a generalization of attribution from the pixel domain to the space-scale domain using wavelet transforms. Attribution in the wavelet domain reveals where *and* on what scales the model focuses, thus enabling us to assess whether a decision is reliable.

Kasmi, G., Dubus, L., Blanc, P. & Saint-Drenan, Y.-M.. "*Can We Reliably Improve the Robustness to Image Acquisition of Remote Sensing of PV Systems?*". In *Tackling Climate Change with Machine Learning Workshop at NeurIPS 2023*.
DOI: https://doi.org/10.48550/arXiv.2309.12214

Abstract Photovoltaic (PV) energy is crucial for the decarbonization of energy systems. Due to the lack of centralized data, remote sensing of rooftop PV installations is the best option to monitor the evolution of the rooftop PV installed fleet at a regional scale. However, current techniques lack reliability and are notably sensitive to shifts in the acquisition conditions. To overcome this, we leverage the wavelet scale attribution method (WCAM), which decomposes a model's prediction in the space-scale domain. The WCAM enables us to assess on which scales the representation of a PV model rests and provides insights to derive methods that improve the robustness to acquisition conditions, thus increasing trust in deep learning systems to encourage their use for the safe integration of clean energy in electric systems.

Kasmi, G., Dubus, L., Blanc, P. & Saint-Drenan, Y.-M.. "*Towards unsupervised assessment with open-source data of the accuracy of deep learning-based distributed PV mapping*". In *MACLEAN: MAChine Learning for EArth ObservatioN Workshop co-located with the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD 2022)*. URL : https://sites.google.com/view/maclean22.
DOI: https://doi.org/10.48550/arXiv.2207.07466

Abstract Photovoltaic (PV) energy is rapidly growing and key to mitigating the energy crisis. However, distributed PV generation, which amounts to half of the PV installed capacity, is typically unavailable to transmission system operators (TSOs), making it increasingly difficult to balance the load and supply and avoid grid congestions. To assess distributed PV generation, TSOs need precise knowledge regarding the metadata of distributed PV installations. Many remote sensing-based approaches have been proposed to map these installations in recent years. However, to use these methods in industrial processes, assessing their accuracy over the mapping area, i.e., the area covered by the model during deployment, is necessary. We define the downstream task accuracy (DTA) as the accuracy over the mapping area, automatically computed using publicly available data sources and the model's outputs and expressed in an interpretable way for operators. We benchmark existing models for distributed PV mapping and show how they perform in terms of DTA. We show that the accuracy computed on the test set overestimates by about 30 percentage points the accuracy on the mapping area. Our approach paves the way for safer integration of deep-learning-based pipelines for remote PV mapping.

3 Communications in conferences

Kasmi, G., Touron, A., Blanc, P. & Saint-Drenan, Y.-M., Fortin, M., Dubus, L.. "*Enhancing regional PV power estimation using physics-based models, solar irradiance data and deep learning*". International Conference in Energy and Meteorology

(ICEM), 2023, Padova, Italy.

URL : https://www.wemcouncil.org/wp/icem2023/

Abstract Photovoltaic (PV) power generation is growing rapidly and is key to mitigating the energy crisis. However, the safe integration of PV into the grid necessitates high-quality information about PV power generation in real-time and for forecasts. Transmission system operators (TSOs) usually have access to real-time measurements for the largest plants but lack information for smaller plants and rooftop PV generation. This lack of precise measurements is known as the lack of observability of PV power generation. Recently, various methods improved the estimation and forecast of PV power generation when measurements are lacking. Physics-based probabilistic methods have shown to be successful in accurately estimating regional PV power generation (Saint-Drenan et. al., 2019). These methods rely on meteorological data (irradiation and air temperature) and information on the spatial distribution of the PV installed capacity. However, as the number of rooftop PV systems is expected to strongly increase, two challenges arise. First, accurate estimation of rooftop PV generation with physics-based models requires accurate information on its characteristics, which are different from large plants. Second, reference production data stemming from these installations is not accessible, even to operators such as TSOs. The lack of ground-truth PV power production data results in the impossibility of assessing the accuracy of PV forecasting methods. This work addresses the following question: can regional PV models improve the observability of PV generation when considering both unobservable PV plants and the high penetration of distributed PV generation? We first introduce new regional models to estimate PV power generation and compare them to existing approaches. We then leverage deep learning to accurately map the rooftop PV installed capacity and show how regional models benefit from more detailed information on the PV rooftop capacity. Second, we compare these models to ground-truth data and show how regional PV models benefit from accurate data on rooftop PV installations.

Kasmi, G., Blanc, P., Saint-Drenan, Y.-M. & Dubus, L.. "*Looking for a frequency-based principle to predict the sensitivity of convolutional neural networks to Gaussian image perturbations*". PhD Forum at ECML/PKDD, 2022, Grenoble, France.

URL : https://ecmlpkdd.org/2022/

Kasmi, G., Blanc, P., Saint-Drenan, Y.-M. & Dubus, L.. "*Leveraging Earth observation data and deep learning to estimate the PV output in France*". MACLEAN Workshop @CAp/RFIAP, 2022, Vannes, France.

URL : https://caprfiap2022.sciencesconf.org/resource/page/id/23

4 Submitted works

Kasmi, G., Dubus, L., Blanc, P. & Saint-Drenan, Y.-M.. "*DeepPVMapper: reliable and scalable remote sensing of rooftop photovoltaic installations*".
.

Abstract As photovoltaic (PV) energy grows quickly, transmission system operators (TSOs) need accurate data to ensure its optimal integration into the grid. To this end, deep learning-based remote sensing of rooftop PV systems emerged as a promising approach. However, existing works struggle to scale at the size of countries or power systems and lack reliability, as their test accuracy is not representative of the accuracy during deployment. To bridge this gap, we introduce DeepPVMapper, a deep learning-based algorithm for the remote sensing of rooftop PV installations. We constructed a test bench representative of the operational conditions. We optimized DeepPVMapper to focus only on relevant areas and maximize the accuracy for detecting PV installations. We evaluate DeepPVMapper in a setting representative of the operational conditions and demonstrate that it is 21% faster, 16% more accurate, and 31% more energy efficient than the current state-of-the-art. We successfully deployed DeepPVMapper to map PV installations in France and hope it paves the way towards better integrating rooftop PV energy for more sustainable power systems.

5 Preprints

Trémenbert, Y, **Kasmi, G.**, Dubus, L., Blanc, P. & Saint-Drenan, Y.-M.. "*PyPVRoof: a Python package for extracting the characteristics of rooftop PV installations using remote sensing data* ". arXiv preprint arXiv:2309.07143.
DOI: https://doi.org/10.48550/arXiv.2309.07143

Abstract Photovoltaic (PV) energy grows at an unprecedented pace, which makes it difficult to maintain up-to-date and accurate PV registries, which are critical for many applications such as PV power generation estimation. This lack of qualitative data is especially true in the case of rooftop PV installations. As a result, extensive efforts are put into the constitution of PV inventories. However, although valuable, these registries cannot be directly used for monitoring the deployment of PV or estimating the PV power generation, as these tasks usually require PV systems *characteristics*. To seamlessly extract these characteristics from the global inventories, we introduce `PyPVRoof`. `PyPVRoof` is a Python package to extract essential PV installation characteristics. These characteristics are tilt angle, azimuth, surface, localization, and installed capacity. `PyPVRoof` is designed to cover all use

cases regarding data availability and user needs and is based on a benchmark of the best existing methods. Data for replicating our accuracy benchmarks are available on our Zenodo repository , and the package code is accessible at this URL: https://github.com/gabrielkasmi/pypvroof.

6 Communication in expert groups

Kasmi, G., Blanc, P. & Saint-Drenan, Y.-M., Dubus, L.. "*Assessment of the potential of Earth observation data and deep convolutional neural networks to improve the estimation and forecast of the solar power production in France*". IEA-PVPS Task 16 meeting (2022), Sophia Antipolis, France.

URL : Communication

7 Posters

Kasmi, G., Blanc, P., Saint-Drenan, Y.-M. & Dubus, L.. "*Assessment of the potential of Earth observation data and deep convolutional neural networks to improve the estimation and forecast of the solar power production in France*". 4th Symposium of the MADICS, 2022, Lyon, France.

URL : https://www.madics.fr/event/symposium-madics-4/

Kasmi, G., Blanc, P., Saint-Drenan, Y.-M. & Dubus, L.. "*Solar Array Detection on Aerial Photography Based on Convolutional Neural Networks: Image of the Solar Array Characteristics and Image Backgrounds on the Out-of-domain Generalization*". SophIA Summit, 2021, Sophia Antipolis, France.

URL : https://www.sophia-antipolis.fr/events/sophi-a-summit-2021/

RÉSUMÉ

En novembre 2023, la puissance photovoltaïque (PV) installée en France s'élevait à 18,6 GW_c, et le gestionnaire du réseau de transport d'électricité (GRT) français ne disposait pas de mesures de production pour 20% du parc, correspondant principalement à des systèmes de petite taille sur toitures. Dans le contexte de décarbonisation du mix électrique, la puissance installée PV continuera de croître rapidement, aussi le manque d'observabilité du PV risque-t-il compromettre l'intégration du PV dans le système électrique en raison des incertitudes qu'il engendre. Une meilleure connaissance du parc photovoltaïque en toiture, matérialisée par un registre technique national contenant la localisation et les caractéristiques des installations photovoltaïques, est nécessaire pour améliorer l'observabilité du PV. Cette thèse evalue si l'utilisation d'algorithmes d'apprentissage profond et d'orthoimages est une méthode adaptée à la construction d'un registre technique national d'installations photovoltaiques (PV) sur toiture destiné à améliorer l'observabilité de la production PV en France. La thèse discute d'abord des normes de qualité que le registre technique doit satisfaire et introduit une méthode d'évaluation non supervisée pour contrôler l'exactitude du registre en l'absence de données de référence. Deuxièmement, la thèse introduit une nouvelle méthode d'attribution qui permet d'analyser des décisions du modèle en décomposant ses prédictions dans l'espace des ondelettes. La thèse discute de la pertinence de cette décomposition pour évaluer ce que le modèle voit sur l'image d'entrée, comprendre la sensibilité du modèle à des conditions d'acquisition variables, qui affectent la précision et la fiabilité du modèle, et introduire un algorithme robuste et fiable pour cartographier les installations PV sur toiture. Enfin, la pertinence du registre pour améliorer l'observabilité des installations photovoltaïques sur les toits est établie en montrant que des estimations précises et réplicables à grande échelle de la production issue des installations PV sur toiture peuvent être construites à partir du registre et de données météorologiques. Cette thèse apporte des contributions en énergétique et procédés, en montrant comment améliorer l'observabilité du PV toiture et en apprentissage statistique, en améliorant l'interprétabilité des modèles d'apprentissage profond grâce à une nouvelle méthode d'attribution. Plus généralement, cette thèse souligne les conditions nécessaires à l'utilisation de modèles d'apprentissage profond dans des contextes industriels critiques.

MOTS CLÉS

apprentissage profond, interprétabilité, robustesse, fiabilité, énergie photovoltaique, observabilité

ABSTRACT

In November 2023, the French photovoltaic (PV) installed capacity stood at 18.6 GW_p, and the French electricity transmission system operator (TSO) lacked power measurements for 20% of the fleet, which mostly corresponded to small-scale (rooftop) systems. In the context of decarbonizing the electric mix, the PV installed capacity will continue to experience sustained growth in the coming years, and the so-called problem of poor PV observability threatens its long-term integration into the grid due to the uncertainty it creates. A better knowledge of the rooftop PV fleet, embodied in a nationwide technical registry recording the localization and characteristics of the PV installations, is necessary to improve PV observability. This thesis proposes to assess whether deep learning-based remote sensing on orthoimagery is a suitable method for constructing this technical registry. The thesis first discusses the quality standards the technical registry should satisfy and introduces an unsupervised evaluation method to monitor the accuracy of the registry in the absence of ground truth labels. Second, the thesis introduces a new feature attribution method that enables the auditing of the model's decisions by decomposing its predictions into the space-scale domain. The thesis discusses the relevance of this decomposition for assessing what the model sees on the input image, understanding the model's sensitivity to varying acquisition conditions, which are found to affect the model's accuracy and reliability, and introducing a robust and reliable algorithm for mapping rooftop PV installations. Finally, the relevance of the registry for improving rooftop PV observability is established by showing that accurate and scalable estimations of the rooftop PV power production can be derived from the registry and weather data. This thesis features contributions in power systems by showing how to effectively improve rooftop PV observability and in deep learning by improving the interpretability of deep learning models thanks to a new feature attribution method. More generally, this thesis underlines the necessary conditions for using deep learning in critical industrial contexts.

KEYWORDS

deep learning, interpretability, robustness, reliability, solar energy, observability