Statistical modelling of electric vehicle charging behaviours

Yvenn Amara-Ouali

université PARIS-SACLAY

Statistical modelling of electric vehicle charging behaviours

Modélisation statistique des comportements de charge des véhicules électriques

Thèse de doctorat de l'université Paris-Saclay

École doctorale n° 574, Mathématiques Hadamard (EDMH)
Spécialité de doctorat : Mathématiques aux interfaces
Graduate School : Mathématiques
Référent : Faculté des Sciences d'Orsay

Thèse préparée au sein du **Laboratoire de Mathématiques d'Orsay** (Université Paris-Saclay, CNRS),
sous la direction de **Pascal MASSART**, Professeur,
la co-direction de **Jean-Michel POGGI**, Professeur,
le co-encadrement de **Yannig GOUDE**, ingénieur-chercheur EDF R&D
et de **Hui YAN**, ingénieure-chercheuse EDF R&D

Thèse soutenue à Paris-Saclay, le 22 septembre 2022, par

Yvenn AMARA-OUALI

Composition du jury

Mathilde Mougeot Professeure, Ecole Nationale Supérieure d'Informatique pour l'Industrie et l'Entreprise (ENSIIE)	Présidente
Avner Bar-Hen Professeur, CNAM Paris	Rapporteur & Examinateur
Georges Kariniotakis Directeur de Recherche, MINES Paris	Rapporteur & Examinateur
Badih Ghattas Maître de Conférence, HDR, Université d'Aix-Marseille	Examinateur
Christine Keribin Maître de Conférence, HDR, Université Paris-Saclay	Examinatrice
Jean-Michel Poggi Professeur, Université Paris-Saclay	Co-directeur de thèse

Thèse de doctorat

NNT : 2022UPASM020

université PARIS-SACLAY | ÉCOLE DOCTORALE de mathématiques Hadamard (EDMH)

Titre : Modélisation statistique des comportements de charge des véhicules électriques
Mots clés : Processus de Poisson non-homogène, Modèles additifs, Analyse multi-résolution, Recharge intelligente, Pics de consommation

Résumé : Le développement des véhicules électriques (VEs) est un levier majeur vers un transport bas carbone. Il s'accompagne d'un nombre croissant d'infrastructures de recharge qui peuvent être utilisées comme actifs flexibles de gestion du réseau. Pour permettre cette recharge intelligente, une prévision journalière efficace des comportements de charge est nécessaire. Dans ce contexte, l'objectif de cette thèse est triple : (a) identifier les techniques de modélisation actuelles et les données ouvertes disponibles (b) proposer de nouvelles méthodologies de charge des VE pour caractériser leurs comportements de charge (c) spécifier des techniques innovantes pour la prévision des pics de consommation. Le premier chapitre du manuscrit présente les enjeux industriels et introduit le cadre de la modélisation de la charge des VE. Le chapitre 2 présente un examen approfondi des modèles de charge de VE à l'état de l'art ainsi qu'une exploration de 8 jeux de données ouverts de sessions de recharge trouvés dans cette recherche. Le chapitre 3 propose une étude comparative de 14 modèles de charge et d'occupation des VE sur les 8 jeux de données présentés au cours du chapitre précédent. Le chapitre 4 propose un modèle pour les arrivées des VE aux points de charge sous la forme d'un processus de Poisson non homogène avec des effets additifs projetés sur des bases de splines et d'ondelettes. Enfin, le chapitre 5 présente un modèle pour la prévision journalière des pics de demande électrique avec une approche multi-résolution. Nous montrons que les approches proposées dans nos travaux sont compétitives avec les meilleures alternatives existantes en évaluant leurs performances sur des données réelles.

Title : Statistical modelling of electric vehicle charging behaviours
Keywords : Non-homogeneous Poisson process, Additive models, Multi-resolution analysis, Smart charging, Consumption peaks

Abstract : The development of electric vehicles (EV) is a major lever towards low-carbon transport. It comes with a growing number of charging infrastructures that can be used as flexible assets for the grid. To enable this smart-charging, an effective daily forecast of the charging behaviour is necessary. In this context, the objective of this thesis is threefold : (a) to identify current modelling techniques and open data available (b) to propose new EV charging methodologies to characterise their charging behaviours (c) to specify innovative techniques for daily peak load forecasting. The first chapter of the manuscript presents the industrial issues and introduces the modelling framework for EV charging. Chapter 2 is a review of state of the art EV load models as well as an exploration of 8 open charging session datasets. Chapter 3 offers a comparative study of 14 EV load and occupancy models on the 8 datasets presented in the previous chapter. Chapter 4 introduces a model for EV arrivals as a non-homogeneous Poisson process with additive spline and wavelet effects. Finally, Chapter 5 introduces a model for daily electrical peaks with a multi-resolution approach. We show that the approaches proposed in our work are competitive with the best existing alternatives by evaluating their performance on real-world data.

Table of contents

Remerciements

J'ai l'impression qu'on oublie trop souvent qu'un doctorat est un travail, certes incarné par un unique individu, mais qui résulte avant tout d'une collaboration d'une multitude et c'est important pour moi de prendre le temps ici de remercier les personnes qui ont contribué de près ou de loin à ma thèse.

Je veux commencer par remercier Hui Yan. C'est elle qui m'a proposé un poste de stagiaire au sein de la R&D pour comprendre les enjeux de la prévision de consommation électrique. Elle m'a accompagné jusqu'à aujourd'hui et je m'apprête à emboîter ses pas au moment où je finis cette thèse en prenant un poste à la R&D. C'est au cours des entretiens de stage que j'ai rencontré mon autre encadrant : Yannig Goude. A lui, je veux rendre un chaleureux hommage pour l'impact qu'il a eu dans ma vie professionnelle. C'est quelqu'un qui m'a transmis la passion de son travail et qui brille par sa rigueur et ses innovations. Bien qu'il soit très attaché à la réussite de nos projets, il fait toujours passer l'humain avant tout et s'est toujours enquis de mes sentiments à chaque étape de nos travaux. Merci à toi.

Yannig à son tour m'a permis de rencontrer Pascal et Jean-Michel qui m'ont accueilli à Orsay dans un cadre merveilleux pour faire des mathématiques. Je remercie donc vivement mes deux directeurs de thèse, tous les deux aussi inspirants l'un que l'autre. Pascal, toujours le mot pour me redonner le sourire même quand je lui présentais mes meilleures inepties. Jean-Michel, un homme sérieux et sage qui par son franc-parler emprunt de bienveillance m'a beaucoup fait avancer. Tous deux m'ont émerveillé par l'étendue de leur connaissance mathématique et par leurs idées toutes plus innovantes les unes que les autres.

Je veux aussi remercier les collègues et amis que j'ai rencontrés pendant ce beau voyage. David, pour tes propos toujours bien tranchés sur le sport, Lionel pour nos discussions transcendantales sur le foot, Joseph pour avoir été un camarade très inspirant, Gilles mon ange gardien et la première personne qui m'ait jamais parlé des GAM, Bachir et Matteo avec qui je me suis entendu à merveille pour rédiger mes premiers papiers de recherche. Je n'oublie pas non plus Margaux B. qui a été un exemple et qui a su me conseiller à certains moments importants pour moi, Margaux Z. avec qui on aimait se prendre la tête jusqu'à pas d'heure pour s'assurer qu'on avait envisagé toute les questions possibles de nos étudiants en TD. De façon plus générale je remercie l'ensemble de l'équipe proba/stat qui m'a accueilli au LMO et les équipes R39 et R36 du département OSIRIS avec qui j'ai collaboré pendant cette thèse.

Il est aussi important pour moi de remercier tous les étudiants que j'ai pu avoir pendant mes cours. Grâce à vos questions et votre engagement j'ai pu repousser la limite de mes connaissances sur de nombreux sujets en statistique et *machine learning*.

A 27 ans, j'ai l'impression d'avoir parcouru déjà un long chemin. Et pourtant, je me rappelle des leçons de mes parents comme s'ils me les donnaient encore hier. Mon père m'a appris l'exigence, le respect et l'intégrité. Des principes de vie que je m'emploie à respecter du mieux que je peux dans mon travail. Il m'a transmis sa passion pour l'enseignement et j'ai pu la vivre pleinement au cours des trois dernières années. Ma

mère a toujours cru en moi et m'a toujours valorisé. Elle a tout fait pour que je sois dans un environnement idéal pour m'exprimer au mieux. Pour tout ça et bien plus encore, je vous remercie.

Je voulais également écrire un petit mot pour mes grands-parents et notamment mon grand-père qui a travaillé pour un fameux concessionnaire français et qui aujourd'hui voit son petit fils travailler sur la mobilité de demain.

Et bien évidemment mes amis de toujours qui ont été là pour me changer les idées et pour me donner de la force Adrien, Andrea, Sofiane, David, Jordan, Côme, Romain, Alexis et Manon.

Je n'oublie pas non plus les chercheurs qui ont pris le temps de répondre à mes questions, notamment Patricia et Anestis. Et c'est avec beaucoup de gratitude que j'ai reçu les rapports d'Avner Bar-Hen et Georges Kariniotakis qui m'ont permis d'élever la qualité de ma thèse. J'en profite également pour remercier l'ensemble des membres du jury qui ont accepté de participer à ma soutenance.

Enfin, Emma je te remercie de tout mon coeur pour avoir été un pilier inébranlable sur lequel j'ai pu me reposer et pour tes nombreuses relectures de mes travaux. Je remercie également Lynn et Ronnie qui m'ont accueilli lors de la rédaction finale du manuscrit et qui ont été aux petits soins avec moi pour que je puisse être dans les meilleures conditions pour écrire.

Contributions

Publications

Chapter 2 is based on the following publications :

1. Amara-Ouali, Y., Goude, Y., Massart, P., Poggi, J. M., & Yan, H. (2021). A review of electric vehicle load open data and models. Energies, 14(8), 2233. https://doi.org/10.3390/en14082233
2. Amara-Ouali, Y., Massart, P., Poggi, J. M., Goude, Y., & Yan, H. (2021, June). A review of electric vehicle charging session open data : Poster. In Proceedings of the Twelfth ACM International Conference on Future Energy Systems (pp. 278-279). https://doi.org/10.1145/3447555.3466568

Chapter 3 is based on the following publication :

3. Amara-Ouali, Y., Hamrouche, B., Goude, Y., & Bishara, M. (2022, June). A Benchmark of Electric Vehicle Load and Occupancy Models for Day-Ahead Forecasting on Open Charging Session Data. To appear in Proceedings of the Thirteenth ACM International Conference on Future Energy Systems. https://doi.org/10.1145/3538637.3538850

Chapter 4 is based on a draft paper which proposes the following contributions :

* An algorithm for estimating the intensity function of a non-homogeneous Poisson process with an additive model of spline and wavelet effects
* An application led on EV charging session data to assess the performance of the proposed method in a real-world scenario

Chapter 5 is based on the following publication :

4. Amara-Ouali, Y., Fasiolo, M., Goude, Y., & Yan, H. (2022). Daily peak electrical load forecasting with a multi-resolution approach. International Journal of Forecasting. https://doi.org/10.1016/j.ijforecast.2022.06.001

Implementations and additional resources

During this thesis, many pieces of code were produced to conduct the required experiments. We have worked hard to make this work available to everyone on GitHub [1]. This repository contains 4 folders which correspond to the following 4 chapters of this manuscript : the first one is an open load session data exploration ; the second one is an EV load and occupancy forecasting tool using various approaches and expert aggregation ; the third one is an R package for fitting additive Poisson models using both wavelets and splines ; the fourth and last one is a case study showcasing an innovative approach to modelling peak load and timing.

Notations

The following abbreviations are used in this manuscript :

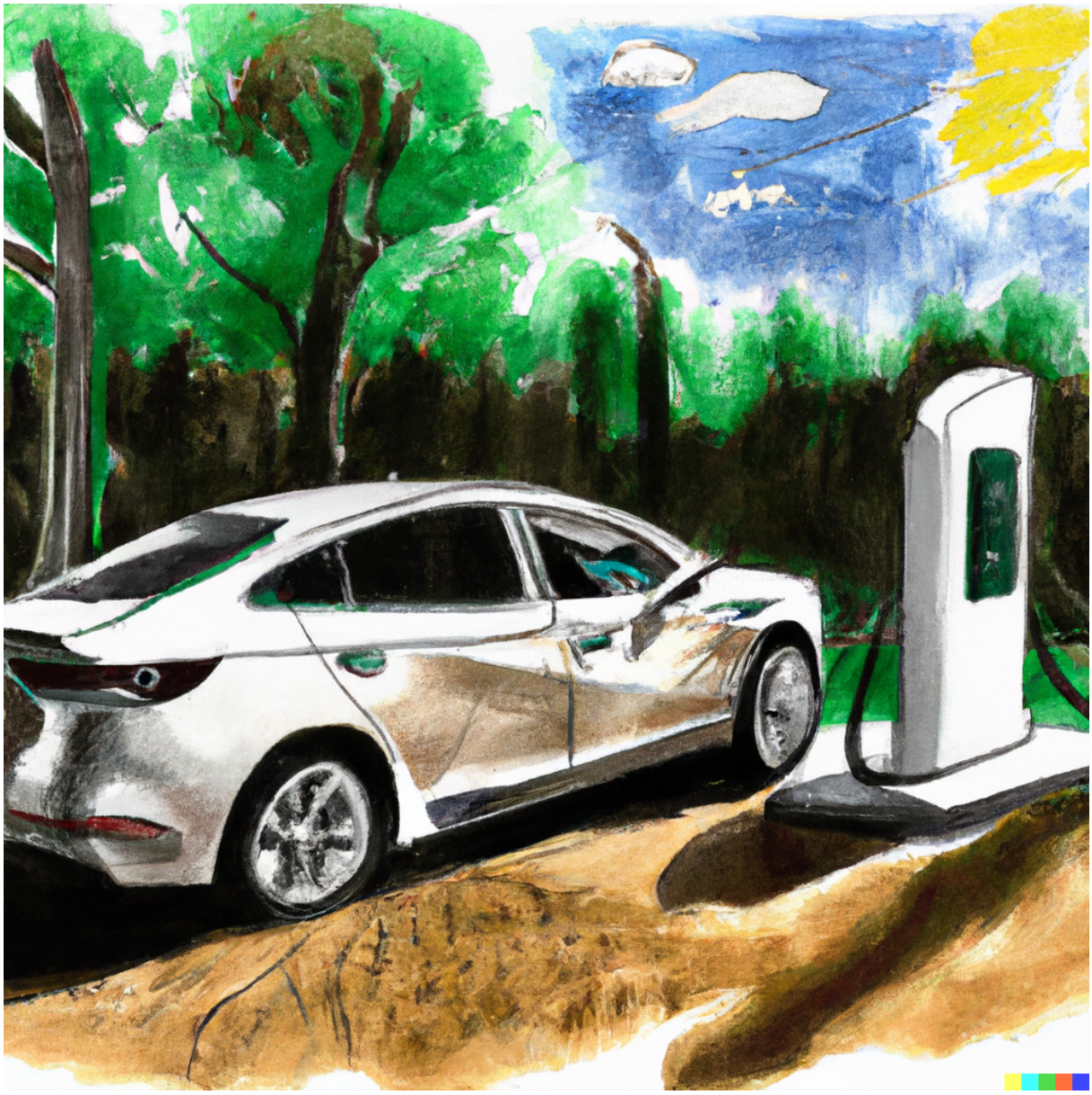
EV	Electric Vehicle
ICEV	Internal Combustion Engine Vehicle
IEA	International Energy Agency
EVI	Electric Vehicle Initiative
EVSE	Electric Vehicle Supply Equipment
V2G	Vehicle-to-Grid
SoC	State of Charge
E	Energy Consumption
C	Battery Capacity
SoC_{init}	Initial State of Charge
D	Distance Travelled
NHTS	National Household Travel Survey
GDP	Gross Domestic Product
NOAA	National Oceanic and Atmospheric Administration
RWI	Rheinisch-Westfälisches Institute
RMSE	Root Mean Squared Error
KDE	Kernel Density Estimator
GKDE	Gaussian Kernel Density Estimator
DKDE	Diffusion Kernel Density Estimator
HKDE	Hybrid Kernel Density Estimator
pdf	probabilty density function
SARIMA	Seasonal Auto-Regressive Integrated Moving Average
LM	Linear Model
SVM	Support Vector Machine
GAM	Generalised Additive Model
MR	Multi-Resolution
RF	Random Forest
ANN or NN	Artificial Neural Network
k-NN	k-Nearest Neighbours
MM	Mixture Model
MPSF	Modified Pattern-based Sequence Forecasting
MLP	Multi-Layer Perceptron
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
GHG	Greenhouse Gases
IOT	Internet of Things
NILM	Non-Intrusive Load Monitoring

Chapitre 1

De l'importance de la charge des Véhicules Électriques

Dans ce premier chapitre introductif, le lecteur trouvera un bref historique de l'émergence des véhicules électriques, quelques généralités sur le marché de l'électricité et certains principes fondamentaux pour une bonne intégration des VE dans le réseau électrique. De plus, certains faits importants sur le cycle de vie des VE sont rappelés. Enfin, les variables ainsi que les modélisations communément utilisées de la charge des VE sont présentées pour aboutir à la problématique générale de ce manuscrit : fournir une prévision journalière efficace des comportements de charge des VE.

Sommaire



1.1 L'émergence du marché des Véhicules Électriques

Cette première section propose un bref historique des Véhicules Électriques (VE) ainsi qu'une classification des différents types de VE, avant de conclure sur les raisons de leur essor récent. Le but de cette section est de donner une introduction aux enjeux liés à la mobilité électrique d'un point de vue historique et chronologique.

1.1.1 Un peu d'histoire

Les véhicules électriques (VE) sont souvent présentés comme l'innovation qui nous permettra de minimiser les problèmes environnementaux posés par les véhicules à combustion interne (ICEV). On imagine alors que d'ici quelques années ces VE empliront nos rues pour la première fois. Pourtant, cette image existait bel est bien il y a plus d'un siècle. En quelque sorte, notre "future" mobilité ne serait pas si surprenante pour un New-Yorkais de la fin du 19ème siècle (cf. Figure 1.1).

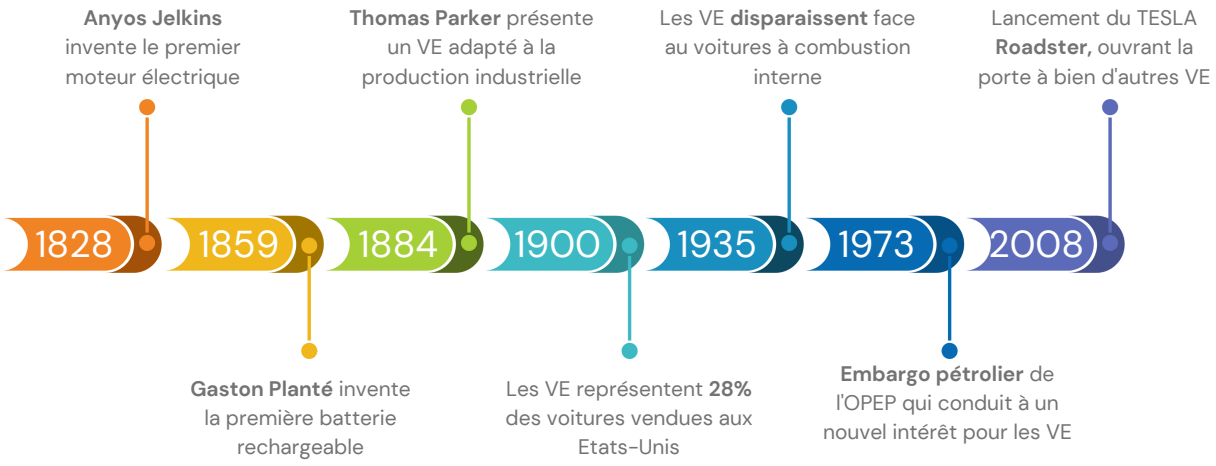


Figure 1.1 – La voiture électrique en 7 grandes dates

A l'origine

L'histoire des VE débute en 1827 avec Anyos Jedlik, un inventeur hongrois. Cette année là, Jedlik commence à travailler sur des dispositifs rotatifs électromagnétiques, et en 1828, il invente le premier dispositif contenant les trois composants principaux des moteurs à courant continu : le stator, le rotor et le commutateur 1.2. Ce premier événement est néanmoins controversé car son inventeur aurait attendu plusieurs dizaines d'années avant de la partager avec le grand public. C'est entre 1832 et 1839 que Robert Anderson, inventeur écossais, présenta le premier modèle de véhicule électrique. Toute l'ingéniosité de ses travaux résidait dans son prototype de batterie qui permettait d'alimenter un moteur électrique qu'il disposa sur un chariot (très similaire à ceux tirés par des chevaux à la même époque). Ce bond technologique était malgré tout accompagné d'une difficulté majeure : la batterie n'était pas rechargeable. Il fallait donc remplacer la batterie à chaque utilisation ce qui pouvait être très coûteux. En parallèle des travaux de Anderson, Thomas Davenport, inventeur américain, dévoile une petite

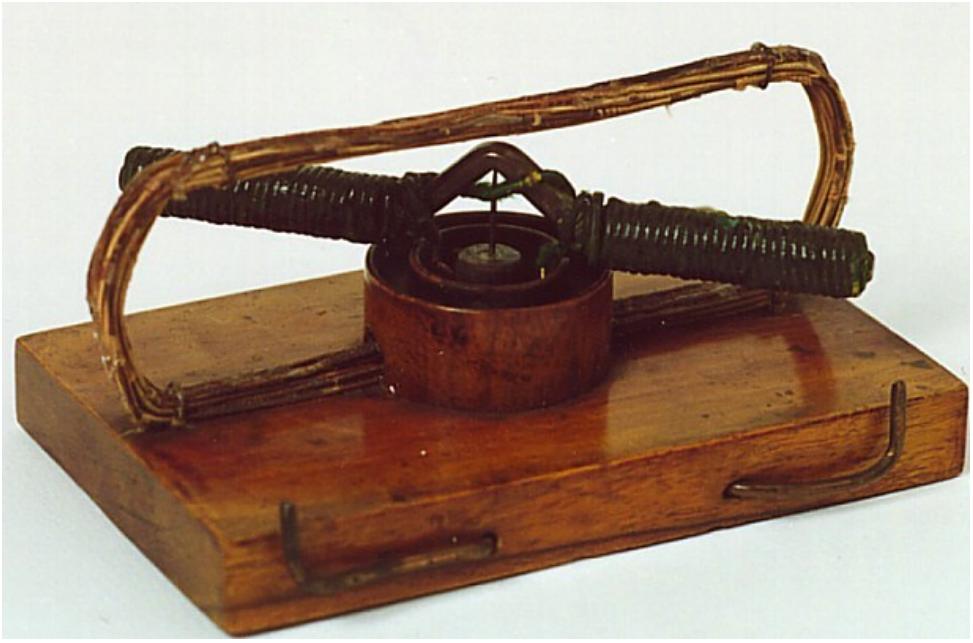


Figure 1.2 – Le premier moteur électrique conçu par Anyos Jedlik en 1828

locomotive propulsée par le premier moteur électrique à courant continu sur le sol américain. Là où les inventions précédentes étaient des modèles réduits ou de petites voitures, la locomotive de Davenport a été le premier véhicule électrique prêt à l'emploi jamais conçu. Le fort engouement autour de ces nouvelles technologies finit par s'essouffler faute de technologie viable au niveau de la batterie (pas rechargeable).

Il faudra attendre Gaston Planté, physicien français, qui inventa la batterie plomb-acide en 1859, connue comme étant la première batterie électrique rechargeable. En 1881, Camille Alphonse Faure, ingénieur français, mettra au point un modèle plus efficace et plus fiable qui connaîtra un grand succès pour relancer l'engouement autour des premières voitures électriques. Les améliorations de Faure augmentèrent considérablement la capacité de ces batteries et conduira directement à leur fabrication à l'échelle industrielle.

L'apogée

L'année 1884 marqua le début de l'âge d'or des VE. Cette année là, Thomas Parker construit la première voiture électrique capable d'être produite à une échelle industrielle à Londres. Ce véhicule utilisait ses propres batteries rechargeables à haute capacité. En 1888, l'ingénieur allemand Andreas Flocken proposa également son modèle de voiture électrique à quatre roues : la *Flocken Elektrowagen*. En 1889-1891, William Morrison introduit un wagon électrique très simple aux États-Unis qui n'était pas en soi une innovation, mais aura un succès retentissant.

Avec cette multitude de voitures électriques qui commençait à voir le jour, les premiers véhicules électriques commerciaux commencèrent à entrer dans la flotte de taxis de New York en 1897. Le constructeur automobile, Pope Manufacturing Co., devient le premier fabricant de VE à grande échelle aux États-Unis. La popularité du VE s'accompagne également de grands records. En effet, la "Jamais contente" a été le premier véhi-

cule routier à dépasser les 100 kilomètres à l'heure. C'était un véhicule électrique belge avec une carrosserie en alliage léger en forme de torpille. Ce record de vitesse terrestre a été établi entre le 29 avril et le 1er mai 1899 à Achères (Yvelines) près de Paris.

En 1900, les voitures électriques deviennent le véhicule routier le plus vendu aux États-Unis, capturant 28% du marché, et le nombre de VE en circulation atteindra un pic historique en 1912 avec environ 30 000 véhicules sur les routes [2].

La chute

On attribue souvent la chute des véhicules électriques à son compétiteur de toujours, le véhicule à combustion interne (ICEV). En particulier, dès 1908 la Ford T de Henry Ford, fonctionnant à l'essence, est introduite sur le marché. A l'époque, il y avait déjà les mêmes points qui séparaient ces deux technologies. En effet, les ICEV permettaient déjà de voyager plus longtemps que les VE. Pourtant les VE rencontraient un franc succès de part leur propreté et leur entretien facilité (moins de pièces). Pour les personnes qui pouvaient se le permettre, c'était la solution privilégiée par les citadins. D'ailleurs il est intéressant de noter que le marketing de ces véhicules était plus tourné vers les femmes justement pour l'aspect immaculé et discret des VE alors que le ICEV, plus sale et plus bruyant était réservé aux hommes. La femme de Henry Ford elle-même possédait une voiture électrique.

En 1912, le démarreur électrique, inventé par Charles Kettering, évite de recourir à une manivelle pour démarrer les ICEV. Cela rend ces véhicules encore plus facile d'accès. La suprématie de ces véhicules ne se fit plus trop attendre, en grande partie grâce à leur prix, bien plus abordable que celui des VE. Le Ford T a été un tel succès qu'Henry Ford en avait vendu 15 millions en 1927, 19 ans seulement après sa sortie. En 1935, les VE ont pratiquement disparu de la circulation en raison de la prédominance des ICEV et de la disponibilité d'essence bon marché.

Le renouveau

Mais les VE n'avaient pas dit leur dernier mot. Pour des raisons environnementales mais aussi politiques, ils préparent leur retour depuis l'après-guerre. Dès 1947 le rationnement du pétrole au Japon mène le constructeur automobile Tama à lancer une voiture électrique de 4,5 ch avec une batterie au plomb de 40V. On reprend donc la technologie pratiquement là où on l'avait laissée au début des années 1900. En 1966, le Congrès américain dépose une loi pour encourager la production et l'adoption des véhicules électriques comme moyen de réduire la pollution de l'air. Mais le point crucial qui a renouvelé l'intérêt des politiciens pour le VE est naturellement l'embargo pétrolier de l'OPEP en 1973 qui provoqua une forte augmentation des prix du pétrole mais aussi de longues files d'attente dans les stations-service. La même année, une percée majeure a été faite lorsque le chimiste britannique M. Stanley Whittingham a inventé les premières batteries lithium-ion rechargeables au monde. Les mêmes batteries qui sont utilisées dans les VE d'aujourd'hui. Suite à la crise pétrolière, en 1976, le gouvernement français lance le programme "PREDIT" pour accélérer la recherche sur les VE. Mais les prix du pétrole redevenant assez bas assez rapidement, l'intérêt autour des VE sera réduit. En 1996, pour se conformer à la politique zéro émission en Californie et les exigences relatives aux véhicules antipollution promulgué en 1990, General Motors

commence à produire et commercialiser la voiture électrique EV1. Considéré comme le premier VE de l'ère moderne, il ne sera pas un franc succès commercial. En revanche, en 1997 au Japon, Toyota lance la commercialisation de la Prius, première voiture hybride au monde qui sera commercialisée. 18 000 exemplaires sont vendus la première année de production et elle continue d'être commercialisée de nos jours ce qui constitue le premier grand retour au premier plan des VE.

En 2008, les prix du pétrole atteignent plus de 145 USD le baril. La même année, TESLA lance son premier VE, fontionnant exclusivement à l'électricité, le Roadster. Lors des tests, le véhicule atteint près de 400 kilomètres sur une seule charge et avait une vitesse de pointe de 200 km/h. La prouesse technologique était retentissante. Il s'agissait maintenant de continuer à améliorer la capacité des batteries et d'arriver d'une manière ou d'une autre à baisser les prix pour rendre les VE plus accessibles. Cette année coïncide également avec le début des sujets de recherches proche de ma thèse comme nous pourrons le voir plus tard dans le manuscrit (Section 2).

En 2010, la Nissan LEAF est lancée et elle gagnera le prix de la voiture européenne de l'année en 2011. Cette année-là, le nombre de VE en circulation dépasse le record de 1912 en atteignant 50 000. Toujours cette année là, le plus grand service d'auto-partage de VE au monde, Autolib, est lancé à Paris avec une cible de 3 000 véhicules électriques. L'état français s'engage à acheter 50 000 véhicules électriques sur quatre ans. Ces grands changements ne sont pas réservés au continent européen. En 2012, la Chevrolet Volt se vend mieux que la majorité des modèles de voitures sur le marché américain. Et cette année le nombre de véhicules en circulation a plus que triplé par rapport à 2011 car il dépasse les 180 000.

1.1.2 Classification

A ce jour, on recense cinq technologies de VE différentes par leur moteur (cf. Figure 1.3). Nous utilisons les abréviations anglophones car ce sont celles qui sont le plus souvent retenues dans la littérature.

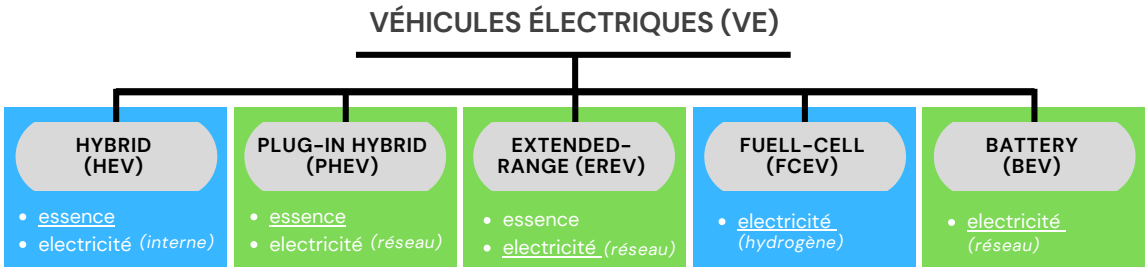


Figure 1.3 – Classification des VE, avec leur source d'énergie principale soulignée

Un véhicule électrique hybride (HEV) combine un système de propulsion à moteur à combustion interne (ICE) conventionnel avec un système de propulsion électrique (transmission de véhicule hybride). Dans les modèles modernes, les batteries peuvent également être chargées grâce à l'énergie générée lors du freinage, transformant l'énergie cinétique en énergie électrique. On parle de freinage régénératif. Le VE hybride re-

chargeable (PHEV) utilise à la fois un moteur électrique et un moteur à essence pour fonctionner. Son moteur électrique utilise des batteries qui se rechargent en se branchant sur une source d'alimentation électrique (une prise murale ou un chargeur de VE). Le moteur à essence peut fonctionner avec le moteur électrique, ou séparément, pour alimenter le groupe motopropulseur. Le VE à autonomie prolongée (EREV) possède une unité d'alimentation auxiliaire (appelée prolongateur d'autonomie) qui augmente l'autonomie de l'EREV. La plupart des prolongateurs d'autonomie sont de petits ICE qui entraînent un générateur électrique fournissant de l'électricité aux batteries électriques et au moteur. Un véhicule électrique à pile à combustible (FCEV) génère de l'électricité pour alimenter le moteur, généralement en utilisant de l'oxygène de l'air et de l'hydrogène comprimé. Enfin, le VE à batterie (BEV) tire toute sa puissance de ses batteries pour alimenter ses moteurs électriques. Il ne contient pas de moteur à combustion interne (ICE). Son moteur électrique utilise des batteries qui se rechargent en se branchant sur une source d'alimentation électrique (une prise murale ou un chargeur de VE). Globalement, seuls deux types de VE fonctionnent exclusivement à l'électricité : le FCEV et le BEV. Dans cette thèse, nous nous intéressons plus particulièrement aux PHEV, EREV et BEV (en vert sur la Figure 1.3) car ce sont ceux qui peuvent se brancher au réseau électrique. Dans la suite, nous appellerons "VE" uniquement ces trois types de VE à part si l'on précise autrement.

1.1.3 Essor récent

Historiquement, nous pouvons retenir deux raisons majeures pour l'essor des VE. La hausse des prix et l'épuisement des ressources fossiles. L'impact de l'homme sur le changement climatique avec le secteur des transports étant souvent montré du doigt à juste titre comme étant responsables d'une part significative des émissions de gaz à effet de serre. Cela a amené les états et les entreprises à travailler de concert pour proposer une mobilité électrique considérée comme étant le bon chemin vers une trajectoire zéro carbone. Nous nuancerons les bienfaits environnementaux des VE dans la section 1.4. Cela s'est traduit par des chiffres en constant progrès depuis les années 2000. En 2021, près de 10% des ventes mondiales de voitures étaient électriques, soit quatre fois plus que la part de marché des VE en 2019. Cela a porté le nombre total de voitures électriques sur les routes à environ 16,5 millions dans le monde, soit trois fois plus qu'en 2018 [3].

1.1.4 Scénarios

Trois scénarios principaux d'adoption des VE sont retenus par l'Agence Internationale de l'Energie (IEA). Le premier est le scénario lié aux lois et réglementations déclarées par les différents gouvernements (Stated Policies Scenario). Ce scénario reflète les politiques et les mesures existantes, ainsi que les ambitions et les objectifs politiques qui ont été légiférés par les gouvernements du monde entier. Il comprend les politiques et réglementations actuelles liées aux véhicules électriques et les développements futurs basés sur les impacts attendus des déploiements annoncés et des plans des parties prenantes de l'industrie. Il vise à dresser un miroir des plans des décideurs politiques et à illustrer leurs conséquences. Le deuxième scénario est le scénario des promesses po-

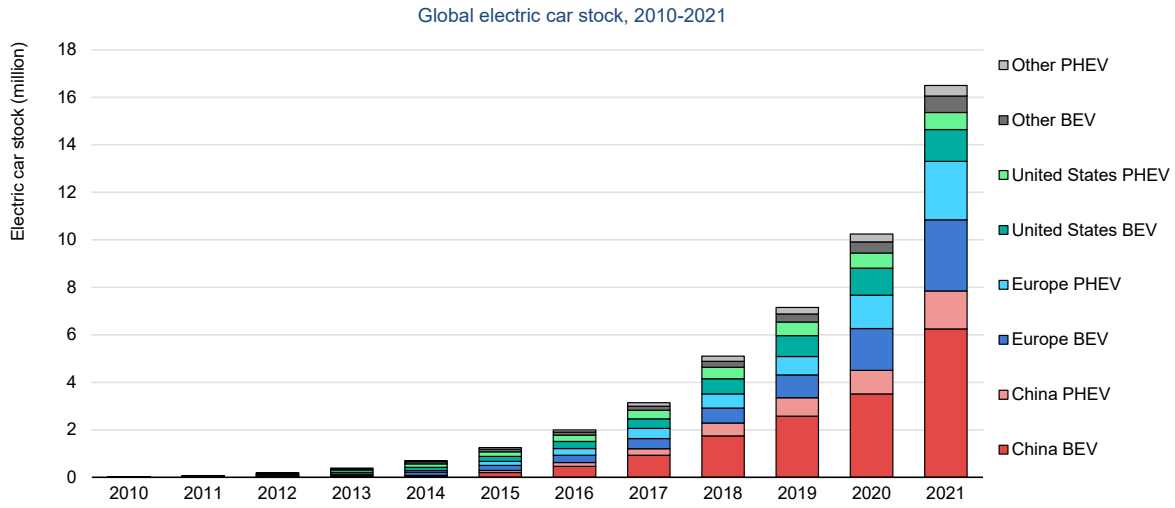


Figure 1.4 – Nombre de véhicules en circulation par type [3]

litiques annoncées (Announced Pledges Scenario). Celui-ci suppose que les ambitions annoncées et les objectifs fixés par les gouvernements du monde entier, y compris les plus récents, sont atteints intégralement et dans les délais. En ce qui concerne les VE, il comprend toutes les annonces majeures récentes d'objectifs d'électrification et d'émissions nettes nulles à plus long terme et d'autres engagements, qu'ils aient ou non été ancrés dans la législation ou dans des contributions déterminées au niveau national mises à jour. Enfin, le scénario zéro émission en 2050 (Net Zero by 2050) est un scénario normatif qui définit une voie étroite mais réalisable pour le secteur mondial de l'énergie pour atteindre zéro émission nette de CO_2 d'ici 2050. Ce scénario est associé avec la limitation de l'augmentation de la température mondiale à 1,5° conformément aux réductions évaluées par le Groupe d'experts intergouvernemental sur l'évolution du climat (GIEC). Il existe de nombreuses voies possibles pour atteindre zéro émission nette de CO_2 dans le monde d'ici 2050 et de nombreuses incertitudes qui pourraient affecter chacun d'entre elles. Par conséquent, le scénario zéro émission n'est qu'une voie et non la voie vers des émissions nettes nulles.

D'après la Figure 1.5, il est clair que quel que soit le scénario, l'adoption des VE suit une évolution exponentielle et continue sur la tendance des dix dernières années. En revanche, il est assez inquiétant de voir qu'en l'état, même en prenant en compte les promesses faites par les différentes entités gouvernementales, nous sommes encore loin de pouvoir atteindre le scénario zéro émission qui nous permettrait de tenir les accords de Paris [4].

1.2 Généralités sur le marché de l'électricité

L'électricité est une denrée locale. Nous entendons par là deux choses fondamentales. La première est que l'électricité ne peut être stockée telle quelle. Il existe bien des batteries stockant l'électricité sous forme d'énergie chimique pour des appareils électroniques ou certains bâtiments, mais il est inenvisageable avec les technologies actuelles d'utiliser des batteries à une plus grande échelle. La seconde chose est que le

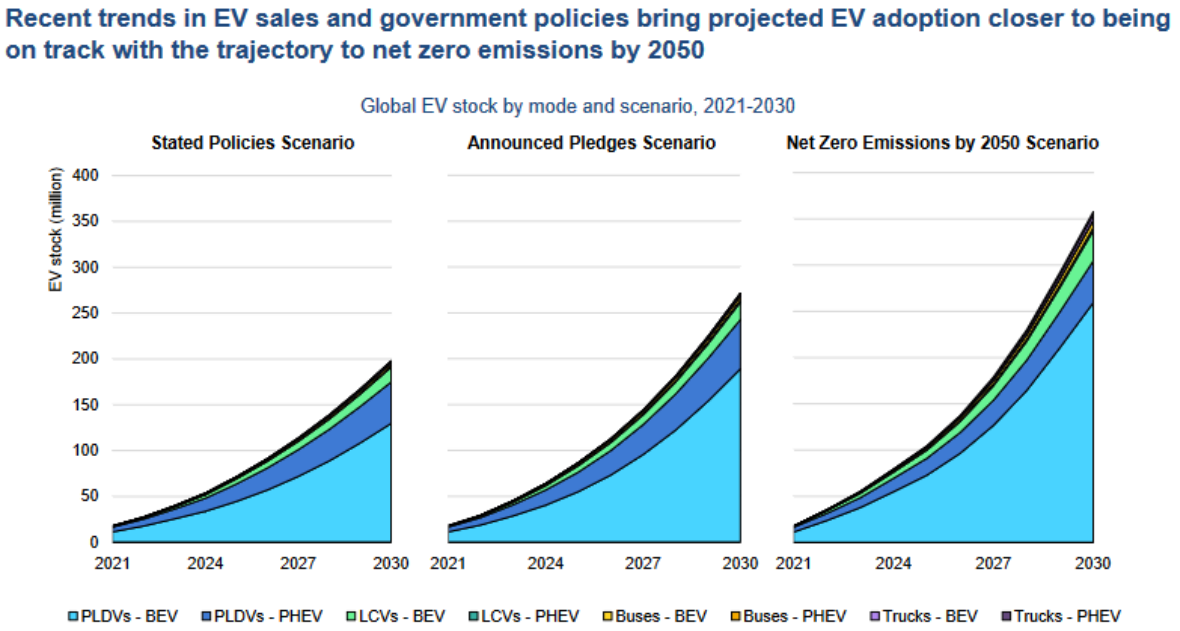


Figure 1.5 – Evolution du nombre de VE jusqu'en 2030 selon trois scénarios majeurs [3]

transport de l'électricité obéit à des lois bien spécifiques (lois de Kirchhoff). Par conséquent, dans un réseau électrique maillé, l'électricité ira d'un point à un autre en utilisant tous les chemins possibles, provoquant ainsi d'éventuelles interférences du flux d'électricité. Cela restreint donc les opportunités d'échanges transfrontaliers qui sont limités par la capacité de transfert disponible. Ce n'est pas le cas d'autres sources d'énergies secondaires comme le baril de pétrole qui peut être relativement aisément transporté partout dans le monde.

A cela s'ajoute une contrainte physique fondamentale. La puissance soutirée doit être égale à chaque instant à la puissance injectée dans le réseau. Cet équilibre s'exprime avec une fréquence constante (e.g., 50 Hz en France). Si la fréquence chute sous sa valeur de référence, la consommation est supérieure à la production. *A contrario*, si la fréquence augmente, la production dépasse alors la consommation. Le principal garant de cet équilibre est le gestionnaire du réseau - *Independant System Operator* (ISO) ou *Transport System Operator* (TSO) en anglais. Il doit maintenir en permanence l'équilibre en s'assurant que la déviation de la fréquence de référence n'est pas trop importante. Dès qu'il y a un décrochage il se doit de réagir au plus vite pour éviter des catastrophes comme le blackout de 2003 en Amérique du Nord [5].

Dans les sections suivantes, nous allons brièvement présenter comment l'électricité est acheminée au consommateur en expliquant les trois étapes fondamentales (génération, transmission, distribution). Puis nous allons rapidement présenter la structure générale du marché de l'électricité. Enfin, nous expliquerons en quoi la prévision de la consommation est un enjeu fondamental pour les différents acteurs du marché de l'électricité.

1.2.1 L'acheminement de l'électricité au consommateur

Depuis la mise en concurrence et l'ouverture du marché en 2007 à l'échelle nationale et européenne la plupart des pays fonctionnent de la manière suivante : trois étapes fondamentales accompagnent le voyage de l'électricité de sa production à sa consommation. Tout commence par la **génération** de l'électricité dans des centrales utilisant une source d'énergie primaire (e.g., soleil, vent, nucléaire, charbon). L'électricité générée est alors envoyée dans des transformateurs pour augmenter la tension du courant afin de lui permettre de parcourir de longues distances tout en minimisant les pertes par effet Joule. C'est l'étape de **transmission** permise par les lignes à haute tension qui s'étendent à travers le pays. Le courant atteint alors une station électrique secondaire où la tension est abaissée afin qu'il puisse être envoyé sur des lignes électriques plus petites destinées à la **distribution**. C'est avec cette dernière étape que l'électricité atteint les différents bâtiments mais aussi les infrastructures de recharge des VE.

Génération

L'énergie électrique est une source d'énergie secondaire qui provient d'énergies primaires comme le vent ou le charbon. Elle est générée à un niveau de basse tension (quelques dizaines de kV) pour un meilleur rapport coût-efficacité dans les centrales. Les deux principaux acteurs de la génération d'électricité en France sont EDF et ENGIE. A eux deux, ils génèrent 95% de l'électricité sur le sol français. Les 5% sont constitués d'autres entreprises et de producteurs indépendants. Ce sont donc eux qui sont responsables de l'injection de la puissance dans le réseau électrique. Depuis l'ouverture du marché en 2007 néanmoins, la puissance injectée par EDF ne couvre pas uniquement des clients mais peut aussi correspondre à des intermédiaires.

Transmission

Une fois l'électricité générée (à basse tension), la tension est ensuite augmentée dans un transformateur pour baisser l'intensité du courant électrique de l'alimentation. La réduction du courant entraînera la réduction des pertes ohmiques dans le système. Cela peut être expliqué en utilisant l'équation de puissance instantanée dissipée par effet joule : $P_j(t) = RI(t)^2$ avec I l'intensité (ou courant), R la résistance et P_j la puissance instantanée dissipée par effet Joule (chaleur). Une tension plus élevée permettra également de distribuer la puissance à une distance beaucoup plus éloignée. La transmission longue distance à basse tension entraîne une résistance plus élevée. Par conséquent, la tension est pompée jusqu'à une valeur beaucoup plus élevée avant la transmission (plusieurs centaines de kV). Il s'agit généralement d'un courant alternatif triphasé. Comme les centrales électriques sont généralement situées loin des consommateurs, l'électricité doit parcourir une grande distance. Pour réduire l'énergie perdue lors d'une transmission à grande distance, l'électricité est transmise à haute tension. En France, c'est le Réseau de Transport d'Electricité (RTE) qui est en charge de la transmission.

Distribution

Une fois que l'électricité se rapproche du consommateur un transformateur abaisse la tension. On repasse ici à une tension aux alentours de quelques dizaines de kV. Deux types de clients sont distingués : les clients primaires (e.g., les bâtiments industriels), les clients secondaires (e.g., les particuliers). A ces deux types de clients sont associés un voltage spécifique. En France, l'un des acteurs majeur de la distribution est ENEDIS.

1.2.2 La structure du marché

L'électricité étant donc une denrée locale, il existe autant de marchés de l'électricité que d'états. La microstructure du marché dépend fortement de la réglementation nationale. Néanmoins, une structure commune émerge, portée par la nécessité d'un équilibre entre consommation et production avec un rôle central de l'ISO, (e.g., RTE en France et ENTSO-E à l'échelle européenne). Il n'y a pas qu'un seul marché mais une séquence de marchés qui peuvent être classés par horizons temporels (cf. Figure 1.6). Le marché *infra-day* (infrajournalier), le marché *day-ahead* (journalier) et le marché *forward* (de quelques jours à plusieurs années). Au moyen/long-terme les contrats bilatéraux sont négociés de gré à gré. Cela peut couvrir des produits standards ou plus spéciaux (e.g., profil de puissance, maturité) : on parle de *forwards* ou de *futures*. Au court-terme (à partir de J-1), les négociations ont lieu sur le marché Spot. On y retrouve des produits standards sur le marché day-ahead et une procédure d'offre/demande en temps réel sur le marché infrajournalier. Enfin des mécanismes d'ajustement mis en place par l'ISO entre en jeu pour rééquilibrer le réseau s'il fait face à une déviation de la fréquence de référence. Les acteurs qui interviennent sur le marché sont : les producteurs d'électricité qui négocient et vendent la production de leurs centrales électriques, les fournisseurs d'électricité qui négocient et s'approvisionnent en électricité et la vendent ensuite aux clients pour leur consommation, les négociants qui achètent pour revendre (ou inversement) et favorisent ainsi la liquidité du marché, les opérateurs d'effacement qui valorisent la consommation de leurs clients à certains instants de la journée.

Marché Spot

Le marché Spot ou marché "physique" fonctionne en deux temps. Tout d'abord en *Day-ahead* où il regroupe les offres de vente et/ou d'achat d'électricité soumises par l'ensemble des acteurs du marché pour chaque heure du lendemain. Une fois l'ensemble de ces offres soumises (jusqu'à midi la veille), les courbes d'offres et de demandes sont croisées pour chaque heure du lendemain et l'intersection de ces deux courbes fixe le prix pour cette heure (ou pour chaque bloc d'heures). Tous les acteurs ayant soumis une offre supérieure au prix fixé, voient leur offre validée et doivent régler, non pas le prix précisé dans leur offre, mais bien celui du prix fixé *a posteriori* pour l'heure en question. C'est en quelque sorte un système d'enchère à l'aveugle où l'on ne paye que le prix fixé *a posteriori*. Ce mécanisme est utile pour combattre la volatilité des prix qui peut être un ennemi de la résilience du réseau. Grâce à cette procédure, deux prix de référence émergent : la moyenne des prix sur toute la journée *Baseload* et la moyenne des prix sur la période définie comme étant les heures de pointe *Peakload* (de 8h à 20h en France). Ils seront utilisés pour la création de contrats futures sur le marché

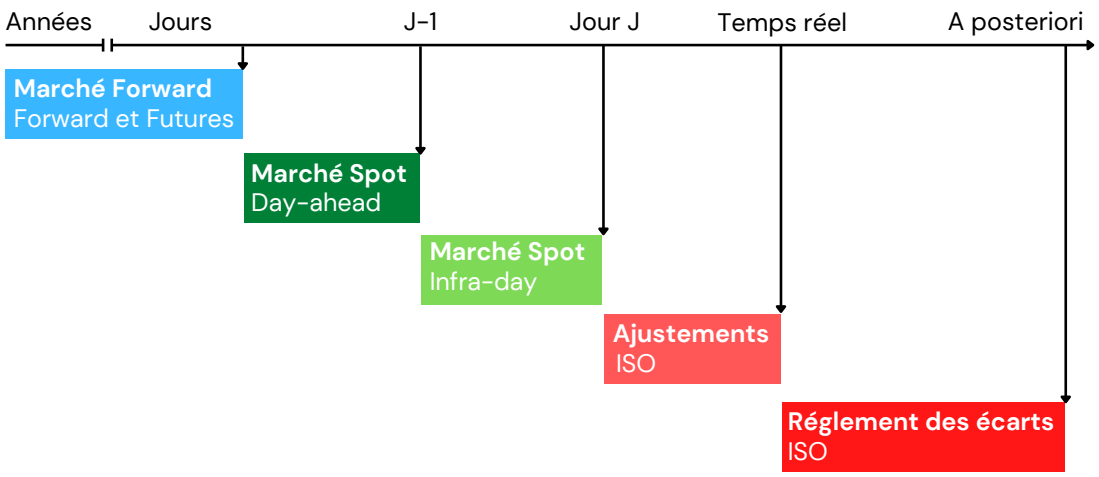


Figure 1.6 – Les différents marchés de l'électricité

Forward (cf. Section 1.2.2). Le marché Spot *Day-ahead* regroupe donc des produits horaires (sur une ou plusieurs heures) avec livraison le lendemain. Les offres de vente se situent généralement légèrement au-dessus des coûts marginaux des actifs de génération. Par conséquent, cette courbe suit le principe du *merit order* où les actifs de générations sont sollicités par ordre croissant de leurs coûts marginaux. Généralement, on retrouve parmi les actifs de génération ayant les coûts marginaux les plus faibles les énergies renouvelables (la matière première étant "gratuite") puis vient le nucléaire et enfin les énergies fossiles. Les moyens du parc énergétique étant assez stables dans le temps (moyennant l'évolution du prix des matières premières, les périodes de maintenance des différentes centrales et l'intermittence des énergies renouvelables) c'est souvent la demande qui est bien plus volatile et qui peut entraîner des pics de prix ou *a contrario* des périodes où les prix sont moins élevés. Dans un second temps, le marché Spot infrajournalier regroupe les transactions concernant des produits demi-horaires, horaires ou par blocs de plusieurs heures, avec livraison le jour même. C'est la possibilité pour les différents acteurs d'effectuer des ajustements de dernière minute après les offres retenues sur le marché textitday-ahead. En France, ce marché ouvre à 15h la veille du jour de livraison. Le fonctionnement de ce marché repose sur l'offre de vente et d'achat d'électricité en temps réel (contrairement au marché Spot *day-ahead*). Dès qu'une offre de vente correspond à une offre d'achat la transaction est opérée. La liquidité du Spot infrajournalier est donc limitée. Enfin, ce marché est le plus volatil car il peut s'avérer très complexe de combler une demande soudaine en électricité en raison des contraintes physiques évoquées plus haut.

Marché Forward

Le marché *Forward* correspond à des produits avec des échéances supérieures à la journée. Cela regroupe principalement deux types de produits : les contrats *forwards* et les contrats *futures*. En Europe, les différents produits *futures* sont standardisés et dis-

ponibles sur l'*European Energy Exchange* (EEX). Il peuvent couvrir, plusieurs semaines, trimestres et ce jusqu'à 6 années. Ces produits se déclinent de trois façons : sur toutes les heures d'une période (*Baseload*), uniquement pour les heures de pointes, de 8h à 20h en France (*Peakload*) et en prenant les blocs de plusieurs heures. Quant aux produits *forwards*, ce sont des contrats qui peuvent être signés directement de gré à gré ou par l'intermédiaire d'un courtier. Ils ne sont donc pas standardisés et les prix sont négociés pendant la création du contrat. Le but de ce marché est de permettre aux acteurs de sécuriser leurs achats et/ou leurs coûts au moyen/long-terme. En effet, les prix de ces produits sont généralement moins volatils que les prix du marché spot. Pour les acteurs responsables de la génération d'électricité, cela les engage à maintenir voire à développer leurs actifs de générations au fur et à mesure du temps. C'est aussi un moyen pour eux de sécuriser des financements pour l'expansion de leurs actifs venant de différentes entités gouvernementales.

Réserves et mécanismes d'ajustement

A la suite des transactions effectuées sur les marchés *Forward* et Spot, l'ISO rentre alors en jeu pour assurer la sécurité du réseau électrique à l'aide de ses réserves et de mécanismes d'ajustement. Le but est toujours de maintenir l'équilibre en garantissant que la puissance soutirée soit égale à la puissance injectée à chaque instant. Cet équilibre peut-être menacé en temps réel par la perte de certaines unités de génération d'électricité, des erreurs dans les prévisions de consommation ou de production ou simplement la congestion du réseau. L'ISO dispose de deux moyens fondamentaux pour rééquilibrer le système. Premièrement, un système de réserves successives. En France, elles sont au nombre de trois : la réserve primaire, avec un temps de réponse inférieur à 30 secondes permet de suppléer le réseau avec 500 MW automatiquement. La réserve secondaire elle aussi s'active automatiquement avec un temps de réponse inférieur à 3 minutes et une puissance de 600 MW. Enfin, la réserve tertiaire elle s'active manuellement avec un temps de réponse de 15 minutes et représente une puissance d'environ 1500 MW. Au-delà de ces réserves, l'ISO peut également mobiliser un mécanisme d'ajustement. Ces mécanismes font rentrer en jeu un nouveau type d'acteurs appelés responsables d'équilibres (RE). Les RE sont des opérateurs du marché qui ont signés un contrat les engageant auprès de RTE pour régler les écarts observés *a posteriori* entre l'électricité injectée et celle soutirée sur leur périmètre de clients. Les RE soumettent alors des offres d'augmentation ou de diminution de leur production ou consommation et l'ISO sélectionne les offres en fonction de la préséance économique dans un délai de 30 minutes.

Règlement des écarts

En France, RTE établit, *a posteriori*, la facture à payer ou à recevoir par un RE, pour les écarts observés sur son périmètre (puissance injectée/soutirée). La formule de règlement de ces écart est basée sur le prix spot et les coûts de production d'électricité. L'idée étant d'inciter un comportement vertueux tant pour les producteurs que pour les consommateurs. Prenons un scénario où le réseau serait en manque de puissance. Si EDF est en excès de puissance sur son périmètre, alors EDF reçoit le prix Spot pour l'instant de la période de la journée considérée. En revanche, si EDF est en manque

de puissance, alors EDF devra payer des pénalités également en fonction du prix Spot pour la période considérée.

1.2.3 Modèles de prévision

Jusqu'à présent nous avons détaillé les enjeux ainsi que les activités spéculatives liées au marché de l'électricité. Minimiser les risques à la fois en terme de sécurité et sur un plan financier est au cœur des activités des RE. Cela se traduit par la production de modèles de prévision des différentes inconnues du marché (e.g., les prix, la demande, la production). Tout comme le marché de l'électricité, ces prévisions peuvent se faire à long-terme, moyen-terme ou à court-terme avec des objectifs différents dans chacun de ces cas. De plus, il ne s'agit pas uniquement de produire une prévision ponctuelle des différentes quantités mais on peut également rencontrer de plus en plus fréquemment des modèles de prévision probabiliste (quantiles) et plus particulièrement des modèles de prévision de pics. L'émergence de ce secteur d'activité n'est pas uniquement due à l'ouverture du marché à la compétition mais est aussi rendue de plus en plus difficile avec l'émergence des énergies renouvelables (intermittentes par nature) et de nouveaux usages de consommation (e.g., les VE). Les travaux présentés dans ce manuscrit s'inscrivent précisément dans le contexte de la prévision de la consommation et en particulier celle du nouveau marché que représente les VE.

1.3 L'intégration des Véhicules Électriques dans le réseau

Dans cette section, nous introduisons certains points fondamentaux pour l'intégration des VE dans le réseau électrique. En particulier, nous détaillons les notions liées à l'infrastructure de recharge et la recharge intelligente ou *smart-charging*.

1.3.1 Infrastructure de recharge

La recharge des VE s'effectue le plus souvent à domicile mais aussi sur des bornes publiques (e.g., municipalités, autoroutes) ou privées (e.g., lieu de travail). Quel que soit le lieu de recharge, nous pouvons trouver différentes normes concernant les infrastructures de recharge, qui sont déterminées, principalement, par la région dans laquelle elles sont appliquées. Plus précisément, trois normes principales existent actuellement (cf. Table 1.1 [6]) : 1. en Amérique du Nord et dans la zone Pacifique (SAE-J1772), 2. en Chine (GB/T 20234) et 3. en Europe (IEC-62196). En résumé, on retrouve différents modes de charge délivrant différents niveaux de puissance maximale. Plus la puissance est élevée plus la charge sera rapide. La charge la plus rapide possible peut être atteinte avec des infrastructures de recharges en courant direct (DC). Souvent très larges et encombrantes, ces bornes ne sont pas disponibles à l'usage domestique. Elles peuvent atteindre plusieurs centaines de kW en puissance maximale ce qui par exemple permet de charger une Tesla en moins d'une demi-heure. Cependant, le recours à ces infrastructures DC doit être parcimonieux au risque de mettre en péril la durée de vie de

Norme	Mode	Voltage (V)	Intensité (A)	Puissance Maximale (kW)
SAE-J1772 (Amérique)	AC Niveau 1	120	16	1.9
	AC Niveau 2	240	80	19.2
	DC Niveau 1	200-500	80	40
	DC Niveau 2		200	100
IEC-62196 (Europe)	Mode 1 (AC)	16	230-240	3.8
			480	7.6
	Mode 2 (AC)	32	230-240	7.6
			480	15.3
	Mode 3 (AC)	32-250	230-240	60
			480	120
	Mode 4 (DC)	250-400	600-1000	400
GB/T-20234 (Chine)	AC	250	10	2,5
			16	4
			32	8
		440	16	7
			32	14
			63	27.7
	DC	750-1000	80	80
			125	125
			200	200
			250	250

Table 1.1 – Les trois principales normes pour les infrastructures de recharges [6]

la batterie. De plus, charger une multitude de véhicules en DC n'est également pas forcément souhaitable à l'échelle du réseau car on pourrait avoir de grosses contraintes et des problèmes en certains endroits.

1.3.2 Impact et contraintes de recharge

Les VE représentent un défi important pour ce qui concerne leur intégration dans le réseau de distribution à grande échelle [7]. Une mauvaise gestion de ce nouveau marché peut avoir un impact très négatif sur la courbe de charge (augmentation des pics de demande). De plus, cette contrainte ajoutée sur le réseau peut mener à la surcharge de certains composants du système liés à des déséquilibres de tension et fréquence. La stabilité du réseau de distribution est donc en danger. Un unique VE est loin d'être un problème pour le réseau électrique. Cependant, une charge simultanée de centaines de VE dans un quartier pourrait grandement dépasser la capacité du réseau. Pour illustrer cela il suffit de regarder la courbe de charge d'un ménage possédant un VE Figure 1.7. Pour autant, si l'intégration des VE représente pour beaucoup une contrainte additionnelle sur le réseau, en modifiant le paradigme du consommateur qui se recharge quand bon lui semble de façon non-contrôlée pour passer à un modèle avec des agré-

gateurs qui ont pour mission d'optimiser la recharge des VE, on peut imaginer que la charge des VE pourrait servir différents desseins. Ce modèle s'appelle la recharge intelligente ou *smart charging*.

1.3.3 Smart Charging

Le *smart charging* consiste à connecter les bornes de recharge de VE avec des agrégateurs. Chaque fois qu'un VE est branché, la station envoie des informations (e.g., le temps de charge, la puissance délivrée) par Wi-Fi ou Bluetooth à une plateforme de gestion centralisée. Des données supplémentaires complètent les informations de la station pour fournir une vision plus générale du réseau aux agrégateurs (e.g., capacité du réseau local). Ces données sont analysées et visualisées en temps réel et peuvent être utilisées pour prendre des décisions automatiques sur comment et quand les véhicules électriques seront chargés. Grâce à cela, les agrégateurs peuvent contrôler et réguler la consommation d'énergie facilement et à distance via une plate-forme, un site Web ou une application mobile. Nous présentons ci-dessous les options de *smart charging* les plus courantes ou les plus étudiées à ce jour.

Partage de la puissance

Le partage de puissance, également parfois appelé équilibrage de charge ou nivellement, permet aux opérateurs de réseau ou aux entreprises disposant de plusieurs chargeurs sur site de répartir la capacité énergétique disponible proportionnellement sur toutes les bornes de recharge actives pour VE. Étant donné que la puissance disponible est limitée sur chaque site, une plus grande demande d'énergie pourrait engendrer des mises à niveau coûteuses de l'infrastructure électrique. Grâce au *smart charging*, la puissance peut être distribuée de manière optimale afin d'éviter de telles mises à niveau. C'est le même principe à une échelle plus grande comme une ville ou une région. Partager la puissance permet d'éviter de surcharger le réseau en acceptant un temps de recharge plus long pour les utilisateurs.

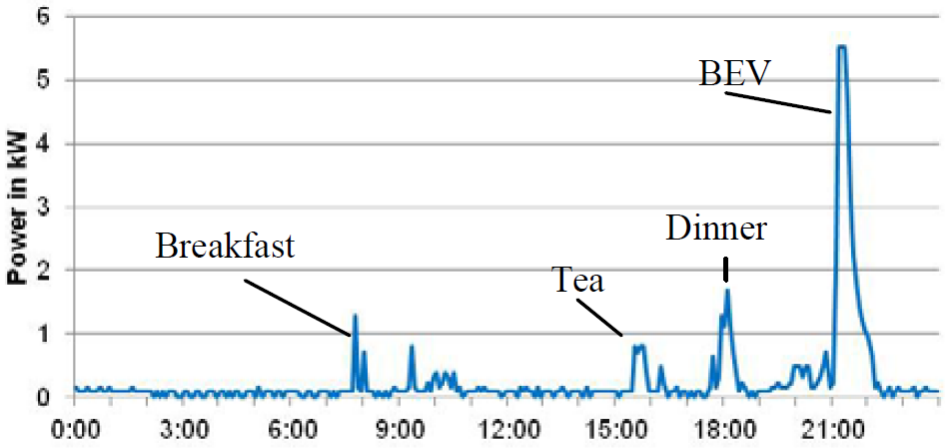


Figure 1.7 – L'impact d'un BEV sur la courbe de charge d'un ménage [8]

Minimisation d'externalités négatives

Le *smart charging* pourrait également contribuer à réduire considérablement les émissions du secteur des transports. Alors que les véhicules électriques sont déjà plus propres que les véhicules ICE, charger 1 million de véhicules électriques au bon moment sur le réseau actuel équivaut à retirer entre 20 000 et 80 000 véhicules ICE supplémentaires de la route avec les technologies actuelles aux États-Unis [9]. Dans ce rapport, plusieurs scénarios de charge de jour et de nuit concernant six RE aux États-Unis sont étudiés. En particulier, la charge non-contrôlée est comparée à la charge optimisée par rapport aux émissions de gaz à effet de serre. Deux facteurs clés conduisent à des émissions maximales : le mix énergétique du réseau local et le comportement de recharge des usagers. Donner la priorité à la charge de niveau 2 (cf. Table 1.1) avec des temps de séjour plus longs pour maximiser la flexibilité des véhicules électriques en tant qu'actifs du réseau. Cela pourrait inciter les conducteurs de VE à se brancher tous les jours ou toutes les nuits de manière plus consistante et maintenir un certain degré de prévisibilité. De plus, le fait de se brancher tous les jours augmente la flexibilité de chaque session de charge, maximisant ainsi les économies potentielles en terme d'émissions. Une autre recommandation de ce rapport est adressée aux services publics. Il exhorte ces acteurs du réseau d'intégrer l'électrification des transports dans la planification énergétique. Par exemple, pour les réseaux avec une importante quantité d'énergie solaire diurne, cela signifie accélérer les programmes de recharge sur le lieu de travail. Pour les réseaux fortement éoliens, le modèle opérationnel de recharge nocturne dans un dépôt centralisé (e.g., pour les bus de la municipalité) s'associe parfaitement à l'augmentation de la production éolienne pendant la nuit.

Vehicle-to-Grid

La transition énergétique place les fournisseurs d'énergie devant un double défi. D'une part, ils doivent pouvoir exploiter de plus en plus des sources d'énergie intermittentes comme l'éolien et le solaire, mais sans pouvoir stocker l'énergie qu'ils produisent à grande échelle. D'autre part, ils doivent pouvoir garantir la stabilité du réseau et répondre instantanément à la demande des consommateurs. C'est dans ce contexte qu'interviennent les technologies vehicle-to-grid (V2G). C'est un changement complet de paradigme où la batterie d'un VE est considérée comme une extension du réseau électrique. Précisément, le paradigme V2G considère la batterie d'un VE comme une réserve d'énergie dans laquelle les fournisseurs d'électricité peuvent puiser si nécessaire. La charge devient alors un processus bidirectionnel, ce qui signifie que le réseau ne se contente plus d'alimenter en électricité la batterie du véhicule : il considère également cette batterie comme une source d'énergie à utiliser pour répondre aux différents besoins de consommation d'énergie. Avec le V2G, un utilisateur de VE peut donc décider de stocker l'électricité lorsque les tarifs sont les plus bas, puis de l'utiliser lorsque le prix augmente. Un conducteur qui rentre chez lui la nuit pourrait par exemple utiliser l'énergie stockée dans la batterie de sa voiture électrique pour alimenter ses appareils électroménagers. Il peut alors recharger cette même batterie plus tard dans la soirée, au moment où le fournisseur d'électricité propose les tarifs les moins chers.

De la même manière, la flexibilité apportée par le V2G permet de recharger une batterie pendant les heures où l'énergie est produite par des sources renouvelables, puis

d'utiliser l'électricité en cas d'indisponibilité de l'énergie solaire ou éolienne. C'est aussi le principe des systèmes de stockage d'énergie stationnaire sur batteries, qui visent à donner une seconde vie aux batteries en créant des réserves d'électricité à l'échelle d'une maison ou d'une borne de recharge. A l'échelle du réseau national, la capacité de stockage d'énergie du réseau mise à disposition par V2G aide les opérateurs à mieux gérer les fluctuations de la demande. Il peut, par exemple, permettre d'absorber un pic de consommation sans dépendre d'une coupure de courant sélective, ou il peut compenser les micro-perturbations qui peuvent survenir lors du basculement de la production d'énergie d'une source à une autre. Dans le cadre de ce modèle, les opérateurs rémunèrent les clients qui mettent leurs batteries à disposition : le V2G aide ainsi le consommateur final à réduire ses dépenses énergétiques.

1.4 Le Cycle de Vie des Véhicules Électriques

Les analyses du cycle de vie (ACV) évaluent l'impact d'un produit ou d'un service sur l'environnement, généralement de sa production à son élimination [10]. Alors que le nombre de VE vendus dans le monde augmente d'année en année, dans le but de réduire les émissions de gaz à effet de serre dans le secteur des transports, il devient de plus en plus important de comprendre dans leur intégralité leurs coûts et avantages d'un point de vue environnemental. Il nous semble très important d'un point de vue déontologique et éthique de détailler ces points trop souvent oubliés ou passés sous silence.

Le cycle de vie complet des VE peut être divisé en trois approches clés : *cradle-to-gate* (du berceau à la porte), *cradle-to-grave* (du berceau à la tombe) et *cradle-to-cradle* (du berceau au berceau), comme le montre la Figure 1.8. Le *cradle-to-gate* fait référence à un cycle de vie partiel du produit, du traitement des matières premières à la production de véhicules, se terminant à la sortie de l'usine. Cette approche ne prend pas en compte l'usage des consommateurs ni la phase de recyclage en fin de vie des véhicules. Le *cradle-to-grave* est une évaluation complète du cycle de vie depuis l'extraction des matières premières, couvrant l'entretien et le ravitaillement tout au long de l'utilisation des véhicules et se terminant par le démontage et le recyclage des composants du véhicule. Enfin, le *cradle-to-cradle* est connu comme un cycle de vie en boucle fermée où les matériaux récupérés en fin de vie des véhicules sont réutilisés comme matières premières dans le cycle de vie suivant. Les approches *well-to-wheel* tiennent compte du cycle de vie du vecteur énergétique, comme les combustibles fossiles ou l'électricité, ainsi que des émissions des véhicules résultant de son utilisation. Elles peuvent être à nouveau subdivisées en *well-to-tank* qui se concentre sur la chaîne d'approvisionnement en carburant et en électricité et *tank-to-wheel* qui correspond aux émissions des véhicules provenant de l'exploitation par le consommateur [11, 12].

Les résultats de l'ACV de différentes études varient en fonction du choix de l'approche et des sources de données ainsi que des spécifications techniques telles que le type de véhicule et le poids qui ont été utilisées dans l'analyse. Différents choix de limites de système et d'hypothèses de modélisation peuvent également conduire à des résultats différents entre les études [13]. Dans la phase de production, la charge environnementale des VE a tendance à être plus élevée que celle des ICEV. Cela est dû aux produits chimiques et aux métaux utilisés lors de la production de la batterie. En par-

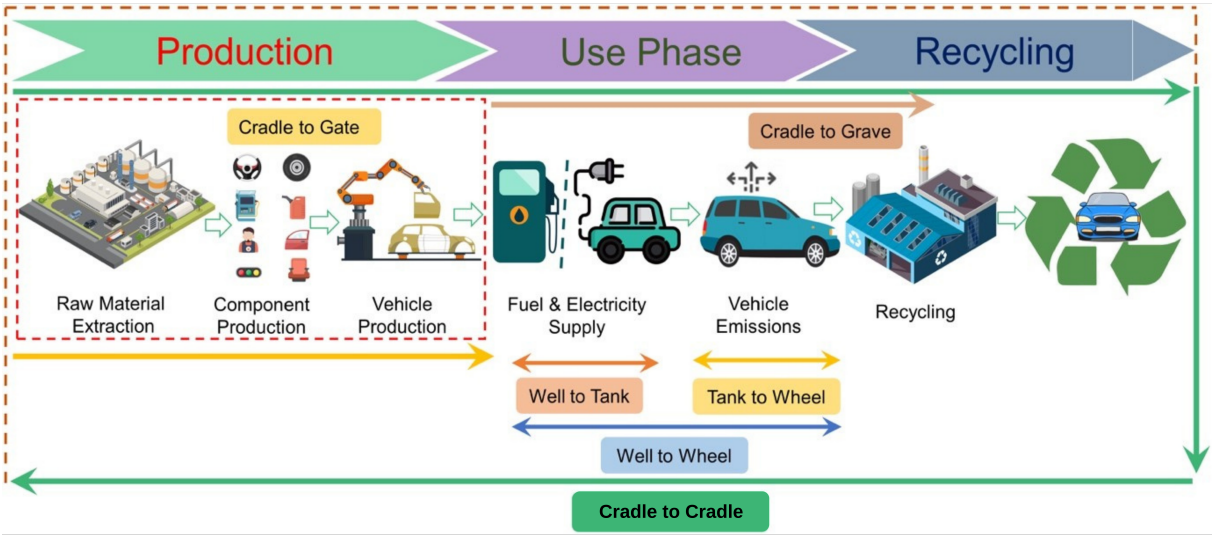


Figure 1.8 – Approches classiques pour l'analyse du cycle de vie des VE [11]

ticulier, l'impact des VE sur le potentiel d'eutrophisation (EP) et le potentiel de toxicité humaine (HTP) a tendance à être plus élevé que celui des ICEV. Dans une étude, l'impact de la toxicité humaine des VE a été calculé comme étant environ cinq fois supérieur à celui des ICEV. Cela est principalement dû aux émissions provenant de l'extraction de matières premières et de la fabrication de produits chimiques et de métaux tels que l'aluminium, le cuivre, le nickel et le platine qui sont utilisés dans la transmission.

Dans la phase d'utilisation, les VE ont tendance à avoir une performance environnementale plus élevée que les ICEV, mais cela dépend fortement du mix énergétique du réseau. L'utilisation de combustibles fossiles pour la production d'électricité peut réduire considérablement les avantages environnementaux des VE par rapport aux ICEV et dans les régions où la production d'électricité est caractérisée par une grande proportion d'énergie fossile, les VE pourraient entraîner une augmentation des émissions des gaz à effet de serre. Par conséquent, les sources d'énergie propres doivent être augmentées et promues aux côtés des véhicules électriques afin d'atteindre les objectifs d'atténuation des gaz à effet de serre [14].

La phase de recyclage offre des opportunités pour compenser la charge environnementale élevée de la phase de production. Les batteries Lithium-Ion peuvent être réutilisées grâce à la remise à neuf et à la réaffectation. La remise à neuf fait référence au processus de réparation ou de reconditionnement des cellules des batteries des EV destinés à être utilisés par les équipementiers. Par rapport aux batteries créées à partir de matériaux vierges, la remise à neuf a le potentiel de réduire la consommation d'énergie et les émissions de gaz à effet de serre [15]. La réaffectation fait référence au processus de reconfiguration des batteries pour une application différente, comme le stockage connecté au réseau, l'alimentation de secours ou les services auxiliaires. L'utilisation de batteries obsolètes pour les systèmes de stockage d'énergie peut faciliter la transition vers les énergies renouvelables en offrant une énergie propre à faible coût hors pointe. Ce type de technologie est nécessaire en raison de l'intermittence des

énergies renouvelables.

L'ACV des batteries dépend des émissions et de la toxicité associées à leur production, de la facilité avec laquelle elles peuvent être recyclées et du mix électrique dans la région où elles sont développées. Par exemple, alors que les batteries Pb-Ac et Ni-Cd peuvent être facilement recyclées avec l'infrastructure actuelle, elles sont plus toxiques que les batteries NiMH. Dans l'ensemble, les batteries Lithium-Ion ont tendance à avoir un impact environnemental plus positif que les batteries Ni-Cd, Pb-Ac et NiMH. Le traitement des matériaux cathodiques et le séchage des électrodes ont tendance à être les processus associés à une consommation d'énergie et à des émissions élevées.

L'extraction des matières premières pour la production de batteries nécessite, en outre, de grandes quantités d'eau et d'énergie. Les matières premières (lithium et cobalt) nécessaires à la fabrication des batteries Lithium-Ion sont souvent extraites par des travailleurs dans des conditions extrêmement dangereuses. Environ un tiers de l'approvisionnement mondial en lithium provient des terres salines d'Argentine et du Chili et environ 70% de l'approvisionnement en cobalt se trouve en République démocratique du Congo [16].

Une étude examinant les structures de pouvoir, le patriarcat et le travail des enfants dans l'exploitation minière artisanale et à petite échelle du cobalt en RDC a conclu que la façon dont le cobalt est extrait exploite actuellement les populations vulnérables tout en contribuant à la dégradation de l'environnement et le changement climatique. Les femmes et les filles finissent souvent par accepter des emplois dans les pires conditions pour un salaire et des conditions inférieurs à ceux des travailleurs masculins et le travail des enfants est dominant dans de nombreuses mines. Les populations locales ont des niveaux élevés de métaux toxiques dans leur corps, entraînant le développement de maladies cardiaques et respiratoires [17].

En résumé, les VE ont le potentiel de réduire les émissions de gaz à effet de serre dans le secteur des transports par rapport aux ICEV. Afin d'atteindre ce potentiel, la transition vers les VE doit aller de pair avec une transition vers des sources d'énergie non fossiles à faible émission de carbone. Cela réduira les émissions associées à la phase d'utilisation des véhicules ainsi que les émissions associées à la création de la batterie dans la phase de production. Par rapport aux ICEV, la production de VE est associée à une toxicité humaine plus élevée et à un épuisement des métaux rares. Cela concerne la phase de production et est associée à l'extraction des matières premières et à la production ultérieure des batteries. Par conséquent, afin de réduire l'impact environnemental des véhicules électriques, des processus de production de batteries hautement efficaces sont nécessaires ainsi que des matériaux éco-efficaces innovants utilisés dans la création de batteries afin de réduire la quantité de matières premières utilisées dans leur production. De plus, recycler les batteries sera essentiel pour garantir que l'impact environnemental des VE est minimal. Enfin, veiller à ce que l'extraction des éléments nécessaires à la production de batteries soit durable et équitable sera la clé de l'avenir de la production des VE. Les fabricants de batteries pourraient viser à améliorer la responsabilité de la chaîne d'approvisionnement liée à l'extraction du lithium et du cobalt afin de garantir que les structures électriques en place puissent être démantelées et que l'exploitation des mineurs soit éradiquée [17, 13].

1.5 Comportements de charge des Véhicules Électriques

1.5.1 Variables à modéliser

Les comportements de charge des VE peuvent être caractérisés de différentes manières. En particulier, il est important de comprendre les différentes variables les plus pertinentes pour la modélisation de la charge des VE ainsi que pour les applications opérationnelles (*smart charging*). Dans les sections qui suivent, nous distinguons deux types d'informations qui sont soit agrégées soit désagrégées. Par agrégé, nous faisons référence à des informations concernant un ensemble de VE tandis que par désagrégé nous faisons référence à des informations concernant un seul VE (ou plus précisément une seule session de charge).

Information agrégée

Au niveau agrégé, plusieurs informations peuvent être envisagées pour la modélisation. La grandeur la plus couramment modélisée est la courbe de charge car c'est celle qui reflète le mieux la contrainte apportée par les VE sur le réseau et elle est au centre de la plupart des défis pour les acteurs du marché de l'électricité. Une grande expertise a été acquise au cours des dernières décennies dans le domaine de la prévision de la charge électrique. Le nouveau défi qui accompagne l'adoption des VE est que ce nouveau type de demande n'est pas animé par la même dynamique ni ne répond aux mêmes contraintes que la demande d'électricité traditionnelle. Il est d'autant plus crucial d'améliorer les prévisions de pics de charge qui feront l'objet de multiples approches innovantes dans ce manuscrit. Une autre information est particulièrement utile pour les applications de *smart charging*. Il s'agit de la courbe d'occupation. Alors que la courbe de charge se concentre sur la puissance soutirée par les VE, la courbe d'occupation reflète l'occupation réelle en termes de nombre de véhicules aux points de recharge. Prévoir l'occupation des points de recharge permet de déployer des stratégies de *smart charging* et contribue également à l'introduction d'un paradigme *V2G* dans le réseau.

Information désagrégée

Au niveau désagrégé, il peut être utile de modéliser une grande variété de variables. Les plus évidentes sont les variables caractérisant une session de charge : l'heure d'arrivée, l'heure de charge, l'heure de départ et la puissance soutirée au cours du temps (ou la demande totale d'énergie) par VE. Ce sont des informations qui sont actuellement facilement collectées aux bornes de recharge. Dans ce manuscrit, nous avons exploré, en particulier, les arrivées des VE aux points de recharge, car un indicateur direct des moments où les contraintes sur le réseau sont les plus importantes. De plus, une autre information désagrégée est particulièrement pertinente mais toujours difficile à collecter : l'état de charge initial (SoC) qui est le pourcentage d'état de la batterie du VE lorsqu'elle est branchée au réseau. Il est actuellement modélisé en utilisant principalement des simulations à partir de distributions statistiques avec des paramètres donnés par des experts ou des analyses globales, mais rarement dérivés de données réelles collectées aux points de recharge. La collecte du SoC initial aiderait grandement

à améliorer les modèles de comportements de charge des véhicules électriques car il contient des informations critiques qui peuvent influencer les autres facteurs influents susmentionnés.

Agréger l'information désagrégée

Enfin, il est important de noter que les informations désagrégées peuvent être utilisées pour reconstruire les courbes de charge et d'occupation à l'aide d'un choix judicieux des variables désagrégées modélisées. En particulier, nous verrons dans le chapitre 3 comment cela peut être fait.

1.5.2 Modélisations proposées

Dans cette section, nous introduisons les différentes modélisations retenues dans ce manuscrit pour modéliser les différents facteurs influents définis dans la section 1.5

Nombre de sessions de recharge journalières

La modélisation du nombre de sessions quotidiennes de recharge de VE est un type de modèle d'occupation où nous nous concentrons sur le nombre d'arrivées de VE aux bornes de recharge tout au long de la journée. Une façon typique d'aborder ce problème consiste à utiliser la modélisation des séries chronologiques. Bien qu'il existe une grande variété de modèles de séries chronologiques, dans notre travail, nous nous sommes concentrés sur deux options. Le premier est le modèle auto-régressif (ARIMA) qui est un modèle statistique conçu pour l'analyse et la prédiction de données de séries chronologiques. Il prévoit les estimations futures sur la base des valeurs prises par la série chronologique dans le passé. D'autres versions du modèle ARIMA appelées ARIMAX peuvent également intégrer des covariables. Ceci sera détaillé plus loin dans le manuscrit. Le deuxième modèle que nous utilisons pour la modélisation des séries chronologiques est celui des réseaux de neurones récurrents. Plus particulièrement, nous nous concentrons sur les architectures Gated Recurrent Units (GRU).

Sessions de recharges au niveau désagrégé

Le fait de disposer d'une prévision du nombre quotidien de sessions de recharge de VE peut être combiné à un modèle multivarié de la distribution des sessions de recharge pour dessiner des sessions de recharge individuelles et fournir des informations désagrégées sur les comportements de recharge des VE. En utilisant un modèle de mélange multivarié des diverses informations désagrégées demandées (par exemple, heure d'arrivée, durée de charge, demande d'énergie), cela peut être réalisé. De plus, en supposant que les arrivées d'EVs sont connues (ou tout autre élément du vecteur aléatoire multivarié modélisé), conditionnellement à cette information, les autres éléments de la session de charge peuvent être dérivés avec une régression mixte (voir chapitre 3).

Arrivées des Véchicules Électriques aux points de charges

En descendant de la granularité journalière, il est possible d'envisager de prévoir les arrivées de VE à une granularité plus fine. En supposant que les arrivées EV sont la réalisation d'un processus ponctuel, le but est d'estimer sa fonction d'intensité de premier ordre λ. Plusieurs types de processus ponctuels ont été pris en compte dans notre travail (par exemple, Poisson, Hawkes), mais nous avons constaté que le processus de Poisson non homogène (NHPP) était le mieux adapté à notre application aux arrivées EV. La flexibilité des NHPP et leur simplicité nous ont permis d'utiliser différentes approches pour estimer λ. En particulier, nous avons exploré des modèles non-paramétriques de l'intensité avec des modèles additifs avec splines et/ou ondelettes (voir chapitres 3 et 4). De plus, nous avons également estimé λ avec des approches utilisant des forêts aléatoires (voir chapitre 3)

Courbes de charge et d'occupation

Enfin, la modélisation agrégée des courbes de charge et d'occupation peut se faire directement à l'aide d'outils statistiques et d'apprentissage automatique courants tels que GAM ou Random Forest. De plus, les informations désagrégées obtenues dans l'approche susmentionnée peuvent être utilisées pour reconstruire les courbes de charge et d'occupation (voir chapitre 3), nous nous intéressons donc maintenant aux approches directes. Quant à la modélisation des pics, primordiale en la matière, nous explorons également différentes approches de modélisation utilisant les GAM et les réseaux de neurones (voir chapitre 5).

1.6 Problématique et plan du Manuscrit

Comme nous l'avons rappelé dans les sections précédentes, le développement des VE est un levier majeur vers un transport bas carbone. Il s'accompagne d'un nombre croissant d'infrastructures de recharge qui peuvent être utilisées comme actifs flexibles de gestion du réseau. Pour permettre la meilleure gestion possible du réseau à court-terme, une prévision journalière efficace des comportements de charge est nécessaire. En particulier, trois enjeux sont au coeur de cette problématique :

1. Quels modèles sont à l'état de l'art pour la modélisation de la charge des VE et quelles sont les données disponibles ?
2. Comment comparer les performances des modèles de prévisions et ainsi définir le meilleur modèle ?
3. Comment prévoir les pics de consommation électrique quotidiens ?

Le but de la thèse est de donner des éléments de réponse à ces trois enjeux. Pour ce faire, le reste du manuscrit s'articule de la manière suivante (cf. Figure 1.9) :

Le Chapitre 2 répond au premier axe de la thèse avec une exploration approfondie des jeux de données ouverts trouvés au cours de nos travaux et propose une taxonomie des modèles de la charge des VE. Pour motiver ce chapitre, de nombreux articles soulignent le manque de données de recharge des VE disponibles dans la littérature. Des données ouvertes sont nécessaires pour construire des modèles reproductibles

et cohérents avec la réalité. Notre hypothèse est qu'avec une adoption croissante des véhicules électriques dans le monde, des données ouvertes peuvent être disponibles en ligne. Par conséquent, l'objectif est de chercher des données ouvertes décrivant la charge des véhicules électriques au niveau des points de charges (e.g., publics, professionnels, résidentiels). On tient également à tenir compte des données exogènes telles que le trafic, les enquêtes sur les déplacements et la qualité de l'air. Le périmètre de l'étude est centré sur 14 pays de l'*Electric Vehicle Initiative* (EVI) classés par part de marché nationale des VE.

Le Chapitre 3 est une étude comparative de différents modèles entraînés sur l'ensemble des jeux de données de sessions de recharge présentées dans le chapitre 2. Bien que la prévision de la recharge des véhicules électriques soit un domaine de recherche en plein essor, aucune étude comparative de référence n'a été proposée sur des jeux de données ouverts. Cela entrave donc la reproductibilité des méthodes définies. Nous proposons un *benchmark* d'une grande variété de modèles adaptés aux comportements de charge les plus courants. L'objectif est d'unifier les prévisions des différents modèles en une ultime prévision à l'aide d'une agrégation d'experts. Pour ce faire, nous avons retenus 3 méthodes innovantes inspirées par l'état de l'art et avons produits 14 prévisions différentes pour chacun des 8 jeux de données ouverts explorés dans le chapitre 2.

Le Chapitre 4 introduit une procédure d'estimation de l'intensité d'un processus de Poisson non-homogène avec un modèle additif d'effets splines et ondelettes pénalisées. Cette méthode est appliquée à l'estimation d'arrivées des VE aux points de charge. Après avoir donné la caractérisation du modèle, nous proposons une procédure d'estimation inspirée du *backfitting* qui est illustrée par une étude de cas sur les arrivées réelles de VE aux bornes de recharge. L'idée derrière cette approche de modélisation est d'évaluer si les pics d'arrivées aux points de charge peuvent être mieux capturés en combinant des effets splines et ondelettes. Précisément, les splines, des fonctions lisses, ont (*a priori*) pour rôle de capturer les basses fréquences du signal. Pendant ce temps, les ondelettes, plus localisées, ont pour but d'intervenir aux instants où les arrivées sont plus erratiques avec des changements drastiques dans la fonction d'intensité du processus d'arrivées. On utilise ici des données de charge résidentielle (Domestics UK) et également des données de charge publique (Palo Alto).

Le Chapitre 5 répond directement au troisième enjeu de la thèse avec l'étude d'une méthode pour prédire l'ampleur des pics de consommation électrique quotidiens ainsi que leur timing en utilisant des informations à différentes échelles temporelles. En effet, lorsque la demande électrique dépasse la capacité du réseau, cela peut entraîner des pannes du système électrique. Le marché émergeant des VE ajoute une forte contrainte au réseau qui pourra être localisée à certains instants de la journée. Avoir une prévision précise de l'ampleur des pics de consommation et du moment de ces pics permet aux fournisseurs d'énergie d'optimiser le planning de leurs moyens de productions. Notre hypothèse est que les estimations du pic de demande et de l'instant de pic peuvent être améliorées en utilisant à la fois des informations à basse résolution et à haute résolution : ce que nous appelons une approche multi-résolution. L'objectif est alors de proposer un cadre intégrant des informations à différentes échelles temporelles dans un unique modèle et d'apprécier les performances de la méthodologie proposée sur des données consolidées. Pour ce faire, nous étudions deux classes de

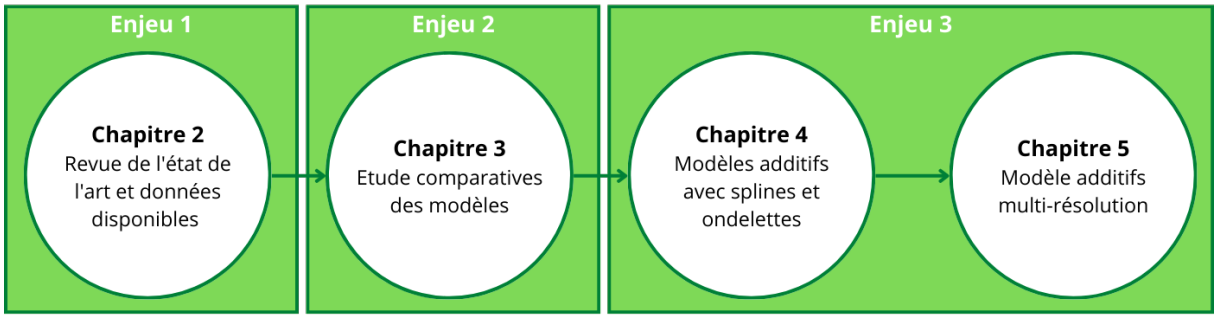


Figure 1.9 – Plan du manuscrit

modèles : les modèles additifs généralisés et les réseaux de neurones. Les expériences de ce chapitre sont menées sur les données du réseau national du Royaume-Uni dans un contexte de prévision avec fenêtre glissante.

1.7 Collaboration Industrielle

Cette thèse académique s'inscrit dans une collaboration industrielle avec EDF R&D et plus particulièrement le département OSIRIS (Optimisation, SImulation, RIsques et Statistiques pour les marchés de l'Énergie). Précisément, cette collaboration a abouti à la production de plusieurs livrables tout au long de la thèse. De plus, certains résultats de la thèse ont été présentés à différents clients internationaux de l'entreprise et directement utilisés par le groupe *Méthodes, Modèles et Outils d'Optimisation*. L'accès aux données EDF ainsi que la collaboration avec différents experts du groupe *Prévision de Consommation Court/Moyen-terme* ont contribué à enrichir les travaux présentés dans ce manuscrit.

Chapitre 2

A taxonomy of EV load models and open data cartography

This chapter is based on a paper published in MDPI - Energies [18].

The field of EV charging load modelling has been growing rapidly in the last decade. In light of the Paris Agreement, it is crucial to keep encouraging better modelling techniques for successful EV adoption. Additionally, numerous papers highlight the lack of charging station data available in order to build models that are consistent with reality. In this context, the purpose of this chapter is threefold. First, to provide the reader with an overview of the open datasets available and ready to be used in order to foster reproducible research in the field. Second, to review electric vehicle charging load models with their strengths and weaknesses. Third, to provide suggestions on matching the models reviewed to six datasets found in this research that have not previously been explored in the literature. The open data search covered more than 860 repositories and yielded around 60 datasets that are relevant for modelling electric vehicle charging load. These datasets include information on charging point locations, historical and real-time charging sessions, traffic counts, travel surveys and registered vehicles. The models reviewed range from statistical characterisation to stochastic processes and machine learning and the context of their application is assessed.

Summary

2.1 Introduction

Assuming a low-carbon energy mix, Electric Vehicles (EVs) are a credible alternative to internal combustion engine vehicles (ICEVs) supporting the transportation sector in its low-carbon transition. A substantial number of governments are heavily investing in electric mobility with more than 5.1 million electric passenger cars on the roads globally in 2018, according to the International Energy Agency (IEA) [19]. Several countries are achieving high rates of EV adoption such as Norway which approached an EV market share of almost 47% in 2019 [19]. This is due in large part to major incentives implemented by governments to foster EV uptake [20]. The EV30@30 Campaign [21] sets a target of 30% EV market share by 2030 for the member countries of the Electric Vehicle Initiative (EVI) [22]. This enthusiasm for EVs comes hand in hand with great concern about how to manage the surge in electricity demand which could greatly disrupt the current schedule [23].

In order to overcome potential pitfalls, businesses and researchers are proposing solutions including pricing strategies [24] and smart charging [25]. The goal of these solutions is to avoid dramatically shifting EV users' behaviours and power plants production schedules. However, their implementation requires a precise understanding of charging behaviours. Thus, EV load models are necessary in order to better understand the impacts of EVs on the grid. With this information, the merit of EV charging strategies can be realistically assessed.

In this article, the term "EVs" refers to small vehicles (e.g., light motorcycles), passenger vehicles (e.g., cars) and goods-carrying vehicles (e.g., trucks) as per the classification from the European Commissions' official report "Mobility & Transport : Vehicle Categories" [26, 27]. Passenger vehicles constitute the majority of EVs. Additionally, all energy system management that can be plugged to the grid are considered : BEV, EREV or PHEV [28]. Furthermore, Electric Vehicle Supply Equipment (EVSE) will be referred to as any type of charging point, be it public or private. Finally, an EV charging session (or transaction) refers to the period of time an EV has spent charging at an EVSE.

2.1.1 Aims and strategies for EV charging schemes

Electricity distribution occurs such that at any point in time and space, the consumption has to be equal to the production in order to avoid severe consequences such as blackouts [29]. A significant rise in the number of EVs in circulation leads to an increase in electricity demand which could cause such a blackout if the balance in the grid is not effectively maintained. Therefore, EVs have an important role to play in maintaining this balance [23]. The purpose of this section is to explore the different aims and strategies required to overcome the potential difficulties caused by increased EV penetration. Figure 2.1 summarises these aims and strategies.

Load flattening

While some studies show minimal impact of EVs on peak load [30, 31], the consensus in the field is that the grid will not be able to sustain its operations with the projected demand from EVs [24, 27, 32, 33, 34, 35, 36, 37].

One of the first articles dealing with the impact of EVs on load management was published in 1983 [38]. In this article, EVs were suggested as a way to minimise the overall grid load factor f. This factor is defined as the ratio of the average load (L) over the maximum load in a given period of time : $f = avg(L)/max(L)$. The maximisation of this quantity results in a more efficient distribution of resources over time. The article proposed that using off-peak recharging of EVs will significantly increase the load factor. This means shifting the EV demand to times when the rest of the demand is low (e.g., night time) in order to flatten the load curve. The flexibility analysis produced in [39] suggests that it is possible to shift the EV charging to the afternoon and night valleys for different clusters of users without changing their behaviours. This could lead to peak reduction and load factor maximisation with little change to users' requirements and lifestyles.

Articles such as [40] strived to estimate the benefits of this kind of controlled or incentivised EV charging. However, these articles do not always account for potential mistakes in load forecasting, therefore the benefits calculated could be inaccurate. Hence, it is critical to improve EV load forecasting models in order to alleviate the risk of unrealistic optimisation schedules for maximising the load factor.

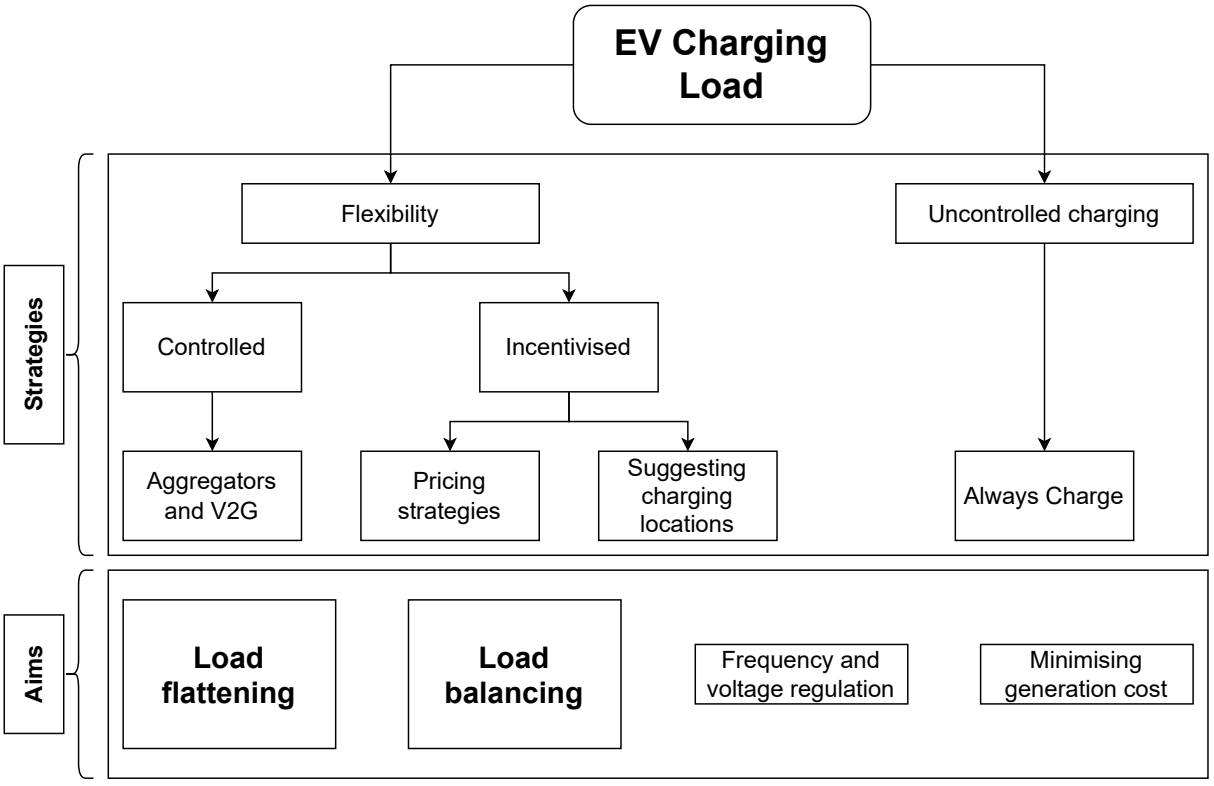


Figure 2.1 – The variety of strategies envisaged for handling EV load and their associated aims. Incentivised flexibility and controlled flexibility are used to achieve specific aims while uncontrolled charging lets the market decide the prioritisation of these aims. Load flattening and load balancing are the most common aims found in the literature and they are the focus of section 2.1.1.

Load balancing

An early article from 1997 [41] considered using EVs as a source of electricity for the grid when demand is high. In other words, using EVs plugged-in to the grid as an ancillary service or as way to bring flexibility to the overall shape of the load. According to a study focused on 400,000 EV charging transactions from 2012 to 2016 in the Netherlands, 75% of EVs connected at public EVSEs are already fully charged [42]. This study therefore supports the strategy of using fully recharged EVs which are still connected, as a source of energy in order to supply the grid. This paradigm shift, using what could be a major constraint and treating it as an opportunity, is called "Vehicle-to-grid" (V2G).

Additionally, integrating renewable energy sources onto the grid is also the focus of numerous studies [25]. Many countries with climate related commitments are aiming to increase the share of renewables in their energy mix. However, the main drawback of renewable energies is their intermittent delivery of supply. Indeed, solar panels and wind farms are highly weather-dependent. In this context, EVs can adequately balance the energy coming from renewable power plants. This strategy consists in considering multiple EVs acting as a large battery or electricity storage system which can be discharged back into the grid when weather conditions do not allow renewable power plants to produce enough energy [37].

Although V2G has many advantages, one drawback is that it reduces battery lifetime by adding unnecessary cycles of charge and discharge to the vehicle [43]. Furthermore, this strategy requires the existence of global and local communication and monitoring channels which do not exist yet. These channels are necessary for the development of EVs in general and particularly for V2G and load balancing [44, 45]. Finally, in order to ensure effective communication, EV load models are critical as they can reduce uncertainty and minimise contradicting signals from what is expected and what is observed by operations management.

2.1.2 Chapter structure and contributions

The purpose of this chapter is to enable a better understanding of EV load data available and models produced in the literature. The main contributions of this chapter are as follows :

- The results of an in-depth open data search with a structured list of datasets available for use
- A comprehensive review of EV load models including their strengths, weaknesses and their application in the literature
- A preliminary study on matching EV load models to six open datasets found in this research and not previously explored in the literature

The rest of the article is structured as follows. Section 2 defines EV load and its most common drivers. Section 3 presents the open data found which can be used to model EV load. Section 4 reviews EV load models comparing the different approaches taken. Section 5 explores charging session data not previously explored in the literature and suggestions are provided on the models reviewed that could be applied to these datasets. Finally, Section 6 highlights the current knowledge gaps and discusses the different options in order to pave the way for future work.

2.2 EV load and its main drivers in the literature

EV load corresponds to the power or energy consumed at EVSEs over time. This information can also be directly derived from other closely related factors. In particular, knowing the arrival time and charge duration of EVs allows a deterministic reconstruction of EV load.

2.2.1 EV load as a model output

EV load can be considered at different levels of aggregation. The total energy demand at all EVSEs can be referred to as the aggregated output of EV load models. The same model output can be envisaged in a disaggregated fashion. Two setups are widely used in practice. The first is vehicle-centric which considers the contribution of each EV member of a fleet to the aggregated load. The second is EVSE-centric which considers the perspective of one or multiple EVSEs. Neither approach is mutually exclusive and the two setups can be combined to model EV load.

Aggregated

The aggregated approach is shown in various articles such as [42] and [46] where the total EV load across multiple EVSEs is modelled. In [42] 1,750 charging stations (2,900 charging points) are used while [46] uses a single station with many charging piles. This kind of approach usually performs well due to the smoothness of the aggregated load curve assuming there are enough EVs or charging stations in scope. While they give a holistic view of the charging load, they can lack detail with regards to the temporal and spatial distribution of the load which is one of the key concerns raised in the literature [23].

Vehicle-centric

In order to explore the finer details of EV load, a vehicle-centric approach can be adopted. In [47] individual EV loads are modelled in order to recover the aggregated load. This approach can be qualified as a vehicle-centric approach as it uses individual outputs of EVs. In this case, it is assessed in terms of aggregated load. The same can be said for [48] and [49] where individual behaviours are modelled. A similar study can be found in [50] where the EV load outputs are separated into urban and rural behaviours while [31] looks at public and residential charging. This can give a better understanding of the spatial and temporal properties of EV load.

Additionally, models which consider the spatial components of EV charge are detailed in [27] and [36]. For instance, in [36] four schedules for EVs are identified which enables one to better distinguish and evaluate their temporal impact on the grid. Furthermore, the spatial dimension is addressed by modelling EV charging locations. Both outputs are brought together in order to reconstruct the aggregated EV load.

EVSE-centric

The EVSE-centric approach is rare in the literature as it usually is superseded by the vehicle-centric approach. However, there are some occurrences of such work for instance in [30] where residential charge is envisaged from each household perspective. The authors used a bottom-up approach to forecast the aggregated EV load using each household individual load. The debate of using EVSE-centric over vehicle-centric approaches is illustrated in [51]. In this article, it was found that both approaches yield comparable prediction errors even though the EVSE-centric approach was slower to compute.

2.2.2 Input data used for EV load modelling

Battery

Battery inputs are variables which closely relate to the charging demand of EV load from a "physical/chemical" perspective. The most common ones used across the literature are the State of Charge (SoC), Energy Consumption (E) and battery capacity (C). Generally speaking, the SoC is the rate at which the battery is charged whether the EV is plugged-in, idle or travelling [52]. The SoC when the EV arrives at an EVSE is a critical influential factor of EV demand. This is referred to as the initial SoC (SoC_{init}) in the literature. By incorporating the distance travelled (D) by the EV, the following formula can be written :

$$SoC_{init} = \frac{C - E \times D}{C} \tag{2.1}$$

with C in kWh, E in kWh/km, D in km and SoC_{init} in %.

On one hand, battery capacity and other engine specifications are usually assumed to be known constants. Based on EU MERGE data, probability functions were derived to characterise EV specifications in [27]. On the other hand, the SoC and energy consumption evolve over time and with vehicle usage. Both are highly correlated and they can be deduced from each other from the formula above or by a set of assumptions. For instance, the initial SoC of EVs is assumed to be equal to 0%, 30%, 60% in [53] to match different scenarios. Similarly, the initial SoC is used as an input in [54] along with D, C and the charging rate of the EV charging model in [55]. Furthermore, in [48] and [56] battery specifications and stochastic characteristics are also part of model inputs. Finally, the EV load itself can be used as an input when considering time series approaches [51].

Travel

From this literature review, it appears that travel behaviours are the most widely used exogenous factor driving EV load models. It is important to distinguish between travel inputs extracted from travel surveys [32, 57] or estimated pattern data [48, 58, 27] (which usually require further statistical treatment to be part of a model), and real-world traffic patterns (which are deduced either from pilot experiments [59] or direct GPS driving data [60]).

The input variables used in most papers (whether they are estimated or recorded) are the daily distance travelled and travel time. In [60] daily travel distance and individual trip distances distributions were extracted from a survey conducted between 2012 and 2013 in Beijing with real-world GPS data collected on 112 volunteer vehicle owners. Likewise, a pilot experiment was put in place for a week in Germany [59] in order to record the evolution of daily trips through GPS data.

When such exact data is not available, researchers use travel surveys instead [61]. These datasets hold valuable general information on drivers behaviours and can be used in order to estimate parameters of statistical distributions for daily distance travelled or travel time. However, they can lack accuracy as the information is usually collected through questionnaires. For instance, in [57] the authors used the 2009 National Household Travel Survey (NHTS) as well as the New-York State Transportation Federation Traffic Data Viewer in order to extract traffic statistics such as EV speed travelling from one charging station to another. In [32] daily trips from a single real-world vehicle from the NHTS is randomly assigned to a fictional EV used in the model. This procedure is applied to the desired number of EVs to obtain a fictional EV traffic. Similarly, [48] used Barcelona's mobility patterns while [62] used the 2008 transportation data from the Dutch Ministry of Transportation in order to extract traffic statistics.

Weather

EV load models have stemmed from electrical load models. They have been developed over 100 plus years [63] and are comprised of some strongly established characteristics. One such characteristic is the thermosensitivity of electrical load [64]. In short, this means that some obvious patterns can be derived from analysing both load demand and temperature. Thus, it is natural that the most frequently used input for EV load models is temperature and its traditionally associated statistics (e.g., average, maximum, minimum) [30]. Even though temperature is used in most electrical load models it is rarely used in EV load models. Nevertheless, there exists reasonable arguments to include weather data in EV load models.

[65] explores the influence of different weather variables on daily EV charging demand. This includes, minimum, maximum and mean daily temperature as well as mean wind speed, maximum gust, rainfall, global radiation and sunny hours. The results of this study showed that temperature and specifically mean air temperature is the most correlated weather input to daily EV load relative to the others reaching a 27% correlation relationship in one of the regions considered.

Similarly, in [24] the authors argue that temperature can be used to model EV load as it is correlated to electricity prices and demand. However, there is no mention of other potential weather factors which could be included.

A relational analysis is used in [66] to assess the impact of weather factors on traffic volume in South Korea. It was found in this case study that maximum and average temperature as well as average humidity are the most influential weather factors on traffic volume. Average wind speed on the other hand is less influential and was discarded in their model.

Finally, it is also argued in [67] that temperature has a great impact on EV charging station load while wind and humidity were discarded.

Economy

Amongst the articles covered only a few include economical factors such as electricity prices [46, 24, 57], Gross Domestic Product (GDP) [48] or trends [65, 58]. While some locations still provide free charging as an incentive to foster EV adoption, most public EVSEs have a charging price based off a subscription or peak/off-peak tariffs. However, China is one of the countries where real-time electricity pricing affects the price consumers pay at EVSEs. Thus, for [46] it is natural to include time of use tariffs as this study was made on EVs in China. In [57], the authors also include electricity prices as it can have an impact on the decision making undertaken by an EV driver when choosing which station to charge their vehicle.

Interestingly, GDP is included as a model input in [48] as it was shown in previous work [68, 69, 70] that GDP and other socio-economical variables such as place of residence and household characteristics have an impact on EV load and can be leveraged using an vehicle-centric approach. This is something worth exploring as these variables are easily accessible in travel surveys and general country statistics. They can be used to better anticipate charging behaviours in various locations of the grid. Global EV trend usage with uptake scenarios [58] or calculated trends [65] can also be used as model inputs.

Calendar

Temporal inputs are used in most model set-ups. They are easy to integrate and bring consistency as well as performance with the strong explanatory power they hold. They require no heavy statistical treatment as opposed to other variables (e.g., travel and battery) which makes them easy to use. For instance, in [49] and [26] day of the week and time of day are used in EV load models and more generally, EV load is derived in most research papers from day of the week, time of day and seasonal variation.

2.3 Open Data Search

Few review articles that deal with related topics to EV load modelling have included information regarding open data with associated references [71]. To the best of the authors' knowledge, there exists no article at the time of writing which has attempted this type of endeavour for EV load models. Indeed, a great majority of articles produced in the EV load modelling domain are based off simulated data or information owned by private entities which are very rarely made available [46, 67]. This prevents reproducible work and slows down research in the field.

Therefore, the objective is to fill this gap by providing the community with a structured and carefully selected list of open datasets ready to be used in order to foster data-driven research in the field. This open data search was possible in great part thanks to the open data inception initiative which gathers links to more than 3,500 open data repositories on their website all across the world [72] [1].

1. Links to the datasets are provided throughout this section and are up to date at the time of writing

2.3.1 Research Criteria

This study focuses on datasets which give information on transactions between EVs and EVSEs. In other words, charging sessions.

Additionally, datasets holding information on exogenous variables such as traffic, travel surveys and air quality have also been considered. These variables are widely used in the domain in order to simulate travel behaviours especially when considering spatiotemporal models. Weather data is also used for EV load modelling and electrical load modelling in general [64]. However this type of information was excluded from this data research as global resources which provide high quality weather data already exist. For example, the *riem* package [73] written in R retrieves data from airport weather stations all over the world via the Iowa Environment Mesonet website. Alternatively, the National Oceanic and Atmospheric Administration (NOAA) also provides extensive weather data [74].

In terms of the perimeter of this research, the top 14 countries active in the EVI during the period covering 2018 to 2019 have been targeted. They are ranked by market share of electric cars according to the IEA [19]. This list includes, Norway, Iceland [2], Sweden, Netherlands, Finland, China, Portugal [3], USA [4], Canada, France, New Zealand, United Kingdom, Germany and Japan [76].

Most of the repositories covered used native language, therefore, the use of direct query search was minimised as it can be approximate, especially in a foreign language. Thus the following standardised process was used for each repository covered : every time a categorical hierarchy was available, datasets under the following categories were searched for : “Environment”, “Natural Resources”, “Infrastructure”, “Transportation”, “Traffic”, “Climate & Weather”, “Urban Development”, “Planning”. If a category search was not enabled, then the following key words were used with their translated variants : “Travel (Survey)”, “Electric Vehicle (or Car)”, “Charge-Charging”, “Traffic”, “Station”, “Air Quality”, “Mobility”.

2.3.2 Open datasets

Overall, more than 860 repositories have been explored and more than 60 relevant datasets have been found that are directly (endogenous) or indirectly (exogenous) useful for modelling EV load. Table 2.1 summarises the results found across all countries covered with the most relevant datasets in each category. Regarding EVSE data, a distinction is made between real-time and historical charging session data. Historical data gives information on charging sessions which occurred in the past. This is the essential type of data sought to model EV load. Real-time data refers to EVSE occupation information which is updated on short time frames (every few minutes) and not stored. It requires regular scraping to be transformed into a historical charging session dataset and only then can it be leveraged for EV load modelling.

2. Iceland is not officially part of the EVI but has the same ICEV ban by 2030 target than other countries in the EVI and is often mentioned alongside them in IEA reports and charts [75]. The magnitude of its EV market share makes it highly relevant to this analysis.

3. As an observer

4. Participation was being assessed at the time and at time of writing the USA are no longer part of the EVI

Countries	EVSE data			Exogenous data		
	Location	Charging Sessions		Traffic Counts	Travel Survey	Registered Vehicles
		Historical	Real-time			
Norway	[77]		[77]	[78] [79]		[80]
Iceland				[81]		[82]
Sweden	[83]		[77]			[84]
Netherlands	[85] [86]	[87]	[85]	[88]	[89]	[90]
Finland	[77]		[77]	[91]		[92]
China	[93]			[94] [95]	[96]	
Portugal	[97]					[98]
USA	[99]	[100] [101] [102] [103]		[104]	[61] [105]	[106]
Canada	[107]	[108]		[109]	[110]	[111]
France	[112]	[113] [114]	[115]	[116] [117]	[118]	[119]
New Zealand				[120]		[121]
UK	[122] [123]	[124] [125]		[126] [127]	[128]	[129]
Germany	[130]		[131]	[132] [133]	[134] [135]	[136]
Japan					[137]	[138]

Table 2.1 – The most relevant open dataset available found in this research with the associated references.

For each country the corresponding EV market share from the IEA [19] is provided as well as the estimated value of the number of EVs to which this market corresponds. The estimated number of EVs sold is calculated by using the number of passenger sales in 2019 given on [139] multiplied by the EV market share from the IEA [19]. In Figure 2.2, the national EV market share and estimated number of EVs sold are shown, coloured by the type of data available for each country. It is interesting to note that countries with the highest market share and number of EVs sold are not the ones for which historical charging session data was found. First of all, countries for which historical charging session data was found will be discussed as it is the most relevant and rarest information to find. Then, the information available from countries without historical charging session data but with real-time charging session data will be outlined. Finally, the countries where only traffic information is available will be presented.

Countries with historical charging session data

Netherlands 6.6% *national EV market share [19] equating to approximately* $29,000$ *EVs sold in 2019 [139]*

23 repositories were covered in the Netherlands with every type of relevant data found. First of all, ElaadNL [87] holds historical charging sessions which were studied in multiple papers [39, 42]. With regards to traffic data, *Onderweg in Netherlands* is the national travel survey published on a yearly basis [89]. While its tables are quite hard to

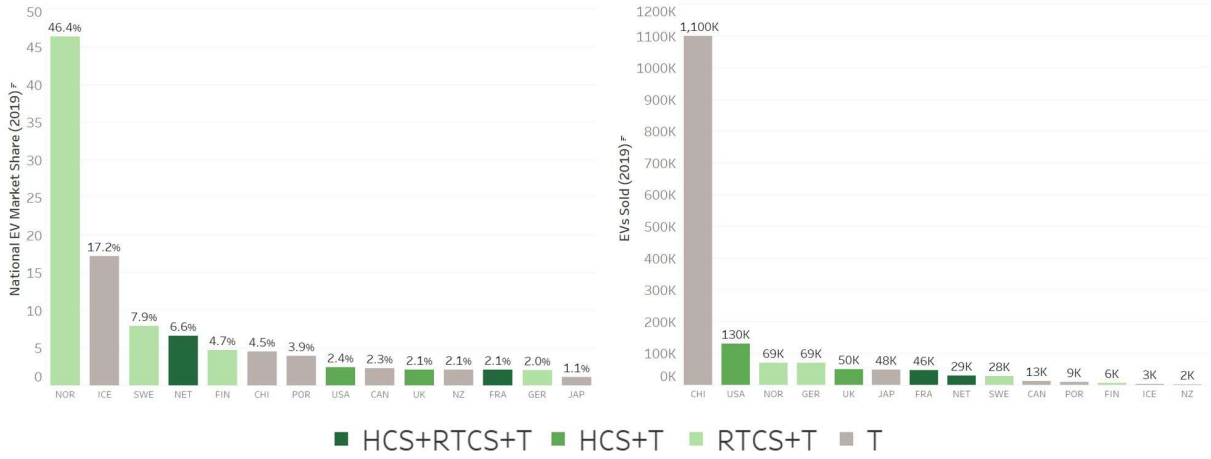


Figure 2.2 – For each of the 14 countries in scope, the national EV market share [19] and the estimated number of EVs sold [139] is shown. HCS refers to historical charging session data, RTCS refers to real-time charging session data and T refers to traffic counts and/or travel survey data. This demonstrates the existing gap between EV penetration in each country and the availability of open charging session data.

study as-is for non-native speakers, they are summarised in another website in English [88]. Real-time data on utilisation and consumption at public EVSEs installed in Rotterdam can be found on the EV-BOX website which is one of the EVSE providers [85]. Registered vehicles [90] and public EVSE locations are also available (e.g., in Eindhoven [86]). Additionally, historical traffic data from 2010 extracted from 24,000 measure points which stores information on vehicles such as speed and travel time [140] was found.

USA 2.4% *national EV market share [19] equating to approximately* $130,000$ *EVs sold in 2019 [139]*

The open data search for the USA was extensive. Around 370 repositories were covered in the analysis. Among them three relevant charging session datasets were found [100, 101, 102]. The first provides a continuous dump of session data from 2018 on EV sessions recorded at city-owned EVSEs in Boulder (Colorado) [100]. The second gives the same information for charging sessions of EVs in the city of Palo Alto (California) from 2011 to 2017 [101]. Finally, the third provides us with an aggregated monthly view of transactions in the city of Evanston (Illinois) between 2016 and 2017 [102]. Furthermore, a charging session open dataset from Caltech, which is continuously updated in collaboration with Power Flex, is available at [103] and an exploration of this dataset was produced in [141]. On top of these charging session datasets, EVSE locations are also available from the Alternative Fuels Data Center [99], as well as many travel surveys including the National Household Travel Survey (NHTS) [61], which are frequently used to simulate EV behaviours from conventional vehicles. In particular, a mobility survey was performed in April 2019 for the City of Boulder on 203 residents. Information extracted from [105] brought together with the EV charging session dataset of the city of Boulder [100] could lead to more consistent and accurate representation of EV load than by using the more general NHTS. Finally, a large proportion of states share traffic volumes

in various municipalities across the country (e.g., the city of Houston [104]).

France 2.1% *national EV market share [19] equating to approximately* $46,000$ *EVs sold in 2019 [139]*

France (mainland) also has a large number of open data repositories. In total, 151 repositories were explored. Among them all kinds of relevant data were found. Firstly, charging sessions were recorded from April to May 2017 on Belib' stations in Paris [113]. Furthermore, the Paris Data website provides the Belib' real time availability of public EVSEs in Paris [115] which can be scraped on a regular basis via an API in order to reconstruct a historical dataset. Regarding private EVSEs, the charging sessions of a fleet of EVs owned by SAP Labs France have been recorded from June 2017 [114]. This dataset is updated every three months. On top of charging session data, registered vehicles across the territory [119], traffic counts in numerous cities [116], real-time traffic [117], and a national travel survey [118] are available in order to perform a spatiotemporal analysis of EV load. Different road traffic open data repositories are gathered on the Cerema website [142].

United Kingdom 2.1% *national EV market share [19] equating to approximately* $50,000$ *EVs sold in 2019 [139]*

72 repositories were covered for the UK mainland which yielded multiple charging session datasets. Two of them are situated in Scotland : Dundee City [124] and Perth & Kinross City Councils [125]. The former gathers two years of charging session data from 2017 to 2018 while the latter covers four years from 2016 to 2019 to the granularity of each session. Additionally, the UK government led an EVSE analysis over the year 2017, with domestics [143], and public [144], [145] chargers. The raw datasets available include charging session data for each type of EVSE. There were also some initial trials led by the UK power networks in 2013-2014 which can yield useful information [146]. Public EVSE locations are available in numerous municipalities of the UK [123] as well as a national charging point registry [122] with real-time [126] and historical traffic counts [127]. Moreover, yearly national surveys are also available [128].

Countries with real-time but no historical charging session data

Norway 46.4% *national EV market share [19] equating to approximately* $69,000$ *EVs sold in 2019 [139]*

Norway is by far the country which has the highest penetration rate of EVs to date. Thus, it is no surprise that some highly relevant data for EV load modelling was found regardless of a relatively small number of repositories available (13). Norway was an early-mover in fostering EV adoption. In 2009, the first large investments were made by cities and the government with Oslo being the major contributor [20]. The most relevant data feed comes from the NOBIL database API [77]. This service provides (after benefiting of an API key from NOBIL) real-time information on EVSEs all across Norway, Sweden, Finland and Denmark (e.g., location, usage, details). Historical dumps do not seem to be available through the API, however a regular scraping may be put in place in agreement with NOBIL in order to reconstruct historical data. Other data sources which describes exogenous variables are available such as traffic volumes [78], [79]

and vehicle registrations by fuel types [80] which gives an overview of the trend in EV adoption.

Sweden 7.9% *national EV market share [19] equating to approximately* $28,000$ *EVs sold in 2019 [139]*

Sweden, with 18 repositories covered, also benefits from the NOBIL API which gathers real-time information on public EVSEs activity across the territory [77]. Some of NOBIL's information is gathered on an external Swedish website which provides historical statistics on EV public charge use [147]. On top of that data source, the map of public EVSEs [83] and statistics on newly registered vehicles per county, town and fuel type on a monthly basis are also available [84]. This latter dataset can be used in load forecasting models as a variable explaining the trend in EV usage particularly thanks to its monthly granularity.

Finland 4.7% *national EV market share [19] equating to approximately* $5,700$ *EVs sold in 2019 [139]*

Finland is also one of the countries which has adopted the NOBIL database API [77]. Amongst the 20 repositories covered, exogenous information with traffic in real time in a few municipalities (e.g., the city of Tampere was found [148]) as well as registered vehicles between 1922 and 2019 [92] and average distance travelled by vehicles between 1980 and 2015 [91]. Even though these sources, provide us with extensive historical data, the most recent years are the most relevant for EV load models. These datasets can give an overall understanding of the overall traffic trends in Finland.

Germany 2% *national EV market share [19] equating to approximately* $69,000$ *EVs sold in 2019 [139]*

With regards to Germany, the most relevant datasets found among the 52 repositories covered were real-time public EVSE usage [131] and real-time traffic data [132] in the city of Bonn. Scraping both sources and associating these can lead to precise EV load models. In addition, travel surveys at fine levels of details are available from the German Mobility Panel [135] as well as the Rheinisch-Westfälisches Institute (RWI) [134]. The RWI dataset was used for a study on mobility patterns in [149]. Furthermore, the number of vehicles registered [136] and traffic counts in several municipalities [133] can give an understanding of the trend in EV usage across the country. Finally, as for most other countries, public EVSE locations are also available [130].

Countries with traffic data and no charging session data

Iceland 17.2% *national EV market share [19] equating to approximately* $3,100$ *EVs sold in 2019 [139]*

As for Iceland, 4 repositories were covered and the most relevant datasets found do not include any charging sessions but descriptive statistics on transports in Reykjavik [81] as well as vehicles distance and fuel consumption between 1995 and 2019 [82]. This can enable an understanding of the trends in EV adoption and high-level travel behaviours. However, limited analysis can be conducted as real charging session data

is unavailable and would have to be simulated from other markets. Additionally, no real-world traffic data or travel survey was found which also limits spatial studies.

China 4.5% *national EV market share [19] equating to approximately* $1,100,000$ *EVs sold in 2019 [139]*

Being the country with the largest volume of EVs, China is at the forefront of EV deployment worldwide. However, this research did not result in finding any charging session data for China. One explanation for this is that more than 90% of EVSEs are owned by private firms [150]. Most of the articles which use data from charging stations on Chinese territory do not make it available as it is usually part of an agreement between the researcher and the entity owning the data. Nevertheless some relevant traffic data [94], [95] for the whole territory was found and travel surveys [96] as well as EVSE locations [93] specifically in Hong-Kong.

Portugal 3.9% *national EV market share [19] equating to approximately* $8,900$ *EVs sold in 2019 [139]*

Regarding Portugal's open data, traffic statistics with the number of vehicles registered by type and fuel was found [98]. Additionally, EVSE locations in Lisbon were also available [97]. No charging session data was found.

Canada 2.3% *national EV market share [19] equating to approximately* $13,000$ *EVs sold in 2019 [139]*

Being the co-lead of the EVI activities along with China [19], Canada is a major player in the field of EV deployment. Around 76 repositories were explored with numerous travel surveys which describe various aspects of drivers' behaviours [110]. Traffic volumes [109] and EV registrations [111] are also available with details on EVSEs available for public use in some municipalities (e.g., the city of Edmonton [107]). Even though no historical nor real-time charging session data was found, there exists an EV Home Charging Program [108] which gathers residential charging session data. However, this dataset is not open at the time of writing but might be accessed with the relevant access grants.

New Zealand 2.1% *national EV market share [19] equating to approximately* $2,300$ *EVs sold in 2019 [139]*

For New Zealand, 22 repositories were covered with successful findings in traffic statistics and vehicle registrations. Several locations in New Plymouth record traffic count [120] and the number of vehicles registered by type across the country is also available [121]. This data as-is is difficult to exploit for EV load modelling as it lacks EVSE locations and charging sessions.

Japan 1.1% *national EV market share [19] equating to approximately* $48,000$ *EVs sold in 2019 [139]*

Finally, with Japan, 14 repositories which did not contain any charging session or station location information were covered. Nevertheless, exogenous data can be extracted with numerous travel surveys [137] and some statistics on registered vehicles [138].

2.4 EV load models

The scope of this review focuses on papers detailing an EV load model as defined in Section 2. Most often, the model output is the power or energy demand at EVSEs but it can also be closely related features (e.g., EVs arrival/departure times, charging durations) from which the load can be reconstructed. In particular, the focus was given on presenting a wide variety of methods to encompass multiple modelling settings. The purpose of this section is to enable an understanding of the strengths and weaknesses of the methodologies proposed to model EV load. From the papers considered for this review, EV load models can be segmented into three categories : statistical characterisation, stochastic processes and machine learning models. The comprehensive list of models considered in this review is presented in Appendix A.

2.4.1 Statistical Characterisation

The goal of statistical characterisation models is to produce a distributional analysis for the outputs shall it be data-driven [42] or entirely deduced from exogenous variables such as travel data and statistical assumptions [32]. The different characterisations of EV load and proxy variables such as charging duration or inter-arrival time are summarised in Table 2.2.

Model	Strengths	Weaknesses	Ref
Gaussian	Particularly suited for large simulations	Unrealistic as negative values have a non-zero probability	[32]
Weibull, Lognormal, Exponential	Rapid to implement while providing an approximation consistent with reality	Fail to capture significantly diverse behaviours in the data	[31]
Mixtures (e.g., Beta, Gaussian)	Captures significantly different users' behaviours in the data and respects real-world constraints	Unsuitable for medium or large dimension problems with numerous covariates	[42] [151] [141] [152]
KDE	Highly versatile model as no explicit prior on the distribution is required	Weak interpretability power in addition to a sensitivity to outliers	[153] [154] [155] [156] [157]

Table 2.2 – Statistical characterisation models for EV load

In [32], the authors did not benefit from any EVSE data. Nevertheless, they used the NHTS [61] ICEV behaviours from 2009 to derive EV travel patterns in order to simulate an EV fleet and characterise their behaviours. In their work, the simulation showed that the power consumption can be seen as a normal distribution without any loss of accuracy.

This can be true in practice, however, it is usually more consistent to assign distributions which are defined on $\mathbb{R}^+$ as it is unrealistic to observe negative power demand in that context. It is however convenient for model conciseness and computational speed.

In [31] a statistical analysis is conducted on data extracted from an EV trial conducted in Victoria (Australia) on 33 EVs on a 3-month period. This article showed that the Weibull distribution was the best fit for charging duration compared to the exponential and lognormal laws. They have also characterised the time to the next charging event as a mixture of two lognormal distributions. This is a vehicle-centric approach which considers the time to next charge from the EV perspective. These characterisations were used on a Monte-Carlo simulation which created 4,000 EVs by random sampling and assessed their overall impact on the grid. While these distributions are more consistent than a Gaussian distribution, they still fail to capture the irregularity of EV drivers' behaviours hidden in the data.

In [42], a dataset provided by Elaad NL [87] has been studied. This paper characterises EV load through a mixture of beta distributions. Its parameters are optimised by minimising the Root Mean Squared Error (RMSE) of the point-wise difference with the empirical distribution. Additionally, Kolmogorov-Smirnov testing was used to assess the goodness-of-fit. From the observations that weekly charging sessions present two peaks (namely a morning and a late afternoon peak) it was reasonable to consider a mixture of distributions to account for the different modes. In [151], 13 different charging session profiles were identified using Gaussian mixture clustering based on data provided by the G4 cities of the Netherlands. Other recent studies complement this work by using Gaussian mixtures to model the triplet (Arrival Time, Charging Duration, Energy Consumed) in order to characterise EV load. In [141] the triplet is modelled by a multivariate Gaussian mixture while in [152] only the couple (Charging Duration, Energy Consumed) is modelled by a Gaussian mixture with the Arrival Time modelled by an exponential distribution. The results produced are more accurate than for elementary distributions. However, they are structurally limited to the joint use of few covariates which keeps from fully integrating exogenous information.

A few articles also modelled EV load with a kernel density estimator (KDE). Two main types have been used in the literature : the Gaussian kernel density estimator (GKDE) and the diffusion kernel density estimator (DKDE). These methods are highly versatile because no prior knowledge over the distribution is hypothesised. Thus, they can reach high accuracy when fitting empirical data at the cost of weak interpretability. Looking at [153], a GKDE is used to estimate daily trip distance and end time of the last trip. Both variables are critical for EV charging schedules and this method improves the accuracy of the distributions compared to parametric methods. A similar conclusion is drawn in [154] from a GKDE estimating the triplet (Arrival time, Charging duration, Charging capacity). In [155] and [156] the authors have compared both the GKDE and DKDE when estimating EV load. Thanks to its optimal bandwidth selection process, DKDE was found to produce better load estimations. Finally, in order to make the best of both GKDE (which is less sensitive to outliers) and DKDE (which has a higher overall accuracy), [157] has proposed a hybrid density estimator (HKDE). This HKDE reached significantly better root-mean square performance in estimating the EV load than the DKDE and GKDE on their own on the dataset used for this study.

2.4.2 Stochastic Processes

In the context of EV load models, three main types of stochastic processes have been detailed in the literature : purely temporal, spatiotemporal and queuing theory viewpoints. The various stochastic processes presented are summarised in Table 2.3.

Table 2.3 – Stochastic Processes for EV load

Model	Strengths	Weaknesses	Ref
Temporal	Adequate for modelling one EVSE or one EV	Generally assumes independence between EVSEs	[33] [158] [55] [159] [62]
Spatiotemporal	Suited for modelling clusters of charging stations simultaneously	Large increase in complexity with scale	[27] [36] [50]
Queues	Easily scalable with strong theoretical grounds	Restricts reality with simplifying assumptions	[47] [160] [34] [57]

Temporal

One of the early works on EV temporal load models was completed in [33] where the authors explored the stochastic nature of EV load by using probabilistic travel patterns to determine initial SoC and starting time of battery charge. In particular, assuming battery type is known, recharge starting time is then assumed to be a random variable with a probability density function (pdf) determined by the tariff structure (scenarios) and patterns of EV usage. Initial SoC is also considered as a random variable dependent on the total distance travelled since last charge. Introducing a lognormal pdf for the daily distance driven, the initial SoC can be derived assuming a linear discharge (also assuming that it was fully recharged originally). Finally, they obtain a discretised version of the stochastic process of the load on half hourly intervals for a single EV which is then extended to an arbitrary number of EVs.

In [158], the authors defined a temporal stochastic process modelling charging patterns at a public EVSE with a Markov Chain comprising three states : unoccupied, charging and plugged-in but not charging. Essentially, the Markov Chains setup assumes that the current state of the process, conditionally to all past states, only depends on the previous state. It simplifies the calculation and has been extensively studied in the literature through many applications [161]. In [158], after initialising the transition probability matrix which drives the path of the process they let the system evolve and assess the revenue made by the charging station.

Auto-regressive integrated moving averages (ARIMA) are a particular type of temporal process. Box and Jenkins [162] formalised a precise methodology to estimate the different orders of ARIMA processes. In [55] the ARIMA process is quantised on hours

of the day. In other words, 24 sub-processes are estimated in their model. The final process obtained is thus a day-ahead hourly forecaster of EV load. In a following paper [159], they improved the performance of their model by forecasting separately conventional load and EV load. The results obtained in this paper reinforces the argument that EV load is structurally different from conventional load and requires specific load forecasting models.

Similarly, in [62] the authors modelled household EV load demand by using stochastic behaviours of three random variables : start-time of trip,end-time of trip and travelled distance. With a vehicle-centric approach, they present a Monte Carlo simulation method to derive overall system load. A particularity of this model is that it used a copula to characterise the multivariate distribution function of model variables. Then, using typical EV charging profiles, they derived the electricity demand at different EV uptake levels while observing the grid impacts.

Purely temporal models are particularly suited for one EVSE or one EV. They are not consistent for modelling cluster of EVSEs which require spatial considerations.

Spatiotemporal

Spatiotemporal models are usually designed for disaggregated approaches. The EV load at different stations is modelled separately using temporal features as well as travel patterns. They are rare in the literature as they require the combination or simulation of both the charging sessions and EV trips. Furthermore, they are limited as they cannot scale to large geographical scopes. Nevertheless, they can explore in fine details the intricacy of the relationship between EVs and EVSEs in specific regions.

In [27], the authors introduced a spatiotemporal model using Monte-Carlo simulation to specifically assess EV load demand in urban areas. The core of this method lies in the origin-destination analysis used to determine daily travel patterns of EVs. Additionally, probability functions to describe EV characteristics were identified. Using both travel patterns and EV characteristics, they ran a Monte-Carlo estimation of EV charging load for each busbar. By construction, this model can also be used for probabilistic assessment which indicates the branches most vulnerable to potential overloading.

In [36], the authors modelled both temporal and spatial stochastic aspects of PHEV owners behaviours to then derive their pdf. They modelled the temporal dimension with a uniform distribution for the start and end of charging time. As for the spatial dimension, they described the number of PHEVs arriving at an EVSE by a Poisson process according to driving behavior and traffic state. Assuming that both dimensions are independent, they derived the joint spatiotemporal pdf by multiplying both individual pdfs for charging times and arrival at EVSE. Ultimately, they expressed the effect on the daily load curve under various number of PHEVs for 150 PHEVs dispersed in the test system.

Finally, [50] proposed another probabilistic approach to characterise the spatiotemporal diversity of EV charging demand specifically on peak load demand. A Monte-Carlo simulation was used to evaluate the impacts of charging demand on the grid in urban and rural environments. It showed that this diversity of location helped the grid handle the demand better.

Queuing theory

Queuing theory models often use Kendall's shorthand notation which describes the arrival (A), the serving time (B) and the number of servers (C) in a compact form : A/B/C. EV load models are a suitable context for this theory as it was detailed in numerous articles [47, 57, 34, 56, 160].

One of the early works on EV load modelling was performed in [47]. This simple theoretical approach proposed to use an $M/M/n_{max}$ queue where the two first components characterises the Poisson processes for the number of EVs arriving at a public EVSE and the number of EVs served while n_{max} refers to the number of maximum parallel charging EVs at charging points. A case study was conducted on the first car produced by Tesla, the Roadster Model, in order to assess the stochastic power demand output from the model.

The same queuing model was also used in [56] and was compared to a Monte-Carlo simulation in order to ultimately fit a distribution for the entire load demand of PHEVs. Additionally, in [160] the authors also used this queuing model and complemented it with a fluid traffic model in order to look at EV charging load on highway charging stations.

In a more general fashion, the authors of [34] have opted for an $M_t/GI_t/\infty$ queue where the number of arrivals follows an inhomogeneous Poisson process (indicating that the intensity function varies over time), the serving time is a general time-dependent distribution with an infinite amount of servers or EVSEs in the EV load context. Using some established results of queuing theory and previous work on estimating non homogenous Poisson process rates, the authors managed to forecast each disaggregated intensity function for day-ahead forecasting. This paper is the only one found for stochastic processes applied to EV load which uses both travel patterns from the NHTS and real charging session data.

Thus, one important advantage of queuing network analysis applied in a spatio-temporal context of EV load is that it can capture interactions among multiple charging stations. In that sense, BCMP networks (named after their inventors : Baskett, Chandy, Muntz and Palacios) introduced in [163] were applied in [57] to produce an EV load model. BCMP networks are a type of queuing network which yield a product-form stationary distribution. This kind of network is commonly used to study interconnected queues. In the EV load context, it means that it enables the model to take into account the potential shift of users from one station to another and control it to envisage different scenarios.

It is clear that queuing models are to be reserved for theoretical considerations rather than for operational implementation. Nevertheless, thanks to their solid mathematical foundations, they bring great insights for understanding EV load behaviours especially when EVSE data is scarce.

2.4.3 Machine Learning

Four machine learning branches have mainly been explored for modelling EV load : Linear Model (LM), Support Vector Machine (SVM), Random Forest (RF) and Artificial Neural Network (ANN). In [49] the authors compare decision trees/tables, SVM and ANN. SVM demonstrated the best performance while the ANN and decision trees are

Table 2.4 – Machine Learning models for EV load

Model	Strengths	Weaknesses	Ref
Linear Model	Easily interpretable with fast implementation	Structurally limited for capturing complex and irregular patterns	[166] [167] [168]
Support Vector Machines	Easy implementation and effective in high-dimensional spaces	Not suitable for large and complex datasets which are not linearly separable even in high-dimensional spaces	[49] [51] [26] [165]
Random Forest	Versatile model with no prior assumptions on the shape of the data	Weak interpretability with no ability to extrapolate from training data	[51] [30] [169] [165]
Neural Networks	Can reach the highest level of performance	Architecture selection process can be laborious with long training time	[49] [51] [59] [67] [46] [170]

ten times quicker to test on new data. A limitation of this work is that the dataset does not come from real charging session data but an aggregated distributional analysis produced by ECOtotality [164]. In [165], SVM, RF, k-Nearest Neighbours (k-NN) and a method called Modified Pattern-based Sequence Forecasting (MPSF) which uses k-means are compared. They found that SVMs and RF reach the best performance with regards to the Mean Absolute Error while k-NN and MPSF achieve better performance with regards to the Symmetric Mean Absolute Percentage Error. Since, k-NN and MPSF are much faster to compute predictions, they concluded that MPSF and k-NN were better suited for operational use. The different machine learning branches studied for EV load models are gathered in Table 2.4.

Linear Model

It is common practice to start addressing a machine learning problem with simple models such as LM. In [166], and [167] LM was chosen as a first step to implement a smart charging strategy. This gives a more realistic operational context as opposed to other articles which skip predictive models before implementing an optimal charging strategy. Furthermore, in [168] an assessment of model inputs is presented using LM. They found that the voltage level of each EV had a critical influence over their model. However, these models are limited as they cannot capture irregular patterns in the data which is expected across EV drivers.

Support Vector Machines

SVM were originally defined by Vapnik [171]. In a nutshell, the idea behind this algorithm is to find the hyperplane which maximises the margin between different sets of populations. It is easy to implement but yields relatively long training times when working with large datasets.

In the context of EV load, SVM were compared to a Monte-Carlo forecasting technique in [26] and showed a better performance on a theoretical charging session dataset. Additionally, in [51] SVMs are used alongside other machine learning algorithms in order to model EV load from a vehicle-centric as well as from an EVSE-centric perspective. Because the EVSE-centric approach requires more data, it demonstrates a significantly longer running time as expected for SVMs. This study used a dataset extracted from UCLA campus parking lots. Thus, it is unlikely that these kinds of models will scale adequately for a larger scope of charging stations. Furthermore, articles using SVMs are now becoming rare as other alternatives with similar or better performances can be found.

Random Forests

RF is a learning algorithm which was popularised by Leo Breiman [172]. In short, it is an ensemble method which uses decision trees as elementary components for its construction.

On top of SVMs, [51] also used RF to model EV load. The few hyperparameters required to be tuned (e.g., number of trees, sampling rates) enables a fast and easy implementation with the possibility to iterate rapidly. In [30], RF demonstrated their ability to forecast day-ahead EV load charging blocks for households in an EVSE-centric fashion. As mentioned in previous paragraphs, the EVSE-centric approach can be difficult to implement as it requires large amounts of data and complex modelling. Thus, the use of RF for this kind of disaggregated approaches is adequate. [169] precisely illustrates the ability of RF to handle the EV load problem from both time and spatial dimensions. This article shows that RF can model both a single station as well as a group of stations considering spatial and temporal inputs. Single station models are more accurate as they have more consistent behaviours while the group of station models is slightly less accurate in terms of the mean absolute percentage error but brings a more holistic view to the problem. Other ensemble methods which stemmed from the same area of machine learning such as gradient boosting could also be considered [167], [173].

Neural Networks

ANNs were initially presented by Frank Rosenblatt in 1958 [174] in their most elementary form in the name of the perceptron. They were extended shortly afterwards to the Multi-Layer Perceptron (MLP). After being forgotten for a few decades, ANNs have experienced a rebound in interest from the end of the 80s in particular with the formulation of the backpropagation algorithm [175] and the breakthroughs in computer vision with Convolutional Neural Networks (CNN) [176].

In [59] an MLP with tilted loss function is used for probabilistic forecasting of EV load. It is compared with a kernel density estimator as well as quantile regression and

it showed the best performance using the same inputs and outputs. It is quite common across various scientific fields that ANNs reach the highest performance on many problems compared to other machine learning or statistical methods. The main drawback is the lack of interpretability of such models which are highly complex [177, 178]. Elaborate ANN architectures such as CNNs [67] and Recurrent Neural Networks (RNNs) [46] have been explored for modelling EV load. [170] compares 12 different architectures including CNNs, RNNs. In this last article, the Long Short-Term Memory (LSTM) architecture showed the best performance on the dataset studied. From all these articles it is challenging to decide which ANN architecture is the best overall for EV load modelling. However, some clear conclusions can be made. RNNs are particularly performant as they take into account historical EV load values. In the current operational context, this is information that is hard to obtain at fine time steps. Thus, until real-time communication channels are available, it is likely that the most useful ANN models will be CNNs or RNNs with larger timesteps.

2.5 Matching EV load models to open datasets : a preliminary study

Six datasets dealing with historical charging session information have been selected for their completeness and accessibility : Boulder [100], Palo Alto [101], Dundee, [124], Perth, [125], Paris [113] and Domestics UK [143]. According to this research, none of these datasets were used in the EV load modelling literature so far. The purpose of this section is to identify the variables available and to enable a high-level understanding of charging behaviours. In addition, an association of the 6 datasets selected with the models reviewed in Section 2.4 is proposed.

2.5.1 Variables and data quality

The fields available in each of the six datasets selected are summarised in Figure 2.3. These six datasets provide us with session start and end times as well as the energy consumed. With the exception of the Domestics UK dataset, the station address (location), and the power level of the charging port are available in these datasets. In addition, the Palo Alto and Boulder datasets contain gasoline and greenhouse gases (GHG) savings as well as the charge duration which represents the amount of time the vehicle was plugged-in and actively charging. This is different to the park duration which also captures the time a vehicle was plugged-in and no longer charging which is a variable only given in the Palo Alto dataset. However, this park duration can be deduced from the session start and end times in the remaining datasets. Finally, for customer specific information, the Paris data provides a unique identifier per customer badge and the Palo Alto dataset gives the post code registered by the driver. Information regarding postcodes is interesting for models that include travel inputs such as the distance between the driver's home and stations nearby.

Additionally, a data quality analysis was conducted on the six datasets. In [170], outliers were identified by using a set threshold from the variability between current and

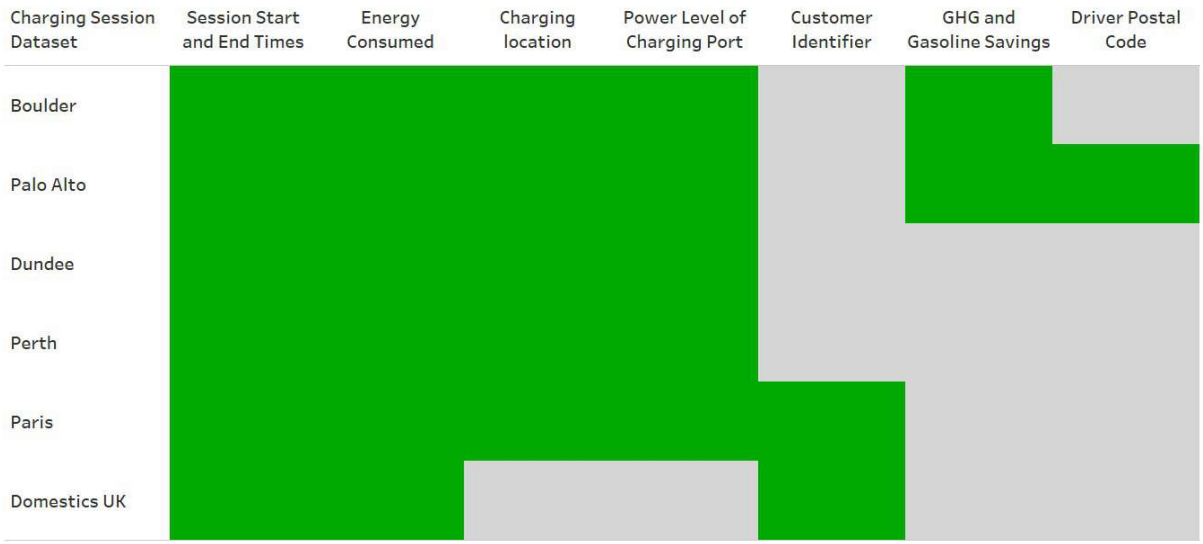


Figure 2.3 – Fields available in the six datasets selected. Fields available are in green and fields not available are in grey for each dataset considered.

previous values. Instead, in this analysis, fixed boundaries were chosen and the following set thresholds were observed :

— Charge and/or Park Duration has to be positive and less than 24 hours
— Energy Consumption needs to be positive and less than 100 kWh

The first criterion is important as some datasets have some obvious errors in the end times column which are set in 1970. This might indicate a manipulation error from the customer which led to a computational mistake along the process of data collection. Also, as most charging sessions last for a few hours, charging sessions that lasted for more than a day were discarded.Similarly, recorded energy consumption values for Perth [125] and Dundee [124] are highly variable, reaching anomalously negative and highly positive values indicative of potential errors. The 100 kWh upper bound was chosen as it is close to the highest capacity of the Tesla Model S which is the EV with the largest battery capacity amongst the most widespread models [179]. If a transaction does not fit these criteria, it is discarded from the following analysis. This preparation had very little impact on the Palo Alto dataset with only 0.17% of transactions discarded, while the Boulder, Dundee, Perth, Paris and Domestics UK datasets have seen 8%, 11%, 4% , 14% and 7% of their data discarded respectively.

2.5.2 Exploratory analysis

Figure 2.4 shows the trend in the total number of transactions per day over each dataset specific time frame. Due to increasing EV uptake [19], an increase in EV charging sessions is expected as illustrated by Palo Alto and Perth. However, this is not the case for Boulder and Dundee. Instead, a decreasing number of charging sessions at the end of each time series can be observed. This could be due to external factors such as an increase in charging session prices. As for the Domestics UK, only one year of data is available which indicates that the plot shown describes the yearly cycle rather than the long-term trend. Similarly, the Paris data cannot be extrapolated as it only represents

two months of data.

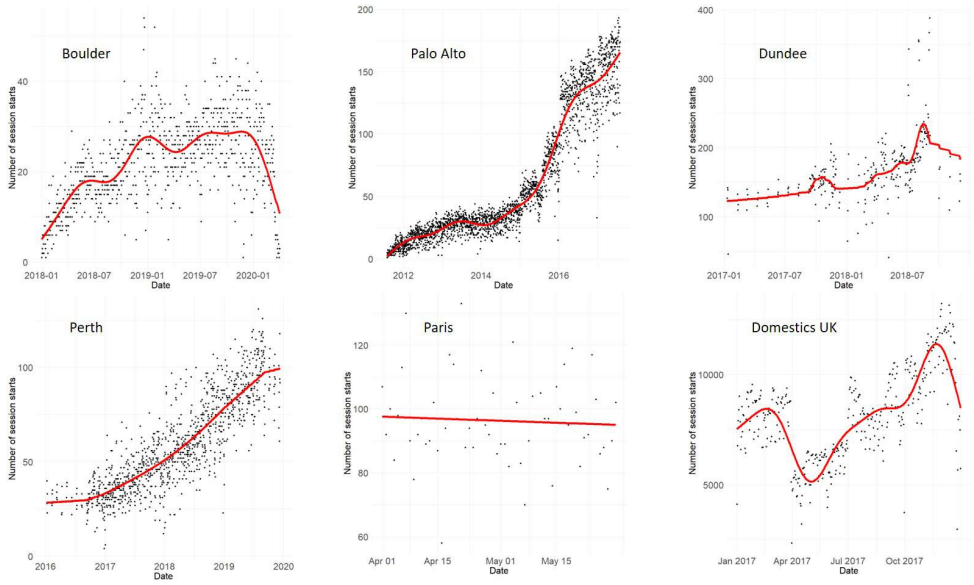


Figure 2.4 – The black dots represent the daily number of sessions while the red curve is the output of a generalised additive model used to smooth the values. Each model was produced using the *mgcv* package in R and its default settings [180].

Overall statistics of the six datasets in scope are provided in Table 2.5. The dataset which covers the largest time frame is from the city of Palo Alto with 6 years, followed by Perth with 4 years, Boulder with a little over 2 years, Dundee with 2 years, Domestics UK with 1 years and Paris with 2 months. In terms of the transactions (or sessions), Domestics UK records the largest number of transactions. Moreover, Dundee and Boulder both cover 2 years of data but Dundee has close to three times more transactions. Naturally, there are consistently more transactions on weekdays than on weekends in total and on average across all datasets. Furthermore, the average park duration and charge duration varies significantly across the datasets. Indeed, while for Palo Alto the average park duration is around 2 hours and 40 minutes, in Perth it is closer to 1 hour and 15 minutes, so less than half of the time. Additionally, the average Park Duration for Domestics UK is greater than 9 hours which is expected as this dataset describes residential charge instead of public charge for the others. The Charge duration on the other hand is relatively close for Boulder and Palo Alto which are both American cities (located in Colorado and California). Finally, the average energy demand is consistently between 8 to 11 kWh across all datasets.

2.5.3 Suggested matching of EV load models with the datasets considered

From this exploratory analysis, some suggestions can be given on how to match the EV load models reviewed in Section 4 with the six datasets presented above. The first criterion identified for this selection is whether the dataset describes public charging or residential charging. Only two of the models reviewed deal with residential charging and will thus be assigned to the Domestics UK dataset. They would benefit from the

Table 2.5 – Overall statistics and high-level information for the six datasets in scope.

Charging Session Dataset		Boulder	Palo Alto	Dundee	Perth	Paris	Domestics UK
First Transaction Date		2018-01-01	2011-07-29	2017-01-09	2016-01-09	2017-04-01	2017-01-01
Last Transaction Date		2020-03-31	2017-07-31	2018-12-05	2019-12-08	2017-05-30	2017-12-31
Total Transactions	All	18,052	133,329	47,051	63,936	4,225	2,956,198
	Weekdays	13,487	101,486	34,434	46,607	4,907	2,208,695
	Weekends	4,565	31,843	12,617	17,329	1,555	747,503
Average Transactions	Weekdays	23	65	173	60	101	8,495
	Weekends	20	51	160	55	86	7,119
Park Duration (h)	Mean	-	2.7	2.29	1.24	1.7	9.16
	Standard Deviation	-	2.41	4.56	2.13	2.93	6.49
Charge Duration (h)	Mean	1.81	2.05	-	-	-	-
	Standard Deviation	1.34	1.39	-	-	-	-
Energy Demand (kWh)	Mean	8.42	8.18	9.16	11.01	8.51	8.88
	Standard Deviation	7.08	6.76	6.53	8.49	6.71	7.55

large number of records and the customer identifier provided in this dataset. Thus, it is a good setup for vehicle-centric approaches [62] and machine learning models [30].

Looking now at public charging sessions, Boulder and Palo Alto are the only two datasets which gathered GHG and gasoline savings. These fields are rather uncommon across charging session datasets and they can enable an environmental impact analysis of EVs. However, it would be limited to EV usage rather than a holistic environmental impact with lifecyle assessments [181]. Thus, no mention of this kind of analysis was found in the articles reviewed.

Nevertheless, Palo Alto also possesses the driver's post code. With this knowledge, fine spatiotemporal processes can be derived as proposed in [50]. Additionally, the large amount of records available is suited to test the scalability of queuing models [34, 57] and spatiotemporal processes [27, 36] which require travel information. It also provides an ideal setup for deep learning models which require large training sets [51, 59, 67, 46, 170].

As discussed in Section 3, Boulder not only holds a charging session dataset but also a travel survey led in 2018 with a focus on EVs [105]. It is a rather qualitative survey and can be used in combination with the NHTS [61] to address the more specific behaviours inherent to the city of Boulder. With both travel and charging session data, this is also a favourable setup to apply spatiotemporal models [27, 36, 50]. Considering that this dataset is continuously updated and holds recent data, it would also be interesting to apply models which were precisely built for operational use such as [165] and [169] in the hope of taking consistent conclusions with real-world applications.

The Paris dataset holds customer identifier information which encourages vehicle-centric approaches. However, the small amount of data deters the use of models which leverage numerous parameters. Instead, statistical characterisation techniques with unimodal distributions could yield a sufficient approximation of the phenomenon as proposed in [31] and [32] along with LM [166, 167, 168]. The remaining statistical characterisation models (mixtures [42, 31, 141, 152, 151] and KDEs [153, 154, 155, 156, 157]) can capture diverse patterns and thus could be applied to medium-sized datasets. The Paris dataset could also be used to verify the consistency of simple queuing models as they usually struggle to find concrete applications [47, 160].

As Perth and Dundee are two neighbouring cities of the UK, it would be interesting to compare the difference in charging behaviours between them. Considering that they have the same fields, it would be interesting to independently compare their behaviours as despite their closeness, it is unlikely that there is a significant spatial impact between these for public charging. Thus, temporal processes produced in [33, 158, 55, 159] would be well suited. Additionally, thanks to the medium-size of both datasets, SVM models described in [51, 26] could also be a good option here.

Table 2.6 – Charging session datasets with associated EV load models suggested.

Charging session dataset	Specificities	EV load models	Models references
Boulder [100]	Public Charging Sessions Medium-sized Dataset GHG and Gasoline Savings 2 years	Spatiotemporal RF Mixtures KDE	[50] [27] [36] [165] [169] [42] [152] [141] [151] [153] [154] [155] [156] [157]
Palo Alto [101]	Public Charging Sessions Large Dataset Driver's Post Code 6 years	Spatiotemporal Queues ANN	[50] [27] [36] [34] [57] [51] [59] [67] [46] [170]
Dundee [124] & Perth [125]	Public Charging Sessions Medium-sized Datasets 2 years (Dundee) & 4 years (Perth)	Temporal SVM Mixtures KDE	[33] [158] [55] [159] [51] [26] [42] [152] [141] [151] [153] [154], [155] [156] [157]
Paris [113]	Public Charging Sessions Small dataset Customer Identifier 2 months	Unimodal distributions LM Queues	[31] [32] [166] [167] [168] [47] [160]
Domestics UK [143]	Residential charging sessions Large dataset Customer Identifier 1 year	Temporal RF	[62] [30]

2.6 Discussion and future work

The purpose of this section is to highlight and discuss the current gaps and limitations from both open data and EV load models perspectives (Table 2.7).

2.6.1 Data usage and privacy issues

With this article, the community has clear visibility on a carefully selected list of open datasets useful for modelling EV load [191]. In most research papers, datasets obtained

Table 2.7 – Gaps and future work for EV load data and models. Sections 6.1 and 6.2 deal with data prospects and limitations while Sections 6.3 and 6.4 describe new ways of modelling and tie EV load models to optimisation of charging schedules.

	Section	Keywords	References
Data	6.1 Usage and privacy issues	Variables Accessibility Regulations	[74] [73] [59] [182]
	6.2 Other types of relevant data	General Electrical Load Non-Intrusive Load Monitoring Synthetic data	[71] [183] [184] [185] [152]
Models	6.3 Composite Approaches	Stochastic Processes Machine Learning	[186] [187] [188]
	6.4 Link with optimisation	Smart Charging Probabilistic Approach Reinforcement Learning	[24] [189] [190]

from system operators or other entities were not explicitly made available and no clear indication was given on how to retrieve them if it was possible. Our hope is to foster the practice of sharing supplementary materials with both the data used and code produced in order to encourage reproducible work in the field.

While the open data search provides visibility on charging session and traffic data, no repositories merging both was found. Thus, it is likely that the standard will remain to manipulate separate datasets for charging and traffic data as it is already the case in the literature [57, 66, 59] at least in the near future. And as such, different locations and different grains of data will still need to be leveraged in order to perform a complete data-driven spatiotemporal description of EV load.

Battery inputs are intrinsically complex to obtain. It would involve establishing an Internet of Things (IOT) between EVs, charging stations and controllers when considering a smart charging scenario [44, 45]. This type of work is currently in progress [59] and the community could benefit from new types of information for EV load models in the near future. However, so far, the articles which include these variables simulate them from prior statistical distributions.

Some of the datasets presented in Section 3 provide unique identifiers for vehicles and even driver's registered post codes [101]. However, data regulations (e.g., GDPR in Europe [182]) may prevent spreading battery and travel inputs openly. Thus, more elaborate and complex models will be required in order to capture hidden information for disaggregated approaches.

The other variables of interest pinpointed in this review are easier to retrieve. For example, weather information can be obtained from the R package *riem* for a wide range of locations [73] or on the NOAA website [74]. If finer information is required, meteorological grid models can be used for that purpose. Economical and calendar variables can be tailored for each analysis depending on the grain chosen.

2.6.2 Other types of relevant data

The open data search presented in Section 3 mainly focuses on charging session and traffic data. However, it is also possible to consider general electrical load open data [71]. Indeed, if a region switches to EVs in a given time period, this change can reflect on the regional load curve. In this context, EV load would be a latent or hidden variable contributing to the general electrical load.

In addition, grid networks data [192] and big cities' electrical load data [183] are becoming more and more available. Combining them with charging session and traffic data may lead to models which have a holistic and data-driven understanding of the reality.

To model the load of specific appliances with general electrical load data, Non-Intrusive Load Monitoring (NILM) methodologies have received a lot of interest in the related literature [193, 194]. The question addressed is whether it is possible to identify and characterise EV load within a general electrical load curve [184], [185].

Synthetic data can also be used in order to produce EV load models. In most research papers, simulators rely heavily on assumptions derived from travel surveys and not so much on real charging session data [158, 33, 32]. Nevertheless, a data-driven simulator has been recently proposed in [152] which was trained on real-world charging sessions and thus can represent more accurately real world charging behaviours.

Finally, there are semi-open or closed data. Most of these closed datasets are related to residential load [108, 195] as it is less feasible to retrieve them without raising data privacy concerns.

2.6.3 Composite Approaches

From a methodology perspective, it is interesting to note that very few stochastic processes approaches used real data [34]. These models are usually theoretical and can be useful for mid-term or long-term scenarios but less relevant for short-term forecasting. Alternatively, the machine learning and statistical characterisation approaches presented were highly data-driven.

In the corpus of articles considered in this review, there exists no article that deals both with stochastic processes and machine learning algorithms in the context of EV load models. Thus, it would be interesting to compare them in terms of performance but also to assess what they can bring to each other in a composite model [188, 186, 187].

Furthermore, it was shown in many articles reviewed in Section 4 that the influx of vehicles at EVSEs is highly time dependent. Consequently, homogenous poisson processes used in articles from the corpus will not be enough to capture the reality of drivers' behaviours [36]. More elaborate processes such as inhomegenous poisson or self-exciting point processes [196] have to be considered to account for this time dependence. Using these stochastic processes hand in hand with machine learning algorithms will foster consistency, conciseness and performance of EV load models.

Finally, another gap brought to light in this review is the lack of work on stacking models or bottom-up approaches [30] which are indeed more costly from a computational perspective but can bring a deeper understanding of EV load.

2.6.4 Link with optimisation

As mentioned in the introduction, EV load models are part of a two-step process. Firstly, behaviours relating to EV load demand must be understood and then current schedules optimised depending on the aim (e.g., load flattening or load balancing). The articles introducing methodologies for optimising charging schedules usually assume a clear knowledge of the future short-term demand. It is less common to see articles which account for the potential uncertainty of EV load models. This is also due to the fact that there has been less focus given to probabilistic EV load models which could yield confidence intervals for evaluating risks of surpassing the energy supply at a given time. Additionally, probabilistic forecasting proposes a more exhaustive representation of the demand as it does not solely focus on the mean demand.

Solutions which include both forecasting and optimisation aspects in the same model or process are required [24], [197]. Again, using the same data for this purpose is essential, as it enables the development of solutions by researchers specialised in different fields such as forecasting and optimisation. To unify both, methodologies can be also developed using reinforcement learning [190]. In addition, specific losses related to the exploitation of probabilistic forecasts in smart charging strategies could be relevant [189].

2.7 Conclusion

In this paper, the reader is provided with a comprehensive list of open data that can be used to model EV load. Additionally, an organised review of EV load models is presented. Finally, six datasets are explored to provide recommendations on how they can be matched to the EV load models reviewed. The open data search focused on the top 14 countries of the EVI ranked by national EV market share. A total of 860+ open data repositories was covered which yielded more than 60 open datasets relevant for modelling EV load. Across the literature, a wide spectrum of EV load models were reviewed. This includes statistical characterisation models from parametric (unimodal distributions and mixtures) to non-parametrical estimation (KDE). Furthermore, stochastic processes with purely temporal models, spatiotemporal models and queues were also included. Finally, machine learning models including LM, SVM, RF and ANN were reviewed.From the open data research, six datasets which have not been previously studied in the literature were considered. Recommendations were provided on how the models reviewed could be matched to each dataset. We hope that this article will encourage the use of the open datasets and models reviewed in order to foster reproducible work and breakthroughs in the field of EV load modelling.

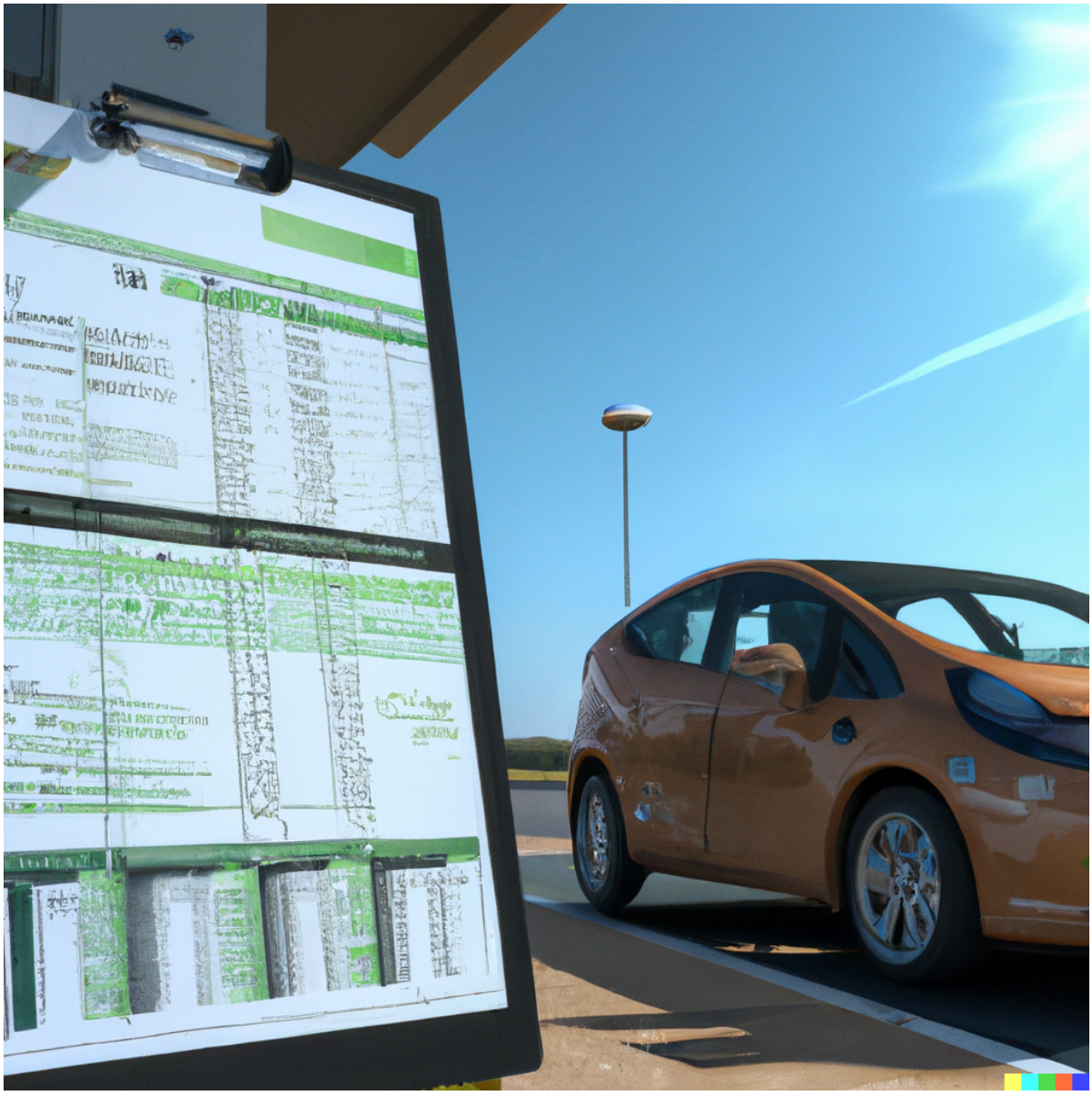
Chapitre 3

A benchmark of EV load and occupancy models on open data

This chapter is based on a paper published in the proceedings of ACM e-Energy 2022 [198].

We propose an extensive benchmark of 14 models for both load and occupancy day-ahead forecasts, covering 8 open charging session datasets of different types (residential, workplace and public stations). Two modelling approaches are compared : direct and bottom-up. The direct approach forecasts the aggregated load (resp. occupation) directly of an area/station whereas the bottom-up approach models each individual EV charging session before aggregating them. This second approach is key to the effective implementation of smart charging strategies. We consider both machine learning models (Random Forests and Gated Recurrent Units) and statistical models (Generalised Additive Models, Poisson Regression, Mixture Regression, Auto-Regressive) in order to maximise the spectrum of our benchmark. We finally propose an adaptive aggregation strategy to assess the variety of forecasts at hand. Overall, we demonstrate that direct approaches reach better performances than bottom-up approaches across all datasets considered. We further show that the different approaches used can lead to an improved performance of direct approaches when using an adaptive aggregation strategy. In fact, our best model produces a forecast which is more than 5 times better relative to the persistence on residential data.

Summary



3.1 Introduction

Electric Vehicles (EV) represent an important lever for a transition to low-carbon transport. By carefully managing EV charging load, it can become a flexible asset to the grid in various ways (e.g., load balancing, load flattening). Operating a system relying on smart charging requires a thorough understanding of the mechanisms underpinning charging behaviours. Therefore, over the last decade, numerous papers have been published to address the question of EV charging load models, and in particular, EV demand forecasting (see [18] or [199] for extensive reviews of the literature).

As the EV market is still in its infancy, most papers rely on simulated or private data to train EV charging load models. The lack of studies using open data has been hindering reproducible research and the establishment of sensible benchmarks. In the meantime, many open datasets have been made available and are still underused by the community [200]. Only recently, [201] proposed a deep learning graphical network model to forecast the daily energy demand of public charging stations in Palo Alto (USA). In addition, [202] proposed a long short-term memory (LSTM) neural network to forecast occupancy of public charging stations in Dundee (UK) at an intraday horizon. Finally, [203] is a data centric statistical study which demonstrates the importance of geospatial information on charging behaviours.

Benchmarking statistical and machine learning approaches on multiple existing open datasets is of crucial interest for the industrial and academic communities. To fill this gap, we propose an investigation of different state of the art and innovative models trained on the open datasets considered. These datasets cover all common charging behaviours : residential, public, and workplace charging. Two main approaches have been retained for this benchmark : a *direct approach* where the load or the occupancy is forecasted at the aggregated level ; a *bottom-up approach* where the load or the occupancy can be derived from individual charging sessions simulated by the model.

The *direct approach* uses the traditional electricity load forecasting set-up. If the EV station is sufficiently large, aggregating the occupancy or the load of each EV will regularize the signal by the law of large numbers. Therefore, we can expect to achieve good results with similar methods used for aggregated electricity consumption data. Generalised Additive Models (GAMs), state of the art models for electrical load forecasting [204] as well as Random Forests (RF) [205] will be considered.

The *bottom-up approach* is highly relevant for assessing the benefits of smart charging strategies. Forecasting individual charging sessions has many advantages : it allows for the simulation of different charging schemes naturally and can incorporate individual information (e.g., personal charging constraints, traffic, habits). A charging session i is characterised by three variables to be estimated : the arrival time (a_i), the charge/park duration (c_i/p_i) and the energy demand (e_i). Charging sessions are modelled by methods including non-homogeneous Poisson processes [206], time series, and multivariate density estimation.

As a result, an in-depth analysis of models' performances on 8 charging session open datasets is provided. A rolling-origin forecasting procedure is used to replicate the operational setting for these models. We hope that this will foster reproducible research in the field and may be used as a baseline reference for future modelling work. The key contributions of this paper are (a) a unique framework for comparing EV

load or occupancy forecasting models, (b) a discussion of 14 model performances on 8 open charging sessions datasets and (c) a demonstration of the benefits of leveraging the variety of bottom-up and direct forecasts by aggregating them into one forecasting model. The rest of the paper is structured as follows : Section 3.2 describes the problem at hand and the datasets in scope. Section 3.3 gives the detail on the methodologies considered in this benchmark. Finally, Section 3.4 presents the forecasting setting, the model selection and validation procedure and compares model performances.

3.2 Problem Formulation

The aim of this work is to produce day-ahead EV load and occupancy forecasting models at a one-minute time resolution. A wide variety of models have been presented in the literature to address forecasting problems in relation to EVs [18, 199]. These models either characterise EV charging sessions to then derive information on the load/occupancy or directly forecast the load/occupancy. This is not surprising as most data used is set at a charging session level. This is a particularity of EVs as in the electrical load forecasting literature it is common to use meter readings data which are still scarce for EVs. In the following sections, we define a procedure for reconstructing a load/occupancy curve with charging session information. We also briefly present the 8 open datasets used in this benchmark.

3.2.1 From charging sessions to load and occupancy curves

To reconstruct the load curve, three variables are required for each charging session i : the arrival time (a_i), the charge duration (c_i) and the energy drawn (e_i). With these three variables, an elementary load curve can be reconstructed in the shape of a step function :

$$l_i(t) = \begin{cases} \frac{e_i}{c_i} & \text{if } t \in [a_i, a_i + c_i] \\ 0 & \text{otherwise.} \end{cases}$$

The overall load curve is obtained by summing all elementary load curves : $L(t) = \sum_i l_i(t)$ We assume that the electric vehicle supply equipment (EVSE) charges the EV at a constant power. This assumption is accurate for EVSEs which are not subject to smart charging. There are some well-known transitional regimes at the beginning and the end of each charging session in terms of the power delivered due to electrical constraints but common practice is to assume that this phenomenon is negligible. When smart charging is involved, the load curve reconstructed with that process may be unrealistic if it involves drastic and/or frequent changes to the power delivered. This is not the case for the datasets in the scope of this paper.

As for the occupancy curve, the process is even simpler. Only two charging session variables are required : the arrival time (a_i) and the park duration (p_i). Again, an elementary occupancy curve can be reconstructed for each session i :

$$o_i(t) = \begin{cases} 1 & \text{if } t \in [a_i, a_i + p_i] \\ 0 & \text{otherwise.} \end{cases}$$

and the overall occupancy curve is obtained by summing all elementary occupancy curves : $O(t) = \sum_i o_i(t)$

3.2.2 Datasets

As highlighted in [141], EV load/occupancy models which are not built on open data are very common. This drastically hinders reproducible research and a fair comparison of models. For this benchmark, we consider 8 open charging session datasets which were identified and thoroughly presented in [18] and [200]. The quality of the datasets considered and their time span are detailed in Appendix B.1. The datasets used in this paper cover most common charging types : public, workplace and residential charge. All datasets record a_i and p_i for each session i. However, c_i is recorded by 4 of the 8 datasets. Therefore, the load curve will be reconstructed and modelled only for these while the occupancy curve will be analysed for all datasets (Table 3.1). Note that weekends are excluded from our analysis due to significantly different charging behaviours compared to working days [42] and a low number of charging sessions observed in most of the public charging stations in scope.

In terms of the covariates used we want to be consistent and exploit the same ones across all models and datasets. In the literature, traffic data is often used as it has a natural relationship with EVs. However, this data is mostly simulated because of the lack of information focusing on EVs. Simulating traffic data goes against the data-driven vision of this paper. In some rare cases, weather data is also used [65] but it was shown on multiple occasions that it does not yield any significant improvement to model performances. Therefore, we used calendar information and lags as inputs of our models. As for the training procedure, we have adopted a rolling-origin forecasting framework as it is consistent with the day-ahead forecasting set-up we chose for this benchmark. We suppose that the models are re-estimated every 2 weeks, a reasonable frequency for practical applications and a good trade off between accuracy and computational complexity. The test period of all datasets is taken over 8 weeks except for Paris where the test period is only 4 weeks because of the short time window of the dataset. The 8-weeks testing period is shown on Figure 3.1 for Palo Alto. In particular, we can observe the weekly and daily patterns of both load and occupancy.

3.3 Methodologies considered

The methodologies considered in this paper can be divided into two approaches : direct and bottom-up. Figure 3.2 presents a holistic view of the approaches taken. Before going into the details of each method we briefly present them and explain how they fit into each approach. When there is no need to distinguish between load and occupancy, we just refer to both interchangeably as the "target".

3.3.1 Direct Approach

With a direct approach, the target is modelled without any intermediary. Therefore, we can use state of the art methods for electrical load forecasting such as Generalised Additive Models (GAMs) as defined by [207] with the implementation from [208] and

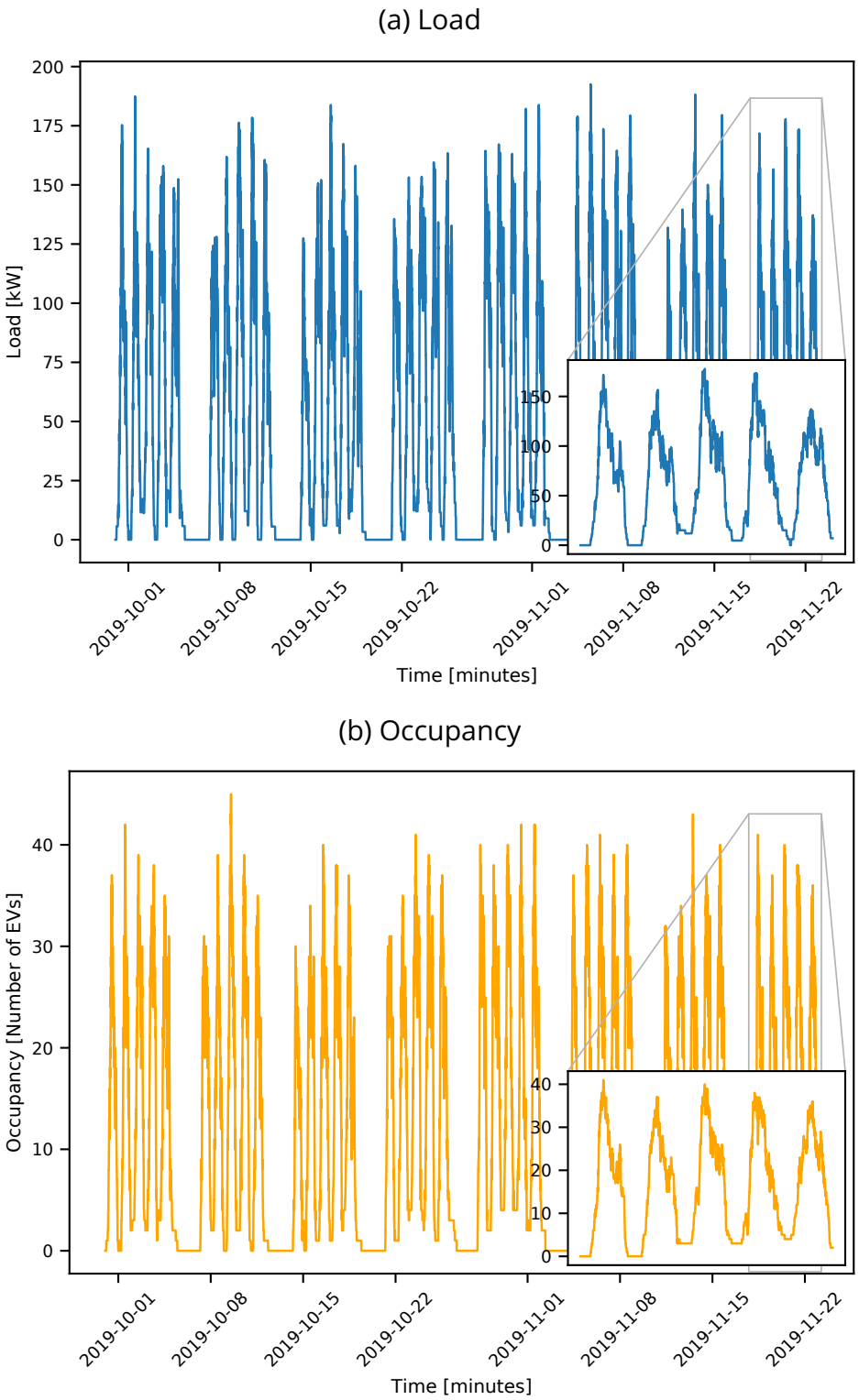


Figure 3.1 – Palo Alto load and occupancy curves over the test period

Table 3.1 – Open datasets considered for the benchmark and relevant fields for charging session i.

Dataset	Relevant fields	Type
Boulder	a_i, c_i, p_i and e_i	Public
Caltech	a_i, c_i, p_i and e_i	Public
Domestics UK	a_i, p_i and e_i	Residential
Dundee	a_i, p_i and e_i	Public
Palo Alto	a_i, c_i, p_i and e_i	Public
Paris	a_i, p_i and e_i	Public
Perth	a_i, p_i and e_i	Public
SAP	a_i, c_i, p_i and e_i	Workplace

Random Forest (RF) as defined by [172]. In addition, we introduce a baseline persistence model which in our case consists of using the observed value of the target 24 hours before. As we only study weekdays, the persistence forecast for Monday 12 :00PM is the observed value on the previous Friday at 12 :00PM. In the following sections, X refers to the vector of covariates and Y is the target variable (load or occupancy).

Generalised Additive Models

State of the art models for electrical load forecasting [209], GAMs are semi-parametric models which can be written in the following fashion : $\mathbb{E}[g(Y)|X] = X^T\boldsymbol{\beta} + \sum_{l=1}^{L} f_l(\boldsymbol{X})$, with g a link function depending on the law of the data, $\boldsymbol{\beta}$ a vector of coefficients to be estimated, and $(f_l)_{l=1}^{L}$, $L \in \mathbb{N}^*$, unknown smoothed functions which we commonly estimate by projection on a spline basis. The two great advantages of GAMs are the flexibility enabled by the non-parametric part of the model and their additive formulation which enables interpretability. The downside is that they usually require a more careful choice of predictors compared to machine learning methods. Since the occupancy is represented as count data, we use a Poisson distribution and a logarithmic link function. The load forecasts are derived from a classical Gaussian distribution hypothesis.

Random Forest

Breiman [172] proposed the Random Forest (RF) method and algorithms. Since then it became one of the most popular algorithms in data science. Among many reasons for this are their automatic computation, their high adaptation to many different problems and their straightforward tuning. Given a sample of observations $(X_t, Y_t)_{1\leq t\leq n}$ drawn from random variables (X, Y), the objective is to fit the very generic model $Y_t = f(X_t) + \epsilon_t$ where the error ϵ_t is such that $E[\epsilon_t|X_t] = 0$. RFs estimate the regression function by computing an ensemble of regression trees [210] using bagging [211] and random sampling of covariates. The intuition behind RF is to reduce the high va-

riance of regression trees by generating a diverse set of trees and aggregating them by a simple mean.

3.3.2 Bottom-up Approach

The bottom-up approach models a vector of variables which defines an EV charging session. For the load it is a triplet (a_i, c_i, e_i) made of the arrival time a_i, the charge duration c_i, and the energy demand e_i. For the occupancy it is a pair (a_i, p_i) made of the arrival time a_i and the park duration p_i. In the following, we just refer to both vectors interchangeably as the "target vector". Finally, the target curve is reconstructed using the target vectors forecasted (procedure detailed in section 3.2.1). To model these vectors, two strategies are considered (see Figure 3.2) :

1. Forecasting the number of daily charging sessions to then sample target vectors from a multivariate distribution
2. Simulating arrivals of a NHPP with a thinning procedure [212] to then forecast the expected value of target vectors conditionally to the arrivals simulated

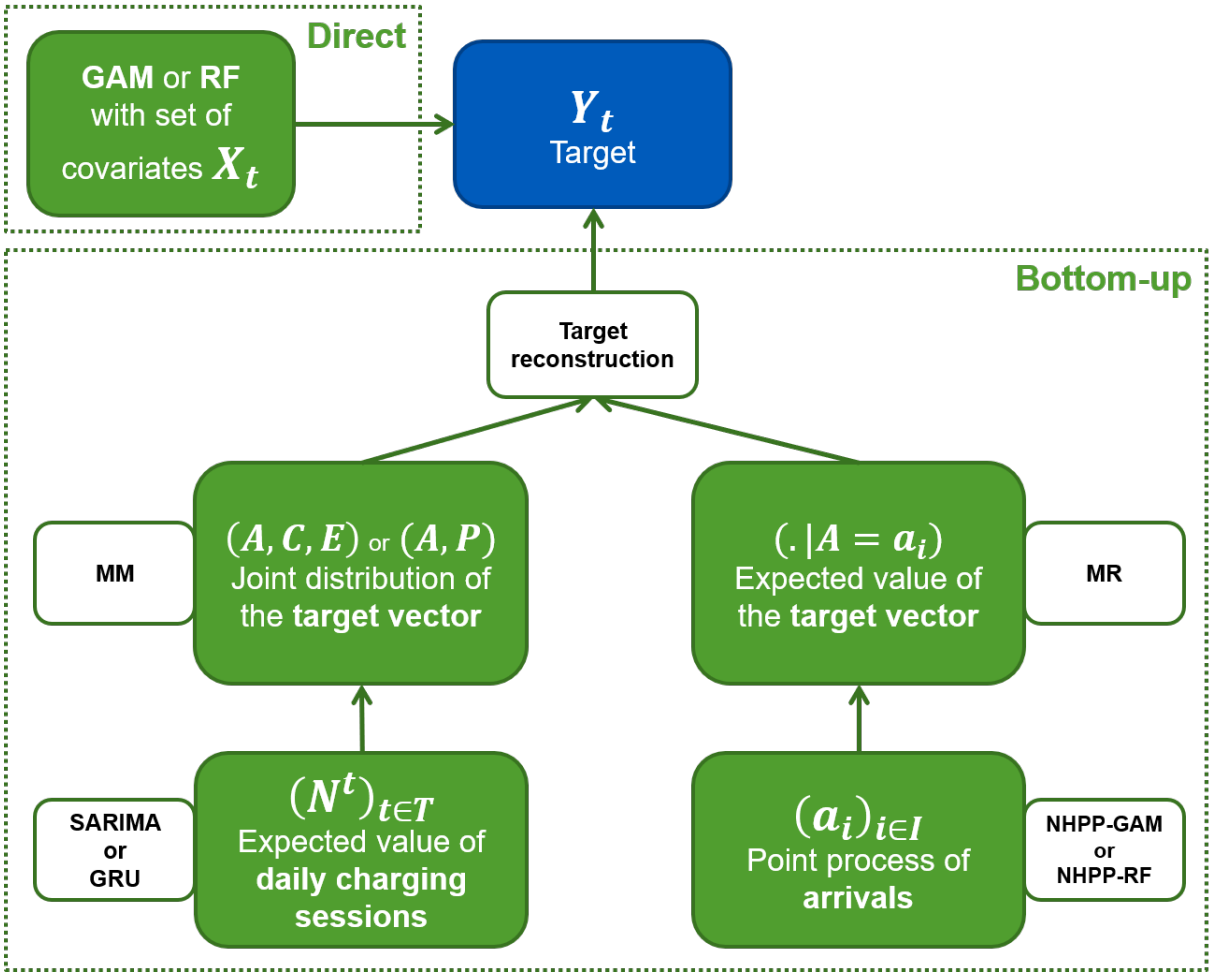


Figure 3.2 – Direct and Bottom-up Approaches

Two methods fit under the first strategy : (a) Gated Recurrent Units as defined in [213] and Mixture Model (GRU-MM) (b) Seasonal Auto-Regressive Integrated Moving Average [214] and Mixture Model (SARIMA-MM). The GRU or SARIMA part predicts the number of daily charging sessions while the MM models the distribution from which the sessions will be sampled.

As for the second strategy, two methods are also studied : (c) Non-Homogeneous Poisson Process estimated with GAM defined in [206] and Mixture Regression (NHPP-GAM-MR) (d) Non-Homogeneous Poisson Process estimated with RF and Mixture Regression (NHPP-RF-MR). In both cases, the NHPP part enables the simulation of arrivals while the MR provides the remaining variables of the target vector conditionnaly to each arrival. In the following sections, we detail the specifics of each of these methods.

Seasonal Auto-Regressive Moving Average

In 1976, [214] introduced and formalised ARIMA models. With their strong theoretical background and easy implementation, they are one of the first-choice models in time series modelling, particularly when patterns can be captured by using past values of the time series. Here, we are trying to model the number of sessions per day which is presented for Palo Alto on Figure 3.3 for instance. Finally, to account for weekly seasonal patterns we also include weekly seasonal orders. Apart from the day of the week which is already taken into account by the seasonal orders, no exogenous variable is available to the SARIMA model. Therefore, we do not use a SARIMAX model which could be well suited in a situation where exogenous variables are relevant and available.

Gated Recurrent Units

When addressing time series forecasting problems with neural networks, the most common architecture used is recurrent neural networks (RNN). It differs from the traditional feed-forward structure by making the information flow in a loop through recurrent cells. In other words, hidden layers have access to their previous state as an input as well as inputs given by the previous layers. This is particularly suited to sequential data such as time series. It is different than just using the lags of the time series as inputs as the hidden units will produce latent time series which will also be used in the network.

As far as RNNs are concerned, the most widely used structures are LSTM [215] and gated recurrent units (GRU) [213]. They both overcome the two major issues of standard RNNs which are exploding and vanishing gradients by extending the memory of the recurrent cells to more than the immediate past. We also want the RNN to forecast the total number of sessions per day. Therefore, we have opted for the GRU architecture in this benchmark as it requires less parameters to estimate than the LSTM and was shown to perform as well on small datasets [213].

Non-Homogeneous Poisson Processes

A NHPP is a time-dependant Markovian stochastic process. It is uniquely determined by its intensity function $\lambda(t)$, the infinitesimal rate at which an event is expected to occur at time t. This type of model has already been used successfully in the case of residential

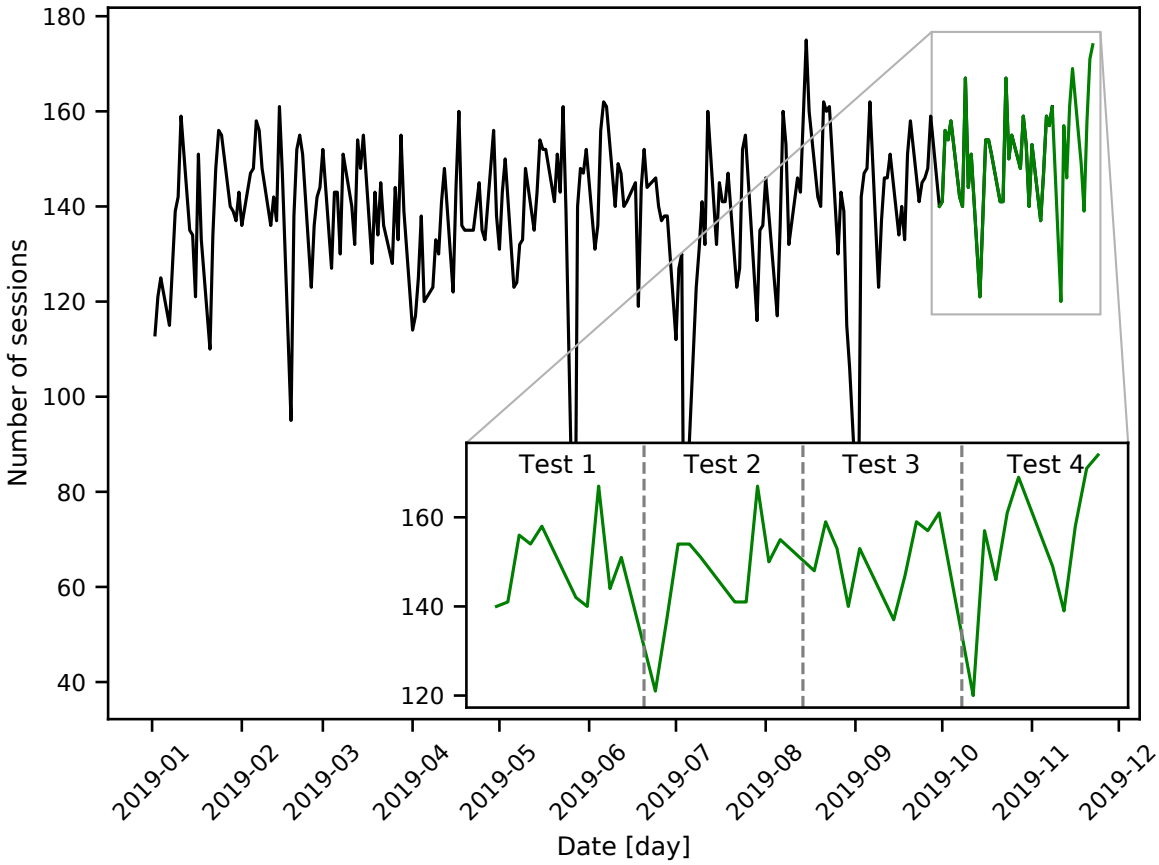


Figure 3.3 – Palo Alto daily number of sessions with the first training period in black and the four test periods in green

EV charging demand [216] with a piecewise constant intensity depending only on the hour of the day. Here, we also assume that the arrivals of EVs at the EVSEs (electric vehicle supply equipment) in scope are the realisation of a NHPP.

NHPP estimated with GAM Following [206], we model $\lambda(t) = \lambda_{\boldsymbol{\beta}}(t)$ with a GAM to capture the smooth variations of the rate of arrivals with time :

$$\log(\lambda_{\boldsymbol{\beta}}(t)) = \boldsymbol{X}^T \boldsymbol{\beta}_0 + \sum_{k=1}^{K} \sum_{d=1}^{D_k} \beta_{kd} b_{kd}(t) \tag{3.1}$$

with basis functions $b_{kd}(.)$ to be specified (typically spline) and coefficients $\boldsymbol{\beta} = \left(\boldsymbol{\beta}_0, (\beta_{kd})_{(k=1,d=1)}^{(K,D_k)}\right)$ to be estimated. K is the number of smooth effects we want to have representing λ and D_k is their dimension. Estimating the $\boldsymbol{\beta}$ coefficients can be done via maximum likelihood. In general, the likelihood of a point process defined on $[0, T]$ can be written in the following fashion [217] :

$$L(\boldsymbol{\beta}) = \left[\prod_{i=1}^{N(T)} \lambda_{\boldsymbol{\beta}}(t)\right] \exp\left(-\int_0^T \lambda_{\boldsymbol{\beta}}(u) du\right) \tag{3.2}$$

with $N(T)$ the total number of arrivals observed on $[0, T]$. This general form of the likelihood is intractable. To solve this issue, we can assume that the intensity function is piecewise constant. Therefore, the integral term of the likelihood can be transformed into a sum on all timesteps : $\int_0^T \lambda(u)du = \sum_{t\in\mathbb{N}\cap[1,T]} \lambda(t)$. We assume that λ is expressed in the same unit as the timestep so that the summation index $\mathbb{N}\cap[1, T]$ does not induce any multiplication by the timestep within the sum (implicitly multiplied by 1).

Smooth estimates can be obtained by adding the following penalty to the log-likelihood : $\frac{1}{2}\sum_{k=1}^{K} \rho_k \boldsymbol{\beta}_k^T \boldsymbol{S}_k \boldsymbol{\beta}_k$ with $\boldsymbol{\beta}_k = (\beta_{kd})_{d\in\{1,\dots,D_k\}}$, ρ_k is the parameter which controls the smoothness of the k^{th} spline basis and S_k is the smoothing penalty matrix composed of the basis functions of the k^{th} spline basis, $\forall k \in \{1, \dots, K\}$

Finally, we obtain the following penalised log-likelihood :

$$l_p(\boldsymbol{\beta}) = \log L(\boldsymbol{\beta}) - \frac{1}{2}\sum_{k=1}^{K} \rho_k \boldsymbol{\beta}_k^T \boldsymbol{S}_k \boldsymbol{\beta}_k \tag{3.3}$$

NHPP estimated with RF A more direct way to estimate the intensity function of a Poisson Process is to perform an approximation of $\lambda(t)$ before estimating it by minimising a loss function (here the square loss). We have chosen to use a centered moving average of the number of arrivals per minute. This approximation of the intensity function which we write $\hat{\lambda(t)}$ can then be estimated via virtually any machine learning model. Here, we chose a RF model as it requires few parameters to be tuned and represents a good candidate for a benchmark.

Mixture Model

A mixture model represents a population as a set of sub-populations. Each sub-population is modelled by a specific component (or distribution). The target vector characterising the individuals are known, however the components to which they belong is unknown. We assume that the random target vector follows a mixture of multivariate lognormal or Gaussian distribution. Formally the density of the target vector $\boldsymbol{Y}$ is defined as follows :

$$p(\boldsymbol{Y}) = \sum_{k=1}^{K} \pi_k \phi_k(\boldsymbol{Y}, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{3.4}$$

with K the number of components, π_k the mixture component weight and ϕ_k the probability density function of component k with parameters $\boldsymbol{\theta}_k = (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$. The parameters of the joint distribution $\pi_k, \boldsymbol{\theta}_k$ are estimated via the EM (Expected-Maximisation) algorithm. The only hyperparameter to calibrate is the number of components K of the mixture.

Mixture Regression

A mixture regression uses a mixture model to derive the parameters of the distribution of a subset of variables conditionally to another (the arrival times in our case). From the mixture model estimated and in the case of Gaussian and lognormal mixtures, a closed form of the conditional distribution density can be directly used [218]. In

particular, we are interested in $\mathbb{E}[(D,E)|A]$ for the load and $\mathbb{E}[P|A]$ for the occupancy. Let us focus on the Gaussian load case for simplicity. For the k-th component of the mixture, we separate the mean vector $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma}_k$ as follows :

$$\boldsymbol{\mu}_k = [\boldsymbol{\mu}_{k,(D,E)}, \mu_{k,A}] \text{ and } \boldsymbol{\Sigma}_k = \begin{pmatrix} \sigma_{k,A} & \Sigma_{k,(D,E)A} \\ \Sigma_{k,A(D,E)} & \Sigma_{k,(D,E)} \end{pmatrix} \tag{3.5}$$

with the mean and covariance matrices associated to each subset of variables. Therefore, and building up on [219], we can now write an explicit expression for the conditional expected value :

$$\mathbb{E}[(D,E)|A] = \sum_{k=1}^{K} h_k(A)(\boldsymbol{\mu}_{k,(D,E)} + \Sigma_{k,(D,E)A}\Sigma_{k,A}^{-1}(A - \mu_{k,A})) \tag{3.6}$$

with, $h_k(A) = \frac{\pi_k \phi_k(A,\mu_{k,A},\sigma_{k,A})}{\sum_{k'=1}^{K} \pi_{k'}\phi_{k'}(A,\mu_{k',A},\sigma_{k',A})}$. It is this expression that we use to get the predicted value of the target vector.

3.3.3 Aggregation of experts

Various EV load/occupation models have been proposed in the previous sections. Each of these models rely on specific hypotheses, producing a set of diverse forecasts. This is a situation where it can be beneficial to produce an ensemble forecast [220] where the idea is to aggregate these forecasts into a single one, which is hopefully better than each individual forecast. As our data are observed sequentially, we consider the framework of online aggregation of experts presented in [221, 222]. The objective is to forecast a bounded sequence of observations $Y_1, \ldots, Y_T \in [0, B], B > 0$ (here the occupancy and/or the consumption). At each time t, N experts provide forecasts of Y_t, denoted $\left(\hat{Y}_t^1, \ldots, \hat{Y}_t^N\right) \in [0, B]^N$. These experts are classically the outputs of a statistical forecasting model or a numerical model. In our study it corresponds to the different prediction models previously presented. Based on past expert forecasts and observation, the aggregation algorithm computes weights $\hat{p}_{j,t} \in \mathbb{R}^N$, and returns as forecast for Y_t a weighted average $\hat{Y}_t = \sum_{j=1}^{N} \hat{p}_{j,t}\hat{Y}_t^j$ of the N forecasts. Then, Y_t is revealed and instance $t+1$ begins. Performance of experts and aggregation forecasts are evaluated according to a convex loss function, here the square loss $\ell_t(x) = (Y_t - x)^2$ -note that another option could have been a Poisson loss for the occupancy. Each time t, expert k suffers loss $\ell_t(\hat{Y}_t^k) = (Y_t - \hat{Y}_t^k)^2$ and the aggregation $\ell_t(\hat{Y}_t) = (Y_t - \hat{Y}_t)^2$. The purpose of aggregation algorithms is to minimise the total loss $\sum_{t=1}^{T}(Y_t - \hat{Y}_t)^2$ over the forecasting period that can be expressed $\frac{1}{T}\sum_{t=1}^{T}(Y_t - \hat{Y}_t)^2 = \frac{1}{T}\sum_{t=1}^{T}(Y_t - \hat{Y}_t^*)^2 + R_T$, where R_T is the regret term corresponding to the error suffered by our algorithm relatively to the error of the oracle $\hat{Y}_t^*$, an optimal but unknown before the forecasting run forecast. A lot of algorithms are proposed to achieve low regrets [223]. In our study we use the ML-Poly algorithm proposed in [224] and implemented in the `R` package `opera` [225]. This algorithm tracks the best expert or the best convex aggregation of experts by giving more weight to an expert that will generate a low regret. This makes this algorithm particularly interesting as no parameter tuning is needed.

3.4 Experiments

In the following section, the experimental procedure is described and the associated results are shown. We first describe how the data was preprocessed before training the models, then we detail which procedures were used for model selection and validation before comparing the performance of each model as well as the aggregation of all forecasts.

3.4.1 Data preprocessing

Following the data quality assessment of the datasets performed in [200] we have only kept charging sessions which respect the following criteria : the charge duration has to be positive and shorter than 24 hours (if the charge duration is not available then the park duration is used instead); energy demand also needs to be positive and under 100 kWh. Charging sessions lasting for more than a day are likely to be data collection errors as an EV is usually charged in a few hours. As for the 100 kWh upper bound, it corresponds to the battery capacity of the Tesla Model S which is the largest amongst top selling EVs.

Furthermore, as mentioned in Section 3.2, we have only kept working days in the data as behaviours are significantly different on weekends and there are less charging sessions for some of the datasets in scope. Finally, as shown on Figure 3.4, there are periods during which arrivals are rare or non-existent. Therefore, we have set a threshold value for each dataset under which the corresponding times are discarded from the analysis. More precisely, we keep the largest continuous time interval where all values are above the threshold. This time interval is shown as a grey shaded rectangle on Figure 3.4. The threshold is set at 7.5% of the maximum number of arrivals and is shown on the plots by a red dotted horizontal line. We found that this preprocessing helps our models to better learn the intensity function of the NHPP.

3.4.2 Model selection

Both statistical and machine learning models considered in this benchmark require fine tuning in order to exploit them at their best but also to avoid one common pitfall : overfitting. Therefore, we have chosen to have a careful selection procedure for each dataset in scope because of the wide variety of EV charging behaviours they represent.

Information criterion

To automate the process of finding the optimal orders of the SARIMA models, we fit all models with a maximum of 10 total orders to finally keep the model which minimises the Bayesian Information Criterion (BIC) introduced in [226]. This routine is implemented in the *forecast* package detailed in [227]. The BIC helps to keep the model complexity to a minimum while also minimising its prediction error.

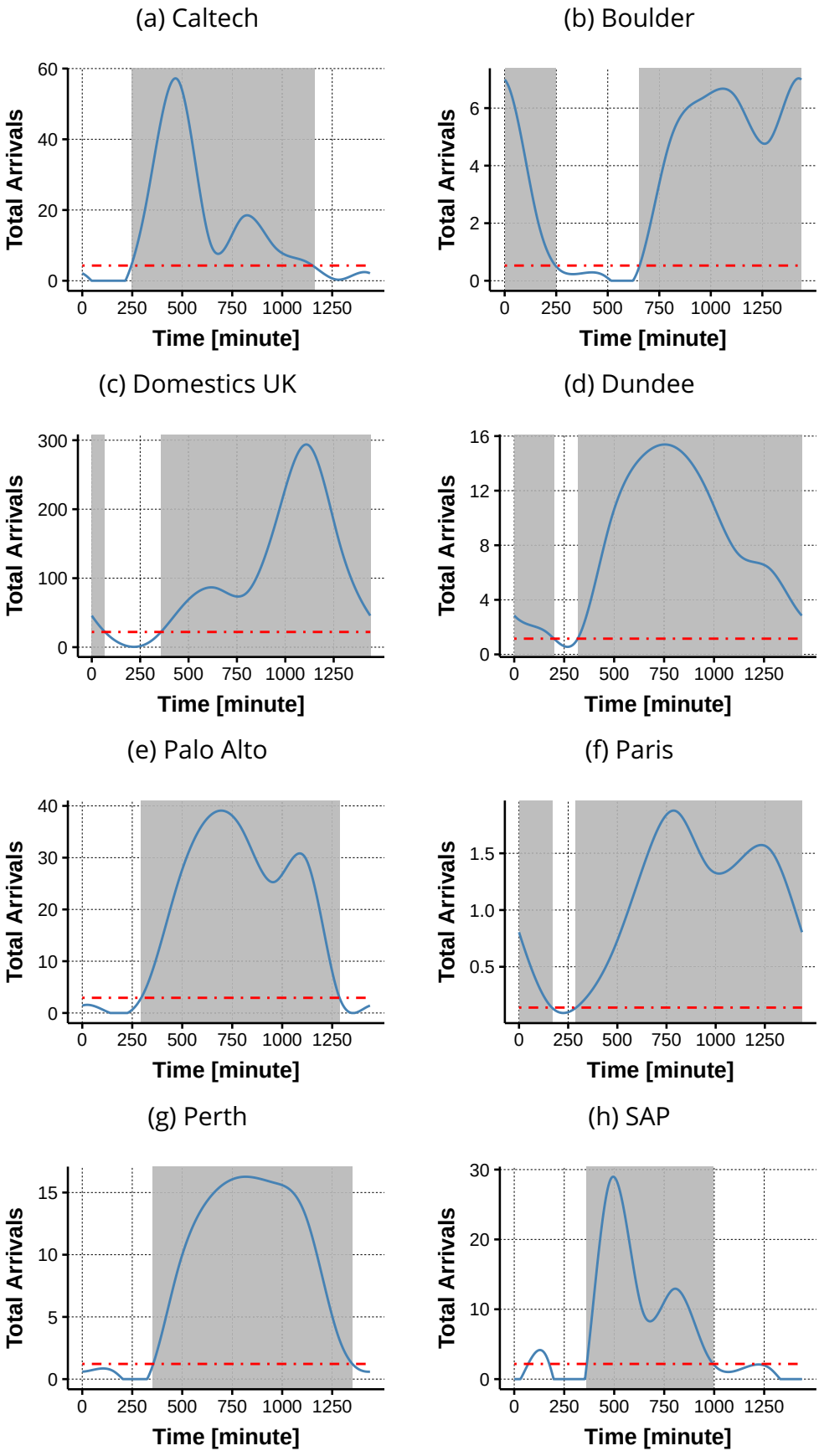
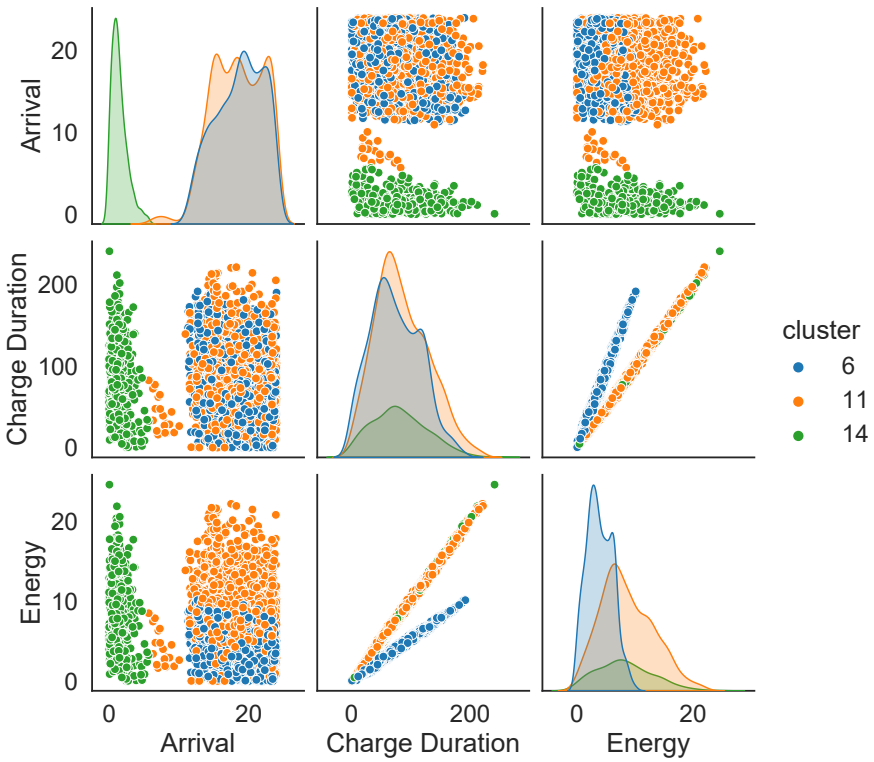


Figure 3.4 – Smoothed total arrivals per minute of the day over the first training period for all datasets considered



Cluster	Weight	$\bar{a}$ [h]	$\bar{c}$ [min]	$\bar{e}$ [kWh]
6	0.2	18.75	72.93	3.87
11	0.24	18.4	83.98	8.48
14	0.07	1.51	80.14	8.12

Figure 3.5 – Gaussian MM plots for components above 5% (Boulder)

With regards to the MM, the only hyperparameter to tune is the number of components k of the mixture. Therefore we fit all MM for $k \in \{2 \dots 30\}$ on the training data for each dataset and we keep the one that minimises the BIC. 30 components might seem a lot for a mixture model but we've found that for most datasets the optimal number of components is between 20 and 30. That can be explained by the fact that we are looking at multivariate mixtures for an EV demand that is still in its infancy and subject to significant variations in behaviours over time. Also, it is important to note that even if many components are kept with this procedure, it usually takes only 10 components to make up for most of the weights.

An example of this is shown on Figure 3.5 with the Gaussian MM fitted on the Boulder data. Here we only keep the clusters that have a weight above 5%. This leaves us with three clusters that make up more than 50% of the components weights. In particular, these three clusters can be clearly distinguished. Cluster 6 and 11 correspond to late afternoon/evening arrivals with similar charge durations but with lower average energy for cluster 6 ($\bar{e} = 3.87$) than for cluster 11 ($\bar{e} = 8.48$). This can be interpreted as vehicles being connected to EVSEs at different power levels. In fact, this corresponds to the two most widely spread alternative current chargers delivering respectively maximum po-

wers of 3kW and 7kW. Cluster 6 and 11 already make up 44% of the MM. Finally, cluster 14 is also clearly distinguishable from the two other clusters as its mean arrival time is at around 1.30AM ($\bar{a} = 1.51$). With a weight of 7%, it represents a minor sub-population of users which charge later at night and benefit from higher power levels as for cluster 11.

Grid search

For each dataset, the first training period is split into a training and validation set. In particular, the first 80% observations are kept in the training set while the last 20% form the validation set. That way, the time dependence of the data is respected. The best hyperparameters are found via grid search and we keep the model which minimises the loss incurred on the validation set. GRU, NHPP-RF, NHPP-GAM, GAM and RF model hyperparameters are optimised through that process.

The GRU hyperparameters tuned are the batch size, width/depth of the hidden layers and the learning rate. For the NHPP-RF and RF models the number of trees as well as the number of predictors used to determine each split are optimised. In addition, the GAM and NHPP-GAM models are also tuned through this grid search process. In particular, the choice of predictors as well as the number of spline functions for the smooth effects are calibrated. The final sets of hyperparameters obtained are available in appendix B.3.

Error correction

Until now, even though some of the bottom-up approaches used lags of the number of arrivals no model used lagged information of the target (load or occupancy). To account for this lack of information in the models and by taking advantage of the day-ahead forecasting set-up of this study, an error correction model is proposed to adjust each model forecast. It consists of a SARIMA model on the residuals of the forecast. More precisely, one SARIMA model is trained on each minute of the day to respect the day-ahead forecasting set-up. In other words, a total of 1440 SARIMA models are trained (one per minute of the day) and each of these model has access to the information of the previous days at the same minute. Again, the orders of the SARIMA model are chosen by minimising the BIC. As the hypothesis of the SARIMA models used are to have Gaussian residuals, the load and occupancy forecasts obtained after this correction can be negative and decimal numbers. Therefore, the error correction forecast is rounded for the occupancy and put to zero if negative for both the occupancy and the load to ensure consistency with the phenomenons described. The error correction forecasts will be identified in the following paragraphs as *Model*-err. For instance, "GAM-err" is the forecast produced by the GAM model combined with the error correction model we have detailed.

3.4.3 Model validation

A model validation procedure is also performed to check the statistical assumptions made for statistical models in scope. Namely, NHPP, SARIMA and GAM models.

A student test for NHPP

Following the NHPP validation procedure described in [216], we use a t-test to check whether the proposed NHPP models are valid. We split the data in training and testing sets as in Section 3.4.2. 100 versions of the process are simulated on the validation set from the intensity function estimated on the training set. The validation set is separated in J equal time windows. For the k-th simulation we define

$$\bar{G}_k = \frac{1}{J}\sum_{j=1}^{J} \hat{D}_{k,j} - D_{k,j},\ k \in \{1 \dots 100\} \tag{3.7}$$

with J the total number of time windows over the validation set, $\hat{D}_{k,j}$ the total number of arrivals forecasted in the j-th time window and $D_{k,j}$ the total number of arrivals observed in the j-th time window. Our model is valid if the sample $\bar{G}_k,\ k \in \{1 \dots 100\}$ is normally distributed with a mean $\mu = 0$. First, we check the sample normality with a Shapiro-Wilk normality test. Second, a t-test is formulated to check whether μ is significantly non-zero :

$$\begin{cases} H_0 & \mu = 0 \\ H_1 & \mu \neq 0 \end{cases} \tag{3.8}$$

We define the following statistic $t_{N-1} = \frac{\sqrt{N}(\bar{G}-\mu)}{s_g}$ with $\bar{G} = \frac{1}{N}\sum_{k=1}^{N} \bar{G}_k$, s_g the sample variance and $N = 100$ the number of processes simulated. Under the null hypothesis, t_{N-1} follows a Student distribution with $N-1$ degrees of freedom. We chose a significance level of 5% to assess whether the value obtained for the sample mean $\bar{G}$ is beyond the acceptable range and therefore whether H_0 is rejected. We take a 15-minute time interval for aggregating the arrivals as it is becoming the standard time step for electrical signals meter readings. However, please note that we led the same validation procedure for windows of 1 hour and 1 day and the results remained the same. For all models considered the sample mean was well within the acceptable range of the test and therefore all NHPP models which came out of the model selection step were valid. The values of the sample means as well as the acceptable range of the Student test for each relevant model and for all datasets can be found in Appendix B.4.

Ljung-Box test for SARIMA

The Ljung-Box test is commonly used for assessing whether there is a lack of fit for time series models [228]. The null hypothesis is defined as $H_0 : \forall\, l \in \{1 \dots L\}\, \rho(l) = 0$ with ρ the auto-correlation function of the time series model residuals. We define the following statistic for a sample size n :

$$Q(L) = n(n+2)\sum_{l=1}^{L} \frac{r_l^2}{n-l} \tag{3.9}$$

where $(r_l)_{l \in \{1 \dots L\}}$ are the sample auto-correlations and L the maximum time lag. Under H_0, $Q(L)$ asymptotically follows a $\tilde{\chi}^2_L$ distribution. Common practice is to set L to be twice the natural seasonality of the series. Therefore, we chose $L = 10$ to consider a two-week period for the test. If the p-value of the test is above 5%, then we cannot

reject the null hypothesis of the test and we consider that our model does not show lack of fit. The models that came out of the model selection procedure all had a p-value of above 5% at the exception of the model trained on the Perth dataset. Thus we cannot reject the null hypothesis for 7 out of the 8 datasets in scope. The low p-values obtained for some of the datasets can be explained by a strong trend in the data as the market adoption is evolving at a fast pace which prevents the time series modelled from being stationary. Nevertheless, the great majority of the models passed this test and it is reasonable to compare their performances. The p-values of the Ljung-Box test for the SARIMA models and for all datasets are shown in appendix B.4.

Residual analysis for GAM

In addition to selecting the GAM formula via grid search on different candidates, an analysis of the residuals was applied to check whether the assumptions of the model were satisfied. A residual analysis applied to the first training period of the Palo Alto load dataset is shown in Figure 3.6. From left to right, the first plot shows the evolution of the residuals over time, the second one shows a histogram of the residuals and the last one, the load values against the fitted values.

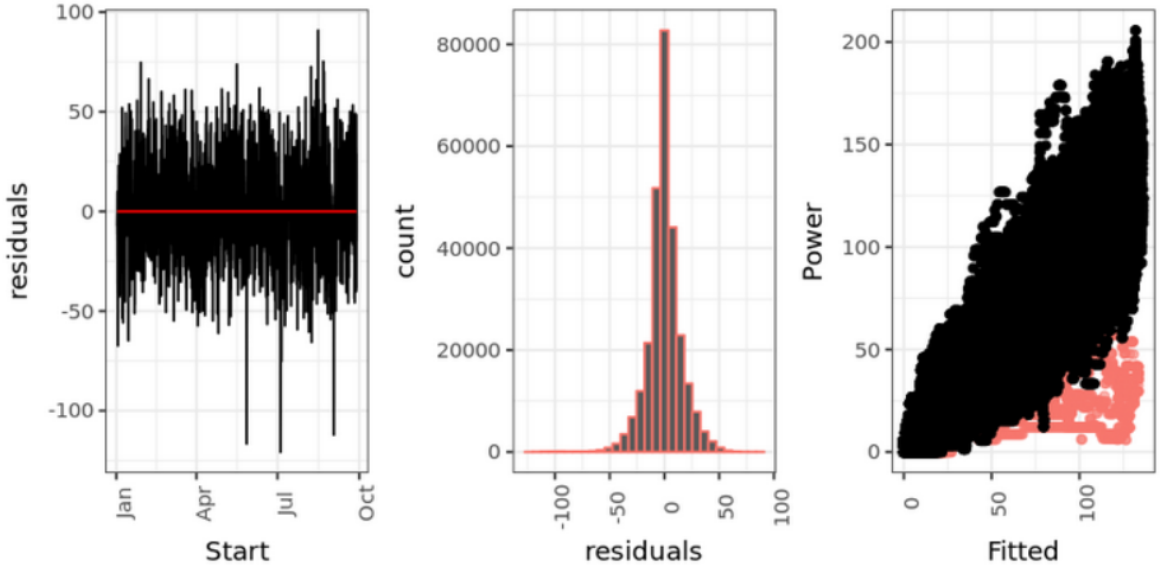


Figure 3.6 – Palo Alto residual analysis for the load forecasting using GAM

The outliers, visible in the three plots, and more evidently in red in the third one, are holidays forecasts. These days are mainly overestimated and are responsible for skewing the histogram to the left. Their removal did not impact the overall performance and the choice was made to keep them in the datasets. Other than the outliers, the model assumptions seem to hold as the residuals have a normally shaped distribution with a mean equal to zero.

3.4.4 Model performance

Traditional metrics for evaluating the forecasting performances of load forecasting models are the Mean Absolute Error (MAE), Root-Mean Squared Error (RMSE) and Mean

Absolute Percentage Error (MAPE). The latter is not an option in our case as the target can often be zero which creates notorious problems for the MAPE [229]. The MAE and RMSE could have been used as zeros are not a problem for these metrics. However, with 8 datasets in scope recording various charging behaviours across multiple countries, the scale of each dataset varies tremendously.

Comparing these metrics across datasets would require scaling the data as performed in [230]. Here, we define a metric which will be easily interpreted across all datasets. This metric is the Percentage Improvement from Persistence (PIP). As its name suggests, the PIP is a metric expressed in percentage which quantifies the forecasting improvement of a proposed model over the persistence. More formally we can write the PIP for a given forecast signal $\hat{y}$ as follows :

$$\text{PIP} = 100 \times \left(\frac{l(y, y^{pers})}{l(y, \hat{y})} - 1 \right) \tag{3.10}$$

with y the observed signal, y^{pers} the persistence forecast signal and l the loss function. Here, we take l as the MAE or the RMSE. Essentially, the percentage improvement is expressed in terms of a ratio between the MAE (resp. RMSE) of the persistence forecast and the MAE (resp. RMSE) of the forecast signal considered. When the PIP is positive, the proposed forecast is more accurate than the persistence on the chosen metric. For simplicity, we now refer to the MAE (resp. RMSE) as the PIP metric based on the MAE loss function (resp. RMSE loss function).

Following a similar procedure than the one described in [231] we have used block-bootstrap resampling to quantify the variability of the PIP metric across all datasets. This a robust procedure that enables a more holistic comparison of model performances. The test set of size n is separated into equal chunks of data of size $S = 60 * 24$ (equivalent to one day). We sample with replacement these chunks until we get a set of size n. Then we can calculate the PIP on this newly created dataset. By running this procedure $K = 250$ times, K metrics are therefore calculated. The sample mean, median and variance of this sample provides more information to assess the quality and robustness of the model forecasts.

Boxplots of the block-bootstrap forecast errors of each methods are presented on Figures 3.7 and 3.8, means on Table 3.2. It shows that the best forecast is produced by the GAM-err for the load and Aggregation of Experts (AGG) for the occupancy (note that AGG followed GAM-err very closely for the Load). This is really in favour of AGG as the weights of the aggregation are estimated sequentially from the beginning of the test set and it is likely that better results can be achieved on longer forecasting runs (see Table B.8 in Appendix B.5 where we show the MAE of AGG on the last test period is the best on block-bootstrap average).

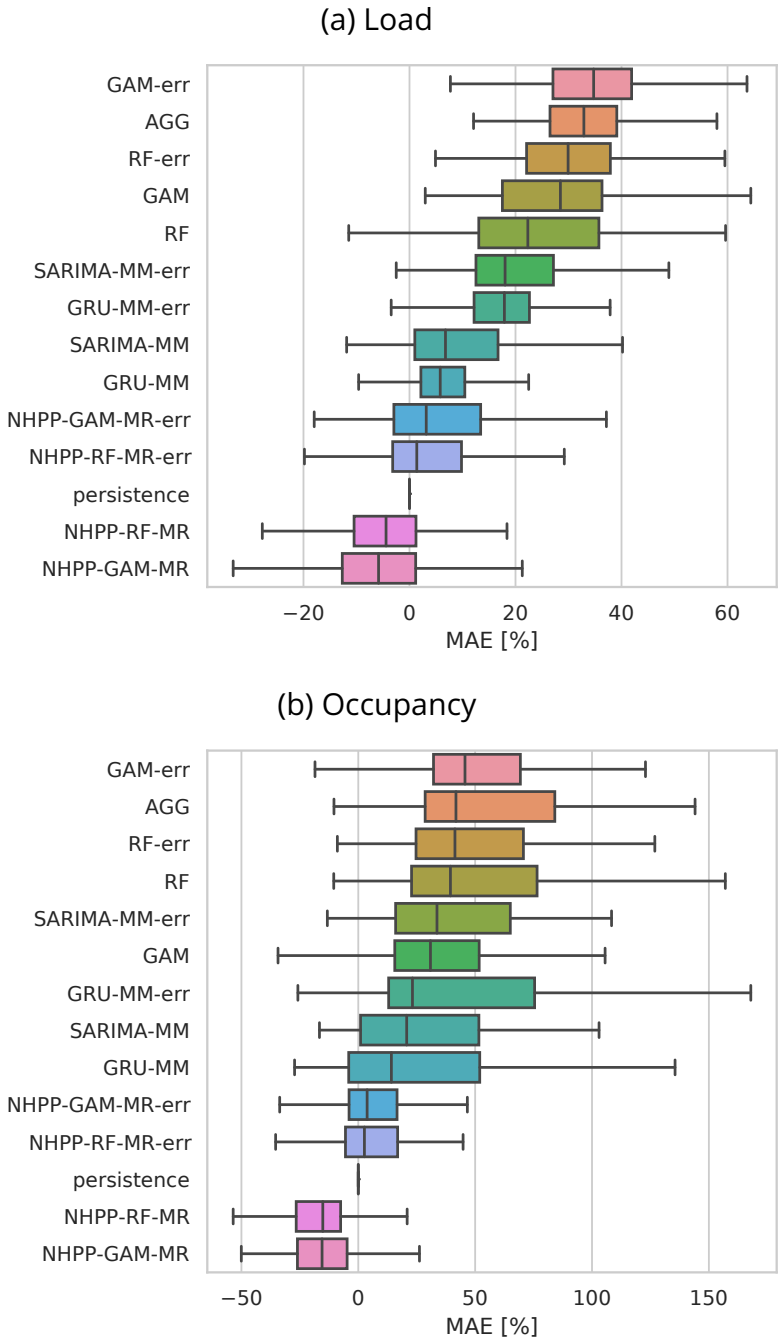


Figure 3.7 – Boxplots of the (PIP) MAE block-bootstrap performances on all datasets and for all models considered

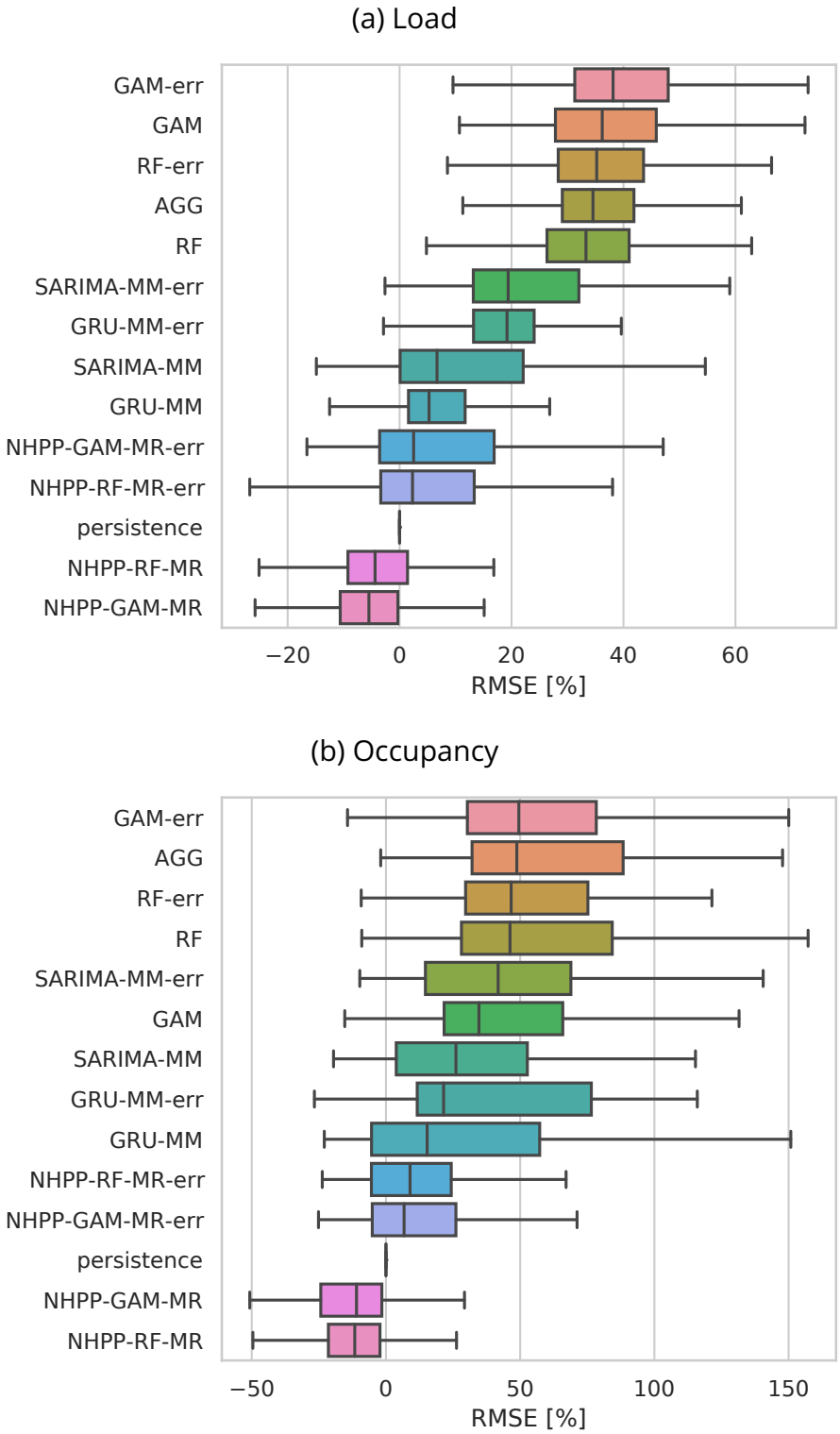


Figure 3.8 – Boxplots of the (PIP) RMSE block-bootstrap performances on all datasets and for all models considered

Overall, the direct approaches obtained better results than bottom-up ones. As expected, the direct approach benefits from a "law of large numbers effect" where sum-

mation of individual events smooths the signal which becomes easier to forecast. Nevertheless, considering the difficulty induced by the bottom-up approach paradigm, we can note the relatively good performances of MM based ones. The metrics obtained by the SARIMA-MM on the occupancy makes it the best model before error correction and aggregation (see Table 3.2 (b)). Error correction based on ARIMA models always brings significant improvement. NHPP methods do not perform well for neither load nor occupancy and we believe it could come from the MR part which connects arrival to the other components of the target vectors. In fact, bottom up approaches perform better for occupancy forecasting than load forecasting.

Table 3.2 – Average block-bootstrap model performances for the PIP (MAE and RMSE)

(a) **Load**

Approach	**Model**	**MAE [%]**	**RMSE [%]**
Direct	GAM-err	35.5	40.9
	RF-err	30.4	37.1
	GAM	28.4	37.8
	RF	23.7	34.4
Bottom-up	SARIMA-MM-err	20.6	23.2
	GRU-MM-err	17.5	18.6
	SARIMA-MM	9.83	11.4
	GRU-MM	6.75	7.24
	NHPP-GAM-MR-err	5.63	6.98
	NHPP-RF-MR-err	3.99	5.89
	NHPP-RF-MR	-5.21	-3.33
	NHPP-GAM-MR	-6.45	-4.73
	AGG	33.3	36.4

(b) **Occupancy**

Approach	**Model**	**MAE [%]**	**RMSE [%]**
Direct	RF-err	71.1	97.7
	GAM-err	56.4	68.2
	RF	55.2	74.6
	GAM	34.1	43.5
Bottom-up	SARIMA-MM-err	68.0	97.6
	SARIMA-MM	56.0	84.0
	GRU-MM-err	51.2	67.0
	GRU-MM	28.9	40.0
	NHPP-GAM-MR-err	23.3	36.9
	NHPP-RF-MR-err	20.5	35.8
	NHPP-GAM-MR	-10.0	-0.53
	NHPP-RF-MR	-14.2	-0.86
	AGG	92.6	125

AGG performances are good due to the diversity of these approaches and all of the different models have a contribution in terms of weights. This is shown in Table 3.3 where the average weights (over time and data sets) for each method are presented. For load forecasting (resp. occupancy), the direct approaches (including the persistence benchmark) achieve a total weight of 0.4 (resp. 0.46) and bottom up approaches contribute to 0.6 (resp. 0.54) in the aggregation.

Figure 3.9 shows the average weight (in time) of each forecast for each data set. We can see a particular distribution of the weights for Domestics UK where the NHPP-RF-MR-err achieves significantly more important weights than for other data sets. Inversely, GAM and GAM-err weights are very low for this data set. A zoom of individual performances on Domestics UK is presented on Table 3.4. We see that the improvement brought by AGG is of more than 5 times compared to the persistence.

Table 3.3 – Average weight in the aggregation for all relevant models and across all datasets

(a) **Load**

Model	**Weight**
GAM-err	0.12
GAM	0.11
RF	0.10
RF-err	0.10
GRU-MM-err	0.07
NHPP-RF-MR-err	0.07
SARIMA-MM-err	0.07
persistence	0.07
GRU-MM	0.06
NHPP-RF-MR	0.06
SARIMA-MM	0.06
NHPP-GAM-MR	0.05
NHPP-GAM-MR-err	0.05

(b) **Occupancy**

Model	**Weight**
RF	0.11
GAM	0.10
GAM-err	0.10
RF-err	0.09
GRU-MM	0.08
GRU-MM-err	0.08
SARIMA-MM	0.08
SARIMA-MM-err	0.08
NHPP-RF-MR-err	0.07
NHPP-RF-MR	0.06
persistence	0.06
NHPP-GAM-MR	0.05
NHPP-GAM-MR-err	0.05

Also, bottom-up approaches and MM in particular are better than direct ones here. Domestics UK being the dataset with the most observations, MM has access to more data for fitting the multivariate distribution and adjusting the clusters than for other datasets. Also, this dataset is set on a fixed population of individuals in the UK, which means that we will definitely observe multiple charging sessions from the same individual. For the public charging station datasets this is not at all guaranteed. Having a fixed population removes the noise brought by one-off users who might have peculiar behaviours and prevent reaching the best possible fit for MM.

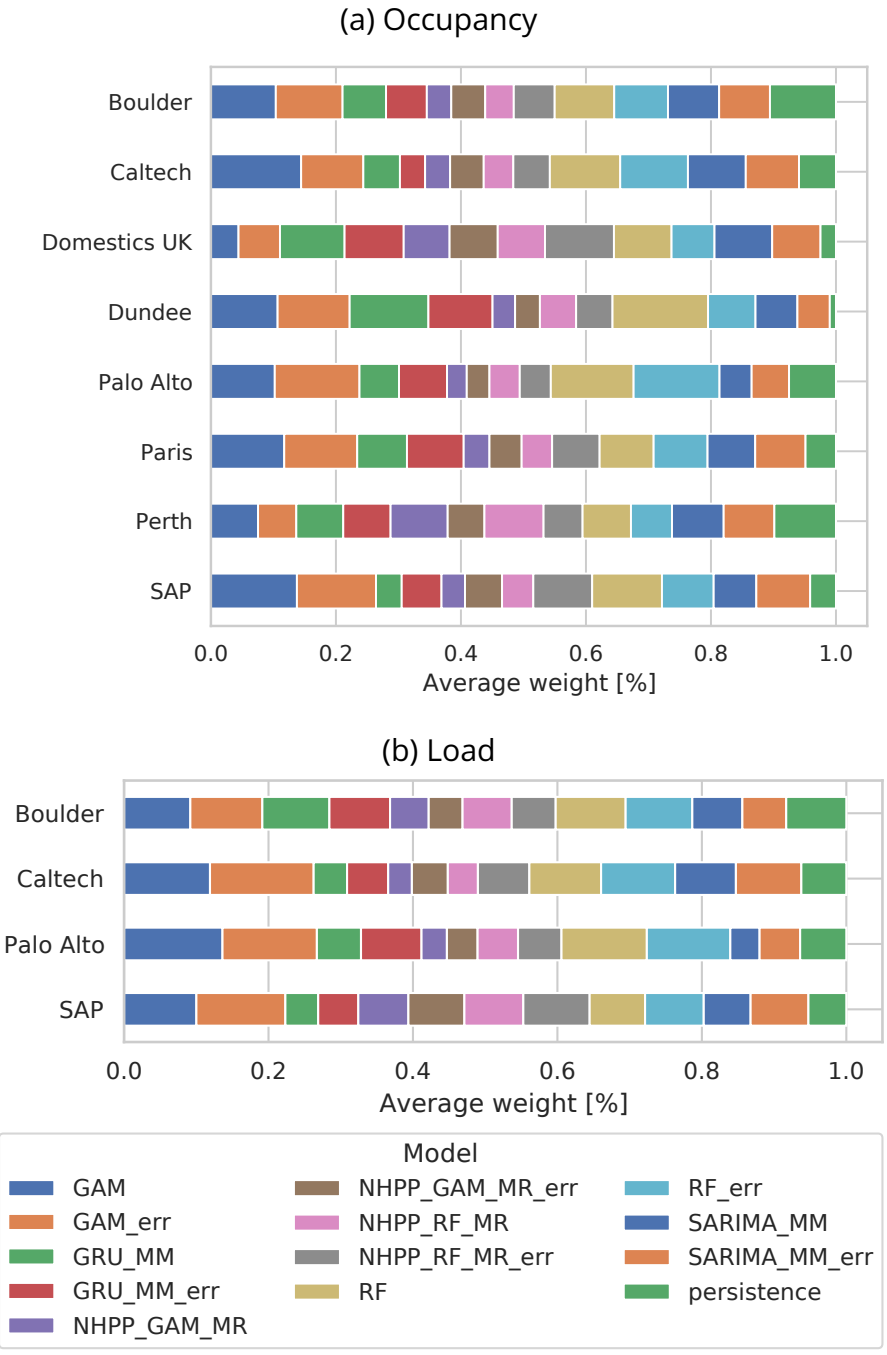


Figure 3.9 – Average aggregation weights over the test periods for all datasets

Table 3.4 – Average block-bootstrap model performances for the PIP (MAE and RMSE) on the Domestics UK dataset

Model	MAE [%]	RMSE [%]
AGG	402	628
SARIMA-MM-err	309	524
SARIMA-MM	301	508
RF-err	291	463
GRU-MM-err	176	310
NHPP-GAM-MR-err	162	247
GAM-err	149	214
NHPP-RF-MR-err	148	242
RF	146	258
GRU-MM	105	194
NHPP-GAM-MR	48	99
NHPP-RF	24	95
GAM	23	62
persistence	0	0

3.5 Conclusion

This paper presented an extensive benchmark of recent methods to forecast load and occupancy of EV charging infrastructures. We cover 8 open datasets chosen for their quality and to represent the diversity of charging behaviours (residential, workspace, public stations). We showed significant differences in performance between the direct approaches and the bottom up ones. Direct approaches perform better for most of the datasets and this is particularly true for load forecasting. Amongst bottom-up methods, mixture models coupled with time series model were the best ones. We exhibit the particularity of Domestics UK for which bottom-up approaches outperformed direct ones.

We proposed an aggregation method which takes advantage of the diversity of the different approaches developed. We showed that all methods contributed to the aggregation and that even if their individual performances were not the best, bottom-up approaches contribute largely in the aggregation. Future work includes improving NHPP models performances by refining either the MR part or the estimation of the intensity. For the MR part, we believe that adding more covariates to the regression might help in better capturing the dynamics between the arrival times and the other components of the target vector. With regards to the intensity of the NHPP, wavelet projection could help capture drastic changes that can occur at this fine resolution. As for the aggregation of experts, we could try other types of losses (e.g., Poisson loss for the occupancy). Also, other popular machine learning models such as boosting trees could be tested instead of RF.

In addition, this benchmark put an emphasis on point forecasts. Because of the probabilistic nature of the bottom-up approaches, another track for future work would be to assess probabilistic forecasts in the context of EV load and occupancy. In particular,

estimating the load/occupancy peaks with quantile forecasts would be possible with the methods detailed in this paper and highly valuable for various stakeholders of the EV industry (e.g., EVSE manufacturers, independent system operators). In the hope of fostering reproducible research in the field, the R and Python codes used to produce the benchmark will be made available.

Chapitre 4

Modelling the arrivals of EVs with non-homogeneous Poisson processes

In this chapter, we introduce an additive model using both spline and wavelet effects for fitting the intensity of a non-homogeneous Poisson process (NHPP). After giving some background on the model, we propose a novel estimation procedure inspired from backfitting which is illustrated by a case study on real-world EV arrivals at charging points. The idea behind this modelling approach is to assess whether arrival peaks at charging points can be better captured by combining both spline and wavelet effects.

Summary

4.1 Introduction

Non-homogeneous Poisson Processes (NHPPs) have been used to model a diverse range of phenomenons in the literature : climate change (e.g., number of excesses over threshold [232]), seismology (e.g., earthquakes [233]), imaging (e.g., tomography [234]), extreme weather events (e.g., seasonal rainfall extremes [235]), energy (e.g., electric vehicle charging demand [216]) and many more. Unlike other point processes counterparts (e.g., Hawkes and Aalen [236]) which are more restrictive in their specification, the NHPP lets its intensity function $\lambda(t)$ depend on time quite freely. The intensity function uniquely determines a point process.

To learn the intensity function of such processes, various methods ranging from parametric models to non-parametric models such as kernel density estimators (KDEs) have been considered. Recently, additive semi-parametric methods have raised ample interest for estimating the intensity of NHPPs. In particular, models inspired from the abundant literature of additive models have been applied to NHPP intensity estimation. An important focus has been given to spline basis for additive models and many theoretical and experimental results are available. In the meantime, progress has also been made on regression with additive wavelet effects. An overview of these methods is proposed in Section 2.

In this chapter we consider an additive model of the intensity function of a NHPP of the following form :

$$\log \lambda(t) = \beta_0 + \sum_{l=1}^{L_p} \beta_l x_l^p(t) + \sum_{l=1}^{L_s} s_l(x_l^s(t)) + \sum_{l=1}^{L_w} w_l(x_l^w(t)) \tag{4.1}$$

with, β_0 the intercept, $\beta_\cdot$ the coefficients of the linear component, $s_\cdot$ and $w_\cdot$ respectively spline and wavelet basis expansions. In addition, $\boldsymbol{x(t)} = (\boldsymbol{x}^p(t), \boldsymbol{x}^s(t), \boldsymbol{x}^w(t))$ the vector of covariates evaluated at time $t \in [0, T]$ ($T \in \mathbb{R}^+$ being the final time at which we observed the NHPP). $\boldsymbol{x}^p(t) = \{x_l^p(t)\}_{\{1,\dots,L_p\}}$, $\boldsymbol{x}^s(t) = \{x_l^s(t)\}_{\{1,\dots,L_s\}}$ and $\boldsymbol{x}^w(t) = \{x_l^w(t)\}_{\{1,\dots,L_w\}}$ are the vectors of covariates respectively used for the linear, spline and wavelet effects. The idea behind this model is to decompose the signal into linear, smooth and irregular components. Also we chose the additive structure to have an interpretable model with the contribution of each component made clear for analysis.

The rest of the paper is divided as follows : firstly, Section 2 presents a review of the related work which inspired our model. In Section 3. we introduce the specificities of the model and the algorithmic procedure proposed for the estimation. Finally, in Section 4. we propose a case study on real data for electric vehicle arrival at charging points.

4.2 Related Work

This section provides an overview of the literature of both NHPP intensity estimation and additive regression models with spline and wavelet effects. It was conducted to identify gaps in the field and led us to the proposed model. Historically, point process estimation has been thoroughly detailed in [237] including NHPP. The likelihood

of a NHPP process with n observed arrival times $\{\hat{t}_i\}_{i\in\{1\dots n\}}$ can generally be written as follows :

$$L = \left(\prod_{i=1}^{n} \lambda\left(\hat{t}_i\right)\right) \exp\left(-\int_0^T \lambda\left(s\right) ds\right) \tag{4.2}$$

Estimators of the intensity function can thus be obtained by maximising L. Another idea worth mentioning is the introduction of a least-square contrast as thoroughly shown in [236] in the case of the Aalen, Hawkes and Poisson processes. In this paper we focus on maximum likelihood estimates of the first-order intensity function for NHPPs. In [238], a semi-parametric estimator of the conditional intensity of temporal point processes was introduced with a Tukey shrinkage procedure. Applications led on neurophysiological and seismological data were conducted. According to the authors, the wavelets were used for their smoothing capability. In our work, we want to leave the smoothing components for the spline basis to then witness more irregular components of the intensity thanks to high order wavelet coefficients. The theoretical properties of wavelet coefficient estimators were later studied in [239] for the first-order intensity of multi-dimensional NHPPs and extended in [240] for second-order intensity functions of non-homogeneous point processes. In particular, the probability density functions of the wavelet coefficient estimators are derived in [239]. Also, an unbiased estimator of the second order moment of the intensity was proposed under both linear and hard thresholding setups.The main results obtained in these two papers are useful in practice for wavelets with a compact support and with an analytical formulation. Essentially, they are only applicable to Haar wavelets in practice. In [241], Meyer wavelets are preferred to Haar wavelets for estimating the intensity function for n independent realisation of a NHPP. While keeping a closed form, these wavelets do not have a compact support. In particular, an adaptive estimator with a hard thresholding procedure was proposed. The authors have shown that when n goes to infinity, a near-minimax rate of convergence can be derived for the proposed estimator. A wider class of wavelets is studied in [242] with biorthogonal wavelets. The main specificity of this class of functions is that they have more degrees of freedom than traditional orthogonal wavelets and allow for different multi-resolution analysis. Precisely, instead of having one mother and one father wavelet, there are actually two of each. In [242], the particular case of biorthogonal spline wavelet basis is developed. As in [241], the authors have shown that the proposed adaptive estimator achieves minimax convergence rate up to a logarithmic term. The thresholding strategy adopted is inspired from the universal threshold proposed in [243]. More recently, [244] and [206] have proposed fitting procedure for the intensity of non-homogeneous point processes with splines basis. Precisely, an additive Poisson process is proposed in [244] which focuses on high-order interactions splines. The idea is to propose an efficient and performant model for correlated stochastic processes. The authors shown that it outperforms its counterparts particularly when facing extremely sparse samples while being more computationally efficient. Finally, in [206] the intensity is modelled as a smooth function of time and space depending on a set of covariates. The formulation proposed by the author is a Generalised Additive Model (GAM) of the intensity and is fitted using the penalised iterative least-square (PIRLS) procedure described in [208]. Their model is applied to windstorm peaks with covariates expanded on thin-plate, cubic and cyclic regression

spline basis. This type of model enhances interpretability with additive effects which is one of our main concerns for industrial applications.

Moving on from point processes, let us now look at semi-parametric regression. With the emergence of backfitting procedures, many modelling framework have been proposed since the 1980s, [245], [207]. They all have in common the projection of covariates on adequate function basis, usually splines as the proof of backfitting convergence was originally proposed for cubic spline smoothers [246]. Partially linear models (PLMs) where only one functional effect is added to a traditional linear predictor have been thoroughly reviewed in [247]. Partially linear additive models (PLAMs) which include more than one functional effect have also been thoroughly reviewed and [248]. A general procedure for fitting semi-parametric regression models with additive functional effects waw proposed in [207] with backfitting. In parallel, a General Cross Validation (GCV) criterion first defined in [249] was used for estimating a single smoothing spline and extended in [250] to multiple smoothing parameters. This result was used to propose a PIRLS fitting procedure for generalised additive models (GAMs) in [251]. Finally, it was further extended and implemented in the R package `mgcv` [208]. While splines have been the basis of choice for additive semi-parametric models, wavelet basis have gained more and more interest over time. [252] proposes to integrate wavelets into semi-parametric regression in a similar way as it is done for splines in generalised additive models [208]. While splines estimated with PIRLS contain a ridge-like penalty, wavelets are given a lasso-type penalty. It is mentioned that other penalties can be applied such as bridge, hard thresholding, smoothly clipped absolute deviation (or SCAD) and minimax concave penalties. The close relationship of the lasso penalty with soft thresholding established in [253] is what makes it our choice in our work. Many papers have studied wavelets for Gaussian PLMs such as [254], [255] and [256]. Also worth mentioning is the work of [257] which presents a maximum likelihood estimation procedure for Poisson regression using wavelet model selection. The first paper to expand wavelet PLMs to the exponential family of distributions and which obtains theoretical guarantees is [258]. In this paper, a PLM is proposed with the functional effect expanded on a wavelet basis. A backfitting algorithm is used to fit both the linear and non-parametric parts. The maximum penalised likelihood estimators proposed for the parametric and non-parametric part achieve near minimax convergence rates. The author established that the Lasso penalty also leads to an adaptive estimation. Finally, they compared their approach with a similar PLM estimated with spline and kernel methods. One key takeaway for our work is that their simulation study found that their proposed wavelet procedure led to worse quality estimates than splines procedure in a Poisson PLM.

On an additional note, we wanted to highlight some of the current implementations. For Poisson regression, We have already mentioned `mgcv` for GAMs in R. There are also the implementations of `glm` in R fitted by iterative least-square (IRLS) and penalised glms implemented in `glmnet` fitted using coordinate gradient descent (CGD) [259]. More details on these algorithms are given in appendix C. These three implementations and the theory behind them are the stepping stones of our proposed algorithm for fitting the model presented in equation 4.1. Also, several packages have been introduced for directly fitting NHPP intensity functions such as [260] for GLMs and [206] for GAMs.

Overall, this literature review showcases that on one hand, papers detailing NHPP

intensity with PLAMs or GAMs are rare and often distinguish usage between splines and wavelets. A hybrid model proposed in [261] did combine backfitted splines and kernel methods. However, very rarely splines and wavelets are combined in the same model specification. The only occurrence of this kind of work we have found is in [248] where a hybrid approach for Gaussian regression PALMs is proposed with an application of this model to electricity demand in [262]. Therefore, in this chapter we want to extend these papers and study more precisely a PLAM of the first-order intensity function of NHPPs as presented in equation 4.1.

4.3 Problem Formulation

In practice, the likelihood presented in equation 4.2 is intractable unless approximations are made. In particular, we can consider that the intensity function is piecewise constant provided that we take a small enough timestep relative to the phenomenon modelled. With that approximation, the integral term becomes a discrete sum over the number of timesteps. Therefore, equation 4.2 becomes :

$$L = \left(\prod_{i=1}^{n} \lambda \left(\hat{t}_i \right) \right) \exp \left(- \sum_{t \in [0,T] \cap \mathbb{N}} \lambda \left(t \right) \right) \tag{4.3}$$

Here, we express the intensity function in the unit of the timestep chosen so that there is no need to multiply each term of the sum by the timestep. This approximation is widely used as it is enough to make the likelihood tractable under reasonable assumptions on the intensity function. The intensity function of a Poisson process is often confused with the rate (also denoted by λ in most cases), even though they are conceptually different. These two quantities coincide exactly when we assume that the intensity function is piecewise constant. Adopting a Poisson regression approach leads to another version of the likelihood presented in equation 4.3 which depends on time through different temporal signals $\boldsymbol{x}(t)$. In the context of our problem, we observe a sample $\{(Y_i, \boldsymbol{x}_i),\ i \in \{1 \dots n\}\}$ of size $n \in \mathbb{N}$. Therefore the likelihood of the equivalent regression formulation of the NHPP likelihood from equation 4.3 can be rewritten as follows :

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} \frac{\exp\left(-\lambda_i(\boldsymbol{\theta})\right) \lambda_i(\boldsymbol{\theta})^{y_i}}{y_i!} \tag{4.4}$$

with $\boldsymbol{theta}$ the vector containing the parameters for all effects. Finally, the log-likelihood is as follows :

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n} Y_i \log\left(\lambda_i(\boldsymbol{\theta})\right) - \sum_{i=1}^{n} \lambda_i(\boldsymbol{\theta}) - \sum_{i=1}^{n} \log\left(Y_i!\right) \tag{4.5}$$

with the last term of this log-likelihood independent of λ which makes it irrelevant for maximising the log-likelihood.

4.3.1 Background

In this section we will introduce the different algorithms used as stepping stones for our proposed approach to maximise this likelihood based on the model presented in equation 4.1.

Generalised Linear Model

A generalised linear model (GLM) extends the traditional Gaussian linear regression to a wider range of statistical distributions for the response variable while also adding a non-linear relationship between the response and the covariates as a link function. In our case, it directly corresponds to the parametric part of model 4.1 which we can write as follows :

$$\log \lambda(t) = \beta_0 + \sum_{l=1}^{L_p} \beta_l x_l^p(t) \tag{4.6}$$

this parametric model is generally estimated using an iteratively (reweighted) least-square (IRLS) algorithm. While for the Gaussian linear regression setting, the maximum likelihood estimator has a closed form, in general this is not true for GLM. Therefore, assuming a least-square objective, IRLS is used to iteratively update the parameters of the model until convergence. Essentially, the IRLS consists in updating the parameters estimate with a gradient descent algorithm. The update slightly differs from a traditional gradient descent with each observation receiving a weight depending on the magnitude of the residuals for that particular observation. It was found that this procedure handles outliers in non-Gaussian distributions better.

Generalised Additive Model

Now for the non-parametric part of the model we start off with the generalised additive model (GAM) which can be written as

$$\log \lambda(t) = \sum_{l=1}^{L_s} s_l(x_l^s(t)) \tag{4.7}$$

with the same notations as for equation 4.1. It has been formalised in [207] with a fitting procedure called backfitting. This proof of convergence to an optimal solution can be found in [263]. More recently, a penalised IRLS (or PIRLS) procedure was proposed in [251] which led to faster implementation while ensuring the smoothness of estimates with a penalty term on the integral of second order derivatives of the smooth functions estimated.

Penalised Wavelet Additive Model

The last stepping stone of our proposed model is the following :

$$\log \lambda(t) = \sum_{l=1}^{L_w} w_l(x_l^w(t)) \tag{4.8}$$

It is very similar to the GAM apart from the basis functions which are wavelets instead of splines. It can be referred to as a wavelet additive model. Because of the sparse nature of wavelets, we have chosen to adopt a LASSO-penalisation approach which can be written as follows :

$$l_l^p(\boldsymbol{\theta_l^w}) = l_l(\boldsymbol{\theta_l^w}) + \gamma_l ||\boldsymbol{\theta_l^w}||_1,\ l \in \{1 \dots L_w\} \tag{4.9}$$

with $l_l(\boldsymbol{\theta_l^w})$ the log-likelihood defined in equation 4.5 when it is restricted to to only one wavelet effect $l \in \{1 \dots L_w\}$ of model 4.8. Also, because of the nature of the problem at hand where we are trying to model a function supposed to be piecewise constant, we have chosen Haar wavelets but other wavelet basis can be used in practice. To maximise the LASSO-penalised likelihood defined in 4.9 we have chosen a CGD approach [259]. Furthermore, a choice needs to be made for the coefficient multiplying the penalty that we write γ_l for some $l \in \{1 \dots L_w\}$. As the optimal γ_l is unknown, we adopt a cross validation procedure by crafting a grid of values. A good choice is to start with the maximum γ for which all parameters to be estimated are put to zero. Indeed, above that γ_l^{max}, we should always have a model with all parameters being null. Then we take a value for γ_l^{min} depending on γ_{max} (usually $\gamma_l^{max}.10^{-3}$). And with these two boundaries we can create a grid on the log-scale for the values of γ to try out. One way to find optimal parameters under this LASSO setting is to use Coordinate Gradient Descent (CGD). This procedure consists in optimising sequentially the objective function (penalised likelihood) with regards to each coordinate of the parameter vector. In this work, we used the implementations proposed in [264] with the package `glmnet`. The parameter update step in this procedure can be written as follows :

$$\theta^w_{l_j,(t+1)} \leftarrow \mathcal{T}(\theta^w_{l_j,(t)} - \frac{\partial l_l(\boldsymbol{\theta}_l^w)}{\partial \theta^w_{l_j}}, \gamma_l),\ j \in \{1, \dots 2^J - 1\} \tag{4.10}$$

with $2^J - 1$ the dimension of the wavelet basis for effect l and with the term $\theta^w_{l,(t)} - \frac{\partial l_l(\boldsymbol{\theta})}{\partial \theta_l^w}$ being the coordinate-wise update without penalisation (and with a step size equal to 1) of parameter $\theta^w_{l_j}$. In particular, the soft-thresholding operator $\mathcal{T}$ resulting from the LASSO penalty procedure optimised by CGD is as follows [264] :

$$\mathcal{T}(a,b) = sign(a)(|a| - b)_+ = \begin{cases} a - b & \text{if } a > 0 \text{ and } b < |a| \\ a + b & \text{if } a < 0 \text{ and } b < |a| \\ 0 & \text{if } a < 0 \text{ and } b \geq a \end{cases}$$

4.3.2 Proposed approach

Using the stepping stones introduced in Section 4.3.1, we propose two algorithms which can be used to successfully fit model 4.1. The first one is referred to as `OBO` which stands for "One-By-One". As its name suggests, the purpose of this first algorithm is to fit each part of the model (only once) sequentially and in a particular order. The second algorithm which will be referred to as `BAC` is a version of backfitting applied to this model. The following sections unravel the details of each algorithm.

OBO

The OBO algorithm consists in fitting sequentially the different components of the model. Firstly, we start by fitting the linear part, then the splines components to finally end with the wavelet effects. Apart from the linear component which is fitted all at once, each non-parametric component is fitted separately. The idea behind this algorithm is to move from the lowest frequency (linear part and splines with not too many degrees of freedom) to the highest frequency of the signal (wavelet basis of relative high-order). Algorithm 1 formally presents the implementation of this approach with the notations introduced in the introduction. In addition, we note $\eta_{-l}(\theta_{-l})$ the additive model without the $l-th$ effect which can also be extended to components. For instance, $\eta_{-s}(\boldsymbol{\theta_{-s}}) = \beta_0 + \sum_{l=1}^{L_p} \beta_l x_l^p(t) + \sum_{l=1}^{L_w} w_l(x_l^w(t))$, which is simply the same model as in equation 4.1 without the spline component.

Algorithme 1 : OBO

Target : $\boldsymbol{Y}$;
Number of components to be fitted separately : $L = 1 + L_s + L_w$;
Covariates : $\boldsymbol{X} = (\boldsymbol{X}^p, \boldsymbol{X}^s, \boldsymbol{X}^w)$;
Parameters : $\boldsymbol{\theta} = (\boldsymbol{\theta}^p, \boldsymbol{\theta}^s, \boldsymbol{\theta}^w)$;
Objective : $\underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, l(\boldsymbol{\theta})$;
$\boldsymbol{\theta} \leftarrow \mathbf{0}$;
$\boldsymbol{\theta}^p = \underset{\boldsymbol{\theta}^p}{\operatorname{argmax}}\, l(\boldsymbol{\theta}, \boldsymbol{Y}, \boldsymbol{X})$, (IRLS) ; `/* Linear component */`
$\tilde{\boldsymbol{Y}} = \boldsymbol{Y} - \exp\left(\eta_{-s}(\theta_{-s})\right)$;
$\boldsymbol{\theta}^s = \underset{\boldsymbol{\theta}^s}{\operatorname{argmax}}\, l(\boldsymbol{\theta}, \tilde{\boldsymbol{Y}}, \boldsymbol{X})$, (PIRLS) ; `/* Spline component */`
$\tilde{\boldsymbol{Y}} = \boldsymbol{Y} - \exp\left(\eta_{-w}(\theta_{-w})\right)$;
$\boldsymbol{\theta}^w = \underset{\boldsymbol{\theta}^w}{\operatorname{argmax}}\, l(\boldsymbol{\theta}, \tilde{\boldsymbol{Y}}, \boldsymbol{X})$, (CGD) ; `/* Wavelet component */`
Result : $\boldsymbol{\theta}$

BAC

The other algorithm which we want to propose here is inspired from backfitting. Unlike OBO, the BAC algorithm does not involve an a priori on the order in which the different effects should be fitted. In practice, effects are fitted in a random order. Each time an effect is fitted, the rest of the model fitted up until this iteration is subtracted from the target response. So only the residuals of the current model iteration are fitted at each step. Like backfitting, this procedure is repeated multiple times until convergence. Convergence is reached when the L^2-norm of the difference between the parameters estimate at the previous and current iterations for each effect is less than a certain tolerance threshold ϵ. In fact, OBO could be seen as one iteration of BAC however set in a particular order. Algorithm 2 summarises this procedure. For BAC, we have observed that two additional steps were required to be added for the algorithm to converge. The first one is that the intercept has to be fitted beforehand and outside any of the components. The second one is that we need to recenter the functional components

fitted (splines and wavelets). This is actually common for practical implementations of backfitting as it guarantees convergence as well as improving its speed.

Algorithme 2 : BAC

Target : $\boldsymbol{Y}$;
Number of components to be fitted separately : $L = 1 + L_s + L_w$;
Covariates : $\boldsymbol{X} = (\boldsymbol{X}^p, \boldsymbol{X}^s, \boldsymbol{X}^w)$;
Parameters : $\boldsymbol{\theta} = (\boldsymbol{\theta}^p, \boldsymbol{\theta}^s, \boldsymbol{\theta}^w)$;
Objective : $\underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, l(\boldsymbol{\theta})$;
$\boldsymbol{\theta}_{(0)} \leftarrow \mathbf{0}$;
$\epsilon \leftarrow 10^{-3}$;
do
    $\text{index} = \text{shuffle}(\{1 \dots L\})$;
    for l *in index* **do**
        $\tilde{\boldsymbol{Y}} = \boldsymbol{Y} - \exp(\eta_{-l}(\theta_{-l}))$;
        if $l = 1$ **then**
            $\boldsymbol{\theta}^p_{(t)} = \underset{\boldsymbol{\theta}^p}{\operatorname{argmax}}\, l(\boldsymbol{\theta}, \tilde{\boldsymbol{Y}}, \boldsymbol{X})$, (IRLS) ; `/* Linear component */`
        else if $1 < l <= 1 + L_s$ **then**
            $\boldsymbol{\theta}^s_{(t)} = \underset{\boldsymbol{\theta}^s}{\operatorname{argmax}}\, l(\boldsymbol{\theta}, \tilde{\boldsymbol{Y}}, \boldsymbol{X})$, (PIRLS) ; `/* Spline component */`
        else
            $\boldsymbol{\theta}^w_{(t)} = \underset{\boldsymbol{\theta}^w}{\operatorname{argmax}}\, l(\boldsymbol{\theta}, \tilde{\boldsymbol{Y}}, \boldsymbol{X})$, (CGD) ; `/* Wavelet component */`
        end
    end
    $t \leftarrow t + 1$;
while $\forall l \in \{1, \dots, L\},\ ||\boldsymbol{\beta}_l^{(k+1)} - \boldsymbol{\beta}_l^{(k)}||_2 > \epsilon ||\boldsymbol{\beta}_l^{(k)}||_2$ *and* $t < 100$;
Result : $\boldsymbol{\theta}$

4.4 Case Study : EV arrivals at charging points

In this section we propose a case study on EV charging session data explored at a high level in [18] and [198]. The following sections introduce the experimental protocol and the results associated.

4.4.1 Data collected

The dataset in the scope of this case study gathers charging session information in the United Kingdom (UK) during 2017 [143]. It concerns domestic chargers ranging from 3kW to 22kW. However, it is expected that the great majority of EV chargers in this dataset are 3kW or 7kW chargers. This data was collected by the UK department of transport. One key finding on arrival times (or plug-in times) of EVs was that domestic charging events were more frequent with different patterns on weekdays than on weekends. We have also observed that in the data and that is why we have chosen to focus

solely on weekdays for this case study. The circumstances of EV uptake in 2017 in the UK was strong. The total number of plug-in cars on UK roads passed 130,000 that year. The best BEV seller was the Nissan LEAF with more than 13,000 registrations. Across the whole country, the largest sales were made in London and Eastern England with Scotland and South West garnering the fastest growth [143]. While public EV charging infrastructure is increasing at a fast pace, domestic charging remains the first choice for a majority of EV users.

In addition to this dataset, we have gathered temperatures in the UK from 8 of the top 10 cities in terms of population from the Iowa Environment Mesonet website [73]. That is, London (8.9 million), Birmingham (1.15 million), Glasgow (612 thousand), Liverpool (579 thousand), Bristol (572 thousand), Manchester (554 thousand), Leeds (503 thousand), Edinburgh (508 thousand). The temperature for each city is not particularly recorded at the same time nor at a regular timestep. Therefore, we have interpolated all these temperature with cubic splines. In order to have a more compact model, we propose a weighted average version of the temperature as follows :

$$\text{temp}(t) = \frac{1}{\sum_{a=1}^{8} pop_s} \sum_{s=1}^{8} pop_s T_{s,t} \tag{4.11}$$

where $T_{s,t}$ is the temperature recorded at time t by station s and $\text{temp}(t)$ is the weighted mean temperature which will be used in the modelling experiments.

4.4.2 Experimental setting

The experimental protocol we have chosen to try out our methods is close to the operational setting. Indeed, it is a rolling-forecasting origin procedure which trains the model on all data available up until a certain date to then forecast the following week. The first training set runs from the 1st January to the 1st September 2017. Therefore, it comprises 17 test weeks from the 4th September to the 31st December 2017 (the 2nd and 3rd September are a weekend). The two algorithms considered in our experiments are OBO and BAC as defined in section 4.3.2. Each algorithm have been tested with three different variations. All of them include a linear part which is simply the indicator of the day of the week. The first variation only takes into account splines components for the hour of the day and the weighted temperature defined in equation 4.11 (OBOs and BACs). The second variation is the same with only wavelet components (OBOw and BACw). And finally, the third variation includes both splines and wavelet components (OBOsw and BACsw). Our hope was that the linear and splines components would capture most of the variations in the intensity function while the wavelets would enhance the performance during peak times.

To assess the efficiency of the various approaches, we have retained two types of metrics : performance and parcimony metrics. For performance, 4 metrics have been chosen : mean absolute error (MAE), root-mean squared error (RMSE), peak RMSE, and deviance. The MAE and RMSE are performance metrics widely used which we have already defined in previous chapters. The peak RMSE only focuses on the daily peak. Therefore, the better the approach captures daily peaks, the lower the peak RMSE should be. Finally, deviance is used as a goodness of fit test. In the context of Poisson regression, the deviance formula can be written as follows :

$$D = 2\sum_{i=1}^{n}\left(Y_i \log\left(\frac{Y_i}{\hat{Y}_i}\right) - (Y_i - \hat{Y}_i)\right) \tag{4.12}$$

with $\hat{Y}_i$ being the model prediction. This formula can be derived from the likelihood ratio test comparing the proposed model and the saturated model. The latter predicts exactly the observed value (like an oracle model). Again, the purpose is to have minimal deviance.

As for parcimony metrics, we have chosen to take into account the estimated number of degrees of freedom (d.o.f). In addition, we also collected the number of iterations (particularly relevant for the BAC algorithm), the number of parameters to be estimated and the number of non-zero parameters kept and the runtime.

Due to the fact that only one year of data is available and that there is an obvious yearly cycle, we have attempted to detrend the data by dividing the number of arrivals by the trend estimated by thin plate regression splines over the whole year. The trend is divided instead of subtracted in order to ensure that all values are still positive. Precisely, we perform a euclidian division to make sure that the target is a natural number. However this did not improve the results on both training and testing sets so we have finally decided to keep the data in its original form.

4.4.3 Results

Figure 4.1 represents a fitted `BACsw` approach for a random week of the first training set. On this fit it is interesting to note that the benefit of the wavelet effect can be clearly observed on four out of the five peak estimates. The benefit of wavelets is marginally seen on the ascendant and descendant part of the curve. This reinforces our a priori which was that wavelets could help better capture peaks. Essentially, we can see that most of the work is done by the linear and splines part but the wavelet really seem to do what we wished for them to do which is enhancing the performances on peak estimates. The grey curve only shows the difference in level for every day estimated by the linear part. As the department for transport reported and as we observed in the data, there is marginal variation between the level of different days. Therefore, even though there is a slight change in value every day of the week, it is not easy to conclude whether is it significant. The estimate for the intercept is probably enough for the linear part in our case.

On Figure 4.2 we show the wavelet basis expansion of the hour of day effect. We have chosen to present this effect because it does not have too many levels and is easy to understand. However, the idea is obviously to use basis of much higher order like it is the case for the temperature for example or other potential covariates. Here, we have chosen to use the Haar basis because we assume that the intensity is piecewise constant and also for its ease of implementation. However, the proposed procedure is not restricted to a certain type of wavelet. It is interesting to compare what happens to this basis after fit. In particular, the soft-thresholding procedure obtained by the LASSO fit described in section 4.3.1 is illustrated by Figure 4.3. It shows the same wavelet functions than in Figure 4.2 which are multiplied by the coefficients estimated in the `BACsw` approach.

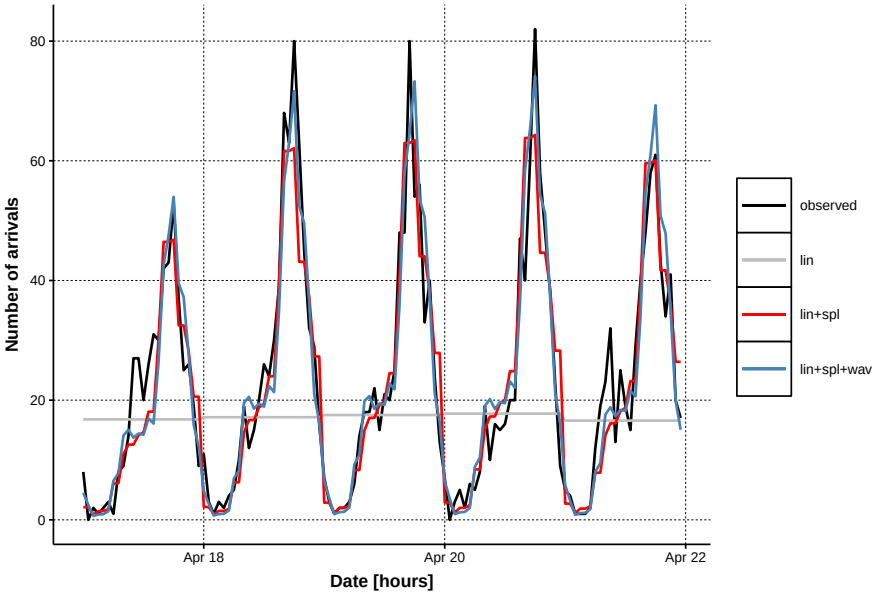


Figure 4.1 – Random week taken in the training set

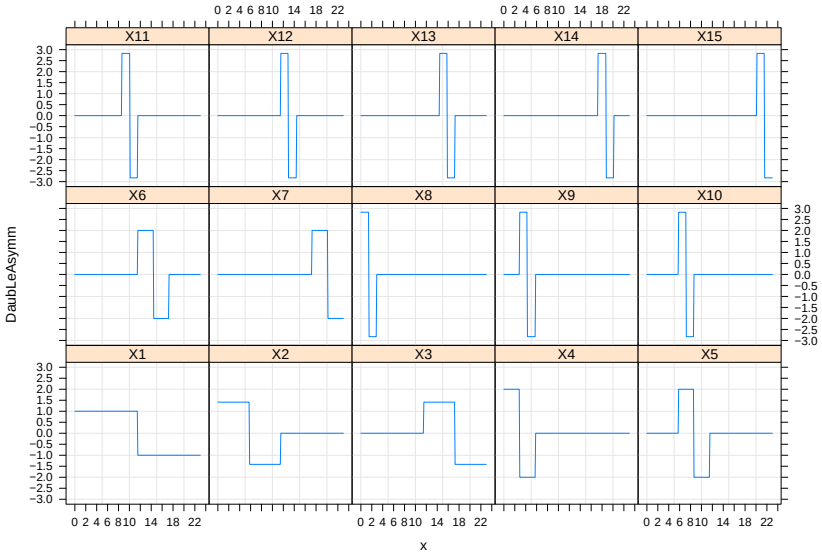


Figure 4.2 – Before fit

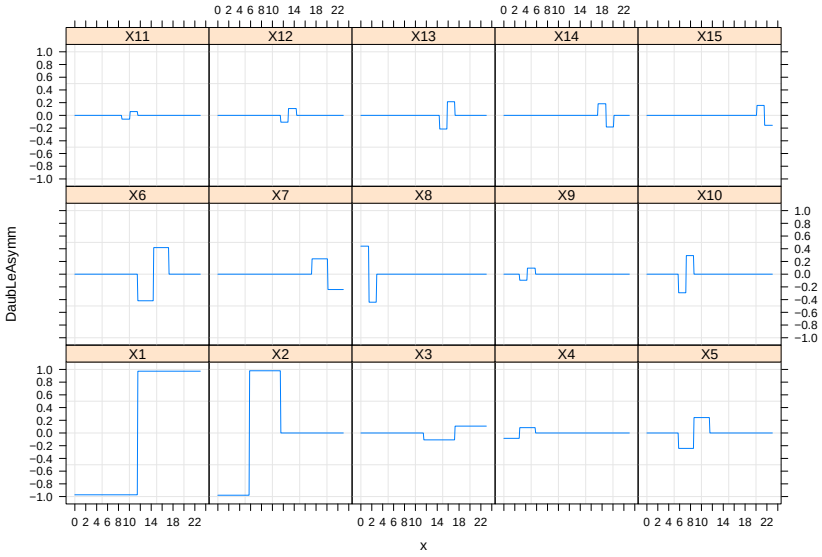


Figure 4.3 – After fit

Figure 4.4 shows the sum of all these individual functions which results in the time of day effect for the `BACsw` approach. We can observe that the evening peak is clearly identified and localised between 4 and 7pm.

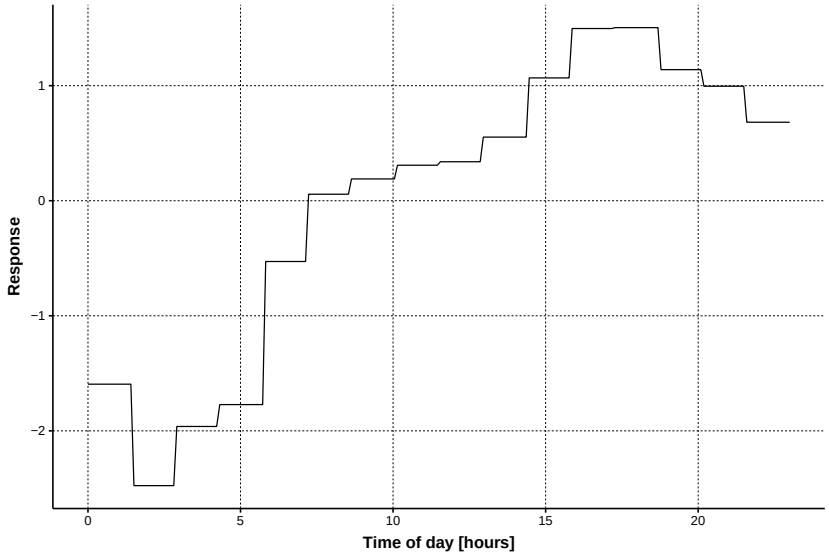


Figure 4.4 – Time of day wavelet effect fitted

In terms of performances on the training sets, Figure 4.5 represents the boxplot of the 17 training sets for the 4 performance metrics retained in this analysis. The 4 metrics seem to agree that the `BACsw` performs best. Also it seems clear that the BAC algorithm performs better than the OBO. This can be due to the fact that with multiple iterations the model can be further refined. Also, it may also mean that the order in which the components are fitted does not actually matter as the nature of each component might be enough for the model to use their behaviour as we expect it to.

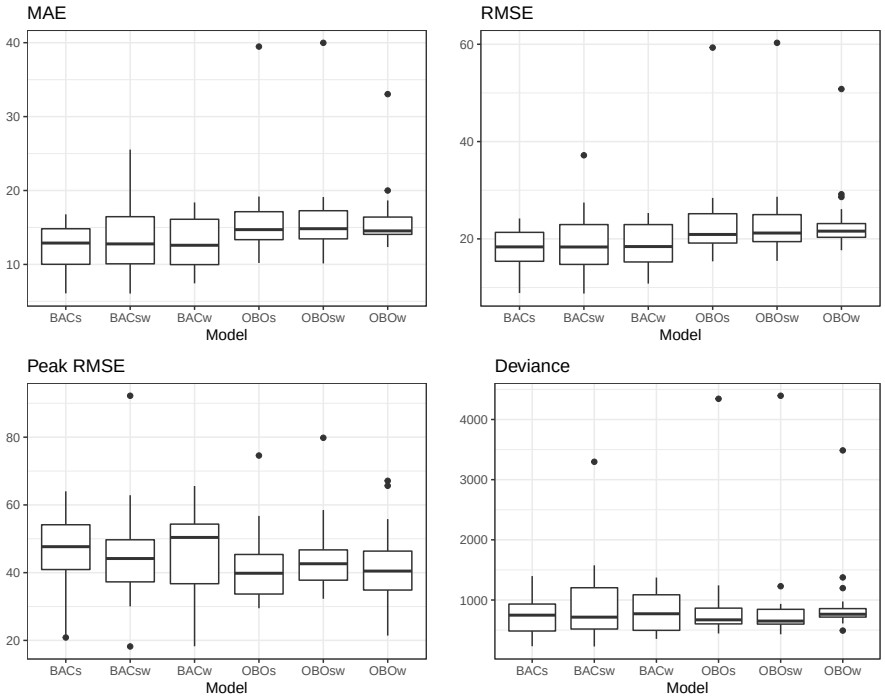


Figure 4.6 – Performances on the testing sets (Domestics UK)

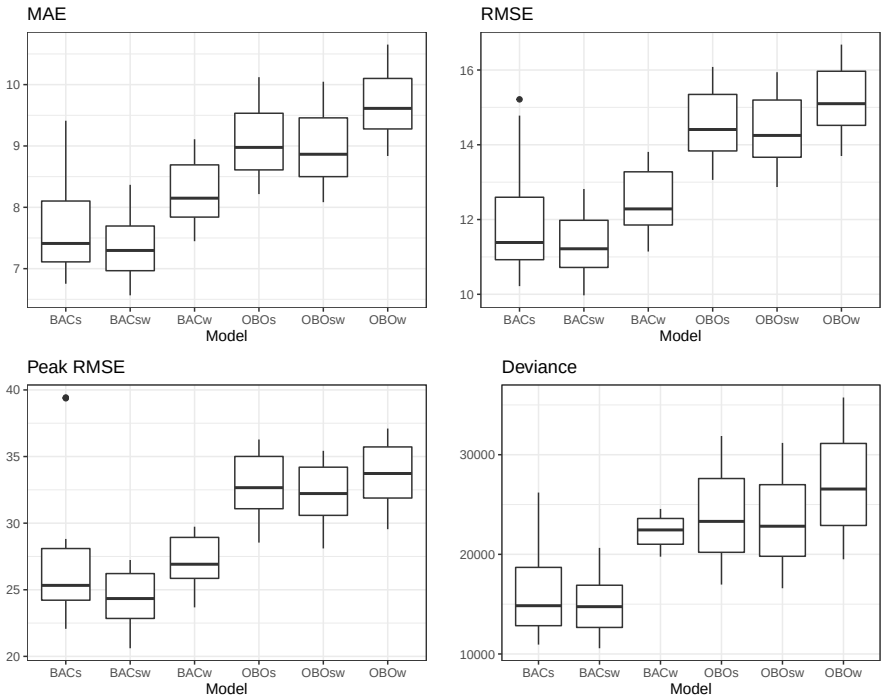


Figure 4.5 – Performances on the training sets (Domestics UK)

The performances on the testing sets are harder to distinguish probably because we are lacking some information to properly generalise the model. In particular, the metrics are much worse on the testing sets compared to the training sets which could indicate an overfitting issue. However, because this is the case for all approaches, it is

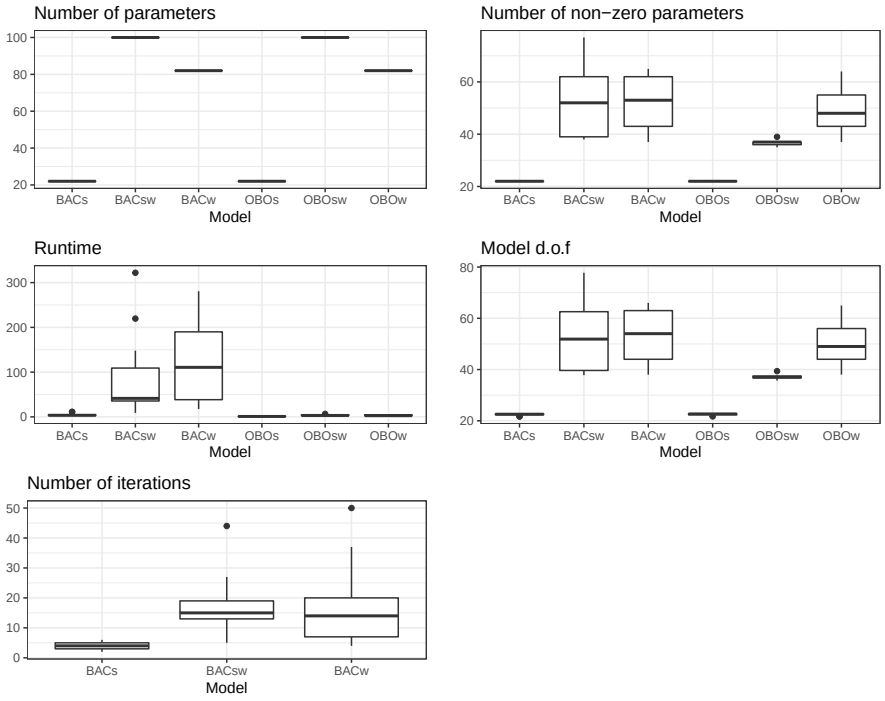


Figure 4.7 – Parcimony measures on training sets (Domestics UK)

more likely that we are missing other critical information such as traffic. Also, because we only have one year of data, the yearly cycle cannot be estimated. Even if we made an attempt at removing the trend, the yearly cycle still seems to have a significant impact.

Finally, when looking at parcimony measures (Figure 4.7), it is interesting to note that the estimated d.o.f for the `BACsw` et `OBOsw` is smaller than their wavelet only counterparts. Mixing both wavelets and splines leads to a more performant and parcimonious model than just with wavelets. Also, even though there are more parameters to be estimated initially, the number of non-zero parameters is relatively equivalent and the `BACsw` runs quicker than the `BACw` (apart from certain outliers which are probably due to some limitations of our implementation). Overall, the variation of `BAC` using wavelets take longer to fit inherently because of the penalised wavelet implementation with the cross validation to find the optimal penalisation parameter γ.

4.4.4 Additional study on Palo Alto data

As the results on the testing sets were not conclusive on the Domestics UK dataset, we have conducted an additional study on Palo Alto (USA, California) data [265]. The experimental set up is similar to the one used for the Domestics UK experiments except that thanks to a much larger time period covered by the datasets, we can train our models on more data. The first training set runs from Monday 4th January 2016 to Friday 4th January 2019. Then the testing sets covers the 20 consecutive weeks starting from Monday 7th January 2019. Also, the yearly cycle trend is now included by using a time of year covariate which ranges from 0 to 1. The rest of the covariates remain the same. The results of these experiments in terms of performances obtained are shown on Figure 4.8.

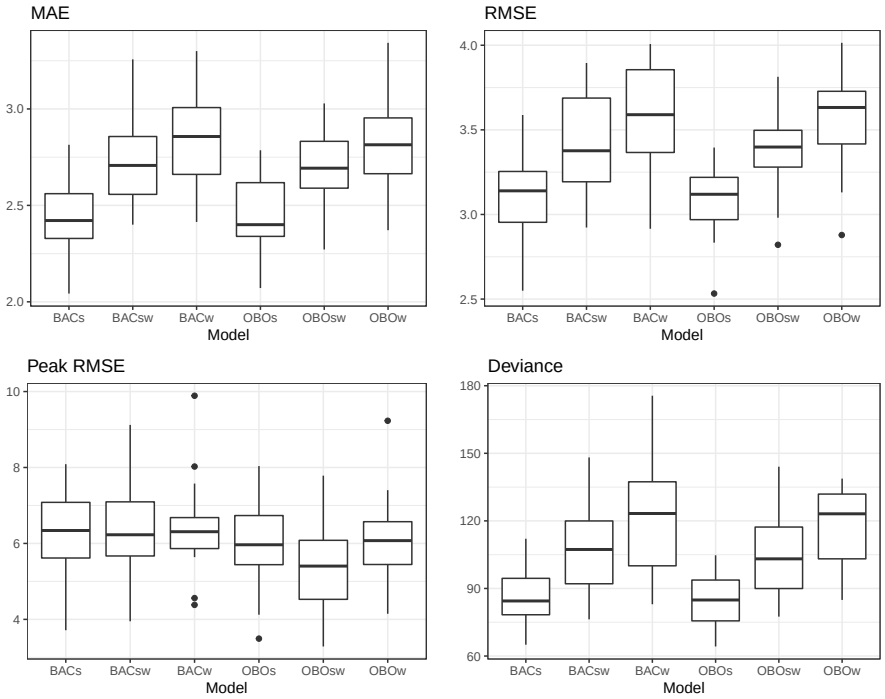


Figure 4.8 – Performances on the testing sets (Palo Alto)

It is interesting to note that unlike what we previously observed, the OBO approaches seems to be equivalent to their BAC counterparts for the RMSE and MAE metrics and are even marginally better for the peak RMSE and Deviance. Also, it seems that the OBOsw approach is able to capture peaks better than the other approaches which also indicates that the combination of spline and wavelets is particularly suited to this application. However, it is still unclear whether this improvement is significant as we can see that for the BAC approaches, all methods seem to be relatively equivalent.

4.5 Conclusion

In this chapter we have studied an additive model with both wavelet and spline components for estimating the first-order intensity function of NHPP. Two algorithms were proposed, OBO and BAC. We have shown that BAC performs better on fitted residential data (Domestics UK) but the approaches seems to be fairly equivalent on testing data. However, we observed that there could be some benefit to use a model with additive spline and wavelet effects to capture peaks when looking at the peak RMSE metric on both datasets used in this case studys. The methodology proposed can be extended to any timestep (as long as it is constant), other wavelet basis (e.g., Daubechies, Meyer). The convergence speed of the algorithm could be accelerated by fitting first the linear component to then perform the backfitting of the non-parametric part for instance. Better performances could be obtained by giving more degrees of freedom (knots) to splines but it can sometimes lead to overfitting. One limitation of the BAC algorithm proposed is that despite the empirical evidence of convergence of the algorithm, there is no unicity of the parameter vector estimated. That is in great part due to the random path taken to fit the various components. Finally, other metrics could have been used

to assess peak estimates as it can be argued from an operational perspective that it is more important to obtain a better estimate of the peak magnitude not too far from peak time than a worse estimate exactly at peak time. Therefore, it might be interesting to compare the daily time series forecasted with the observed one with a measure of similarity thanks to dynamic time warping for instance.

Chapitre 5

Multi-resolution peak load forecasting

This chapter is based on a paper published in International Journal of Forecasting [266].

In the context of smart grids and load balancing, daily peak load forecasting has become a critical activity for stakeholders of the energy industry. An understanding of peak magnitude and timing is paramount for the implementation of smart grid strategies such as peak shaving. The modelling approach proposed in this paper leverages high-resolution and low-resolution information to forecast daily peak demand size and timing. The resulting multi-resolution modelling framework can be adapted to different model classes. The key contributions of this paper are a) a general and formal introduction to the multi-resolution modelling approach, b) a discussion on modelling approaches at different resolutions implemented via Generalised Additive Models and Neural Networks and c) experimental results on real data from the UK electricity market. The results confirm that the predictive performance of the proposed modelling approach is competitive with that of low- and high-resolution alternatives.

Summary

5.1 Introduction

The electric daily peak load is the maximum of the electricity power demand curve over one day. Having an accurate forecast of the daily peak enables independent system operators (ISOs) and energy providers to better deliver electricity and optimise power plant schedules. The importance of such a forecast is increasing as the integration of intermittent renewable production sources progresses. In particular, renewable energy sources are at the bottom of the merit order curve which makes them (currently) the most economical source of energy used to serve the market. However, they are intermittent and provide time-varying levels of power generation, which are only partially under human control. If electricity demand is high and renewables cannot provide for it alone, ISOs have to deliver electricity from sources with higher marginal costs (e.g., gas-fired plants) for the stakeholders as well as for the environment in terms of CO_2 emissions. In such a context, accurately forecasting the peak demand magnitude and timing is essential for determining the generation capacity that must be held in reserve.

Electrical equipment is tailored to support a specific peak load. If the demand comes close or exceeds the network capacity, it can lead to distribution inefficiencies and ultimately power system failures, such as blackouts. With the increasing number of electric vehicles (EV) in circulation, a further source of stress is added to the electricity system. For instance, 46% of vehicles sold in Norway in 2019 were EVs [19]. The challenge posed by the additional EV demand must be met by more tailored management systems and policies, if expensive infrastructural works are to be avoided. Dynamic electricity pricing schemes, for example, the Triads in the UK or the Global Adjustment in Ontario, Canada, have been developed to reduce the system peak load. Consumers who can correctly estimate and cut their use during peak events can unlock great savings. Peak demand forecasts will thus be key for the development of such policies.

To account for the increasing demand for electricity and to prevent system failures, smart grid technologies and policies are being implemented to foster communication between the various stakeholders of the electricity supply chain to achieve a more efficient use of energy. One major objective is to maximise the load factor. The load factor is the average load over a specific time period divided by the peak load over the same period. Maximising it leads to a more even use of energy through time, thus preventing system failures and surges in electricity prices. One of the most common ways to achieve load factor maximisation is peak shaving, which refers to the flattening of electrical load peaks. Three major strategies have been proposed for peak shaving, namely integration of Energy Storage System (ESS), integration of Vehicle-to-Grid (V2G) and Demand Side Management (DSM) [267]. ESS and V2G integration provide ancillary sources to balance the grid through batteries while DSM shifts consumer demand to flatten the peak. To be activated adequately, all these strategies require accurate forecasts of the demand peak magnitude (DP) and of the instant at which it occurs (IP).

This article proposes novel methods to forecast the DP and the IP by leveraging information at different time resolutions. In particular, the multi-resolution approach proposed here is illustrated in the context of two model classes : Generalised Additive Models (GAMs) and Neural Networks (NNs). Both are state of the art predictive models, widely used to forecast electrical load in industry and academia. The performance of the multi-resolution framework under both model classes is assessed using aggregate

UK electricity demand data from the National Grid [268].

The rest of the paper is structured as follows : Section 2 presents a literature review of daily peak forecasting methodologies. Section 3 introduces multi-resolution modelling using GAMs and neural networks. Section 4 explains how the different models were set up in the high-resolution, low-resolution and multi-resolution settings. Section 5 analyses the results of the models described in Section 4, using UK demand data.

5.2 Related work

This section provides an extensive literature review of peak forecasting methods and was conducted to identify gaps in the field. It includes methods ranging from probabilistic approaches to deep learning.

Probabilistic forecasts have been widely adopted in the context of load forecasting applications (e.g.,[63] for an overview), but little has been done on probabilistic peak demand forecasting. Two probabilistic set-ups, commonly used for peak load forecasting, were outlined by [269]. The first is block maxima (BM), where data is separated into time chunks of equal lengths and the maximum of each chunk is assumed to approximately follow a generalised extreme value (GEV) distribution. The second is peaks over threshold (POT), which approximates the distribution of the excess load over a threshold by a generalised Pareto distribution. While the POT and BM settings can be unified via point processes [270], in this work we are mainly interested in the BM case.

In a long-term forecasting setting, [271] used demand data at the daily resolution to forecast the magnitude and timing of the yearly peak (i.e., the day characterised by the largest total demand). They considered a forecasting lead time of one full year and obtained a probabilistic forecast by simulating year-long trajectories for the weather variables and plugging them into a deterministic linear regression model. Similarly, [272] considered a long-term forecasting application, where the aim was to forecast the probability distribution of the annual and weekly peak electricity demand. They used semi-parametric additive models to capture the effect of covariates, such as temperature, on the demand and obtained a probabilistic forecast by adopting a simulation and scenario-based approach. [273] used quantile regression methods to forecast the DP one day ahead. Even though they used quantile regression to obtain an upper bound on demand, quantile estimates at several probability levels could be used to estimate the full peak demand distribution. Also [274] modelled the DP via a quantile regression model, but their objective was post-processing daily estimates to forecast the annual demand peak, rather than modelling the DP probabilistically.

Multivariate regression models using multivariate adaptive regression splines (MARS) were proposed by [275] to forecast the DP in South Africa. Explanatory variables including meteorological variables are aggregated at the daily resolution (e.g., average, minimum and maximum temperature). The model outperforms piecewise polynomial regression models with an autoregressive error term. [276] studied time series of the DP and illustrated its heteroscedastic structure. A SARIMA–GARCH errors model and a regression-SARIMA–GARCH model are then proposed to forecast it at a short-term horizon. Results show that SARIMA-like models produce forecasts with an accuracy around 1.4 in mean absolute percentage error on a testing period.

[277] proposed a hybrid model to forecast whether the following day will be a peak

load day for the billing period for customers subject to demand charge structure. They apply their model to optimise the electricity bill of an American University. Load data is provided every five minutes from January 2013 to April 2016. Here, the POT set-up was used with a threshold depending on a monthly average and variance of the daily load. An original combination of 4 forecasts was proposed. First, a linear model is used to forecast the maximum daily load at a monthly horizon which is then coupled to short-term load forecasting models (NN and ARIMA) to provide two forecasts. Two other forecasts were computed using binary classifiers (logistic regression and NN) and a synthetic minority over-sampling technique (SMOTE) was used to balance the classes. The authors demonstrated that their methods led to better statistical accuracy and to reduced electricity bills.

NNs are one of the most popular algorithms for peak load forecasting tasks because of their strong performance in non-linear modelling. Their flexibility is remarkable, but it is difficult to pick the right architecture and hyperparameters for a specific problem. One of the first papers proposing a NN peak load forecasting method was produced by [278]. According to the authors, NNs performed well on load forecasting problems, but they were much less performant on peak load forecasting tasks. A fuzzy NN was found to be more robust and accurate than a traditional NN structure. It involved an additional layer of fuzzification of the inputs before entering the only hidden layer of the network.

In a more traditional set-up, [279] tested a Fully Connected Neural Network (FCNN) with different variants of back-propagation algorithms where training was conducted separately in four periods of time during a year. Their work was further developed by [280], where numerous weather variables were included (e.g., temperature, rainfall, wind speed, evaporation per day, sunshine hours and associated statistics). Similarly, different optimisation procedures were considered and it was found that an adaptive learning method based on the learning rate and momentum was the most performant. [281] combined a self-organising map with a NN to find better clusters of training data to improve forecasting performance. Some authors considered other form of networks. For instance, [282] adopted abductive networks with the aim of obtaining a better intuition and a more automated way to address peak load forecasting. In particular, these networks split the overall problem into smaller and simpler ones along the network with abductive reasoning. It is based on an automated procedure which organises the data available into different chunks and deals with them separately.

More recently, recurrent Neural Networks (RNNs) have been used by [283] in the form of Gated Recurrent Units (GRU). In particular, a dynamic time warping (DTW) analysis was used to produce the GRU inputs. The DTW distance was used to find the most similar load curve to the one observed before the targeted load curve. Assuming that subsequent load curves are also very similar, they used the subsequent load curve from the training data to encode the inputs of the GRU network. A Long Short-Term Memory (LSTM) architecture has been used by [284] and was found to be more computationally efficient compared to FCNNs and other RNNs. Three statistical metrics were used to evaluate model performance : Mean Absolute Percentage Error (MAPE), Root-Mean Squared Error (RMSE) and mean bias error. In our work, statistical metrics including MAPE and RMSE will also be used to avoid introducing any bias towards a particular operational application.

The literature on deep learning peak load forecasting is sparse, but deep learning probabilistic load forecasting is much more common (e.g., [285], [286] and [287]). Such models do not explicitly focus on the DP or the IP as the objective functions used to estimate their parameters are based on demand observed at a higher frequency (intra-day). The high-frequency forecasts thus obtained can be post-processed to produce a forecast for the DP.

Support Vector Regression (SVR) is another popular class of load forecasting method, based on structural risk minimisation instead of empirical risk minimisation as in NNs. [288] used SVR in a local prediction framework. Recently, [289] used an ensemble forecasting approach with other Machine Learning algorithms such as boosting machines, tree-based methods and bagging techniques. A compensation process based on an isolation forest is later added by analysing the predicted values of the ensemble models to detect outliers in the peak data. SVR are compared to NNs by [290] for a control strategy of peak load and frequency regulation. LSTM NNs were used to forecast power load and improve the control strategy considered in this particular use case.

From this literature review, it can be concluded that a wide range of methodologies have been adopted in peak load forecasting applications. In most short-term applications, model inputs are manually chosen features that are defined at the same (daily) time resolution as the peak demand, which is the variable to be forecasted. Conversely, in long-term applications, weather variables are simulated at the original (high) resolution to produce demand forecasts at the same resolution, which are then post-processed to obtain low resolution (e.g., yearly) peak forecasts. Hence, to the best of our knowledge, the existing literature on peak forecasting has not explored methods that are able to integrate both low- and high-resolution signals in a single model. However, in the field of functional data analysis, hybrid approaches have been used for clustering and forecasting functional data (e.g., [291] and [292]). Therefore, this paper aims to exploit functional methods to tackle multi-resolution problems. From a feature engineering point of view, the goal is to automate feature extraction of high-resolution signals, that is to let the model decide which hidden features to extract from the signal. This can be done with signal processing procedures such as tensor product decomposition, wavelets or Fourier transforms [293].

The literature review also suggests that not much effort has been directed towards forecasting the IP, which is surprising because forecasting the IP is at least as important as forecasting the DP, for the purpose of short-term smart grid management and operational planning [294]. To fill this gap, the performance of multi-resolution methods will be illustrated in this paper on both a DP and an IP forecasting problem.

5.3 Multi-resolution modelling

In this section, the multi-resolution modelling approach is introduced with its general principles. It is then developed formally and illustrated with GAMs and NNs.

5.3.1 General idea

The main idea behind multi-resolution modelling is to build a parsimonious model that is able to handle input and output variables that are available at different resolutions. In the context of DP load forecasting, low-resolution variables (e.g., day of the week, maximum daily temperature) are observed daily, while high-resolution variables (e.g., temperatures or raw demand) are updated every hour or half-hour. Such problems are usually handled by manually placing all variables at the same resolution. In particular, one option is to take a high-resolution approach, which consists in doing the modelling at the highest available resolution, which might require interpolating some of the low-resolution variables. Such an approach often lacks in parsimony, as the low-resolution variables are brought to the higher resolution, thus increasing the size of the data that needs to be processed, while adding no extra useful information. Another option is to take a low-resolution approach, that is to transform the high-resolution variables into a set of manually chosen daily summaries or features. In this approach, the size of the data is reduced, but feature engineering is time consuming and some of the information contained in the high-resolution variables is lost in the process.

The multi-resolution approach proposed here aims at capturing all the information contained in the high-resolution variable, while avoiding explicit feature engineering and retaining the parsimony of the low-resolution approach. To describe the multi-resolution idea more formally, let us consider $\mathbf{y}_i = \{y_i(t)\}_{t\in\{1,\dots,T\}}$ the vector of electricity demand at each time step $t > 0$ of the day $i \in \mathbb{N}$. T is the total number of daily steps (e.g., T=48 for half-hourly steps). Then, the DP of day i is $\mathrm{DP}_i = max(\mathbf{y}_i)$ and IP_i is the time step corresponding to DP_i. Let $\mathbf{x}_i^{low}$ be the i-th vector of covariates observed daily and let $\mathbf{x}_i^{high}$ be the corresponding vector of covariates containing information at the intra-day resolution. The multi-resolution approach exploits both sets of covariates as model inputs to obtain the forecasts of the $\hat{\mathrm{DP}}_i$ or the $\hat{\mathrm{IP}}_i$, that is

$$\hat{\mathrm{DP}}_i = \psi_1(\mathbf{x}_i^{low}, \mathbf{x}_i^{high}) \tag{5.1}$$

$$\hat{\mathrm{IP}}_i = \psi_2(\mathbf{x}_i^{low}, \mathbf{x}_i^{high}) \tag{5.2}$$

where ψ_1 and ψ_2 represent the model for, respectively, the DP and the IP. This general definition does not specify how the high-resolution inputs should be dealt with in practice. Several approaches could be considered, the aim being to process the information contained in a (possibly high-dimensional) signal vector, while avoiding information loss and retaining computational efficiency. In this paper, two options are considered. In particular, a description of how high-resolution covariates can be handled within GAMs and NNs is given below.

5.3.2 Particular instances of the multi-resolution approach

The multi-resolution approach is detailed firstly for GAMs which, due to their performance and interpretability [262], are widely used in industry for load forecasting. Then, the multi-resolution approach is extended to NNs, which often perform well on load forecasting problems and enable the flexible handling of heterogeneous model inputs [295].

Generalised Additive Models

First introduced by [296], GAMs are a semi-parametric extension of generalised linear models (GLMs) where the response variable, y_i, is assumed to follow a parametric probability distribution. That is, $y_i \sim \text{Dist}(\mu_i, \boldsymbol{\theta})$ where μ_i and $\boldsymbol{\theta}$ are model parameters. While the elements of $\boldsymbol{\theta}$ do not depend on i, parameter μ_i is modelled as follows [208] :

$$g(\mu_i) = \mathbf{x}_i^T \boldsymbol{\gamma} + \sum_j f_j(\mathbf{x}_i), \tag{5.3}$$

where g is a monotonic transformation, which is simply the identity function in this paper. Two separate terms can be distinguished on the right-hand side of this equation : a parametric part $\mathbf{x}_i^T \boldsymbol{\gamma}$, where $\mathbf{x}_i$ is a vector of covariates while $\boldsymbol{\gamma}$ is a vector of regression coefficients, and a non-parametric part $\sum_j f_j(\boldsymbol{x}_i)$ which is a sum of smooth functions of covariates. The smooth effects are built via linear combinations of K_j basis functions, while the corresponding basis coefficients are penalised via generalised ridge penalties. The strength of the penalties is controlled via smoothing hyperparameters, which are selected using criteria such a generalised cross-validation.

In the context of forecasting DP_i, it is interesting to consider for $\text{Dist}(\mu_i, \boldsymbol{\theta})$ a generalised extreme value (GEV) distribution. In fact, the GEV model is asymptotically justified for block-maxima as $T \to \infty$ [269]. Thus, when enough steps are available throughout the day, the GEV distribution is particularly attractive for modelling the DP. The scaled-T (a scaled version of Student's t) distribution provides an alternative, which is particularly suited for heavy tailed data such as peak load. The Gaussian distribution can be used as a baseline model. As for the IP, an ordered categorical (ocat) distribution based on a logistic regression latent variable is used. All of these distributions as well as GAM building and fitting methods are implemented in the *mgcv* R package [297].

Within the additive structure of GAMs, $\mathbf{x}_i^{low}$ and $\mathbf{x}_i^{high}$ can be treated as inputs for different smooth functions. The elements of $\mathbf{x}_i^{low}$ can be handled via separate standard smooth effects, which take scalars as inputs, while the joint effect of several elements of $\mathbf{x}_i^{low}$ can be captured via standard multivariate smooth effects. However, the $\mathbf{x}_i^{high}$ covariates have to be treated via functional smooth effects. The latter are smooth functions which take the vectors of high-resolution covariates as inputs and output a scalar. Therefore, functional GAMs permit the handling of each covariate at its original resolution, thus avoiding interpolation and guaranteeing parsimony.

In addition to the principle of parsimony, the goal is also to retain the time dependence of the covariates. In fact, it is important to ensure that the model is aware that each element of the high-resolution covariates has a different impact on the peak load distribution, as it belongs to a different time of day. However, modelling the effect of each element separately would ignore temporal dependencies and might lead to overfitting or lack of interpretability. A way to achieve a compromise when modelling high-resolution covariates is to make them interact with the time of day sequence via tensor product effects. Such effects can easily be integrated in GAMs, as explained in the following.

In continuous time, the smooth effect for a high-resolution (functional) covariate,

$x_i(u)$, can be written as follows :

$$f(x_i) = \int_0^T \phi(x_i(u), u)du \tag{5.4}$$

where ϕ is the time-dependent effect of the covariate, which needs to be estimated, while u is the time of day. In practice, on the i-th day, $x_i(u)$ is observed at F discrete instants $0 \leq t_1 \leq \cdots \leq t_F \leq T$ and the corresponding values of $x_i(u)$ are stored in the vector $\boldsymbol{x}_i$. Hence, approximating the integral with a summation and constructing ϕ via a tensor product expansion leads to :

$$\begin{aligned} \hat{f}(\boldsymbol{x}_i) &= \sum_{r=1}^{F} \hat{\phi}(x_i(t_r), t_r) \\ &= \sum_{r=1}^{F}\sum_{k=1}^{K}\sum_{l=1}^{L} \beta_{kl} a_k(x_i(t_r)) b_l(t_r) \end{aligned} \tag{5.5}$$

where $\{a_k\}_{(k)\in\{1,\ldots,K\}}$ and $\{b_l\}_{(l)\in\{1,\ldots,L\}}$ are known spline basis functions and $\{\beta_{kl}\}_{(k,l)\in\{1,\ldots,K\}\times\{1,\ldots,L\}}$ are parameters to be estimated. By using such effects, high-resolution information can be parsimoniously incorporated into the model, while retaining the temporal information contained in the covariates.

Neural Networks

NNs are convenient machine learning algorithms to implement a multi-resolution model. In fact, common architectures such as Convolutional Neural Networks (CNN) and RNNs already make use of inputs from different scales. Recent work was undertaken to make tensor inputs available for multi-layer perceptrons with MatNet [295] which further shows their versatility. From scalars to tensors, the flexibility of NNs is hard for other machine learning models to compete with.

A FCNN or CNN architecture, without its output layer, can be generally written as follows :

$$H_k(\mathbf{x}, \Theta) = h_k(\ldots h_3(h_2(h_1(\mathbf{x}, \theta_1), \theta_2), \theta_3) \ldots, \theta_k) \tag{5.6}$$

where k is the number of hidden layers of the NN, $h_{i,i\in\{1\ldots k\}}$ are the transformations made by the hidden layers (e.g., linear operation, activation and dropout) and $\Theta = \{\theta_i\}_{,i\in\{1\ldots k\}}$ is the sequence of parameter vectors (weights and biases). In a multi-resolution approach, one part of the architecture will contain low-resolution information feeding a FCNN branch and the other one will contain the reshaped high-resolution data feeding a CNN or RNN branch. In this paper, only CNNs were considered in depth for this latter branch, with the lags of the response provided as model inputs. The CNN enables a very close replication of the tensor product construction used for GAMs, thus creating a consistent set-up for comparing both algorithms.

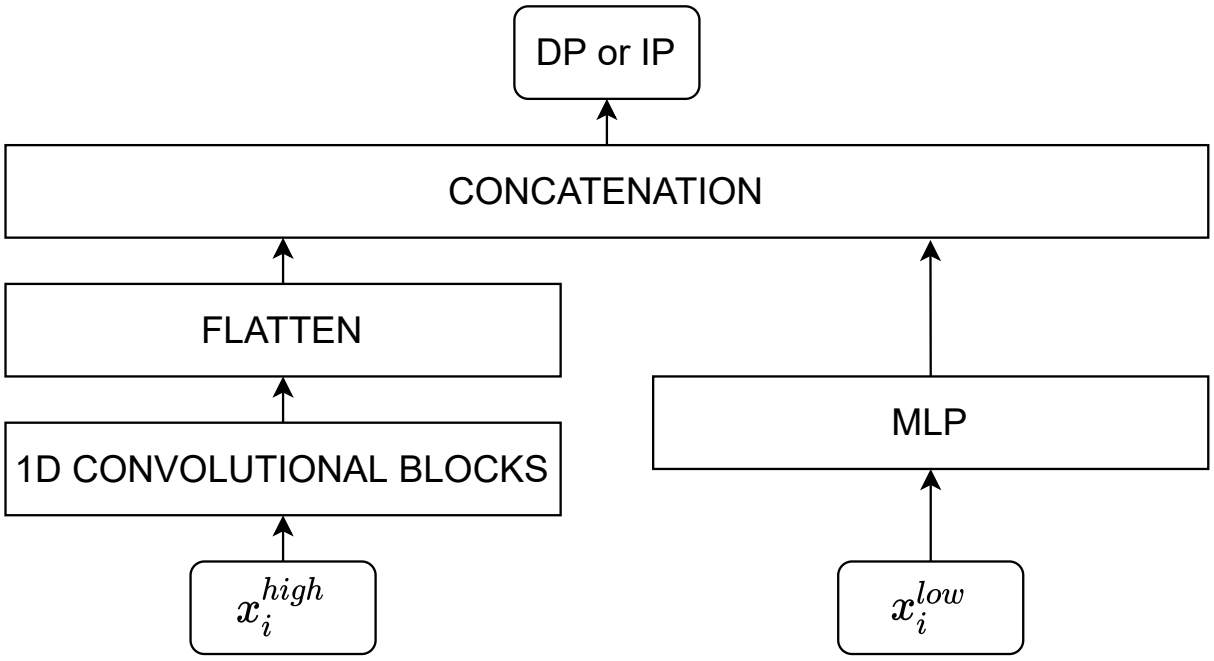


Figure 5.1 – Multi-resolution architecture for NNs with a Multi-Layer Perceptron (MLP) taking low-resolution inputs, and a CNN with high-resolution inputs

Even though the CNN and FCNN branches do not have similarly shaped inputs and outputs, the unit shapes can be transformed along the network to interact and be brought together without losing consistency. This process consists in flattening the tensor shapes in order to bounce back on vectorial inputs within some layer of the network. It is precisely this flexibility that can be leveraged to build a multi-resolution architecture (Figure 5.1). More precisely, the CNN branch contains one convolutional block for each of the high-resolution time series. In this way, each tensor product of the GAM formula can find its equivalent in the CNN branch of the network. In fact, the multi-resolution NN architecture can be concisely written as follows :

$$\mu_i = F_j(H_k(\mathbf{x}_{low}, \Theta), H'_l(\mathbf{x}_{high}, \Theta')) \tag{5.7}$$

In (7), H_k is the FCNN which handles low-resolution terms while H'_l is the CNN which deals with the high-resolution information. Then, in the final part of the network, both outputs are concatenated (after flattening the CNN branch) and enter another FCNN F_j which can be reduced to the output layer when $j = 1$. Here, μ_i is the mean of the random output variable considered. This multi-resolution architecture is summarised in Figure 5.1.

5.4 Experiments

On the DP and the IP forecasting tasks, the multi-resolution approach is compared to two alternative modelling approaches : a high-resolution approach and a low-resolution approach (Figure 5.2). The low-resolution approach uses inputs aggregated at the daily level (e.g., maximum daily temperature, day of the week) to forecast the

DP and the IP separately. The high-resolution approach uses inputs at the half-hourly level to forecast the half-hourly demand and then it extracts the DP and the IP by taking the maximum of the half-hourly forecasted values and the corresponding time of day. Therefore, the high-resolution approach leverages all the information available by taking half-hourly inputs and outputs while the low-resolution approach directly models the variables of interest (DP and IP) with less parameters to be estimated. The multi-resolution approach can be seen as a compromise, aimed at integrating the advantages of both approaches, and the following experiments are designed to assess whether it can outperform them.

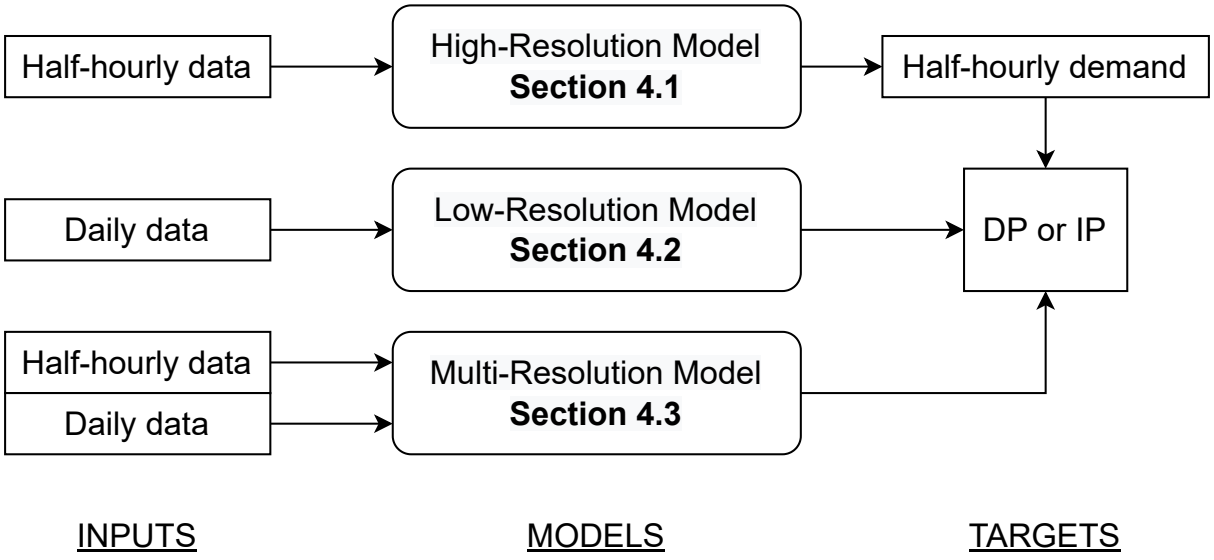


Figure 5.2 – The different modelling settings compared in this work

The comparison includes baseline models : a naive persistence model, which simply consists of forecasting the DP and the IP based on the value taken by the target variable on the previous day; a low-resolution ARIMA (on daily peaks with horizon 1); a high-resolution ARIMA aggregated forecast composed of 48 ARIMA models, each fitted on the half-hourly load of a specific time of day with horizon 1. That is, the high-resolution ARIMA produces 48 forecasts at horizon 1 instead of one forecast at horizon 48. All ARIMA models are fitted using the [298] algorithm without using exogenous information.

The performance metrics chosen for DP models are the mean absolute percentage error (MAPE) and the root mean squared error (RMSE). As for IP models, the RMSE is also used but the MAPE is substituted with a relaxed accuracy (R-Accuracy) metric in the form of a binary loss function (equal to 1 if the IP forecasted is more than 2 instants away from the observed IP and 0 if it is within 2 instants of the observed IP). While the R-Accuracy metric is also relevant in operational settings where it is crucial to know the IP within a small time window, the RMSE penalises forecasts proportionally to their distance from the observed IP.

A rolling-origin forecasting procedure is used to replicate a realistic short-term load forecasting set-up. Model parameters are updated on a monthly basis with consolidated data since, in an operational setting, threats to data validity and computational

constraints can emerge when refitting a model too often using real-time data.

The data used in the experiments is the half-hourly load consumption (total national demand) between 2011-07-01 00 :00 :00 and 2016-06-30 23 :30 :00, available via the UK [268] website. Temperature data at different locations (London, Sheffield, Manchester, Leeds, Cardiff, Bristol, Birmingham, Liverpool, Crosby and Glasgow) was downloaded from the [299] website. The temperature data is at an hourly resolution. It is interpolated (natural cubic spline interpolation) to obtain half-hourly data. Furthermore, if pop_s is the population of the nearest city to station s, a weighted mean temperature is calculated as follows :

$$\text{temp}(t) = \frac{1}{\sum_{a=1}^{10} pop_s} \sum_{s=1}^{10} pop_s T_{s,t} \tag{5.8}$$

where $T_{s,t}$ is the temperature recorded at time t by station s and $\text{temp}(t)$ is the weighted mean temperature which will be used in the modelling experiments. An exponentially smoothed version of the weighted mean temperature will also be included in the model features : $\text{temp95}(t) = \alpha \times \text{temp95}(t-1) + (1-\alpha) \times \text{temp}(t)$. It was computed using a smoothing parameter $\alpha = 0.95$, based on expert knowledge.

5.4.1 High-resolution approach

Forecasting the electricity hourly or half-hourly demand is a problem that has been extensively studied in the literature [300]. It is well known that a common driver of electrical load is weather and in particular temperature. In addition, calendar information can be used to explain the seasonal variation of the demand. Finally, lagged demand values are highly informative for the subsequent values. These variables are summarised in Table 1.

Table 5.1 – High-resolution model inputs

Type	Name	Unit	Description
Weather	temp	[C°]	Half-hourly temperature
	temp95	[C°]	Half-hourly smoothed temperature
Calendar	dow	Categorical	Day of the week
	toy	None	Time of year (between 0 and 1)
	t	Categorical	Time of day (between 0 and 47)
Lag	load24	[10^1 GW]	Half-hourly load on the previous day
Output	load	[10^1 GW]	Half-hourly load

The GAM chosen to implement this approach is $y_i(t) \sim N(\mu_i(t), \sigma^2)$ where the mean of the Gaussian distribution is modelled by :

$$\begin{aligned}\mu_i(t) = &\psi_1(\text{dow}_i) + \psi_2(\text{t}) + f_1^{20}(\text{toy}_i(t)) + f_2^{20}(\text{temp}_i(t)) + f_3^{24}(\text{temp95}_i(t)) \\ &+ \text{ti}_1^{5,5}(\text{temp}_i, t) + \text{ti}_2^{5,5}(\text{temp95}_i(t), t) + \text{ti}_3^{5,5}(\text{load24}_i(t), t) \\ &+ \text{ti}_4^{5,5}(\text{toy}_i(t), t)\end{aligned} \tag{5.9}$$

In (9), the ψ functions are parametric effects, while the f functions are univariate smooth effects and the ti functions are bivariate tensor product smooth interactions. The number of basis functions used is indicated in the exponents. For instance, f_1^{20} uses 20 basis functions and $\mathrm{ti}_1^{5,5}$ uses 5 basis functions for each marginal. Thin-plate spline bases are used to build all smooth effects [301]. The model structure (9) was decided on the basis of previous experience in the field and the statistical significance of each effect.

There are many NN architectures which could be considered for this problem. We want an architecture with the minimum number of layers possible and using the same model inputs as the GAM. Adding too many layers would lead to a drastic difference in degrees of freedom between the NN and the GAM which is not realistic in a short-term load forecasting scenario. Furthermore, as we are not in the big data regime, adding too many layers may actually worsen the performance of the network.

Given that the universal approximation theorem ([302] and [303]) guarantees that a two-layer FCNN can approximate any measurable function on a compact support, a FCNN carefully built can approximate any non-linear function of the input variables with only one hidden layer. Therefore, a FCNN architecture was used to build an NN analogue of the high-resolution GAM baseline model.

In practice, there is no bound for the number of hidden units, which can lead to poor generalisation of the model when assessed on the test set. Therefore, a dropout layer was added after the hidden layer to foster the network generalisation. The outcome of the optimisation of hyperparameters led to an architecture which contains 50 neurons in the hidden layer and a dropout layer with a 10% dropout rate. The loss optimised is the mean squared error (MSE) with a Nesterov-accelerated Adaptive Moment (NADAM) optimiser [304]. In addition, the learning rate is 0.001, the number of epochs is 2000 and the batch size is 1024.

After obtaining the half-hourly demand forecast for the GAM and the NN, $\hat{\mathrm{DP}}_i$ is estimated as the maximum daily value forecasted and $\hat{\mathrm{IP}}_i$ is estimated as the half-hour of the day during which $\hat{\mathrm{DP}}_i$ occurred.

5.4.2 Low-resolution approach

Table 5.2 – Low-resolution model inputs

Type	Name	Unit	Description
Weather	tempMax	[C°]	Daily maximum temperature
	temp95Max	[C°]	Daily maximum smoothed temperature
	tempMin	[C°]	Daily minimum temperature
	temp95Min	[C°]	Daily minimum smoothed temperature
Calendar	dow	Categorical	Day of the week
	toy	None	Time of year (between 0 and 1)
Lag	DP24	[10^1 GW]	Previous day peak demand
	IP24	Categorical	Previous day instant of peak
Output	DP or IP	[10^1 GW] or Categorical	Daily demand peak or Daily instant of peak

In the low-resolution approach, all input variables are at the daily resolution (Table 2). Here several distributions could be considered for GAMs. In particular, the scaled-T distribution, which is particularly suited for heavy tailed data, as well as the GEV family, which encompasses several extreme value distributions (Weibull, Gumbell and Fréchet), are used to model the DP. For the IP forecasting task, the ordered-logit model implemented in the *mgcv* R package [297] is used. The low-resolution GAM can be written as follows :

$$\begin{aligned}\mu_i =& \psi_1(\mathrm{dow}_i) + f_1^{10}(\mathrm{IP24}_i) + f_2^{20}(\mathrm{toy}_i(t)) + f_3^{20}(\mathrm{DP24}_i) \\ &+ f_4^{20}(\mathrm{tempMax}_i(t)) + f_5^{20}(\mathrm{temp95Max}_i(t)) \\ &+ f_6^{20}(\mathrm{tempMin}_i(t)) + f_7^{20}(\mathrm{temp95Min}_i(t))\end{aligned} \tag{5.10}$$

For the DP, $\mu_i(t)$ is the location parameter of the distributions estimated, the other parameters are assumed to be constants. For the IP, $\mu_i(t)$ is also the location parameter of a latent logistic distribution. Cut-off points are estimated in the course of model fitting and do not depend on the covariates. See [305] for details.

The same FCNN architecture as for the high-resolution approach was used (50 neurons in the hidden layer followed by a dropout layer). The optimal set of hyperparameters found is also equivalent. The difference between them lies in the inputs used (Table 2) and the response variable modelled which here is directly the DP or the IP. The response structure for the DP is 1 neuron with a ReLU activation while 48 neurons are used for the IP. Instead of the traditional softmax output used in classification problems, an ordinal output structure, more suited to model the IP, is implemented as formalised by [306]. The observed response is structured as a vector of 1 and 0. If the peak was observed at $t \in \{1, \ldots, T\}$ all neurons before and including the t-th one will be 1 and all neurons after will be 0. Therefore, sigmoidal activation functions are used.

5.4.3 Multi-resolution approach

The multi-resolution GAMs leverage the same level of information for model inputs as in the high-resolution GAMs. In addition, the directly targets the DP response variable as in the low-resolution approach.

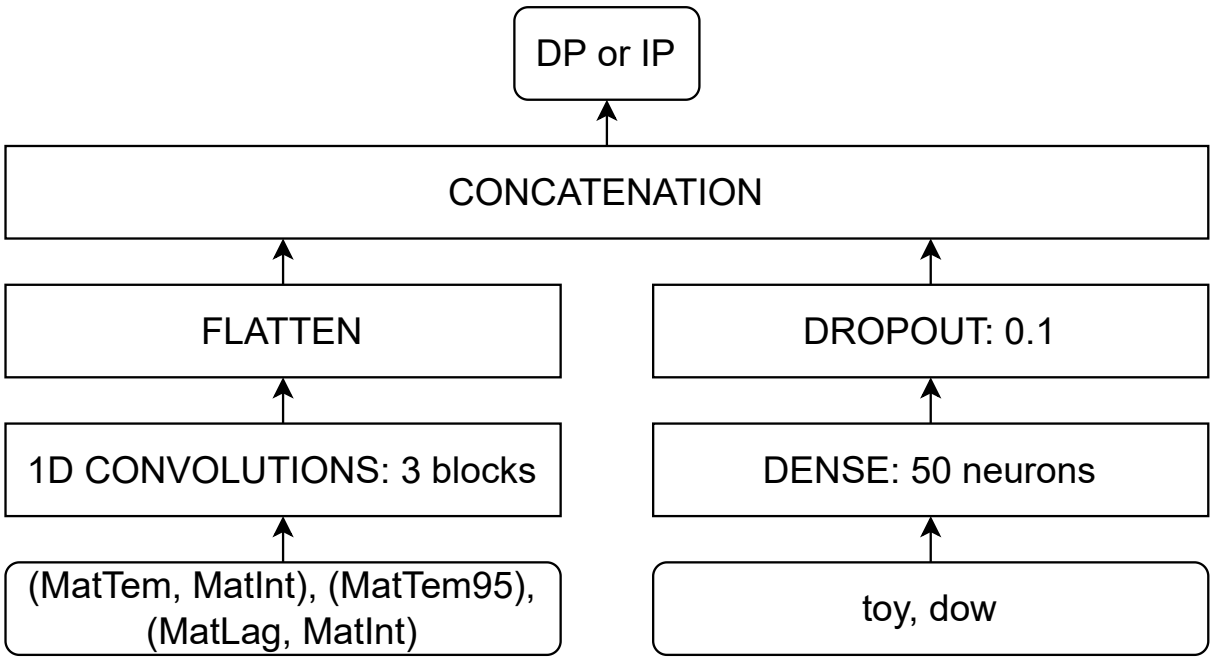
Table 5.3 – Multi-resolution model inputs

Type	Name	Unit	Description
Weather	matTem	[C°]	Vector of half-hourly temperatures
	matTem95	[C°]	Vector of half-hourly smoothed temperatures
Calendar	dow	Categorical	Day of the week
	toy	None	Time of year (between 0 and 1)
	matInt	Categorical	Vector of time steps (between 0 and 47)
Lag	matLag	[10^1 GW]	Vector of half-hourly load from previous day
Output	DP or IP	[10^1 GW] or Categorical	Daily demand peak or Daily instant of peak

Tensor products defined in Section 3.2.1 are used to capture high-resolution information. The *mat* covariates presented in Table 3 are matrices of dimension $(N \times 48)$,

N being the number of observations of the response variable DP. The multi-resolution GAM model is :

$$\begin{aligned}\mu_i = &\psi_1(\mathrm{dow}_i) + f_1^{20}(\mathrm{toy}_i) + ti_1^{15,10}(\mathrm{matTem}_i, \mathrm{matInt}_i) \\ &+ ti_2^{5,5}(\mathrm{matTem95}_i, \mathrm{matInt}_i) + ti_3^{5,5}(\mathrm{matLag}_i, \mathrm{matInt}_i)\end{aligned} \tag{5.11}$$



Learning rate: 0.0001; **Epochs**: 300; **Batch size**: 16;
Loss: MSE; **Optimiser**: Nadam

Figure 5.3 – Multi-resolution CNN architecture (input variable names are detailed in Table 3)

Unlike previous approaches, IP and DP lags are not directly included as they can be captured by the model through the ti_3 tensor interaction. As for the low-resolution approach, Gaussian, scaled-T and GEV distributions are considered for the DP and the ordered categorical distribution for the IP.

For the multi-resolution NN, the tensor product interactions will be replaced by convolution layers. The mechanism looked for through these convolution layers is essentially the same as for tensor products : extracting high-resolution information to directly model the DP or the IP. The high-resolution (half-hourly) data will be passed on to the convolution layers while the low-resolution (daily) data will go through the same FCNN architecture used in the previous approaches. As shown in Figure 5.3, these two sections of the architecture are then concatenated to produce the final forecast of the DP load. The output structure for the DP and the IP are the same as detailed in Section 4.3 with one neuron for the DP and 48 neurons for the IP.

The convolutions used for the high-resolution information are 1D convolutions on two channels. Usually, only one convolution funnel is used to capture interactions between all inputs. Here, each tensor product interaction will be replicated as a unique

convolutional block. Thus, three convolution blocks will independently extract the three high-resolution terms : matTem, matTem95 and matLag. The second channel of each block is the matrix containing the vectors of time steps matInt.

5.5 Results

The performance of the models for the DP and the IP forecasting tasks is evaluated using two statistical metrics. As a rolling-origin forecasting procedure was chosen, a transitional regime can be observed in the first few iterations, particularly for NNs, which usually perform better with a large amount of training data. Therefore, Table 4 (DP) and Table 5 (IP) present the models' performances on the last year of data, that is, from 2015-07-01 to 2016-06-30 included.

Table 5.4 – Performance on the last year of data for the DP (best model and associated metrics are in **bold**)

Resolution	Model	Metrics	
		MAPE [%]	RMSE [MW]
NA	Persistence	4.38	34.3
High	ARIMA	4.08	27.8
	Gaussian GAM	2.43	15.5
	FCNN	1.47	10.3
Low	ARIMA	3.85	26.7
	Scat GAM	1.92	12.9
	GEV GAM	2.67	16.9
	Gaussian GAM	2.26	14.4
	FCNN	2.11	14.4
Multi	GEV GAM	1.52	10.3
	Scat GAM	**1.41**	**9.59**
	Gaussian GAM	1.42	9.63
	CNN	1.56	10.5

With the exception of the high-resolution FCNN, the multi-resolution models perform better than the alternatives across all metrics (Table 4). The relative strong performance of the high-resolution FCNN can be explained by the large amount of high-resolution data available, which suits the needs of NNs. Further, the FCNN contains more parameters to estimate and is thus more flexible than the high-resolution GAMs, which require the user to manually specify how the effect of each input variable should be modelled. Nevertheless, the best model on all metrics is the scaled-T GAM, built using the multi-resolution approach. The GEV GAM performed worse than the other

distributions, which is surprising given that the GEV distribution is asymptotically justified for BM. Interestingly, the shape parameter estimated was found to be close to 0, under which value the GEV model is simply a Gumbel distribution.

Table 5.5 – Performance on last year of data for the IP (best model and associated metrics are in **bold**)

Resolution	Model	Metrics	
		R-Accuracy [%]	RMSE [half-hour]
NA	Persistence	79.4	5.36
High	Gaussian GAM	82.6	4.59
	FCNN	81.8	4.39
Low	Ocat GAM	79.1	4.22
	FCNN	83.2	4.40
Multi	Ocat GAM	79.4	4.08
	CNN	**83.5**	**3.85**

IP multi-resolution models have a similar or better performance than high- and low-resolution alternatives within the same model class on the RMSE metric (Table 5) and the multi-resolution CNN is the best model under all metrics. However, the metrics are affected by high sampling variability. The reasons for this are detailed later in this section, where we also argue that the mediocre performance of ocat GAMs for IP forecasting is not fundamental, but attributable to the insufficient flexibility of the specific ocat parametrisation adopted here.

To quantify the variability of the performance metrics considered so far, we used block-bootstrap resampling. As described by [231], for a test set of size N, we sample with replacement data blocks of fixed size $B = 7$ (i.e., one week) to obtain an evaluation sets of size N. Repeating this procedure K times creates K metric samples, which can be used to estimate the metric's sampling variability. In particular, Figure 5.4 shows block-bootstrapped boxplots for all metrics and models on the last year of data. Figures 5.4 (a-c) clearly demonstrate that the improvement obtained by adopting a multi-resolution approach is substantial and robust within the GAM model class. The HR-FCNN is competitive in terms of prediction but, as we discuss below, it is not easily interpretable and does not have the computational advantages of multi-resolution GAMs. For the IP problem, Figures 5.4 (e-d) make clear that the sampling variability is substantial (reasons for this are discussed below).

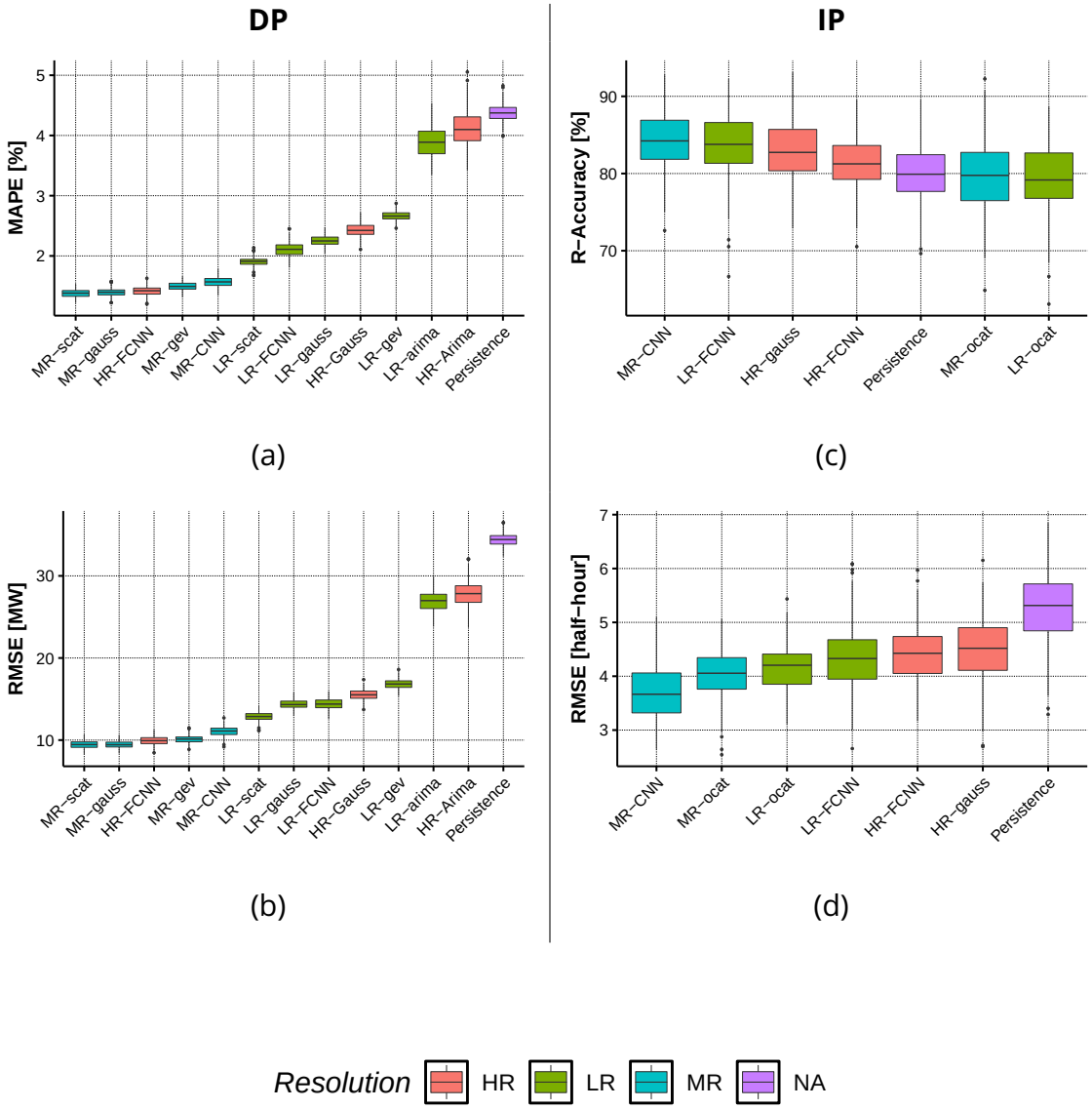


Figure 5.4 – Block-bootstrap boxplots of the two metrics considered for the DP models (a), (b) and IP models (c), (d) on the last year of data

As mentioned above, the rolling-origin forecasting setting may present a transitional regime during the first few training iterations. Figure 5.5 and 5.6 show the evolution of the different cumulative metrics calculated on the prediction signal updated on a monthly basis. Interestingly, the multi-resolution CNN for the DP (Figure 5.5) starts off with a very bad prediction error on the first months. With more data, its performance rapidly improves across all metrics. The other models have a less dramatic performance trend, with the multi-resolution GAMs consistently performing better than the other models. The prediction error of these models oscillates during the first few months, which can be explained by the fact that the models did not have enough information to adequately estimate the yearly cycle, because they were fitted to only one year of data.

After a year, the prediction errors has stabilised.

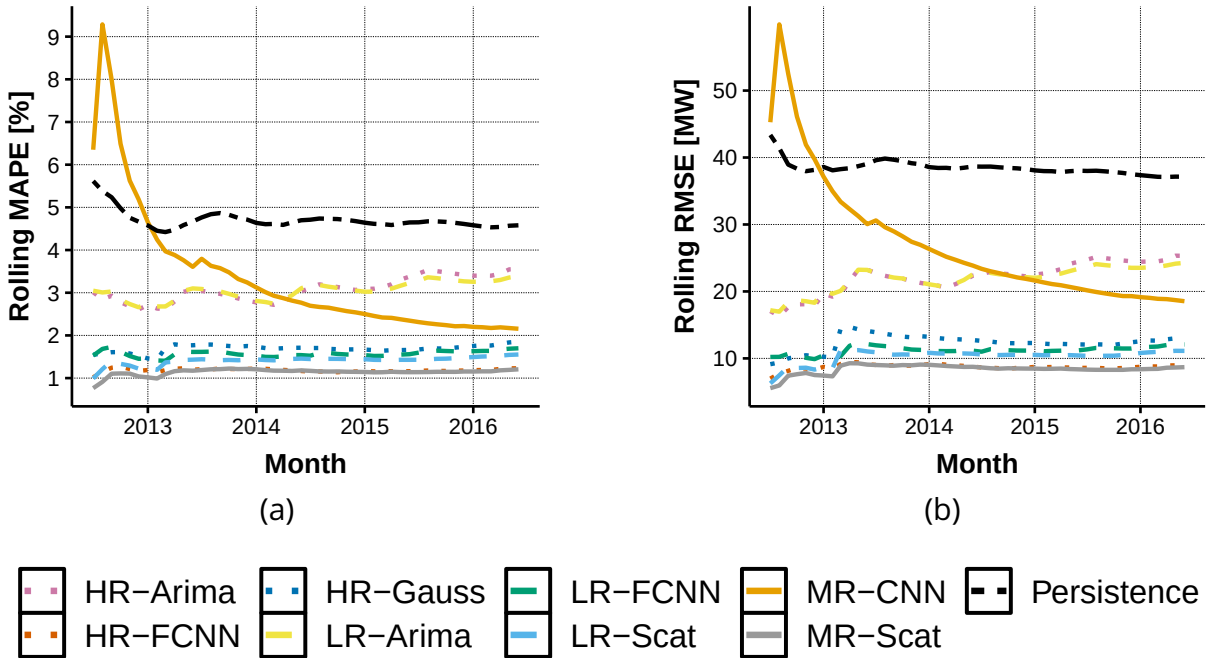


Figure 5.5 – Cumulative forecasting metrics evolution for each of the monthly updated DP models : (a) MAPE, (b) RMSE

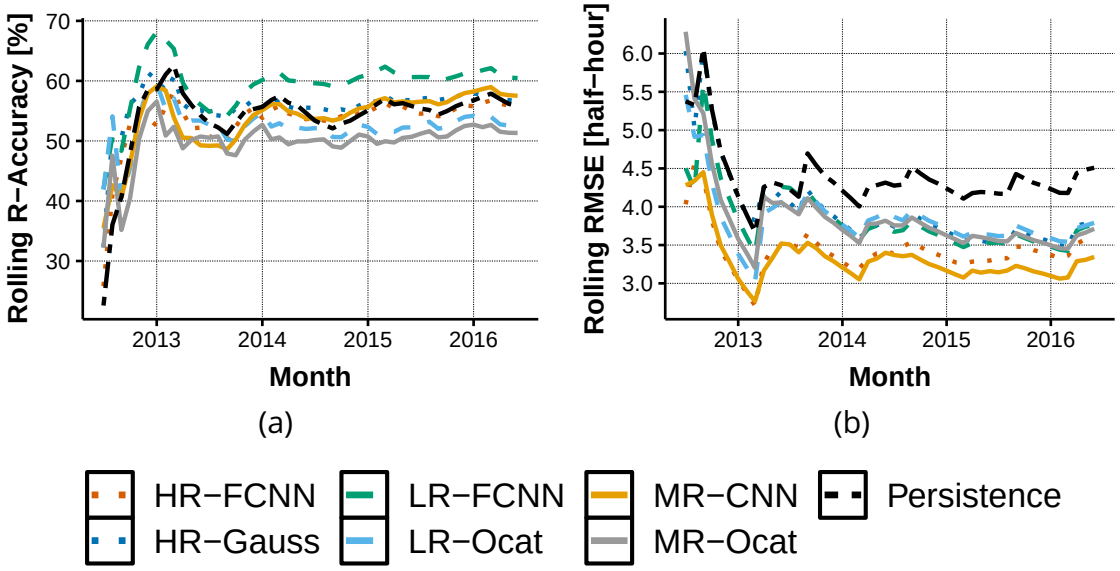


Figure 5.6 – Cumulative forecasting metrics evolution for each of the monthly updated IP models : (a) R-accuracy, (b) RMSE

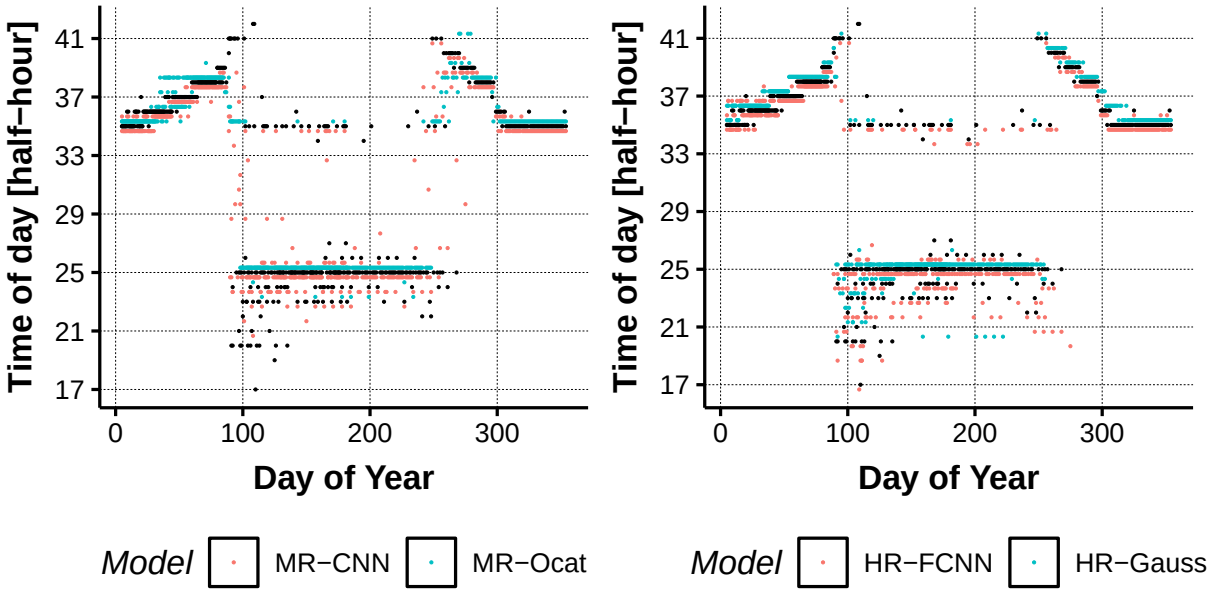


Figure 5.7 – Left : observed IP as a function of the day of year (black) and corresponding predictions from MR-CNN (red, shifted downward for visibility) and MR-ocat (blue, shifted upward). Right : same plot for HR-FCNN (red, downward) and HR-Gauss (blue, upward).

For the IP forecasting task, the different metrics evolve with similar patterns (Figure 5.6), but the seasonal oscillations in performance persist beyond the first year. Precisely, the performance slightly worsens after the first few months of each year. Figure 5.7 explains why predicting the IP is harder in summer (April to August) than in winter (September to March). While winter daily demand profiles have a reliable evening peak, summer load profiles are flatter and on some days the peak distribution becomes bimodal. That is, the daily peak might occur on the 25th half-hour (12.30pm) or on the 35th half-hour (5.30pm) with equal probability. This is shown also by the right plot in Figure 5.8. Hence, it is clear that in the summer the IP point estimates might be unfairly penalised under the metrics considered here. This implies that a forecasting model might be better off providing an IP forecast that falls between the two peaks, as MR-CNN is occasionally doing (see Figure 5.7). Such a forecast might improve the metrics but has little value in an operational setting. Note also that the ocat model struggles to capture an IP distribution that is unimodal or bimodal depending on the time of year. In particular, the ocat model used here is based on a standard ordered-logit parametrisation, which involves modelling the mean of a latent logistic random variable via an additive model. It is not possible to transform a unimodal distribution on the ordered categories (here, IP) into a bimodal one, simply by controlling a location parameter. Hence, a more flexible model (e.g., [307]) would be preferable.

It is interesting to verify the performance of each model for IP forecasting via a bespoke metric. In particular, let t_i^{m} be the observed IP on day i and let $\hat{t}_i^{\text{m}}$ be the corres-

ponding forecast. We propose the following metric :

$$\text{d-RMSE} = \left(\frac{1}{n} \sum_{i=1}^{n} (y_{t_i^{\text{m}}} - y_{\hat{t}_i^{\text{m}}}) \right)^{1/2} \tag{5.12}$$

which is based on the difference between the daily peak demand and the demand at the predicted IP (the d stands for demand). This metric is more relevant to operations than MSE. For instance, in peak shaving applications, providing a forecast $\hat{t}_i^{\text{m}}$ very different from t_i^{m} might not be a problem if $y_{t_i^{\text{m}}}$ and $y_{\hat{t}_i^{\text{m}}}$ are similar, which is what d-RMSE quantifies. Figure 5.8 shows a bootstrapped boxplot of d-RMSE for each model. Interestingly, high-resolution methods are best here, by a substantial margin in the case of HR-FCNN.

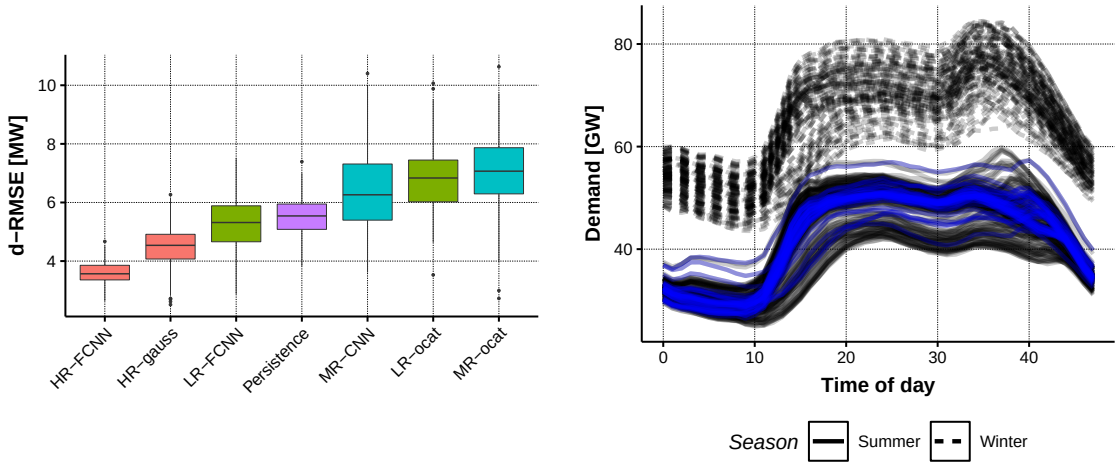


Figure 5.8 – Left : Block-bootstrap boxplots of the d-RMSE metric for the IP problem. Right : daily demand profile curves during winter (shifted upward by 15 GW) and summer. The blue curves are profiles with a small absolute difference between the morning and evening peak ($<$ 50 MW).

The results obtained so far do not provide reliable evidence in favour or against the adoption of a multi-resolution approach for IP forecasting. In fact, the poor forecasting performance of MR-ocat is arguably attributable to the particular ordered-logit parametrisation used here. MR-CNN does well using standard, statistically motivated losses but it is inferior to high-resolution approaches on an operationally relevant one (d-RMSE). It would be interesting to verify whether fitting the MR-CNN model by minimising d-RMSE directly (rather than MSE as done here) would lead to better results. We leave this, and the search for a more flexible distribution for ordered categorical responses, for future work.

Implementing the multi-resolution approach on the DP forecasting problem is more straightforward, hence the results discussed so far are positive and reliable. We further verify their significance by performing [308] (DM) tests on the absolute and squared error losses . The null hypothesis of the tests is : “both forecasts have the same expected loss”. The results of the DM tests are available on Figure 5.9 which confirms that,

within the GAM class, the multi-resolution forecasts are significantly different to the low-resolution and high-resolution approaches under both metrics.

Model	HR-arima	HR-gauss	HR-FCNN	LR-arima	LR-gauss	LR-scat	LR-gev	LR-FCNN	MR-gauss	MR-scat	MR-gev	MR-CNN
HR-arima		0	0	0.116	0	0	0	0	0	0	0	0
HR-gauss			0	0	0.042	0	0	0.001	0	0	0	0
HR-FCNN				0	0	0	0	0	0.729	0.549	0.250	0.029
LR-arima					0	0	0	0	0	0	0	0
LR-gauss						0	0	0.010	0	0	0	0
LR-scat							0	0.122	0	0	0	0
LR-gev								0	0	0	0	0
LR-FCNN									0	0	0	0
MR-gauss										0.063	0	0.001
MR-scat											0	0
MR-gev												0.143
MR-CNN												

(a)

Model	HR-arima	HR-gauss	HR-FCNN	LR-arima	LR-gauss	LR-scat	LR-gev	LR-FCNN	MR-gauss	MR-scat	MR-gev	MR-CNN
HR-arima		0	0	0.161	0	0	0	0	0	0	0	0
HR-gauss			0	0	0.010	0	0.003	0.118	0	0	0	0
HR-FCNN				0	0	0	0	0	0.400	0.362	0.492	0.016
LR-arima					0	0	0	0	0	0	0	0
LR-gauss						0	0	0.938	0	0	0	0
LR-scat							0	0.026	0	0	0	0.002
LR-gev								0	0	0	0	0
LR-FCNN									0	0	0	0
MR-gauss										0.468	0	0
MR-scat											0	0
MR-gev												0.015
MR-CNN												

(b)

Figure 5.9 – P-values from the Diebold-Mariano test for DP forecasts. The test used is from the *multDM* package in R [309]. In black, the null hypothesis is rejected at the 5% threshold and both forecasts are significantly different. In red, the null hypothesis is not rejected at the 5% threshold and both forecasts cannot be significantly differentiated; (a) absolute errors (b) squared errors.

It is interesting to quantify the complexity or parsimony of the models considered so far. AIC can be interpreted as a parsimony measure, but it requires computing the effective number of models parameters and we are not aware of any method that would allow estimating them across all the model classes considered here. Figure 5.10 shows the AICs of low- and multi-resolution GAMs. The multi-resolution approaches consistently have a smaller AIC than the low-resolution approaches. Furthermore, the slopes indicate that with more data the gap continues to increase.

For NNs, parsimony is highly dependent on the chosen architecture. In our case, the low-resolution and high-resolution NNs have a very similar architecture with only one hidden layer and a dropout layer. Only the inputs, outputs and number of observations vary. On the other hand, the multi-resolution NN (Figure 5.3) requires the use of convolutional layers which are leveraged to extract the high-resolution information. The extraction process requires multiple layers which forces the multi-resolution CNN to have a larger number of parameters than the low-resolution and high-resolution NNs.

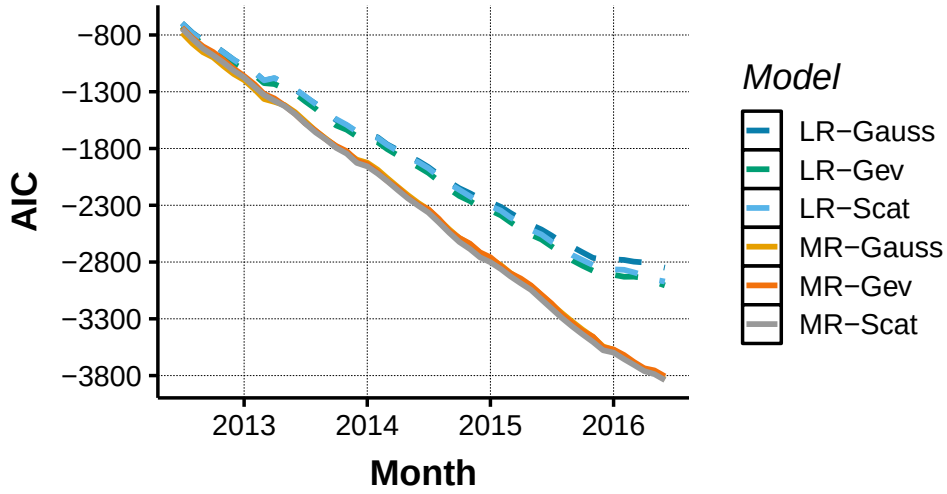


Figure 5.10 – AIC for the low-resolution and multi-resolution DP GAMs

The results discussed in this section show that multi-resolution approaches are superior to low- and high-resolution alternatives for the DP forecasting problem. The forecasting performance of the high-resolution FCNN and the multi-resolution GAMs are not significantly different but, in an operational peak demand forecasting context, the multi-resolution GAM would be preferred because it can be decomposed into additive components, which can be more easily interpreted (and manually adjusted) by operational staff. In addition, note that adopting a multi-resolution approach can bring substantial computational advantages, which are easy to quantify within the GAM model class. In particular, the GAM model matrix $\mathbf{X}$ in the multi-resolution case has T times less rows than in the high-resolution case, where T is the number of daily observations (i.e., $T = 48$ for half-hourly data). Therefore, T times less memory is used, and many computations frequently required during GAM model fitting (such as $\mathbf{X}^T\mathbf{W}\mathbf{X}$, where $\mathbf{W}$ is a diagonal matrix) will take less time.

5.6 Conclusion

This paper proposes a novel modelling approach, which uses both high-resolution and low-resolution information to forecast the daily electrical load peak magnitude and timing. The results demonstrate that this multi-resolution approach is flexible enough to be applied to different model classes and that it provides a competitive predictive performance. In particular, GAMs and NNs with similar input structures were used to implement the multi-resolution approach and to compare its performance that of low-resolution, high-resolution and persistence alternatives. On UK aggregate demand data, the multi-resolution models performed significantly better across all metrics when forecasting peak magnitude. In addition to improved predictions, adopting a multi-resolution approach enables faster computation via data compression and leads to more parsimonious models, as demonstrated by the consistently lower AIC scores achieved by multi-resolution models within the GAM model class.

The results on the peak timing forecasting problem are mixed, but interesting. A multi-resolution neural network does marginally better than the alternatives, when per-

formance is assessed via standard statistical metrics. However, the corresponding forecast is occasionally inappropriate (falling between the morning and evening peaks) and inferior to high-resolution alternatives when assessed via an operationally motivated metric. The results suggest that the multi-resolution neural network should be fitted to data by minimising a problem specific performance metric directly. For instance, one could consider financial metrics on billing periods as done by [277]. The multi-resolution GAM does poorly on the peak timing problem, but this is attributable to the insufficient flexibility of the ordered logit parametrisation used here. Obtaining stronger evidence in favour or against the use of multi-resolution methods for the peak timing problem would require solving the issues just mentioned, which could be the subject of further work.

The forecasting methods presented here could be extended in several ways. Firstly, the set of models described in this paper could be used within an aggregation of experts or ensemble methods, which might lead to more accurate forecasts. Secondly, the benefits of multi-resolution methods have been demonstrated in a context where the covariates were available at different temporal resolutions, but the underlying idea could be generalised to other settings, such as spatio-temporal data or individual customer data (see e.g., [310] for an example application of functional quantile GAMs [311] to residential electricity demand data). Thirdly, the multi-resolution approach could be extended to RNNs and in particular LSTMs, which are known to be powerful NN architecture for time series modelling. Finally, this paper focused on day-ahead daily peak magnitude and time forecasting, but multi-resolution methods could be applied to other short-term windows (e.g., weekly). However, estimating monthly or yearly peaks would require a different approach, because the number of observed demand peaks would be too low.

Conclusions and Perspectives

This research set out to better understand EV charging behaviours. In particular, we focused on forecasting EV charging load. By modelling EV charging behaviours more effectively, we can enable more resilience for the grid. To this end, we reviewed existing methods and proposed innovative statistical modelling approaches to answer three critical questions : 1. Which models are state-of-the-art for EV charging modeling and what data is available ? 2. How can we implement better forecasts ? 3. How can daily electrical consumption peaks be predicted to avoid blackouts and electricity price surges ?.

In Chapter 2, we provide a review of state-of-the-art models spanning the last thirty years. An exploratory method is adopted to identify the most relevant literature and data. A taxonomy of models is proposed ranging from stochastic processes to machine learning. Overall, we covered over 860 databases and explored 8 EV charging session datasets. These datasets were used in the rest of the PhD to craft innovative methods and to assess their performances on real-world scenarios. More data repositories are emerging in relation to EVs and it is crucial to keep collecting them in order to keep proposing reproducible and relevant work for the community. In Chapter 3, we compare 14 models with direct and bottom-up approaches to predict the load and occupancy curves of EV charging points. Overall, we demonstrate that the 14 models provide efficient forecasts at different times, which allows us to obtain a better model with an adaptive aggregation strategy. The 14 models emerged from a discussion over the most promising methods presented in Chapter 2 and novel methods that we introduced throughout the PhD. One advantage of this aggregation strategy is that any new promising method can be added to this mix and may improve the aggregated forecast. Future work involves looking at potential benefits of using all the datasets collected in a holistic way to derive model parameters. Indeed, the models trained in this work were fitted on each dataset separately. However, it would be interesting to assess whether a model can benefit from data collected in various municipalities to learn behaviours that are shared across different locations and thus reach better overall performances. We also see a strong potential in combining the station data collected with traffic information for more accurate models. It is common practice within the electrical load forecasting domain to use weather data to model general electrical consumption. However, when it comes to EVs, weather has little correlation with load consumption relative to traffic. Faced with the scarcity of traffic data at a fine granularity for EVs, and until access to this data is provided, a solution could be to scrape real-time traffic data. For example the city of Bonn in Germany provides both real-time traffic and EVSE usage information at a fine granularity. In Chapter 4, we studied an additive model with both wavelet and spline components for Poisson regression. Two algorithms have been proposed, OBO and BAC. We showed that BAC performs better on one of the training datasets but a significant conclusion could be reached on the test sets for the first case study. However, we observed that there could be some benefit in using a model with additive spline and wavelet effects to capture peaks when looking at performances on peak load. The proposed methodology can be extended to any timestep (as long as it is constant), as

well as any other wavelet basis. The algorithm's convergence speed could be accelerated by fitting first the linear component and then backfitting the non-parametric part. Giving splines more degrees of freedom (knots) can improve performance, but it can also lead to overfitting. One limitation of the proposed `BAC` algorithm is that, despite empirical evidence of algorithm convergence, there is no unicity of the parameter vector estimated. This is due in large part to the random path chosen to fit the various components. Finally, other metrics could have been used to evaluate peak estimates, as it can be argued from an operational standpoint that obtaining a better estimate of the peak magnitude not too far from peak time is more important than obtaining a worse estimate exactly at peak time. As a result, it might be interesting to compare the forecasted daily time series with the observed one using a measure of similarity, such as dynamic time warping. Eventually, in Chapter 5, we rely on high- and low-resolution information to predict the daily peak in consumption. To use them in the same model, we defined a modelling framework called : "multi-resolution". The results confirm that the proposed approach provides better performance on the cases studied. The forecasting techniques described here could be expanded in a number of different ways. First, an aggregation of experts (such as the one proposed in Chapter 3) might be employed with the set of models provided in this paper to produce predictions that may be more accurate. Second, although the advantages of multi-resolution methods have been shown in a setting where the covariates were available at various temporal resolutions, the underlying concept could be applied to other settings, such as spatial and temporal or individual customer data. Thirdly, since RNNs have shown a large success for time series modelling, the multi-resolution technique might be extended to RNNs and in particular LSTMs. Finally, while this chapter focused on daily peak magnitude and timing forecasting for the day ahead, multi-resolution methods could be applied to other short-term windows (e.g., weekly). Estimating monthly or yearly peaks, on the other hand, would necessitate a different approach due to the insufficient number of observed demand peaks.

Ultimately, while forecasting EV load is currently one of the most important tasks to ensure that EV charging behaviours do not disrupt the balance of the grid, it has to be combined with optimisation techniques. In practice, good forecasts of tomorrow's EV demand can be useless if there is no plan of action implemented at charging point to accommodate demand surges. This is the work of optimisers some of who we worked with during this PhD at EDF R&D. They used our forecasts as an input to their optimisation procedure in order to assess the gains with regards to various metrics (e.g., CO_2 emissions, electricity cost). In particular, they used our work on forecasting individual charging sessions which is what we call bottom-up approaches in Chapter 3. With the market expanding, some intermediary level forecast might also be required as some of the methods might not scale as well. In fact, the EV market is still in its infancy which indicates that many challenges are still to come. We will be highly attentive to new developments in the field and we hope that our work will continue to contribute to a resilient electrical network.

Bibliographie

[1] Y. Amara, "Yvenn amara - github." [Online]. Available : https://github.com/yvenn-amara

[2] International Energy Agency, *Global EV Outlook 2013 : Understanding the Electric Vehicle Landscape to 2020*. OECD, Apr. 2013.

[3] IEA, *Global EV Outlook 2022 : Securing supplies for an electric future*. OECD, May 2022.

[4] UNCC, "Report of the Conference of the Parties on its twenty-first session, held in Paris from 30 November to 13 December 2015," UNCC, Paris, Tech. Rep., 2016. [Online]. Available : https://unfccc.int/resource/docs/2015/cop21/eng/10a01.pdf#page=2

[5] E. Francisco and J. Kaiser, "Tales from the Blackout," *Science*, vol. 301, no. 5636, pp. 1029–1029, Aug. 2003. [Online]. Available : https://www.science.org/doi/10.1126/science.301.5636.1029b

[6] J. A. Sanguesa, V. Torres-Sanz, P. Garrido, F. J. Martinez, and J. M. Marquez-Barja, "A review on electric vehicles : Technologies and challenges," *Smart Cities*, vol. 4, no. 1, pp. 372–404, 2021. [Online]. Available : https://www.mdpi.com/2624-6511/4/1/22

[7] H. Das, M. Rahman, S. Li, and C. Tan, "Electric vehicles standards, charging infrastructure, and impact on grid integration : A technological review," *Renewable and Sustainable Energy Reviews*, vol. 120, p. 109618, 2020. [Online]. Available : https://www.sciencedirect.com/science/article/pii/S1364032119308251

[8] S. Abbas, X. Ai, A. Masood, S. Iqbal, and M. Jan, "Electric vehicles and their impacts on integration into power grid : A review," 12 2018.

[9] L. Daniels, B. Groos, L. Christy, and S. Laurie, "More EVs, Fewer Emissions," Mar. 2022. [Online]. Available : https://rmi.org/insight/more-evs-fewer-emissions

[10] M. A. Ilgin and S. M. Gupta, "Environmentally conscious manufacturing and product recovery (ecmpro) : A review of the state of the art," *Journal of Environmental Management*, vol. 91, no. 3, pp. 563–591, 2010. [Online]. Available : https://www.sciencedirect.com/science/article/pii/S0301479709003417

[11] X. Xia and P. Li, "A review of the life cycle assessment of electric vehicles : Considering the influence of batteries," *Science of The Total Environment*, vol. 814, p. 152870, 2022. [Online]. Available : https://www.sciencedirect.com/science/article/pii/S0048969721079493

[12] S. Verma, G. Dwivedi, and P. Verma, "Life cycle assessment of electric vehicles in comparison to combustion engine vehicles : A review," *Materials Today : Proceedings*, vol. 49, pp. 217–222, 2022, international Conference on Advancement in Materials, Manufacturing and Energy Engineering (ICAMME-2021). [Online]. Available : https://www.sciencedirect.com/science/article/pii/S221478532100763X

[13] F. D. Pero, M. Delogu, and M. Pierini, "Life cycle assessment in the automotive sector : a comparative case study of internal combustion engine (ice) and electric car," *Procedia Structural Integrity*, vol. 12, pp. 521–537, 2018, aIAS 2018 international conference on stress analysis. [Online]. Available : https://www.sciencedirect.com/science/article/pii/S2452321618301690

[14] S. Xiong, Y. Wang, B. Bai, and X. Ma, "A hybrid life cycle assessment of the large-scale application of electric vehicles," *Energy*, vol. 216, p. 119314, 2021. [Online]. Available : https://www.sciencedirect.com/science/article/pii/S036054422032421X

[15] S. Xiong, J. Ji, and X. Ma, "Environmental and economic evaluation of remanufacturing lithium-ion batteries from electric vehicles," *Waste Management*, vol. 102, pp. 579–586, 2020. [Online]. Available : https://www.sciencedirect.com/science/article/pii/S0956053X19307081

[16] R. Sharpe, "Scientists should consider how materials can be repurposed as they design them.","," *Nature*, p. 1, Jul. 2021. [Online]. Available : https://www.nature.com/articles/d41586-021-01735-z

[17] B. K. Sovacool, "When subterranean slavery supports sustainability transitions? power, patriarchy, and child labor in artisanal Congolese cobalt mining," *The Extractive Industries and Society*, vol. 8, no. 1, pp. 271–293, 2021. [Online]. Available : https://www.sciencedirect.com/science/article/pii/S2214790X20303154

[18] Y. Amara-Ouali, Y. Goude, P. Massart, J.-M. Poggi, and H. Yan, "A Review of Electric Vehicle Load Open Data and Models," *Energies*, vol. 14, no. 8, p. 2233, Jan. 2021, number : 8 Publisher : Multidisciplinary Digital Publishing Institute. [Online]. Available : https://www.mdpi.com/1996-1073/14/8/2233

[19] International Energy Agency, *Global EV Outlook 2019 : Scaling-up the transition to electric mobility*. OECD, Jun. 2019. [Online]. Available : https://www.oecd-ilibrary.org/energy/global-ev-outlook-2019_35fb60bd-en

[20] H. Kvisle, "The Norwegian Charging Station Database for Electromobility (NOBIL)," *World Electric Vehicle Journal*, vol. 5, no. 3, pp. 702–707, Sep. 2012. [Online]. Available : http://www.mdpi.com/2032-6653/5/3/702

[21] "International Energy Agency, Electric vehicle stock in the EV30@30 scenario," library Catalog : www.iea.org. [Online]. Available : https://www.iea.org/data-and-statistics/charts/electric-vehicle-stock-in-the-ev3030-scenario-2018-2030

[22] "Clean Energy Ministerial, Electric Vehicles Initiative." [Online]. Available : http://www.cleanenergyministerial.org/initiative-clean-energy-ministerial/electric-vehicles-initiative

[23] M. D. Galus, M. G. Vayá, T. Krause, and G. Andersson, "The role of electric vehicles in smart grids : The role of electric vehicles in smart grids," *Wiley Interdisciplinary Reviews : Energy and Environment*, vol. 2, no. 4, pp. 384–400, Jul. 2013. [Online]. Available : http://doi.wiley.com/10.1002/wene.56

[24] K. L. Lopez, C. Gagne, and M.-A. Gardner, "Demand-Side Management Using Deep Learning for Smart Charging of Electric Vehicles," *IEEE Transactions on Smart Grid*, vol. 10, no. 3, pp. 2683–2691, May 2019. [Online]. Available : https://ieeexplore.ieee.org/document/8299470/

[25] R. Fachrizal, M. Shepero, D. van der Meer, J. Munkhammar, and J. Widén, "Smart charging of electric vehicles considering photovoltaic power production and electricity consumption : A review," *eTransportation*, vol. 4, p. 100056, May 2020. [Online]. Available : https://linkinghub.elsevier.com/retrieve/pii/S2590116820300138

[26] E. S. Xydas, C. E. Marmaras, L. M. Cipcigan, A. S. Hassan, and N. Jenkins, "Forecasting Electric Vehicle charging demand using Support Vector Machines," in *2013 48th International Universities' Power Engineering Conference (UPEC)*. Dublin : IEEE, Sep. 2013, pp. 1–6. [Online]. Available : http://ieeexplore.ieee.org/document/6714942/

[27] Y. Mu, J. Wu, N. Jenkins, H. Jia, and C. Wang, "A Spatial–Temporal model for grid impact analysis of plug-in electric vehicles," *Applied Energy*, vol. 114, pp. 456–465, Feb. 2014. [Online]. Available : https://linkinghub.elsevier.com/retrieve/pii/S030626191300826X

[28] S. F. Tie and C. W. Tan, "A review of energy sources and energy management system in electric vehicles," *Renewable and Sustainable Energy Reviews*, vol. 20, pp. 82–102, Apr. 2013. [Online]. Available : https://linkinghub.elsevier.com/retrieve/pii/S1364032112006910

[29] K. Yamashita, J. Li, P. Zhang, and C.-C. Liu, "Analysis and control of major blackout events," in *2009 IEEE/PES Power Systems Conference and Exposition*, Mar. 2009, pp. 1–4.

[30] A. Gerossier, R. Girard, and G. Kariniotakis, "Modeling and Forecasting Electric Vehicle Consumption Profiles," *Energies*, vol. 12, no. 7, p. 1341, Apr. 2019. [Online]. Available : https://www.mdpi.com/1996-1073/12/7/1341

[31] Y. B. Khoo, C.-H. Wang, P. Paevere, and A. Higgins, "Statistical modeling of Electric Vehicle electricity consumption in the Victorian EV Trial, Australia," *Transportation Research Part D : Transport and Environment*, vol. 32, pp. 263–277, Oct. 2014. [Online]. Available : https://linkinghub.elsevier.com/retrieve/pii/S1361920914001187

[32] K. Sun, M. R. Sarker, and M. A. Ortega-Vazquez, "Statistical characterization of electric vehicle charging in different locations of the grid," in *2015 IEEE Power & Energy Society General Meeting*. Denver, CO, USA : IEEE, Jul. 2015, pp. 1–5. [Online]. Available : http://ieeexplore.ieee.org/document/7285794/

[33] K. Qian, C. Zhou, M. Allan, and Y. Yuan, "Modeling of Load Demand Due to EV Battery Charging in Distribution Systems," *IEEE Transactions on Power Systems*, vol. 26, no. 2, pp. 802–810, May 2011. [Online]. Available : http://ieeexplore.ieee.org/document/5535237/

[34] M. Alizadeh, A. Scaglione, J. Davies, and K. S. Kurani, "A Scalable Stochastic Model for the Electricity Demand of Electric and Plug-In Hybrid Vehicles," *IEEE Transactions on Smart Grid*, vol. 5, no. 2, pp. 848–860, Mar. 2014. [Online]. Available : http://ieeexplore.ieee.org/document/6595730/

[35] N. Z. Xu and C. Y. Chung, "Challenges in Future Competition of Electric Vehicle Charging Management and Solutions," *IEEE Transactions on Smart Grid*, vol. 6, no. 3, pp. 1323–1331, May 2015. [Online]. Available : http://ieeexplore.ieee.org/document/6990624/

[36] H. Jiang, H. Ren, C. Sun, and D. Watts, "The temporal-spatial stochastic model of plug-in hybrid electric vehicles," in *2017 IEEE PES Innovative Smart Grid Technologies Conference Europe (ISGT-Europe)*. Torino, Italy : IEEE, Sep. 2017, pp. 1–6. [Online]. Available : http://ieeexplore.ieee.org/document/8260233/

[37] D. B. Richardson, "Electric vehicles and the electric grid : A review of modeling approaches, Impacts, and renewable energy integration," *Renewable and Sustainable Energy Reviews*, vol. 19, pp. 247–254, Mar. 2013. [Online]. Available : https://linkinghub.elsevier.com/retrieve/pii/S1364032112006557

[38] G. T. Heydt, "The impact of electric vehicle deployment on load management stra-regies," *IEEE transactions on power apparatus and systems*, pp. 1253–1259., 1983.

[39] N. Sadeghianpourhamami, N. Refa, M. Strobbe, and C. Develder, "Quantitive analysis of electric vehicle flexibility : A data-driven approach," *International Journal of Electrical Power & Energy Systems*, vol. 95, pp. 451–462, Feb. 2018. [Online]. Available : https://linkinghub.elsevier.com/retrieve/pii/S0142061516323687

[40] E. C. Kara, J. S. Macdonald, D. Black, M. Bérges, G. Hug, and S. Kiliccote, "Estimating the benefits of electric vehicle smart charging at non-residential locations : A data-driven approach," *Applied Energy*, vol. 155, pp. 515–525, Oct. 2015. [Online]. Available : https://linkinghub.elsevier.com/retrieve/pii/S0306261915007059

[41] W. Kempton and S. E. Letendre, "Electric vehicles as a new power source for electric utilities," *Transportation Research Part D : Transport and Environment*, vol. 2, no. 3, pp. 157–175, Sep. 1997. [Online]. Available : https://linkinghub.elsevier.com/retrieve/pii/S1361920997000011

[42] M. G. Flammini, G. Prettico, A. Julea, G. Fulli, A. Mazza, and G. Chicco, "Statistical characterisation of the real transaction data gathered from electric vehicle charging stations," *Electric Power Systems Research*, vol. 166, pp. 136–150, Jan. 2019. [Online]. Available : https://linkinghub.elsevier.com/retrieve/pii/S037877961830316X

[43] G. Saldaña, J. I. San Martin, I. Zamora, F. J. Asensio, and O. Oñederra, "Electric Vehicle into the Grid : Charging Methodologies Aimed at Providing Ancillary Services Considering Battery Degradation," *Energies*, vol. 12, no. 12, p. 2443, Jun. 2019. [Online]. Available : https://www.mdpi.com/1996-1073/12/12/2443

[44] B. Chokkalingam, S. Padmanaban, P. Siano, R. Krishnamoorthy, and R. Selvaraj, "Real-Time Forecasting of EV Charging Station Scheduling for Smart Energy Systems," *Energies*, vol. 10, no. 3, p. 377, Mar. 2017. [Online]. Available : http://www.mdpi.com/1996-1073/10/3/377

[45] M. A. Ahmed, M. R. El-Sharkawy, and Y.-C. Kim, "Remote Monitoring of Electric Vehicle ChargingStations in Smart Campus Parking Lot," *Journal of Modern Power Systems and Clean Energy*, vol. 8, no. 1, pp. 124–132, 2020. [Online]. Available : https://ieeexplore.ieee.org/document/8913668

[46] J. Zhu, Z. Yang, Y. Guo, J. Zhang, and H. Yang, "Short-Term Load Forecasting for Electric Vehicle Charging Stations Based on Deep Learning Approaches," *Applied Sciences*, vol. 9, no. 9, p. 1723, Apr. 2019. [Online]. Available : https://www.mdpi.com/2076-3417/9/9/1723

[47] R. Garcia-Valle and J. G. Vlachogiannis, "Letter to the Editor : Electric Vehicle Demand Model for Load Flow Studies," *Electric Power Components and Systems*, vol. 37, no. 5, pp. 577–582, Apr. 2009. [Online]. Available : http://www.tandfonline.com/doi/abs/10.1080/15325000802599411

[48] P. Olivella-Rosell, R. Villafafila-Robles, A. Sumper, and J. Bergas-Jané, "Probabilistic Agent-Based Model of Electric Vehicle Charging Demand to Analyse the Impact on Distribution Networks," *Energies*, vol. 8, no. 5, pp. 4160–4187, May 2015. [Online]. Available : http://www.mdpi.com/1996-1073/8/5/4160

[49] S. Xydas, C. Marmaras, L. Cipcigan, A. Hassan, and N. Jenkins, "Electric Vehicle Load Forecasting using Data Mining Methods," in *Hybrid and Electric Vehicles Conference 2013 (HEVC 2013)*. London, UK : Institution of Engineering and Technology, 2013, pp. 10.1–10.1. [Online]. Available : https://digital-library.theiet.org/content/conferences/10.1049/cp.2013.1914

[50] M. Neaimeh, R. Wardle, A. M. Jenkins, J. Yi, G. Hill, P. F. Lyons, Y. Hübner, P. T. Blythe, and P. C. Taylor, "A probabilistic approach to combining smart meter and electric vehicle charging data to investigate distribution network impacts," *Applied Energy*, vol. 157, pp. 688–698, Nov. 2015. [Online]. Available : https://linkinghub.elsevier.com/retrieve/pii/S0306261915001944

[51] M. Majidpour, C. Qiu, P. Chu, H. R. Pota, and R. Gadh, "Forecasting the EV charging load based on customer profile or station measurement?" *Applied Energy*, vol. 163, pp. 134–141, Feb. 2016. [Online]. Available : https://linkinghub.elsevier.com/retrieve/pii/S0306261915014348

[52] J. Li, J. Klee Barillas, C. Guenther, and M. A. Danzer, "A comparative study of state of charge estimation algorithms for LiFePO4 batteries used in electric vehicles," *Journal of Power Sources*, vol. 230, pp. 244–250, May 2013. [Online]. Available : https://linkinghub.elsevier.com/retrieve/pii/S0378775312019039

[53] J. Topić, B. Škugor, and J. Deur, "Neural Network-Based Modeling of Electric Vehicle Energy Demand and All Electric Range," *Energies*, vol. 12, no. 7, p. 1396, Apr. 2019. [Online]. Available : https://www.mdpi.com/1996-1073/12/7/1396

[54] M. H. Amini and M. P. Moghaddam, "Probabilistic modelling of electric vehicles' parking lots charging demand," in *2013 21st Iranian Conference on Electrical Engineering (ICEE)*, May 2013, pp. 1–4, iSSN : 2164-7054.

[55] M. Amini, O. Karabasoglu, M. D. Ilic, K. G. Boroojeni, and S. S. Iyengar, "ARIMA-based demand forecasting method considering probabilistic model of electric vehicles' parking lots," in *2015 IEEE Power & Energy Society General Meeting*. Denver, CO, USA : IEEE, Jul. 2015, pp. 1–5. [Online]. Available : http://ieeexplore.ieee.org/document/7286050/

[56] G. Li and X.-P. Zhang, "Modeling of Plug-in Hybrid Electric Vehicle Charging Demand in Probabilistic Power Flow Calculations," *IEEE Transactions on Smart Grid*, vol. 3, no. 1, pp. 492–499, Mar. 2012. [Online]. Available : http://ieeexplore.ieee.org/document/6145673/

[57] H. Liang, I. Sharma, W. Zhuang, and K. Bhattacharya, "Plug-in electric vehicle charging demand estimation based on queueing network analysis," in *2014 IEEE*

PES General Meeting | Conference & Exposition. National Harbor, MD, USA : IEEE, Jul. 2014, pp. 1–5. [Online]. Available : http://ieeexplore.ieee.org/document/6939530/

[58] P. Paevere, A. Higgins, Z. Ren, M. Horn, G. Grozev, and C. McNamara, "Spatio-temporal modelling of electric vehicle charging demand and impacts on peak household electrical load," *Sustainability Science*, vol. 9, no. 1, pp. 61–76, Jan. 2014. [Online]. Available : http://link.springer.com/10.1007/s11625-013-0235-3

[59] J. Huber, D. Dann, and C. Weinhardt, "Probabilistic forecasts of time and energy flexibility in battery electric vehicle charging," *Applied Energy*, vol. 262, p. 114525, Mar. 2020. [Online]. Available : https://linkinghub.elsevier.com/retrieve/pii/S0306261920300374

[60] H. Wang, "Energy consumption of electric vehicles based on real-world driving patterns : A case study of Beijing," *Applied Energy*, p. 10, 2015.

[61] "National Household Travel Survey." [Online]. Available : https://nhts.ornl.gov/

[62] A. Lojowska, D. Kurowicka, G. Papaefthymiou, and L. van der Sluis, "Stochastic Modeling of Power Demand Due to EVs Using Copula," *IEEE Transactions on Power Systems*, vol. 27, no. 4, pp. 1960–1968, Nov. 2012. [Online]. Available : http://ieeexplore.ieee.org/document/6193193/

[63] T. Hong and S. Fan, "Probabilistic electric load forecasting : A tutorial review," *International Journal of Forecasting*, vol. 32, no. 3, pp. 914–938, Jul. 2016. [Online]. Available : https://linkinghub.elsevier.com/retrieve/pii/S0169207015001508

[64] P. Gaillard, Y. Goude, and R. Nedellec, "Additive models and robust aggregation for GEFCom2014 probabilistic electric load and electricity price forecasting," *International Journal of Forecasting*, vol. 32, no. 3, pp. 1038–1050, Jul. 2016. [Online]. Available : https://linkinghub.elsevier.com/retrieve/pii/S0169207015001545

[65] E. Xydas, C. Marmaras, L. M. Cipcigan, N. Jenkins, S. Carroll, and M. Barker, "A data-driven approach for characterising the charging demand of electric vehicles : A UK case study," *Applied Energy*, vol. 162, pp. 763–771, Jan. 2016. [Online]. Available : https://linkinghub.elsevier.com/retrieve/pii/S0306261915013938

[66] M. B. Arias and S. Bae, "Electric vehicle charging demand forecasting model based on big data technologies," *Applied Energy*, vol. 183, pp. 327–339, Dec. 2016. [Online]. Available : https://linkinghub.elsevier.com/retrieve/pii/S0306261916311667

[67] Y. Li, Y. Huang, and M. Zhang, "Short-Term Load Forecasting for Electric Vehicle Charging Station Based on Niche Immunity Lion Algorithm and Convolutional Neural Network," *Energies*, vol. 11, no. 5, p. 1253, May 2018. [Online]. Available : http://www.mdpi.com/1996-1073/11/5/1253

[68] E. Valsera-Naranjo, A. Sumper, R. Villafafila-Robles, and D. Martínez-Vicente, "Probabilistic Method to Assess the Impact of Charging of Electric Vehicles on Distribution Grids," *Energies*, vol. 5, no. 5, pp. 1503–1531, May 2012. [Online]. Available : http://www.mdpi.com/1996-1073/5/5/1503

[69] S. Huang and D. Infield, "The potential of domestic electric vehicles to contribute to Power System Operation through vehicle to grid technology," in *2009 44th International Universities Power Engineering Conference (UPEC)*, Sep. 2009, pp. 1–5.

[70] J. C. Kelly, J. S. MacDonald, and G. A. Keoleian, "Time-dependent plug-in hybrid electric vehicle charging based on national driving patterns and demographics," *Applied Energy*, vol. 94, pp. 395–405, Jun. 2012. [Online]. Available : https://linkinghub.elsevier.com/retrieve/pii/S0306261912000931

[71] Y. Wang, Q. Chen, T. Hong, and C. Kang, "Review of Smart Meter Data Analytics : Applications, Methodologies, and Challenges," *IEEE Transactions on Smart Grid*, vol. 10, no. 3, pp. 3125–3148, May 2019. [Online]. Available : https://ieeexplore.ieee.org/document/8322199/

[72] "Open Data Inception," library Catalog : data.opendatasoft.com. [Online]. Available : https://data.opendatasoft.com/explore/dataset/open-data-sources@public/table/

[73] M. Salmon, "riem : Accesses Weather Data from the Iowa Environment Mesonet," Sep. 2016. [Online]. Available : https://CRAN.R-project.org/package=riem

[74] NOAA, "National Oceanic and Atmospheric Administration, Data Access | National Centers for Environmental Information (NCEI) formerly known as National Climatic Data Center (NCDC)." [Online]. Available : https://www.ncdc.noaa.gov/data-access

[75] "International Energy Agency, Electric car market shares in Electric Vehicle Initiative countries," Nov. 2019, library Catalog : www.iea.org. [Online]. Available : https://www.iea.org/data-and-statistics/charts/electric-car-market-shares-in-electric-vehicle-initiative-evi-countries

[76] "Electric vehicle market statistics 2020 - How many electric cars in UK?" [Online]. Available : https://www.nextgreencar.com/electric-cars/statistics/

[77] NOBIL, "NOBIL Database." [Online]. Available : https://info.nobil.no/eng

[78] DataNorge, "Traffic Volumes." [Online]. Available : https://data.norge.no/data/statens-vegvesen/trafikkmengde-%C3%A5rsd%C3%B8gntrafikk-%C3%A5dt-p%C3%A5-esri-geodatabase-format-20130806

[79] StatisticsNorway, "Road traffic volumes, by main type of vehicle, type of fuel and age of vehicle 2005 - 2019," library Catalog : www.ssb.no. [Online]. Available : https://www.ssb.no/en/transport-og-reiseliv/statistikker/klreg/aar

[80] ——, "Registered Vehicles," library Catalog : www.ssb.no. [Online]. Available : https://www.ssb.no/en/transport-og-reiseliv/statistikker/bilreg/aar/2020-03-31

[81] Reykjavík, "Reykjavík Transport," library Catalog : opingogn.is. [Online]. Available : https://opingogn.is/dataset/reykjavik-samgongur

[82] StatisticsIceland, "Registered Motor Vehicles," library Catalog : www.statice.is. [Online]. Available : https://www.statice.is/statistics/business-sectors/transport/vehicles/

[83] ElbilSverige, "Charging point locations in nordic countries," library Catalog : www.elbilsverige.se. [Online]. Available : http://www.elbilsverige.se/hitta-laddplats/

[84] StatisticsSweden, "Registered vehicles," library Catalog : www.statistikdatabasen.scb.se. [Online]. Available : http://www.

statistikdatabasen.scb.se/pxweb/sv/ssd/START__TK__TK1001__TK1001A/PersBilarDrivMedel/

[85] RotterdamOpenData, "Charging point locations and usage in Rotterdam," library Catalog : rotterdamopendata.nl. [Online]. Available : http://rotterdamopendata.nl/dataset/oplaadpalen-voor-elektrische-auto-s

[86] EindhovenOpenData, "Charging Point locations in Eindhoven," library Catalog : data.eindhoven.nl. [Online]. Available : https://data.eindhoven.nl/explore/dataset/oplaadpalen/information/?disjunctive.adres

[87] "Elaad NL, Data Analytics." [Online]. Available : https://www.elaad.nl/research/data-analytics/

[88] CBS, "Centraal Bureau Voor De Statistiek : Motor vehicles; type, age class." [Online]. Available : https://opendata.cbs.nl/statline/portal.html?_la=en&_catalog=CBS&tableId=82044ENG&_theme=1089

[89] ——, "Centraal Bureau Voor De Statistiek : Onderzoek Onderweg in Nederland - ODiN 2018," 2019, medium : application/pdf,.sav,.por,.dta,.ods type : dataset. [Online]. Available : https://easy.dans.knaw.nl/ui/datasets/id/easy-dataset:156525

[90] ——, "Centraal Bureau Voor De Statistiek : Mobility by person, transport mode and region." [Online]. Available : https://opendata.cbs.nl/statline/portal.html?_la=nl&_catalog=CBS&tableId=84709NED&_theme=400

[91] StatisticsFinland, "Average commuting distance by Year," library Catalog : pxnet2.stat.fi. [Online]. Available : https://pxnet2.stat.fi/PXWeb/pxweb/en/StatFin/StatFin__lii__tyoma/statfin_tyoma_pxt_001.px

[92] ——, "Number of vehicles registered by Year, Vehicle class and Information," library Catalog : pxnet2.stat.fi. [Online]. Available : https://pxnet2.stat.fi/PXWeb/pxweb/en/StatFin/StatFin__lii__mkan/statfin_mkan_pxt_11ib.px

[93] "Data Gov Hong-Kong, Statistics on Passenger, Visitor and Vehicular Traffic." [Online]. Available : https://data.gov.hk/en-data/dataset/hk-immd-set5-statistics-passenger-visitor-vehicular-traffic

[94] StatsGov, "China Statistical Yearbook 2018." [Online]. Available : http://www.stats.gov.cn/tjsj/ndsj/2018/indexeh.htm

[95] "Data Gov Hong-Kong, Details of HK Electric EV charging stations." [Online]. Available : https://data.gov.hk/en-data/dataset/hkelectric-tnd_cs_ci-hkelectric-ev-location

[96] "Data Gov Hong-Kong, Traffic Surveys in Kowloon East." [Online]. Available : https://data.gov.hk/en-data/dataset/hk-devb-traf-traf

[97] DadosGOV, "Charging Stations E-mobility in Lisbon," library Catalog : dados.gov.pt. [Online]. Available : https://dados.gov.pt/pt/datasets/postos-de-carregamento-mobi-e/

[98] PORDATA, "Traffic Satistics in Portugal." [Online]. Available : https://www.pordata.pt/Subtema/Portugal/Rodovi%c3%a1rio-405

[99] USDepartmentofEnergy, "Alternative Fuels Data Center." [Online]. Available : https://afdc.energy.gov/data_download

[100] L. Makram, "Electric Vehicle Charging Stations : Energy Consumption & Savings." [Online]. Available : https://open-data.bouldercolorado.gov/datasets/2d8bad4baf274407a674d2ed3c657951_0

[101] CityofPaloAlto, "Electric Vehicle Charging Station Usage (July 2011 - Dec 2017) · Open Data," library Catalog : data.cityofpaloalto.org. [Online]. Available : http://data.cityofpaloalto.org/dataviews/244892/electric-vehicle-charging-station-usage-july-2011-dec-2017/

[102] CityofEvanston, "City-owned Electric Vehicle Charging Station Usage January 2016 to August 2017 | Open Data," library Catalog : data.cityofevanston.org. [Online]. Available : https://data.cityofevanston.org/dataset/City-owned-Electric-Vehicle-Charging-Station-Usage/nx7w-jb8v/data

[103] "ACN-Data – A Public EV Charging Dataset." [Online]. Available : https://ev.caltech.edu/dataset

[104] CityofHouston, "Traffic Counts in the City of Houston," library Catalog : cohgis-mycity.opendata.arcgis.com. [Online]. Available : http://data.houstontx.gov/dataset/traffic-counts

[105] B. Skerpan, "Resident Mobility Survey." [Online]. Available : https://open-data.bouldercolorado.gov/datasets/ffeccf9676bd495e8c99450243edf0a9_0

[106] NewYorkState, "Electric Vehicles per County | Open Data NY," library Catalog : data.ny.gov. [Online]. Available : https://data.ny.gov/Transportation/Electric-Vehicles-per-County/uu25-czyc

[107] CityofEdmonton, "Public Charging Stations for Electric Vehicles." [Online]. Available : https://data.edmonton.ca/Environmental-Services/Public-Charging-Stations-for-Electric-Vehicles/xzhy-xe8z

[108] OpenGovOntario, "Electric vehicle home charging program applicant data," library Catalog : data.ontario.ca. [Online]. Available : https://data.ontario.ca/dataset/electric-vehicle-home-charging-program-applicant-data

[109] TrafficDataProgram, "Annual Traffic Volumes 2004-2010 in British Columbia," library Catalog : catalogue.data.gov.bc.ca. [Online]. Available : https://catalogue.data.gov.bc.ca/dataset/annual-traffic-volumes-2004-2010

[110] CanadaGov, "Canadian travel survey, travel in Canada, by travel duration," library Catalog : open.canada.ca. [Online]. Available : https://open.canada.ca/data/en/dataset/19f7dc90-a073-4000-8180-a4bc0aa290f8

[111] G. Milot, "Electric and Hybrid Vehicles - CKAN," library Catalog : donnees.ville.montreal.qc.ca. [Online]. Available : http://donnees.ville.montreal.qc.ca/dataset/vehicules-electriques-et-hybrides

[112] IRVE, "Charging point locations for Electric Vehicles," library Catalog : opendata.reseaux-energies.fr. [Online]. Available : https://opendata.reseaux-energies.fr/explore/dataset/bornes-irve/information/?disjunctive.region

[113] DataGouv, "Charging sessions Apr-May 2017 in Paris," library Catalog : www.data.gouv.fr. [Online]. Available : https://www.data.gouv.fr/fr/datasets/belib-reseau-parisien-de-bornes-de-recharges-accelerees-22-kw-ac-dc-pour-vehicules-electric

[114] SAPLabs, "Electric Vehicle charging transactions of SAP Labs France company fleet." [Online]. Available : https://opendata.reseaux-energies.fr/explore/dataset/conso-ve-sap/information/?disjunctive.stop_inactivitystatus

[115] ParisData, "Belib' availability in real-time," library Catalog : opendata.paris.fr. [Online]. Available : https://opendata.paris.fr/explore/dataset/belib-points-de-recharge-pour-vehicules-electriques-disponibilite-temps-reel/information/?disjunctive.statut_pdc&disjunctive.arrondissement

[116] ——, "Historical Traffic Counts in Paris," library Catalog : opendata.paris.fr. [Online]. Available : https://opendata.paris.fr/explore/dataset/comptages-routiers-permanents-historique/information/

[117] RennesMétropole, "Real-time traffic in Rennes," library Catalog : data.rennesmetropole.fr. [Online]. Available : https://data.rennesmetropole.fr/explore/dataset/etat-du-trafic-en-temps-reel/information/

[118] INSEE, "Travel Survey 2015 : home to work," library Catalog : data.grandparissud.fr. [Online]. Available : https://data.grandparissud.fr/explore/dataset/mobilites-professionnelles-en-2015-deplacements-domicile-lieu-de-travailo/information/

[119] ——, "Registered Vehicles in France from 1990 to 2018." [Online]. Available : https://www.insee.fr/fr/statistiques/2045167

[120] NewPlymouthDistricCouncil, "Traffic Counts in New Plymouth." [Online]. Available : https://www.newplymouthnz.com/Residents/Transportation/Maintenance/Traffic-Counts

[121] StatsNZ, "Motor Vehicles Registered." [Online]. Available : http://archive.stats.govt.nz/infoshare/default.aspx?RedirectReason=session_expired

[122] OfficeforLowEmissionVehicles, "National Charge Point Registry," Feb. 2012, library Catalog : data.gov.uk. [Online]. Available : https://data.gov.uk/dataset/1ce239a6-d720-4305-ab52-17793fedfac3/national-charge-point-registry

[123] SunderlandCityCouncil, "Charge your car." [Online]. Available : https://www.sunderland.gov.uk/chargeyourcar

[124] TransportTeam, "Electric Vehicle Charging Sessions Dundee," library Catalog : data.dundeecity.gov.uk. [Online]. Available : https://data.dundeecity.gov.uk/dataset/ev-charging-data

[125] OpenDataTeam, "Electric Vehicle Charging Station Usage in Perth and Kinross," library Catalog : data.pkc.gov.uk. [Online]. Available : https://data.pkc.gov.uk/dataset/ev-charging-data

[126] A. Radford, "Birmingham and West Midlands real-time traffic data," library Catalog : data.birmingham.gov.uk. [Online]. Available : https://data.birmingham.gov.uk/dataset/wm-utmc

[127] DepartmentforTransport, "Traffic Flows, Borough," library Catalog : data.london.gov.uk. [Online]. Available : https://data.london.gov.uk/dataset/traffic-flows-borough

[128] UKDataService, "National Travel Survey 2002-2017." [Online]. Available : https://beta.ukdataservice.ac.uk/datacatalogue/studies/study?id=5340

[129] O. Smith, "How Many Left ?" [Online]. Available : https://www.howmanyleft.co.uk/?utf8=%E2%9C%93&q=&commit=Search

[130] OffenesDatenPortal, "Charging point locations in Moers," library Catalog : www.offenesdatenportal.de. [Online]. Available : https://www.offenesdatenportal.de/dataset/autoladestationen-in-moers

[131] OpenDataBonn, "Charging point locations and usage in real-time in Bonn." [Online]. Available : https://opendata.bonn.de/dataset/standorte-elektrotankstellen-e-lades%C3%A4ulen-realtime-belegung

[132] ——, "Real-time traffic in Bonn." [Online]. Available : https://opendata.bonn.de/dataset/stra%C3%9Fenverkehrslage-realtime

[133] OpenDataBayern, "Traffic Volumes in Bayern." [Online]. Available : https://opendata.bayern.de/detailansicht/datensatz/baysis-verkehrsmengen-wfs?o

[134] "FDZ Ruhr am RWI, Research Data Center." [Online]. Available : https://www.rwi-essen.de/forschung-und-beratung/fdz-ruhr/datenzugang/

[135] L. Ecke, "German Mobility Panel - Startseite," Aug. 2020, archive Location : KIT Publisher : Lisa Ecke. [Online]. Available : https://mobilitaetspanel.ifv.kit.edu/english/index.php

[136] Bochum, "Registered Vehicles in Bochum," library Catalog : www.bochum.de. [Online]. Available : https://www.bochum.de//Open-Data/Datensaetze/Transport-und-Verkehr

[137] MLIT, "Travel Survey in Japan from April 2019." [Online]. Available : https://www.mlit.go.jp/k-toukei/jidousya.html

[138] e Stat, "Registered Vehicles in Japan," library Catalog : www.e-stat.go.jp. [Online]. Available : https://www.e-stat.go.jp/stat-search?page=1&toukei=00600700&bunya_l=10

[139] "Passenger car sales by country, around the world," library Catalog : www.theglobaleconomy.com. [Online]. Available : https://www.theglobaleconomy.com/rankings/passenger_cars_sales/

[140] NationaleDatabankWegverkeersgegevens, "Real-time and historical traffic in the Netherlands." [Online]. Available : http://opendata.ndw.nu/

[141] Z. J. Lee, T. Li, and S. H. Low, "ACN-Data : Analysis and Applications of an Open EV Charging Dataset," in *Proceedings of the Tenth ACM International Conference on Future Energy Systems*. Phoenix AZ USA : ACM, Jun. 2019, pp. 139–149. [Online]. Available : https://dl.acm.org/doi/10.1145/3307772.3328313

[142] Cerema, "Road traffic open data," May 2020, last Modified : 2020-05-27. [Online]. Available : http://trafic-routier.data.cerema.fr/acces-aux-donnees-de-trafic-routier-open-data-r25.html

[143] DepartmentforTransport, "Domestics Chargepoint Analysis 2017," Dec. 2018, library Catalog : data.gov.uk. [Online]. Available : https://data.gov.uk/dataset/5438d88d-695b-4381-a5f2-6ea03bf3dcf0/electric-chargepoint-analysis-2017-domestics

[144] ——, "Rapid Chargepoint Analysis 2017," 2018, library Catalog : www.gov.uk. [Online]. Available : https://www.gov.uk/government/statistics/electric-chargepoint-analysis-2017-local-authority-rapids

[145] ——, "Fast Chargepoint Analysis 2017," 2018, library Catalog : www.gov.uk. [Online]. Available : https://www.gov.uk/government/statistics/electric-chargepoint-analysis-2017-public-sector-fasts

[146] LondonDatastore, "Low Carbon London Electric Vehicle Load Profiles," library Catalog : data.london.gov.uk. [Online]. Available : https://data.london.gov.uk/dataset/low-carbon-london-electric-vehicle-load-profiles

[147] Elbilsstatistik, "Laddinfrastatistik." [Online]. Available : https://www.elbilsstatistik.se

[148] Tampere, "Tampere real-time traffic," library Catalog : data.tampere.fi. [Online]. Available : https://data.tampere.fi/data/fi/dataset/tampereen-kaupungin-liikennetiedoterajapinta

[149] P. Breidenbach and L. Eilers, "RWI-GEO-GRID : Socio-economic data on grid level," *Jahrbücher für Nationalökonomie und Statistik*, vol. 238, no. 6, pp. 609–616, Oct. 2018, publisher : De Gruyter Oldenbourg Section : Jahrbücher für Nationalökonomie und Statistik. [Online]. Available : http://www.degruyter.com/view/journals/jbnst/238/6/article-p609.xml

[150] L. Yuanyuan, "China installed more than 1000 EV charging stations per day in 2019," Jan. 2020, library Catalog : www.renewableenergyworld.com Section : Infrastructure. [Online]. Available : https://www.renewableenergyworld.com/2020/01/13/china-installed-more-than-1000-ev-charging-stations-per-day-in-2019/

[151] J. R. Helmus, M. H. Lees, and R. van den Hoed, "A data driven typology of electric vehicle user types and charging sessions," *Transportation Research Part C : Emerging Technologies*, vol. 115, p. 102637, Jun. 2020. [Online]. Available : https://linkinghub.elsevier.com/retrieve/pii/S0968090X19315414

[152] M. Lahariya, D. Benoit, and C. Develder, "Poster : Defining a synthetic data generator for realistic electric vehicle charging sessions," in *Proceedings of the Eleventh ACM International Conference on Future Energy Systems*, 2020, p. 2.

[153] M. Liang, W. Li, J. Yu, and L. Shi, "Kernel-based electric vehicle charging load modeling with improved latin hypercube sampling," in *2015 IEEE Power & Energy Society General Meeting*. Denver, CO, USA : IEEE, Jul. 2015, pp. 1–5. [Online]. Available : http://ieeexplore.ieee.org/document/7285758/

[154] L. Chen, X. Huang, and H. Zhang, "Modeling the Charging Behaviors for Electric Vehicles Based on Ternary Symmetric Kernel Density Estimation," *Energies*, vol. 13, no. 7, p. 1551, Mar. 2020. [Online]. Available : https://www.mdpi.com/1996-1073/13/7/1551

[155] B. Khaki, Y.-W. Chung, C. Chu, and R. Gadh, "Nonparametric User Behavior Prediction for Distributed EV Charging Scheduling," in *2018 IEEE Power & Energy Society General Meeting (PESGM)*. Portland, OR : IEEE, Aug. 2018, pp. 1–5. [Online]. Available : https://ieeexplore.ieee.org/document/8585744/

[156] ——, "Probabilistic Electric Vehicle Load Management in Distribution Grids," in *2019 IEEE Transportation Electrification Conference and Expo (ITEC)*. Detroit, MI, USA : IEEE, Jun. 2019, pp. 1–6. [Online]. Available : https://ieeexplore.ieee.org/document/8790535/

[157] Y.-W. Chung, B. Khaki, C. Chu, and R. Gadh, "Electric Vehicle User Behavior Prediction Using Hybrid Kernel Density Estimator," in *2018 IEEE International Conference on Probabilistic Methods Applied to Power Systems (PMAPS)*. Boise, ID : IEEE, Jun. 2018, pp. 1–6. [Online]. Available : https://ieeexplore.ieee.org/document/8440360/

[158] P. Sokorai, A. Fleischhacker, G. Lettner, and H. Auer, "Stochastic Modeling of the Charging Behavior of Electromobility," *World Electric Vehicle Journal*, vol. 9, no. 3, p. 44, Oct. 2018. [Online]. Available : http://www.mdpi.com/2032-6653/9/3/44

[159] M. H. Amini, A. Kargarian, and O. Karabasoglu, "ARIMA-based decoupled time series forecasting of electric vehicle charging demand for stochastic power system operation," *Electric Power Systems Research*, vol. 140, pp. 378–390, Nov. 2016. [Online]. Available : https://linkinghub.elsevier.com/retrieve/pii/S0378779616302115

[160] S. Bae and A. Kwasinski, "Spatial and Temporal Model of Electric Vehicle Charging Demand," *IEEE Transactions on Smart Grid*, vol. 3, no. 1, pp. 394–403, Mar. 2012. [Online]. Available : http://ieeexplore.ieee.org/document/5959242/

[161] O. Häggström and O. H. G. M, *Finite Markov Chains and Algorithmic Applications*. Cambridge University Press, May 2002, google-Books-ID : hpLxIJ9LwRgC.

[162] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time series analysis : forecasting and control*, 5th ed., ser. Wiley series in probability and statistics. Hoboken, New Jersey : John Wiley & Sons, Inc, 2016.

[163] F. Baskett, K. M. Chandy, R. R. Muntz, and F. G. Palacios, "Open, Closed, and Mixed Networks of Queues with Different Classes of Customers," *Journal of the ACM (JACM)*, vol. 22, no. 2, pp. 248–260, Apr. 1975. [Online]. Available : http://dl.acm.org/doi/10.1145/321879.321887

[164] "The EV Project - Overview," Nov. 2010. [Online]. Available : https://web.archive.org/web/20101128153912/http://theevproject.com/overview.php

[165] M. Majidpour, C. Qiu, P. Chu, R. Gadh, and H. R. Pota, "A novel forecasting algorithm for electric vehicle charging stations," in *2014 International Conference on Connected Vehicles and Expo (ICCVE)*. Vienna, Austria : IEEE, Nov. 2014, pp. 1035–1040. [Online]. Available : http://ieeexplore.ieee.org/document/7297504/

[166] Y. Xiong, C.-c. Chu, R. Gadh, and B. Wang, "Distributed optimal vehicle grid integration strategy with user behavior prediction," in *2017 IEEE Power & Energy Society General Meeting*. Chicago, IL : IEEE, Jul. 2017, pp. 1–5. [Online]. Available : http://ieeexplore.ieee.org/document/8274327/

[167] O. Frendo, N. Gaertner, and H. Stuckenschmidt, "Improving Smart Charging Prioritization by Predicting Electric Vehicle Departure Time," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–8, 2020. [Online]. Available : https://ieeexplore.ieee.org/document/9082829/

[168] J. Mies, J. Helmus, and R. van den Hoed, "Estimating the Charging Profile of Individual Charge Sessions of Electric Vehicles in The Netherlands," *World Electric Vehicle Journal*, vol. 9, no. 2, p. 17, Jun. 2018. [Online]. Available : http://www.mdpi.com/2032-6653/9/2/17

[169] Y. Lu, Y. Li, D. Xie, E. Wei, X. Bao, H. Chen, and X. Zhong, "The Application of Improved Random Forest Algorithm on the Prediction of Electric Vehicle Charging Load," *Energies*, vol. 11, no. 11, p. 3207, Nov. 2018. [Online]. Available : http://www.mdpi.com/1996-1073/11/11/3207

[170] J. Zhu, Z. Yang, M. Mourshed, Y. Guo, Y. Zhou, Y. Chang, Y. Wei, and S. Feng, "Electric Vehicle Charging Load Forecasting : A Comparative Study of Deep Learning Approaches," *Energies*, vol. 12, no. 14, p. 2692, Jul. 2019. [Online]. Available : https://www.mdpi.com/1996-1073/12/14/2692

[171] V. N. Vapnik, *The Nature of Statistical Learning Theory*. New York, NY : Springer New York, 1995. [Online]. Available : http://link.springer.com/10.1007/978-1-4757-2440-0

[172] L. Breiman, "Random Forest," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. [Online]. Available : http://link.springer.com/10.1023/A:1010933404324

[173] L. Buzna, P. D. Falco, S. Khormali, D. Proto, and M. Straka, "Electric vehicle load forecasting : a comparison between time series and machine learning approaches," in *Proceeding of the International Conference on Energy Transition in the Mediterranean Area*, 2019, p. 5.

[174] F. Rosenblatt, "The perceptron : A probabilistic model for information storage and organization in the brain." *Psychological Review*, vol. 65, no. 6, pp. 386–408, 1958. [Online]. Available : http://doi.apa.org/getdoi.cfm?doi=10.1037/h0042519

[175] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning Internal Representations by Error Propagation," Institute for Cognitive Science, Technical Report ICS-8506, 1985.

[176] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation Applied to Handwritten Zip Code Recognition," *Neural Computation*, vol. 1, no. 4, pp. 541–551, Dec. 1989, publisher : MIT Press. [Online]. Available : https://doi.org/10.1162/neco.1989.1.4.541

[177] S. Chakraborty, R. Tomsett, R. Raghavendra, D. Harborne, M. Alzantot, F. Cerutti, M. Srivastava, A. Preece, S. Julier, R. M. Rao, T. D. Kelley, D. Braines, M. Sensoy, C. J. Willis, and P. Gurram, "Interpretability of deep learning models : A survey of results," in *2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*. San Francisco, CA : IEEE, Aug. 2017, pp. 1–6. [Online]. Available : https://ieeexplore.ieee.org/document/8397411/

[178] Q. Zhang and S.-C. Zhu, "Visual Interpretability for Deep Learning : a Survey," *arXiv :1802.00614 [cs]*, Feb. 2018, arXiv : 1802.00614. [Online]. Available : http://arxiv.org/abs/1802.00614

[179] O. F. Vynakov, E. V. Savolova, and A. I. Skrynnyk, "Modern Electric Cars of Tesla Motors Company," *Automation of technological and business processes*, vol. 8, no. 2, Aug. 2016. [Online]. Available : https://journals.onaft.edu.ua/index.php/atbp/article/view/162

[180] S. Wood, "mgcv : Mixed GAM Computation Vehicle with Automatic Smoothness Estimation," Nov. 2019. [Online]. Available : https://CRAN.R-project.org/package=mgcv

[181] D. Burchart-Korol, S. Jursova, P. Folęga, J. Korol, P. Pustejovska, and A. Blaut, "Environmental life cycle assessment of electric vehicles in Poland and the Czech Republic," *Journal of Cleaner Production*, vol. 202, pp. 476–487, Nov. 2018. [Online]. Available : https://linkinghub.elsevier.com/retrieve/pii/S0959652618325009

[182] "General Data Protection Regulation (GDPR) – Official Legal Text," library Catalog : gdpr-info.eu. [Online]. Available : https://gdpr-info.eu/

[183] eCO2mix, "Electrical consumption in cities open data." [Online]. Available : https://www.data.gouv.fr/fr/datasets/donnees-eco2mix-consommation-des-metropoles-temps-reel/

[184] P. Zhang, C. Zhou, B. G. Stewart, D. M. Hepburn, W. Zhou, and J. Yu, "An Improved Non-Intrusive Load Monitoring Method for Recognition of Electric Vehicle Battery Charging Load," *Energy Procedia*, vol. 12, pp. 104–112, 2011. [Online]. Available : https://linkinghub.elsevier.com/retrieve/pii/S1876610211018418

[185] S. Wang, L. Du, J. Ye, and D. Zhao, "A Deep Generative Model for Non-intrusive Identification of EV Charging Profiles," *IEEE Transactions on Smart Grid*, pp. 1–1, 2020. [Online]. Available : https://ieeexplore.ieee.org/document/9102286/

[186] N. Du, H. Dai, R. Trivedi, U. Upadhyay, M. Gomez-Rodriguez, and L. Song, "Recurrent Marked Temporal Point Processes : Embedding Event History to Vector," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*. San Francisco, California, USA : ACM Press, 2016, pp. 1555–1564. [Online]. Available : http://dl.acm.org/citation.cfm?doid=2939672.2939875

[187] S. Xiao, J. Yan, X. Yang, H. Zha, and S. M. Chu, "Modeling the Intensity Function of Point Process via Recurrent Neural Networks," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, 2017, p. 7.

[188] N. Fallah, H. Gu, K. Mohammad, S. A. Seyyedsalehi, K. Nourijelyani, and M. R. Eshraghian, "Nonlinear Poisson regression using neural networks : a simulation study," *Neural Computing and Applications*, vol. 18, no. 8, pp. 939–943, Nov. 2009. [Online]. Available : http://link.springer.com/10.1007/s00521-009-0277-8

[189] C. Capezza, B. Palumbo, Y. Goude, S. N. Wood, and M. Fasiolo, "Additive stacking for disaggregate electricity demand forecasting," *arXiv :2005.10092 [stat]*, May 2020, arXiv : 2005.10092. [Online]. Available : http://arxiv.org/abs/2005.10092

[190] N. Sadeghianpourhamami, J. Deleu, and C. Develder, "Achieving Scalable Model-Free Demand Response in Charging an Electric Vehicle Fleet with Reinforcement Learning," in *Proceedings of the Ninth International Conference on Future Energy Systems - e-Energy '18*. Karlsruhe, Germany : ACM Press, 2018, pp. 411–413. [Online]. Available : http://dl.acm.org/citation.cfm?doid=3208903.3212042

[191] Y. Amara-Ouali, "EV load Open Data," library Catalog : github.com. [Online]. Available : https://github.com/yvenn-amara/EV-Load-Open-Data

[192] Enedis, "Grid map open data." [Online]. Available : https://data.enedis.fr/pages/accueil/?id=dataviz-cartographie-des-reseaux

[193] S. Makonin and F. Popowich, "Nonintrusive load monitoring (NILM) performance evaluation," *Energy Efficiency*, vol. 8, no. 4, pp. 809–814, Jul. 2015. [Online]. Available : https://doi.org/10.1007/s12053-014-9306-2

[194] H. Salem and M. Sayed-Mouchaweh, "A Semi-supervised and Online Learning Approach for Non-Intrusive Load Monitoring," in *Machine Learning and Knowledge Discovery in Databases*, ser. Lecture Notes in Computer Science, U. Brefeld, E. Fromont, A. Hotho, A. Knobbe, M. Maathuis, and C. Robardet, Eds. Cham : Springer International Publishing, 2020, pp. 585–601.

[195] "Pecan Street Data," library Catalog : www.pecanstreet.org. [Online]. Available : https://www.pecanstreet.org/dataport/about/

[196] P. J. Laub, T. Taimre, and P. K. Pollett, "Hawkes Processes," *arXiv :1507.02822 [math, q-fin, stat]*, Jul. 2015, arXiv : 1507.02822. [Online]. Available : http://arxiv.org/abs/1507.02822

[197] O. Frendo, J. Graf, N. Gaertner, and H. Stuckenschmidt, "Data-driven smart charging for heterogeneous electric vehicle fleets," *Energy and AI*, vol. 1, p. 100007, Aug. 2020. [Online]. Available : https://linkinghub.elsevier.com/retrieve/pii/S2666546820300070

[198] Y. Amara-Ouali, Y. Goude, B. Hamrouche, and M. Bishara, "A benchmark of electric vehicle load and occupancy models for day-ahead forecasting on open charging session data," *e-Energy '22*, 2022.

[199] S. Shahriar, A. R. Al-Ali, A. H. Osman, S. Dhou, and M. Nijim, "Machine Learning Approaches for EV Charging Behavior : A Review," *IEEE Access*, vol. 8, pp. 168 980–168 993, 2020. [Online]. Available : https://ieeexplore.ieee.org/document/9194702/

[200] Y. Amara-Ouali, P. Massart, J.-M. Poggi, Y. Goude, and H. Yan, "A review of electric vehicle charging session open data : Poster," in *Proceedings of the Twelfth ACM International Conference on Future Energy Systems*, ser. e-Energy '21. New York, NY, USA : Association for Computing Machinery, Jun. 2021, pp. 278–279. [Online]. Available : https://doi.org/10.1145/3447555.3466568

[201] F. B. Hüttel, I. Peled, F. Rodrigues, and F. C. Pereira, "Deep Spatio-Temporal Forecasting of Electrical Vehicle Charging Demand," *arXiv :2106.10940 [cs]*, Jun. 2021, arXiv : 2106.10940. [Online]. Available : http://arxiv.org/abs/2106.10940

[202] T.-Y. Ma and S. Faye, "Multistep electric vehicle charging station occupancy prediction using hybrid LSTM neural networks," *Energy*, vol. 244, p. 123217, Apr. 2022. [Online]. Available : https://linkinghub.elsevier.com/retrieve/pii/S0360544222001207

[203] M. Straka, R. Carvalho, G. V. D. Poel, and L. Buzna, "Analysis of Energy Consumption at Slow Charging Infrastructure for Electric Vehicles," *IEEE Access*, vol. 9, pp. 53 885–53 901, 2021, conference Name : IEEE Access.

[204] D. Obst, J. de Vilmarest, and Y. Goude, "Adaptive methods for short-term electricity load forecasting during covid-19 lockdown in france," *IEEE Transactions on Power Systems*, vol. 36, no. 5, pp. 4754–4763, 2021.

[205] B. Goehry, Y. Goude, P. Massart, and J.-M. Poggi, "Aggregation of multi-scale experts for bottom-up load forecasting," *IEEE Transactions on Smart Grid*, vol. 11, no. 3, pp. 1895–1904, 2020.

[206] B. D. Youngman and T. Economou, "Generalised additive point process models for natural hazard occurrence : Generalised additive point process models," *Environmetrics*, vol. 28, no. 4, p. e2444, Jun. 2017. [Online]. Available : https://onlinelibrary.wiley.com/doi/10.1002/env.2444

[207] T. Hastie and R. Tibshirani, "Generalized Additive Models," *Statistical Science*, vol. 1, no. 3, Aug. 1986. [Online]. Available : https://projecteuclid.org/journals/statistical-science/volume-1/issue-3/Generalized-Additive-Models/10.1214/ss/1177013604.full

[208] S. Wood, *Generalized Additive Models : An Introduction with R*, 2nd ed. Chapman and Hall/CRC, May 2017. [Online]. Available : https://www.taylorfrancis.com/books/9781498728348

[209] R. Nedellec, J. Cugliari, and Y. Goude, "GEFCom2012 : Electric load forecasting and backcasting with semi-parametric models," *International Journal of Forecasting*, vol. 30, no. 2, pp. 375–381, Apr. 2014. [Online]. Available : https://linkinghub.elsevier.com/retrieve/pii/S0169207013000800

[210] L. Breiman, J. Friedman, C. Stone, and R. Olshen, *Classification and Regression Trees*, ser. The Wadsworth and Brooks-Cole statistics-probability series. Taylor & Francis, 1984.

[211] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[212] P. A. W. Lewis and G. S. Shedler, "Simulation of nonhomogeneous poisson processes by thinning," *Naval Research Logistics Quarterly*, vol. 26, no. 3, pp. 403–413, Sep. 1979. [Online]. Available : https://onlinelibrary.wiley.com/doi/10.1002/nav.3800260304

[213] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," in *NIPS 2014 Workshop on Deep Learning, December 2014*, 2014.

[214] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time series analysis : forecasting and control*. John Wiley & Sons, 1976.

[215] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, publisher : MIT Press. [Online]. Available : https://www.mitpressjournals.org/doi/10.1162/neco.1997.9.8.1735

[216] X. Zhang and S. Grijalva, "An advanced data driven model for residential electric vehicle charging demand," in *2015 IEEE Power & Energy Society General Meeting*. Denver, CO, USA : IEEE, Jul. 2015, pp. 1–5. [Online]. Available : http://ieeexplore.ieee.org/document/7286396/

[217] D. J. Daley and D. Vere-Jones, *An Introduction to the Theory of Point Processes*, ser. Probability and Its Applications, J. Gani, C. C. Heyde, P. Jagers, and T. G. Kurtz, Eds. New York, NY : Springer New York, 2008. [Online]. Available : http://link.springer.com/10.1007/978-0-387-49835-5

[218] F. Stulp and O. Sigaud, "Many regression algorithms, one unified model : A review," *Neural Networks*, vol. 69, pp. 60–79, 2015. [Online]. Available : https://www.sciencedirect.com/science/article/pii/S0893608015001185

[219] Z. Ghahramani and M. Jordan, "Supervised learning from incomplete data via an em approach," *Advances in neural information processing systems*, vol. 6, 1993.

[220] H. W. Reeve and G. Brown, "Diversity and degrees of freedom in regression ensembles," *Neurocomputing*, vol. 298, pp. 55–68, 2018. [Online]. Available : https://www.sciencedirect.com/science/article/pii/S0925231218302133

[221] N. Cesa-Bianchi and G. Lugosi, *Prediction, Learning, and Games*. New York, NY, USA : Cambridge University Press, 2006.

[222] N. Cesa-Bianchi and F. Orabona, "Online learning algorithms," *Annual Review of Statistics and Its Application*, vol. 8, pp. 165–190, 2021.

[223] Y. Kalnishkan, "Prediction with expert advice for a finite number of experts : A practical introduction," *Pattern Recognition*, p. 108557, 2022. [Online]. Available : https://www.sciencedirect.com/science/article/pii/S0031320322000383

[224] P. Gaillard, G. Stoltz, and T. Van Erven, "A second-order bound with excess losses," in *Conference on Learning Theory*, 2014, pp. 176–196.

[225] P. Gaillard and Y. Goude, "opera : Online prediction by expert aggregation," *URL : https ://CRAN. R-project. org/package= opera. r package version*, vol. 1, 2016.

[226] G. Schwarz, "Estimating the Dimension of a Model," *The Annals of Statistics*, vol. 6, no. 2, Mar. 1978. [Online]. Available : https://projecteuclid.org/journals/annals-of-statistics/volume-6/issue-2/Estimating-the-Dimension-of-a-Model/10.1214/aos/1176344136.full

[227] R. J. Hyndman and Y. Khandakar, "Automatic Time Series Forecasting : The **forecast** Package for *R*," *Journal of Statistical Software*, vol. 27, no. 3, 2008. [Online]. Available : http://www.jstatsoft.org/v27/i03/

[228] G. M. Ljung and G. E. P. Box, "On a measure of lack of fit in time series models," *Biometrika*, vol. 65, no. 2, pp. 297–303, Aug. 1978. [Online]. Available : https://academic.oup.com/biomet/article-lookup/doi/10.1093/biomet/65.2.297

[229] S. Kim and H. Kim, "A new metric of absolute percentage error for intermittent demand forecasts," *International Journal of Forecasting*, vol. 32, no. 3, pp. 669–679, Jul. 2016. [Online]. Available : https://linkinghub.elsevier.com/retrieve/pii/S0169207016000121

[230] M. Zamo, O. Mestre, P. Arbogast, and O. Pannekoucke, "A benchmark of statistical regression methods for short-term forecasting of photovoltaic electricity production, part I : Deterministic forecast of hourly production," *Solar Energy*, vol. 105, pp. 792–803, Jul. 2014. [Online]. Available : https://linkinghub.elsevier.com/retrieve/pii/S0038092X13005239

[231] J. W. Messner, P. Pinson, J. Browell, M. B. Bjerregård, and I. Schicker, "Evaluation of wind power forecasts—an up-to-date view," *Wind Energy*, vol. 23, no. 6, pp. 1461–1481, 2020. [Online]. Available : https://onlinelibrary.wiley.com/doi/abs/10.1002/we.2497

[232] J. A. Achcar and R. P. de Oliveira, "Climate Change : Use of Non-Homogeneous Poisson Processes for Climate Data in Presence of a Change-Point," *Environmental Modeling & Assessment*, vol. 27, no. 2, pp. 385–398, Apr. 2022. [Online]. Available : https://link.springer.com/10.1007/s10666-021-09797-z

[233] S. Aktas, H. Konsuk, and A. Yîğîter, "Estimation of change point and compound Poisson process parameters for the earthquake data in Turkey," *Environmetrics*, vol. 20, no. 4, pp. 416–427, Jun. 2009. [Online]. Available : https://onlinelibrary.wiley.com/doi/10.1002/env.937

[234] R. L. Streit, *Poisson Point Processes*. Boston, MA : Springer US, 2010. [Online]. Available : http://link.springer.com/10.1007/978-1-4419-6923-1

[235] T. Ngailo, N. Shaban, J. Reuder, E. Rutalebwa, and I. Mugume, "Non Homogeneous Poisson Process Modelling of Seasonal Extreme Rainfall Events in Tanzania," *International Journal of Science and Research*, vol. 5, no. 10, p. 12, 2016.

[236] N. R. Hansen, P. Reynaud-Bouret, and V. Rivoirard, "Lasso and probabilistic inequalities for multivariate point processes," *Bernoulli*, vol. 21, no. 1, Feb. 2015. [Online]. Available : https://projecteuclid.org/journals/bernoulli/volume-21/issue-1/Lasso-and-probabilistic-inequalities-for-multivariate-point-processes/10.3150/13-BEJ562.full

[237] D. J. Daley and D. Vere-Jones, *An Introduction to the Theory of Point Processes*, ser. Probability and its Applications. New York : Springer-Verlag, 2003. [Online]. Available : http://link.springer.com/10.1007/b97277

[238] D. Brillinger, "Some wavelet analyses of point process data," in *Conference Record of the Thirty-First Asilomar Conference on Signals, Systems and Computers (Cat. No.97CB36136)*, vol. 2. Pacific Grove, CA, USA : IEEE Comput. Soc, 1997, pp. 1087–1091. [Online]. Available : http://ieeexplore.ieee.org/document/679073/

[239] J. C. S. de Miranda, "Probability Density Functions of the Empirical Wavelet Coefficients of Multidimensional Poisson Intensities," *Brazilian Journal of Probability and Statistics*, vol. 22, no. 2, pp. 157–164, Dec. 2008.

[240] J. C. S. de Miranda and P. A. Morettin, "Estimation of the intensity of non-homogeneous point processes via wavelets," *Annals of the Institute of Statistical Mathematics*, vol. 63, no. 6, pp. 1221–1246, Dec. 2011. [Online]. Available : http://link.springer.com/10.1007/s10463-010-0283-8

[241] J. Bigot, S. Gadat, T. Klein, and C. Marteau, "Intensity estimation of non-homogeneous Poisson processes from shifted trajectories," *Electronic Journal of Statistics*, vol. 7, no. none, Jan. 2013. [Online]. Available : https://projecteuclid.org/journals/electronic-journal-of-statistics/volume-7/issue-none/Intensity-estimation-of-non-homogeneous-Poisson-processes-from-shifted-trajectories/10.1214/13-EJS794.full

[242] P. Reynaud-Bouret and V. Rivoirard, "Near optimal thresholding estimation of a Poisson intensity on the real line," *Electronic Journal of Statistics*, vol. 4, no. none, Jan. 2010. [Online]. Available : https://projecteuclid.org/journals/electronic-journal-of-statistics/volume-4/issue-none/Near-optimal-thresholding-estimation-of-a-Poisson-intensity-on-the/10.1214/08-EJS319.full

[243] D. L. Donoho and I. M. Johnstone, "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, vol. 81, no. 3, pp. 425–455, Sep. 1994. [Online]. Available : https://academic.oup.com/biomet/article/81/3/425/256924

[244] S. Luo, F. Zhou, L. Azizi, and M. Sugiyama, "Additive Poisson Process : Learning Intensity of Higher-Order Interaction in Stochastic Processes," *arXiv :2006.08982 [cs, stat]*, Jun. 2020, arXiv : 2006.08982. [Online]. Available : http://arxiv.org/abs/2006.08982

[245] J. H. Friedman and W. Stuetzle, "Projection Pursuit Regression," *Journal of the American Statistical Association*, vol. 76, no. 376, pp. 817–823, Dec. 1981. [Online]. Available : https://www.jstor.org/stable/2287576

[246] A. Buja, T. Hastie, and R. Tibshirani, "Linear Smoothers and Additive Models," *The Annals of Statistics*, vol. 17, no. 2, Jun. 1989. [Online]. Available : https://projecteuclid.org/journals/annals-of-statistics/volume-17/issue-2/Linear-Smoothers-and-Additive-Models/10.1214/aos/1176347115.full

[247] W. Hardle, H. Liang, and J. Gao, *Partially Linear Models*. Springer, Sep. 2000. [Online]. Available : https://mpra.ub.uni-muenchen.de/39562/

[248] U. Amato and A. Antoniadis, "Estimation and group variable selection for additive partial linear models with wavelets and splines," *South African Statistical Journal*, vol. 51, no. 2, p. 37, Dec. 2017.

[249] P. Craven and G. Wahba, "Smoothing noisy data with spline functions," *Numerische Mathematik*, vol. 31, pp. 377–403, 1979.

[250] C. Gu and G. Wahba, "Minimizing GCV/GML Scores with Multiple Smoothing Parameters via the Newton Method," *SIAM Journal on Scientific and Statistical Computing*, vol. 12, no. 2, pp. 383–398, Mar. 1991. [Online]. Available : http://epubs.siam.org/doi/10.1137/0912021

[251] S. N. Wood, "Modelling and smoothing parameter estimation with multiple quadratic penalties," *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, vol. 62, no. 2, pp. 413–428, May 2000. [Online]. Available : https://onlinelibrary.wiley.com/doi/10.1111/1467-9868.00240

[252] M. Wand and J. Ormerod, "Penalized wavelets : Embedding wavelets into semiparametric regression," *Electronic Journal of Statistics*, vol. 5, no. none, Jan. 2011. [Online]. Available : https://projecteuclid.org/journals/electronic-journal-of-statistics/volume-5/issue-none/Penalized-wavelets-Embedding-wavelets-into-semiparametric-regression/10.1214/11-EJS652.full

[253] A. Antoniadis and J. Fan, "Regularization of Wavelet Approximations," *Journal of the American Statistical Association*, vol. 96, no. 455, pp. 939–967,

Sep. 2001. [Online]. Available : http://www.tandfonline.com/doi/abs/10.1198/016214501753208942

[254] F. Meyer, "Wavelet-based estimation of a semiparametric generalized linear model of fMRI time-series," *IEEE Transactions on Medical Imaging*, vol. 22, no. 3, pp. 315–322, Mar. 2003. [Online]. Available : http://ieeexplore.ieee.org/document/1199633/

[255] X.-W. Chang and L. Qu, "Wavelet estimation of partially linear models," *Computational Statistics & Data Analysis*, vol. 47, no. 1, pp. 31–48, Aug. 2004. [Online]. Available : https://linkinghub.elsevier.com/retrieve/pii/S0167947303002445

[256] J. Fadili and E. Bullmore, "Penalized partially linear models using sparse representations with an application to fMRI time series," *IEEE Transactions on Signal Processing*, vol. 53, no. 9, pp. 3436–3448, Sep. 2005. [Online]. Available : http://ieeexplore.ieee.org/document/1495881/

[257] F. Leblanc and F. Letue, "Maximum Likelihood Estimation in Poisson Regression via Wavelet Model Selection Running title : Poisson regression via model selection," p. 32, 2006. [Online]. Available : https://sites.uclouvain.be/IAP-Stat-Phase-V-VI/PhaseV/publications_2006/TR/TR0607.pdf

[258] I. Gannaz, "Wavelet penalized likelihood estimation in generalized functional models," *TEST*, vol. 22, no. 1, pp. 122–158, Mar. 2013. [Online]. Available : http://link.springer.com/10.1007/s11749-012-0310-6

[259] T. Hastie, J. Qian, and K. Tay, "An Introduction to glmnet," p. 38, Apr. 2022.

[260] A. C. Cebrián, J. Abaurrea, and J. Asín, "**NHPoisson** : An *R* Package for Fitting and Validating Nonhomogeneous Poisson Processes," *Journal of Statistical Software*, vol. 64, no. 6, 2015. [Online]. Available : http://www.jstatsoft.org/v64/i06/

[261] R. Liu, W. K. Härdle, and G. Zhang, "Statistical inference for generalized additive partially linear models," *Journal of Multivariate Analysis*, vol. 162, pp. 1–15, Nov. 2017. [Online]. Available : https://linkinghub.elsevier.com/retrieve/pii/S0047259X17304517

[262] U. Amato, A. Antoniadis, I. De Feis, Y. Goude, and A. Lagache, "Forecasting high resolution electricity demand data with additive models including smooth and jagged components," *International Journal of Forecasting*, vol. 37, no. 1, pp. 171–185, Jan. 2021. [Online]. Available : https://www.sciencedirect.com/science/article/pii/S0169207020300583

[263] C. F. Ansley and R. Kohn, "Convergence of the backfitting algorithm for additive models," *Journal of the Australian Mathematical Society. Series A. Pure Mathematics and Statistics*, vol. 57, no. 3, p. 316–329, 1994.

[264] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *Journal of statistical software*, vol. 33, no. 1, p. 1, 2010.

[265] CityofPaloAlto, "Electric Vehicle Charging Station Usage (July 2011 - Dec 2020) · Open Data," 2021. [Online]. Available : https://bit.ly/3aMpZW7

[266] Y. Amara-Ouali, M. Fasiolo, Y. Goude, and H. Yan, "Daily peak electrical load forecasting with a multi-resolution approach," *arXiv preprint arXiv :2112.04492*, 2021.

[267] M. Uddin, M. F. Romlie, M. F. Abdullah, S. Abd Halim, A. H. Abu Bakar, and T. Chia Kwang, "A review on peak load shaving strategies," *Renewable and Sustainable Energy Reviews*, vol. 82, pp. 3323–3332, Feb. 2018. [Online]. Available : https://linkinghub.elsevier.com/retrieve/pii/S1364032117314272

[268] NationalGrid, "ESO Data Portal : Home | National Grid Electricity System Operator," 2021. [Online]. Available : https://data.nationalgrideso.com/

[269] M. Jacob, C. Neves, and D. Vukadinović Greetham, *Forecasting and Assessing Risk of Individual Electricity Peaks*. Cham : Springer International Publishing : Imprint : Springer, 2020, oCLC : 1141446837. [Online]. Available : https://link.springer.com/book/10.1007/978-3-030-28669-9

[270] J. Boano-Danquah, C. Sigauke, and K. A. Kyei, "Analysis of Extreme Peak Loads Using Point Processes : An Application Using South African Data," *IEEE Access*, vol. 8, pp. 146 105–146 115, 2020. [Online]. Available : https://ieeexplore.ieee.org/document/9163096/

[271] P. McSharry, S. Bouwman, and G. Bloemhof, "Probabilistic Forecasts of the Magnitude and Timing of Peak Electricity Demand," *IEEE Transactions on Power Systems*, vol. 20, no. 2, pp. 1166–1172, May 2005. [Online]. Available : http://ieeexplore.ieee.org/document/1425617/

[272] R. J. Hyndman and S. Fan, "Density Forecasting for Long-Term Peak Electricity Demand," *IEEE Transactions on Power Systems*, vol. 25, no. 2, pp. 1142–1153, May 2010. [Online]. Available : http://ieeexplore.ieee.org/document/5345698/

[273] N. Elamin and M. Fukushige, "Quantile Regression Model for Peak Load Demand Forecasting with Approximation by Triangular Distribution to Avoid Blackouts," *International Journal of Energy Economics and Policy*, vol. 8, no. 5, pp. 119–124, 2018.

[274] C. Gibbons and A. Faruqui, "Quantile Regression for Peak Demand Forecasting," *SSRN Electronic Journal*, 2014. [Online]. Available : http://www.ssrn.com/abstract=2485657

[275] C. Sigauke and D. Chikobvu, "Daily peak electricity load forecasting in South Africa using a multivariate non-parametric regression approach," *ORiON*, vol. 26, no. 2, Dec. 2010. [Online]. Available : http://orion.journals.ac.za/pub/article/view/89

[276] ——, "Prediction of daily peak electricity demand in South Africa using volatility forecasting models," *Energy Economics*, vol. 33, no. 5, pp. 882–888, Sep. 2011. [Online]. Available : https://linkinghub.elsevier.com/retrieve/pii/S0140988311000491

[277] H. Saxena, O. Aponte, and K. T. McConky, "A hybrid machine learning model for forecasting a billing period's peak electric load days," *International Journal of Forecasting*, vol. 35, no. 4, pp. 1288–1303, Oct. 2019. [Online]. Available : https://linkinghub.elsevier.com/retrieve/pii/S016920701930144X

[278] P. Dash, A. Liew, and S. Rahman, "Peak load forecasting using a fuzzy neural network," *Electric Power Systems Research*, vol. 32, no. 1, pp. 19–23, Jan. 1995. [Online]. Available : https://linkinghub.elsevier.com/retrieve/pii/037877969400889C

[279] L. Saini and M. Soni, "Artificial neural network based peak load forecasting using Levenberg–Marquardt and quasi-Newton methods," *IEE Proceedings - Generation, Transmission and Distribution*, vol. 149, no. 5, p. 578, 2002. [Online]. Available : https://digital-library.theiet.org/content/journals/10.1049/ip-gtd_20020462

[280] L. M. Saini, "Peak load forecasting using Bayesian regularization, Resilient and adaptive backpropagation learning based artificial neural networks," *Electric Power Systems Research*, vol. 78, no. 7, pp. 1302–1310, Jul. 2008. [Online]. Available : https://linkinghub.elsevier.com/retrieve/pii/S0378779607002258

[281] M. Amin-Naseri and A. Soroush, "Combined use of unsupervised and supervised learning for daily peak load forecasting," *Energy Conversion and Management*, vol. 49, no. 6, pp. 1302–1308, Jun. 2008. [Online]. Available : https://linkinghub.elsevier.com/retrieve/pii/S0196890408000174

[282] R. Abdel-Aal, "Modeling and forecasting electric daily peak loads using abductive networks," *International Journal of Electrical Power & Energy Systems*, vol. 28, no. 2, pp. 133–141, Feb. 2006. [Online]. Available : https://linkinghub.elsevier.com/retrieve/pii/S0142061505001390

[283] Z. Yu, Z. Niu, W. Tang, and Q. Wu, "Deep Learning for Daily Peak Load Forecasting–A Novel Gated Recurrent Neural Network Combining Dynamic Time Warping," *IEEE Access*, vol. 7, pp. 17 184–17 194, 2019. [Online]. Available : https://ieeexplore.ieee.org/document/8629072/

[284] I. A. Ibrahim and M. J. Hossain, "LSTM Neural Network Model for Ultra-short-term Distribution Zone Substation Peak Demand Prediction," in *2020 IEEE Power & Energy Society General Meeting (PESGM)*. Montreal, QC, Canada : IEEE, Aug. 2020, pp. 1–5. [Online]. Available : https://ieeexplore.ieee.org/document/9281973/

[285] Z. Guo, K. Zhou, X. Zhang, and S. Yang, "A deep learning model for short-term power load and probability density forecasting," *Energy*, vol. 160, pp. 1186–1200, Oct. 2018. [Online]. Available : https://linkinghub.elsevier.com/retrieve/pii/S0360544218313872

[286] Y. Yang, W. Hong, and S. Li, "Deep ensemble learning based probabilistic load forecasting in smart grids," *Energy*, vol. 189, p. 116324, Dec. 2019. [Online]. Available : https://linkinghub.elsevier.com/retrieve/pii/S0360544219320195

[287] Y. Yang, W. Li, T. A. Gulliver, and S. Li, "Bayesian Deep Learning-Based Probabilistic Load Forecasting in Smart Grids," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 7, pp. 4703–4713, Jul. 2020. [Online]. Available : https://ieeexplore.ieee.org/document/8844831/

[288] E. E. El-Attar, J. Y. Goulermas, and Q. H. Wu, "Forecasting electric daily peak load based on local prediction," in *2009 IEEE Power & Energy Society General Meeting*. Calgary, Canada : IEEE, Jul. 2009, pp. 1–6. [Online]. Available : http://ieeexplore.ieee.org/document/5275587/

[289] D.-H. Kim, E.-K. Lee, and N. B. S. Qureshi, "Peak-Load Forecasting for Small Industries : A Machine Learning Approach," *Sustainability*, vol. 12, no. 16, p. 6539, Aug. 2020. [Online]. Available : https://www.mdpi.com/2071-1050/12/16/6539

[290] J. Li, Q. Zhao, H. Wang, W. Wang, Y. Yang, and C. Yan, "Analysis of Deep Learning Control Strategy about Peak Load Regulation and Frequency Regulation with Distribution Thermal Storage Electric Boiler," in *2018 5th IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS)*. Nanjing, China : IEEE, Nov. 2018, pp. 461–464. [Online]. Available : https://ieeexplore.ieee.org/document/8691145/

[291] A. Antoniadis, E. Paparoditis, and T. Sapatinas, "A functional wavelet-kernel approach for time series prediction," *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, vol. 68, no. 5, pp. 837–857, Nov. 2006. [Online]. Available : http://doi.wiley.com/10.1111/j.1467-9868.2006.00569.x

[292] H. Cho, Y. Goude, X. Brossat, and Q. Yao, "Modeling and Forecasting Daily Electricity Load Curves : A Hybrid Approach," *Journal of the American Statistical Association*, vol. 108, no. 501, pp. 7–21, Mar. 2013. [Online]. Available : http://www.tandfonline.com/doi/abs/10.1080/01621459.2012.722900

[293] H. U. Amin, A. S. Malik, R. F. Ahmad, N. Badruddin, N. Kamel, M. Hussain, and W.-T. Chooi, "Feature extraction and classification for EEG signals using wavelet transform and machine learning techniques," *Australasian Physical & Engineering Sciences in Medicine*, vol. 38, no. 1, pp. 139–149, Mar. 2015. [Online]. Available : http://link.springer.com/10.1007/s13246-015-0333-x

[294] A. Soman, A. Trivedi, D. Irwin, B. Kosanovic, B. McDaniel, and P. Shenoy, "Peak Forecasting for Battery-based Energy Optimizations in Campus Microgrids," in *Proceedings of the Eleventh ACM International Conference on Future Energy Systems*. Virtual Event Australia : ACM, Jun. 2020, pp. 237–241. [Online]. Available : https://dl.acm.org/doi/10.1145/3396851.3397751

[295] J. Gao, Y. Guo, and Z. Wang, "Matrix Neural Networks," in *Advances in Neural Networks - ISNN 2017*, F. Cong, A. Leung, and Q. Wei, Eds. Cham : Springer International Publishing, 2017, pp. 313–320.

[296] T. Hastie and R. Tibshirani, *Generalized additive models*. Boca Raton, Fla : Chapman & Hall/CRC, 1999.

[297] S. Wood, "mgcv : Mixed GAM Computation Vehicle with Automatic Smoothness Estimation," Aug. 2020. [Online]. Available : https://CRAN.R-project.org/package=mgcv

[298] R. J. Hyndman and Y. Khandakar, "Automatic time series forecasting : the forecast package for R," Clayton VIC, Australia : Monash University, Department of Econometrics and Business Statistics, Tech. Rep., 2007.

[299] NOAA, "National Oceanic and Atmospheric Administration | Data Discovery Portal," 2021. [Online]. Available : https://data.noaa.gov/datasetsearch/

[300] C. Kuster, Y. Rezgui, and M. Mourshed, "Electrical load forecasting models : A critical systematic review," *Sustainable Cities and Society*, vol. 35, pp. 257–270, Nov. 2017. [Online]. Available : https://linkinghub.elsevier.com/retrieve/pii/S2210670717305899

[301] S. Wood, "Thin plate regression splines," *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, vol. 65, no. 1, pp. 95–114, Feb. 2003. [Online]. Available : http://doi.wiley.com/10.1111/1467-9868.00374

[302] G. Cybenko, "Approximation by superpositions of a sigmoidal function," *Mathematics of Control, Signals, and Systems*, vol. 2, pp. 303–314, 1989.

[303] K. Hornik, "Approximation capabilities of multilayer feedforward networks," *Neural Networks*, vol. 4, no. 2, pp. 251–257, 1991. [Online]. Available : https://linkinghub.elsevier.com/retrieve/pii/089360809190009T

[304] T. Dozat, "Incorporating nesterov momentum into adam," 2016.

[305] S. Wood, N. Pya, and B. Säfken, "Smoothing Parameter and Model Selection for General Smooth Models," *Journal of the American Statistical Association*, vol. 111, no. 516, pp. 1548–1563, Oct. 2016. [Online]. Available : https://www.tandfonline.com/doi/full/10.1080/01621459.2016.1180986

[306] Jianlin Cheng, Zheng Wang, and G. Pollastri, "A neural network approach to ordinal regression," in *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*. Hong Kong, China : IEEE, Jun. 2008, pp. 1279–1284. [Online]. Available : http://ieeexplore.ieee.org/document/4633963/

[307] B. Peterson and F. E. Harrell Jr, "Partial proportional odds models for ordinal response variables," *Journal of the Royal Statistical Society : Series C (Applied Statistics)*, vol. 39, no. 2, pp. 205–217, 1990.

[308] F. X. Diebold and R. S. Mariano, "Comparing Predictive Accuracy," *Journal of Business & Economic Statistics*, vol. 13, no. 3„ pp. 253–263, 1995. [Online]. Available : http://www.jstor.org/stable/1392185

[309] K. Drachal, "multDM : Multivariate Version of the Diebold-Mariano Test," Jun. 2020. [Online]. Available : https://CRAN.R-project.org/package=multDM

[310] M. Fasiolo, S. N. Wood, M. Zaffran, Y. Goude, and R. Nedellec, "qgam : Smooth Additive Quantile Regression Models," Feb. 2020. [Online]. Available : https://CRAN.R-project.org/package=qgam

[311] M. Fasiolo, S. N. Wood, M. Zaffran, R. Nedellec, and Y. Goude, "Fast Calibrated Additive Quantile Regression," *Journal of the American Statistical Association*, pp. 1–11, Mar. 2020. [Online]. Available : https://www.tandfonline.com/doi/full/10.1080/01621459.2020.1725521

[312] State Government of Victoria, "Department of Transport, Creating a market : Victorian Electric Vehicle Trial Mid-Term Report," State Government of Victoria, Tech. Rep., 2013. [Online]. Available : https://apo.org.au/sites/default/files/resource-files/2013-06/apo-nid34464.pdf

[313] "Engineering Products and Training Courses | EA Technology." [Online]. Available : https://eatechnology.com/

[314] GOV.UK, "National Travel Survey." [Online]. Available : https://www.gov.uk/government/collections/national-travel-survey-statistics

[315] PJM, "Markets & Operations, District of Columbia." [Online]. Available : https://www.pjm.com/markets-and-operations

[316] A. D. Dominguez-Garcia, G. T. Heydt, and S. Suryanarayanan, "Implications of the Smart Grid Initiative on Distribution Engineering," Power Systems Engineering Research Center, Tech. Rep., 2011.

[317] "Ministerie van Verkeer en Waterstaat, Rijkswaterstaat, Dienst Verkeer en Scheepvaart, Mobiliteitsonderzoek Nederland 2008," 2008. [Online]. Available : https://easy.dans.knaw.nl/ui/datasets/id/easy-dataset:34107

[318] "Tesla Motors, Electric Cars, Solar & Clean Energy." [Online]. Available : https://www.tesla.com/

[319] K. Leung, W. Massey, and W. Whitt, "Traffic models for wireless communication networks," *IEEE Journal on Selected Areas in Communications*, vol. 12, no. 8, pp. 1353–1364, Oct. 1994. [Online]. Available : http://ieeexplore.ieee.org/document/329340/

[320] UCDavis, "Plug-In Hybrid & Electric Vehicles Research Center," institute of Transportation Studies. [Online]. Available : https://phev.ucdavis.edu/

[321] NewYorkState, "Traffic Data Viewer." [Online]. Available : https://www.dot.ny.gov/tdv

[322] J. Huber, J. Höffer, J. Thumm, and C. Weinhardt, "Parking events derived from trip data from MiD2008," 2019, type : dataset. [Online]. Available : https://publikationen.bibliothek.kit.edu/1000098024

[323] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977. [Online]. Available : http://www.jstor.org/stable/2984875

[324] R. A. Rigby and D. M. Stasinopoulos, "Generalized additive models for location, scale and shape," *Journal of the Royal Statistical Society : Series C (Applied Statistics)*, vol. 54, no. 3, pp. 507–554, 2005.

[325] H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. 19, no. 6, pp. 716–723, Dec. 1974. [Online]. Available : http://ieeexplore.ieee.org/document/1100705/

[326] B. McGinty, "Wavelet transforms." [Online]. Available : https://www.continuummechanics.org/wavelets.html

[327] M. Misiti, Y. Misiti, G. Oppenheim, and J.-M. Poggi, *Les ondelettes et leurs applications*. Paris : Hermes Science Publications : Lavoisier, 2003, oCLC : 895130533.

Annexe A

About Chapter 2

Exhaustive list of models presented in Section 4 with the input dataset(s) used (if applicable), the approach taken (Aggregated, Vehicle-centric or EVSE-centric) and the output variable(s) modelled. For the input datasets, the data repositories or data reports are provided as references when they were clearly made available by the authors of the paper.

Study	Model(s)	Output variable(s)	Approach	Input dataset(s)
[32]	Gaussian Distribution	Power Demand	Vehicle-centric	NHTS daily trips [61]
[31]	Weibull and Lognormal Distributions	Charge Duration, Daily Charge Frequency, Energy Demand, Time to next charge	Vehicle-centric	Charging sessions from the Victorian EV Trial [312]
[42]	Beta Mixture	Number of plugged-in and plugged-out EVs, Charge Duration	Aggregated and vehicle-centric	Charging sessions from ElaadNL [87]
[151]	Gaussian Mixture	Clusters of EV user profiles	-	Charging sessions provided by the G4 cities of the Netherlands
[141]	Gaussian Mixture	Arrival Time, Charge Duration, Energy Demand	Vehicle-centric	Charging sessions from the ACN data [103]
[152]	Poisson Distribution and Gaussian Mixture	Arrival Time, Charge Duration, Energy Demand	Vehicle-centric	Charging sessions from ElaadNL [87]
[153]	GKDE	Daily Trip Distance and end time of last trip	Vehicle-centric	NHTS daily trips [61]

[154]	GKDE	Arrival Time, Charge Duration, Charge Capacity	Vehicle-centric	Charging sessions from an EV charging service company platform in Nanjing
[155]	GKDE and DKDE	Charge Duration and Energy Demand	Vehicle-centric	Charging sessions from UCLA campus
[156]	GKDE and DKDE	Charge Duration, Energy Demand	Vehicle-centric	Charging sessions from UCLA campus and from My Electric Avenue Project trial [313]
[157]	HKDE	Charge Duration and Energy Demand	Vehicle-centric	Charging sessions from UCLA campus and from My Electric Avenue Project trial [313]
[33]	Temporal stochastic process with scenarios	Power Demand	Vehicle-centric	EV charging load for lead-acid and lithium-ion batteries and UK National Travel Survey [314]
[158]	3-state temporal Markov Chain	Energy Demand	Vehicle-centric	-
[55]	ARIMA and rule-based probabilistic model	Power Demand (via the SoC after charge)	Vehicle-centric	PJM historical load data [315], EV drivers simulated data from the Power Systems Engineering Research Center [316]
[159]	ARIMA and rule-based probabilistic model	Power Demand (via the SoC after charge)	Vehicle-centric	PJM historical load data [315]
[62]	Monte-Carlo Simulation with copulas and scenarios	Energy Demand (via initial SoC)	Vehicle-centric	Transportation data from the Dutch Ministry of Transportation [317]

[27]	Monte-Carlo Simulation with origin destination analysis	Power Demand	Vehicle-centric	EU merge EV database [1]
[36]	Spatiotemporal stochastic process with uniform and poisson distributions on different schedules and state transitions	Power Demand	Vehicle-centric and aggregated	BMW Mini E (PHEV) battery characteristics from the smart grid integration project (Shanghai expo garden)
[50]	Spatiotemporal monte-carlo simulation based on EV trial data	Power Demand	Vehicle-centric	Charging sessions and EV journeys from the Switch EV trial
[47]	$M/M/n_{max}$ queue	Power Demand	Vehicle-centric and EVSE-centric	Tesla Roadster EV characteristics [318]
[56]	$M/M/n_{max}$ queue	Power Demand	Vehicle-centric and EVSE-centric	EV drivers simulated data from the Power Systems Engineering Research Center [316]
[160]	$M/M/n_{max}$ queue	Power Demand	Vehicle-centric and EVSE-centric	Traffic fluid model [319]
[34]	$M_t/GI_t/\infty$ queue	Power Demand	Vehicle-centric and EVSE-centric	NHTS [61], charging sessions from the UC Davis PH&EV center [320]
[57]	BCMP network	Power Demand	Vehicle-centric and EVSE-centric	NHTS [61], New York State Transportation Federation Traffic Data Viewer [321] and Ontario Electricity prices [2]

[49]	Decision trees/tables, SVM, ANN	Power Demand	Aggregated	Residential charging sessions provided by ECOtatility [164], public EVSE data pilot EV project in France
[165]	MPSF, kNN, SVM and RF	Energy Demand	EVSE-centric	Charging sessions from UCLA campus data
[166]	Mean estimation and Linear Regression	Energy Demand (Charging Profile)	Vehicle-centric	Charging sessions from UCLA campus data
[167]	Linear Regression, Gradient Boosting, ANN	Departure time	Vehicle-centric	EV charging records from SAP Walldorf
[168]	Linear Regression	Power Demand (Charging Speed)	Aggregated	Charging sessions from EVNET and NUON energy providers in Amsterdam
[51]	MPSF, SVM and RF	Energy Demand (Charging Profile)	Vehicle-centric and EVSE-centric	Charging sessions from UCLA campus data
[26]	SVM and Monte-Carlo	Power Demand	Vehicle-centric	Beijing Olympic Games EV charging station
[30]	RF	Power Demand	Vehicle-centric and EVSE-centric	Charging sessions from Pecan Street [195]
[169]	RF, SVM and Decition tree	Energy Demand (Charging capacity)	EVSE-centric	Charging sessions in Shenzhen
[59]	Quantile regression, MLP and KDE	Trip Distance and Parking Duration	Vehicle-centric	Parking events derived from trip data from MiD2008 [322]
[67]	Lion algorithm with Niche Immunity for a CNN	Power Demand	Aggregated	Charging sessions in one charging station in Beijing

[46]	ANN, RNN, LSTM and Gated Recurrent Units (GRU)	Energy Demand	Aggregated	Charging sessions in Shenzhen
[170]	ANN, RNN, GRU, Stacked auto-encoders and LSTM	Power Demand	Aggregated	Charging sessions in Shenzhen

Annexe B

About Chapter 3

B.1 Data Quality

This section was firstly published in [200]

A data quality check was run on these datasets to evaluate their reliability :

1. Charge and park duration have to be positive and less than 24 hours
2. Energy consumption needs to be positive and less than 100 kWh

The first criterion discards obvious timestamp errors. In addition, as a passenger EV is fully charged in a few hours maximum, charging sessions lasting for more than a day were discarded. Negative records for the energy consumption were discarded as no EV leaves with a battery less charged than when it arrived. The 100 kWh upper bound was chosen as it is the battery capacity of the Tesla Model S, the largest amongst top selling EVs at the time of writing this paper. All datasets retained more than 85% of transactions after discarding irrelevant observations (Table B.1). For instance, only 63 transactions were discarded from the Palo Alto dataset which covers more than 9 years of data.

Table B.1 – Transactions retained after data preprocessing

Dataset	Timespan	Retained	Discarded
Bou	01-2018/03-2021	21,569 (89.57%)	2,512 (10.43%)
Pal	07-2011/12-2020	259,352 (99.98%)	63 (0.02%)
Dun	01-2017/12-2018	47,051 (89.19%)	5,701 (10.81%)
Per	01-2016/12-2019	63,936 (95.91%)	2,728 (4.09%)
Par	04-2017/05-2017	5,780 (85.72%)	963 (14.28%)
Cha	01-2017/12-2017	2,956,198 (93.06%)	220,605 (6.94%)
Cal	04-2018/04-2021	56,976 (94.44%)	3,357 (5.56%)
Sap	06-2017/01-2021	26,434 (98.55%)	389 (1.45%)

B.2 Mixture Models and Regression

From the literature and our own exploration of real data, we were led to believe that unimodal distributions were not enough to capture the variety of profiles of EV users.

Therefore, one way to characterise the multivariate distribution of the target vector thereafter noted (A, D, E) which is used for the load curve reconstruction is to use a Gaussian Mixture Model (GMM).

B.2.1 Gaussian Mixtures

The finite mixture model of probability distributions consists in assuming that the data comes from a source containing several homogeneous subpopulations called components. The total population is a mixture of these subpopulations. The resulting model is a finite mixture model (which is Gaussian when all subpopulations follow a Gaussian distribution). Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a sample of independent and identically distributed (iid) random variables of finite mixture law with K components, of density f of which the general form is :

$$f(x) = \sum_{k=1}^{K} \pi_k f_k(x) \tag{B.1}$$

with π_k the respective proportions of the subpopulations such that $0 < \pi_k \leq 11$ and $\sum_{k=1}^{K} \pi_k = 1$, and f_k the pdf of the the kth component. The mixture model is a model with missing data. If we sampled in a population formed of K subpopulations, we should have the pairs (X_i, Z_i) where X_i represents the measurement made on the ith individual and $Z_i = k$ indicates the subpopulation to which this individual belongs. However Z_i is not observed. To estimate this kind of model, we require an algorithm called Expected-Maximisation (EM). The EM algorithm is an iterative algorithm originally proposed in [323]. It is a parametric estimation method falling within the general framework of maximum likelihood. When there is missing (or latent variables) and/or the expression of the likelihood is analytically impossible to maximize, the EM algorithm can be a solution. The EM algorithm takes its name from the fact that at each iteration it operates two distinct steps :

1. the “Expectation” step, often referred to as the “E-step”, performs the estimation of unknown data, knowing the observed data and the value of the parameters determined at the previous iteration
2. the “Maximization” step, or “M-step”, maximises the likelihood, which is now possible thanks to the estimation of the unknown data in the E-step. It updates the value of the parameter(s) for the next iteration

In short, the EM algorithm proceeds according to an extremely natural mechanism : if there is an obstacle to apply the maximum likelihood method, we simply jump this obstacle then we actually use this method. To formalise this procedure let us detail the steps of the algorithm at iteration $m \in \mathbb{N}$:

1. Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a sample of iid variables of likelihood $P(X|\boldsymbol{\theta})$, with θ the parameter vector of the mixture model
2. Maximising $P(\boldsymbol{X}|\boldsymbol{\theta})$ cannot be directly achieved
3. We consider hidden data $\boldsymbol{Z} = (Z_1, \ldots, Z_n)$ whose knowledge would make it possible to maximise the “completed likelihood” : $P(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$

4. As $\boldsymbol{Z}$ is unknown, we estimate the completed likelihood by taking into account all known information : the natural estimator is $\mathbb{E}_{Z|X,\boldsymbol{\theta}_m}[P(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta})]$ (E-step)
5. Finally, we maximise this estimated likelihood to determine the new value of the parameter (M-step).

Thus, the transition from iteration m to iteration $m+1$ of the EM algorithm consists in determining :

$$\boldsymbol{\theta}_{m+1} = \underset{\boldsymbol{\theta}}{arg\,max} \left\{ \mathbb{E}_{Z|X,\boldsymbol{\theta}_m}[P(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta})] \right\} \tag{B.2}$$

B.2.2 Mixture Regression

Assuming that we will have access to the intensity function of the arrival process, we will need the Conditional Gaussian Mixture Model (CGMM) characterising the distribution of : $(D,E)|A$. Essentially, a two-step approach will be used to calculated the CGMM. Firstly, the GMM will be fitted using the Expected-Maximisation (EM) algorithm in sci-kit learn. Then, using the gmr package from Alexander Fabisch (Journal of Open Source Software, 2021), we will be retrieving the CGMM from the joint distribution estimated. This is called a gaussian mixture regression which gave the name to the package `gmr`.

From the GMM estimated, each component $k \in \{1 \dots K\}$ can be written :

$$\mathcal{N}_k \left(A, D, E \mid \boldsymbol{\mu_k}, \boldsymbol{\Sigma_k}\right) \tag{B.3}$$

with :

$$\boldsymbol{\mu_k} = \begin{pmatrix} \mu_k^A \\ \mu_k^D \\ \mu_k^E \end{pmatrix}, \quad \boldsymbol{\Sigma_k} = \begin{pmatrix} \sigma_k^{AA} & \sigma_k^{AD} & \sigma_k^{AE} \\ \sigma_k^{DA} & \sigma_k^{DD} & \sigma_k^{DE} \\ \sigma_k^{EA} & \sigma_k^{ED} & \sigma_k^{EE} \end{pmatrix} \tag{B.4}$$

Therefore, the conditional distribution of each component can be defined as follows :

$$\mathcal{N}_k \left(D, E \mid A = a, \boldsymbol{\mu_k^{DE}}, \boldsymbol{\Sigma_k^{DE}}\right) \tag{B.5}$$

with :

$$\begin{aligned} \boldsymbol{\mu_k^{DE}} &= \begin{pmatrix} \mu_k^D \\ \mu_k^E \end{pmatrix} + \frac{1}{\sigma_k^{AA}}(a - \mu_k^A) \begin{pmatrix} \sigma_k^{DA} \\ \sigma_k^{EA} \end{pmatrix} \\ \boldsymbol{\Sigma_k^{DE}} &= \begin{pmatrix} \sigma_k^{DD} & \sigma_k^{DE} \\ \sigma_k^{ED} & \sigma_k^{EE} \end{pmatrix} - \frac{1}{\sigma_k^{AA}} \begin{pmatrix} (\sigma_k^{DA})^2 & \sigma_k^{DA}\sigma_k^{EA} \\ \sigma_k^{DA}\sigma_k^{EA} & (\sigma_k^{EA})^2 \end{pmatrix} \end{aligned} \tag{B.6}$$

With all the individual components having an explicit expression, the CGMM priors can be calculated as follows :

$$\pi_k^{DE}(a) = \frac{\pi_k \mathcal{N}_k \left(A = a \mid \mu_k^A, \sigma_k^A\right)}{\sum_{l=1}^K \pi_l \mathcal{N}_l \left(A = a \mid \mu_k^A, \sigma_k^A\right)} \tag{B.7}$$

and we eventually get an expression for the CGMM density and the GMR model :

$$p(D, E \mid A = a) = \sum_{k=1}^{K} \pi_k^{DE}(a) \mathcal{N}_k \left(D, E \mid A = a, \boldsymbol{\mu_k^{DE}}, \boldsymbol{\Sigma_k^{DE}} \right) \tag{B.8}$$

B.3 Grid Search

The following tables present the predictors and hyperparameters found with the grid search procedure detailed in Section 3.4.2. The abbreviations of the predictors that were retained for at least one model in the grid search procedure are detailed in Table B.2. Table B.3 covers NHPP-GAM and GAM models, Table B.4 covers NHPP-RF and RF models and Table B.5 covers the GRU models for all datasets in scope. Modelling the number of daily sessions does not leave space for many covariates to use for the GRU. Therefore, we made the choice that all GRU models would have access to the following predictors : dow and LAG1 to LAG5.

Table B.2 – Abbreviations of the predictors

Abbreviation	Meaning	Range
mod	Minute of the day	$1 \dots 1440$
hour	Hour	$1 \dots 24$
tod32	45-minute period	$1 \dots 32$
winter	October to March	0 or 1
dow	Day of week	$1 \dots 5$
trend	Trend	$\mathbb{N}$
r_lags	Rolling sum of lagged arrivals	$\mathbb{R}^+$
a_lags	Aggregated lagged arrivals	$\mathbb{N}$
LAGi	Daily number of sessions for day$-i$	$\mathbb{N}$

Table B.3 – Optimal hyperparameters and predictors for the NHPP-GAM and GAM models

Model	Dataset	Predictors
NHPP-GAM	Boulder	mod, tod32
	Caltech	mod, tod32
	Domestics UK	mod, tod32
	Dundee	mod, tod32
	Palo Alto	mod, tod32, winter
	Paris	mod
	Perth	mod, tod32
	SAP	mod, tod32
GAM (Occupancy)	Boulder	mod, dow
	Caltech	mod, dow
	Domestics UK	mod, dow, hour, trend, winter
	Dundee	mod, dow, hour, trend, winter
	Palo Alto	mod, dow
	Paris	mod, dow
	Perth	mod, dow, hour, trend, winter
	SAP	mod, dow, hour, trend
GAM (Load)	Boulder	mod, dow, hour trend, winter
	Caltech	mod, dow
	Palo Alto	mod, dow, trend, winter
	SAP	mod, dow, hour, trend, winter

Table B.4 – Optimal hyperparameters and predictors for the NHPP-RF and RF models

Model	Dataset	Predictors	Hyperparameters
NHPP-RF	Boulder	mod, tod32, dow, winter, trend, r_lags	num.trees : 500 mtry : 1
	Caltech	mod, tod32, dow, winter	num.trees : 100 mtry : 2
	Domestics UK	mod, tod32 dow, winter, trend, a_lags	num.trees : 500 mtry : 2
	Dundee	mod, tod32, dow, winter, trend, r_lags	num.trees : 100 mtry : 1
	Palo Alto	mod, tod32, dow	num.trees : 100 mtry : 2
	Paris	mod, tod32	num.trees : 200 mtry : 1
	Perth	mod, tod32, dow, winter	num.trees : 200 mtry : 2
	SAP	mod, tod32, dow, winter	num.trees : 500 mtry : 2
RF (Occupancy)	Boulder	mod, hour, dow, winter	num.trees : 700, mtry : 2
	Caltech	mod, hour, dow, winter	num.trees : 1000, mtry : 3
	Domestics UK	mod, hour, dow, winter	num.trees : 100,mtry : 3
	Dundee	mod, hour, dow, winter	num.trees : 500, mtry : 2
	Palo Alto	mod, hour, dow, winter	num.trees : 500, mtry : 2
	Paris	mod, hour, dow, winter	num.trees : 1000, mtry : 2
	Perth	mod, hour, dow, winter	num.trees : 500, mtry : 2
	SAP	mod, hour, dow, winter	num.trees : 500, mtry : 2
RF (Load)	Boulder	mod, hour, dow, winter	num.trees : 1000, mtry : 2
	Caltech	mod, hour, dow, winter	num.trees : 700, mtry : 2
	Palo Alto	mod, hour, dow, winter	num.trees : 100, mtry : 2
	SAP	mod, hour, dow, winter	num.trees : 100, mtry : 3

Table B.5 – Optimal hyperparameters and predictors for the GRU models

Dataset	Hyperparameters
Boulder	batch_size : 10 hidden_dim : 32 layer_dim : 2 learning_rate : 0.001 n_epochs : 4
Caltech	batch_size : 10 hidden_dim : 64 layer_dim : 6 learning_rate : 0.001 n_epochs : 975
Domestics UK	batch_size : 10 hidden_dim : 16 layer_dim : 4 learning_rate : 0.01 n_epochs : 25
Dundee	batch_size : 10 hidden_dim : 32 layer_dim : 6 learning_rate : 0.01 n_epochs : 698
Palo Alto	batch_size : 5 hidden_dim : 16 layer_dim : 2 learning_rate : 0.001 n_epochs : 297
Paris	batch_size : 2 hidden_dim : 64 layer_dim : 6 learning_rate : 0.01 n_epochs : 566
Perth	batch_size : 30 hidden_dim : 16 layer_dim : 2 learning_rate : 0.01 n_epochs : 2
SAP	batch_size : 5 hidden_dim : 16 layer_dim : 4 learning_rate : 0.01 n_epochs : 877

B.4 Statistical Testing

In this section, the results of the model validation procedure detailed in Section 3.4.3 are given. In particular, Table B.6 presents the 95% confidence interval $[b_{inf}, b_{sup}]$ for the sample mean $\hat{G}$. If the sample mean falls within the bounds, we cannot reject the null hypothesis and therefore the model is considered to be valid. Similarly, Table B.7 gathers the p-values of the Ljung-Box test led for SARIMA models. If the p-value is above the 5% threshold, we cannot reject the null hypothesis and the model is therefore considered valid.

Table B.6 – Bounds of the 95% confidence interval and sample mean $\hat{G}$ of the Student's test for validating NHPP models

Dataset	Model	b_{inf}	b_{sup}	$\hat{G}$
Boulder	NHPP-GAM	-0.17	0.17	-0.07
	NHPP-RF	-0.17	0.17	-0.05
Caltech	NHPP-GAM	-0.71	0.71	-0.16
	NHPP-RF	-0.7	0.7	-0.16
Domestics UK	NHPP-GAM	-4.48	4.48	-2.9
	NHPP-RF	-4.47	4.47	-2.5
Dundee	NHPP-GAM	-0.69	0.69	-0.57
	NHPP-RF	-0.75	0.75	-0.33
Palo Alto	NHPP-GAM	-0.8	0.8	-0.1
	NHPP-RF	-0.8	0.8	-0.1
Paris	NHPP-GAM	-0.45	0.45	0.04
	NHPP-RF	-0.44	0.44	0.02
Perth	NHPP-GAM	-0.37	0.37	-0.21
	NHPP-RF	-0.36	0.36	-0.20
SAP	NHPP-GAM	-0.41	0.41	-0.1
	NHPP-RF	-0.42	0.42	-0.05

Table B.7 – P-values for the 5% confidence threshold of the Ljung-Box test for validating SARIMA models

Dataset	p-value
Boulder	0.54
Caltech	0.17
Domestics UK	0.78
Dundee	0.88
Palo Alto	0.12
Paris	0.68
Perth	0.02
SAP	0.1

B.5 Additional Metrics

With online aggregation, a transitional regime inevitably occurs in the first testing weeks. To determine whether it is responsible for GAM-err being more performant than AGG on Table 3.2(a) we have decided to calculate the block-bootstrap metrics only on the last testing period for all datasets. This should give enough time for AGG to learn optimal weights. In fact, on Table B.8 we can see that we were correct and AGG is better on block-bootstrap average than GAM-err for both load and occupancy forecasting.

Table B.8 – Average block-bootstrap model performances for the PIP on the last testing period

(a) **Load**

Model	**MAE [%]**
AGG	45.7
GAM-err	43.5
RF-err	42.9
RF	38.1
GAM	37.5
SARIMA-MM-err	34.4
GRU-MM-err	29.8
NHPP-GAM-MR-err	25.3
SARIMA-MM	22.0
GRU-MM	19.7
NHPP-RF-MR-err	16.3
NHPP-GAM-MR	7.9
NHPP-RF-MR	5.9
persistence	0

(b) **Occupancy**

Model	**MAE [%]**
AGG	116
GAM-err	109
GAM	107
RF-err	80.0
SARIMA-MM-err	76.0
SARIMA-MM	70.0
RF	57.4
GRU-MM-err	55.4
NHPP-GAM-MR-err	34.4
GRU-MM	31.4
NHPP-RF-MR-err	23.8
persistence	0
NHPP-GAM-MR	-0.7
NHPP-RF-MR	-7.3

Annexe C

About Chapter 4

In the following sections, we introduce the algorithms which are used in Chapter 4 as stepping stones for our proposed estimation procedure. In particular, we recall the Iteratively Reweighted Least Squares (IRLS) for Generalised Linear Models (GLM) and its penalised version (PIRLS) for Generalised Additive Models (GAM). In addition, we also recall some general wavelet properties before explaining how the wavelet basis are built for the BAC and OBO estimation procedures.

C.1 Iteratively reweighted least squares

The basic GLM structure is the following :

$$g(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta} \tag{C.1}$$

with $\boldsymbol{\mu} = \mathbb{E}(\mathbf{Y})$ while g is the 'link' function. The only hypothesis made on this function is that it needs to be smooth and monotonic. So far, the extension seems rather narrow. But the second and most important point is that we assume that $\mathbf{Y}$ can follow any distribution from the exponential family. The weighted log-likelihood of a GLM can be written in the following fashion :

$$l(\boldsymbol{\beta}) = \sum_{i}^{n} \omega_i l_i(\boldsymbol{\beta}) \tag{C.2}$$

with $l_i(\boldsymbol{\beta}),\ i \in \{1 \dots n\}$ the log-likelihood of the model with only one observation. The goal is to maximize this log-likelihood which we can do by using Newton's method. Skipping the heavy calculations, we move directly to the expression of the Newton update :

$$\beta^{k+1} = (\mathbf{XWX})^{-1}\mathbf{X}^{\intercal}\mathbf{Wz} \tag{C.3}$$

with $z_i = g'(\mu_i)\frac{y_i - \mu_i}{\alpha(\mu_i)} + \mathbf{X_i}\boldsymbol{\beta}$, $\alpha(\mu_i) = 1 + (y_i - \mu_i)(\frac{V'(\mu_i)}{V(\mu_i)} + \frac{g''(\mu_i)}{g'(\mu_i)})$, V is a function of μ_i and $W = diag(w_i)$, $w_i = \frac{\alpha(\mu_i)}{g'(\mu_i)^2 V(\mu_i)}$.

The trick here is to understand that the right term of equation C.3 exactly coincides with the least square estimate which is the outcome of the minimisation of the following least square objective :

$$\|\mathbf{z} - \mathbf{X}\boldsymbol{\beta}\|_W^2 = \sum_{i=1}^{n} w_i (z_i - \mathbf{X}_i\boldsymbol{\beta})^2 \tag{C.4}$$

As a result, we can estimate GLMs by the **Iteratively Re-Weighted Least Square** (IRLS) algorithm :

1. **Initialize** : $\hat{\mu_i} = y_i + \delta_i$ and $\mathbf{X_i}\hat{\boldsymbol{\beta}} = g(\hat{\mu_i})$ with δ_i a number close to 0 which ensures that $\mathbf{X_i}\hat{\beta}$ is well defined
2. **Compute** : $z_i := g'(\hat{\mu_i})\frac{y_i - \hat{\mu_i}}{\alpha(\hat{\mu_i})} + X_i\hat{\beta}$ and $w_i = \frac{\alpha(\hat{\mu_i})}{g'(\hat{\mu_i})^2 V(\hat{\mu_i})}$
3. **Minimize** : the quantity C.4 referred to as the weighted least square objective in order to get the new estimate $\hat{\beta}$ (which is equivalent to a step of Newton's method)
4. **Update** : $\hat{\mu_i} = g^{-1}(\hat{\eta_i})$
5. **Iterate** : from step 2 until convergence

And with that algorithm we can obtain our parameter vector estimate $\hat{\boldsymbol{\beta}}$ which can then be used to make predictions with the model and assess its performance.

C.2 Penalised iteratively reweighted least squares

Generalized Additive Models (GAMs) are an extension of the GLM presented previously. Firstly introduced by Hastie and Tibshirani [207], the main particularity of a GAM with regards to its formulation lies in the smoothing terms presented in the linear predictor. Essentially we can write a GAM as follows :

$$g(\mu_i) = \mathbf{A_i}\boldsymbol{\gamma} + \sum_j f_j(x_{ji}), y_i \sim EF(\mu_i, \phi) \tag{C.5}$$

Thus, we can distinguish two parts within this equation when thinking of estimation. A parametric part $\mathbf{A_i}\boldsymbol{\gamma}$, $\mathbf{A_i}$ being a design matrix while $\boldsymbol{\gamma}$ the parameter vector. A non-parametric part $\sum_j f_j(x_{ji})$ which is essentially a sum of smooth functions of some predictor variables. As for the response variable, it follows typically an exponential family distribution just as in the GLM case. However, this scope of distribution can also be extended with the use of GAM for location scale and shape formalised by Rigby and Stasinopoulos [324].

There are two main questions that are fundamental with GAMs : how to determine the scheme of smooth functions and how to control their smoothness. For the former point, the answer is relatively concise : GAMs use basis expansion in order to represent the different f of the model. Thus, a general formulation for the smooth functions is the following one :

$$f(x) = \sum_{l=1}^{k} b_l(x)\beta_l \tag{C.6}$$

In the GAM vernacular, these basis functions b_l are called splines. The most common ones are cubic splines or cyclic splines. We can notice that by inputting equation (C.6) in the general GAM formulation (C.5) we bounce back on a linear formulation in the parameters of the model. Then, we have to estimate the unknown parameters β_l For

instance, k being fixed, the cubic basis being chosen, it is a special case of generalized polynomial regression which is linear in the parameters.

The second point which requires our attention is the smoothness or wiggliness of these functions. In other words how to choose k, the number of basis functions in the representation (knots), in order to ensure that we do not use too simple or too complex functions to represent the data. With GAMs, the knots of the splines are evenly spaced. It means that different choices of k lead to a variety of grids of the predictor space and as such, cannot be considered as different nested configurations. Thus, backward or forward selection is not an option for choosing k.

In effect, one solution to this problem is not to attempt to select k, but to introduce a measure of wiggliness of our functions for a given k. Essentially, the principle lies into a reformulation of the least square objective. Instead of trying to minimise :

$$\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2$$

we introduce λ to control wiggliness with the newly formed objective function to minimise :

$$\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \sum_{l=2}^{k-1} f(x^*_{j-1}) - 2f(x^*_j) + f(x^*_{j+1})$$

with $x^*_j,\ j \in \{1 \dots k\}$ one of the evenly spaced k knots. That way, providing that k is large enough to ensure enough flexibility, searching for a good λ is now the only thing to do to optimise wiggliness. Indeed, the choice of k and the fact that the knots x^*_j are evenly spaced has little impact on the final model fit in comparison to λ [208]. λ is called the **smoothing parameter**. In a nutshell, we need to find a way to find both an estimate for the unknown parameters b_l and the "best" possible λ according to a defined methodology . Hence, GAMs can be considered as GLMs with a smoothing penalty. The general formulation of GAMs (C.5) introduces a pitfall in terms of identifiability for the non-parametric side of the formula. Indeed, each function can compensate the constant part of another function. Thus, it is essential to add some identifiability constraints such as $\sum_j f_j(x_{ji}) = 0$. Now that we have sorted model identifiability requirements we can now define the model matrix. It is made of several sub-matrices from the parametric and the semi-parametric part of the model. $\mathbf{A}$ is the sub-matrix from the parametric part as defined in (C.5) and we now introduce $\mathbf{X}^{[j]}$ as the sub-matrices for each smooth predictor term f_j. Hence, by gathering these sub-matrices we get the global model matrix as such : $\mathbf{X} = (\mathbf{A}, \mathbf{X}^{[1]}, \mathbf{X}^{[2]}, \dots)$. The model parameter vector $\boldsymbol{\beta}$ which is linked to this model matrix contains the parameter vector $\boldsymbol{\gamma}$ and the b_{jl} of the f_j stacked together. Furthermore, we can also define penalisation matrices $\mathbf{S}^{[j]}$ associated to each f_j in such a way to allow that the term : $\lambda_j \boldsymbol{\beta}^\intercal \mathbf{S}^{[j]} \boldsymbol{\beta}$ is exactly the penalty for f_j. With this new formalism, the model defined in (C.5) can now be rewritten in the form of a GLM under additional constraints :

$$g(\mu_i) = \mathbf{X}_i \beta,\ y_i \sim EF(\mu_i, \phi) \tag{C.7}$$

with the following objective function to maximise :

$$l(\boldsymbol{\beta}) - \frac{1}{2\phi}\sum_j \lambda_j \boldsymbol{\beta}^\intercal S^{[j]} \boldsymbol{\beta} \tag{C.8}$$

having the same $l(\boldsymbol{\beta})$ as in (2.7) for the GLM. Finally the optimisation problem can be written as :

$$\underset{\{\boldsymbol{\beta},\lambda_1,\lambda_2,\ldots\}}{\arg\max} \{l(\boldsymbol{\beta}) - \frac{1}{2\phi}\sum_j \lambda_j \boldsymbol{\beta}^\intercal S^{[j]} \boldsymbol{\beta}\}$$

we have just multiplied the penalisation part by a constant term $\frac{1}{2\phi}$ for normalisation. This objective is identical to the GLM case (2.9) to which we add a penalisation term. Consequently, this optimisation problem can be solved with the **Penalised Iteratively Reweighted Least Square** algorithm (PIRLS) which is very similar to the IRLS one defined for the GLM case at the end of section 2.2.3.

Assuming a given λ the PIRLS algorithm is identical to the IRLS version except for step 3. which will be looking at minimising the following quantity :

$$\|\mathbf{z} - \mathbf{X}\boldsymbol{\beta}\|_W^2 + \sum_j \lambda_j \boldsymbol{\beta}^\intercal \mathbf{S}^{[j]} \boldsymbol{\beta}$$

This is a traditional quadratic optimisation problem under constraint which can be solved efficiently with various methods.

We are now able to estimate model parameters for a given $\boldsymbol{\lambda}$. But how do we choose a valid $\boldsymbol{\lambda}$? We will not dive into too much details on that matter. Let us just note that there are two main scenarios depending on whether the scale parameter is known. If it is known, one way to address the choice of λ is the Un-Biased Risk Estimator (UBRE) defined by Craven and Wahba [249]. It is an adapted version of the Akaike Information Criterion [325]. If it is unknown, then two methods can be considered. The first one relies on cross validation. By watching prediction error, we can perform Ordinary Cross Validation (OCV) and compare different values for λ on a certain grid. For computational efficiency, the `mgcv` package created by Simon Wood [180] in R implements a different type of cross validation called General Cross Validation (GCV) which is more computationally efficient than OCV. The second one relies on Bayesian considerations. From a probabilistic perspective it is possible to consider a Bayesian view of smoothing. In that case, the smoothing parameter λ will be associated with a Gaussian prior on model coefficients. Thus, we define the $\log$ of the Bayesian Marginal likelihood as follows :

$$ML(\lambda) = \log \int f(y|\boldsymbol{\beta}) f(\boldsymbol{\beta}) d\boldsymbol{\beta} \tag{C.9}$$

In other words, it is the $\log$ of the joint density of the response vector and model coefficients $\boldsymbol{\beta}$. Then we maximise this quantity with regards to λ in order to get the optimal value for the smoothing parameter.

C.3 Penalised wavelet additive model

In this section, we recall some general wavelet properties before explaining the procedure we have retained for wavelet basis construction in the case of the model proposed in chapter 4.

C.3.1 Wavelet Properties

A wavelet is an oscillating function ψ localised in time and frequency which verifies the following condition :

$$\int_{\mathbb{R}^+} \frac{|\hat{\psi}(\omega)|^2}{|\omega|} d\omega = \int_{\mathbb{R}^-} \frac{|\hat{\psi}(\omega)|^2}{|\omega|} d\omega < +\infty \tag{C.10}$$

with $\hat{\psi}$ the Fourier transform of ψ. In particular, ψ has zero mean and it is usual to have additional conditions on moments of higher-order to be also null to add smoothness to the wavelet. It is frequent for these functions to be defined with a finite support which helps with the localised properties of wavelets. However not all wavelets benefit from a finite support (e.g., Meyer wavelets [254]). In addition, numerous wavelets do not have an analytical formulation and need to be calculated recursively or through algorithms in practice. Several examples of wavelet functions are shown in Figure C.1.

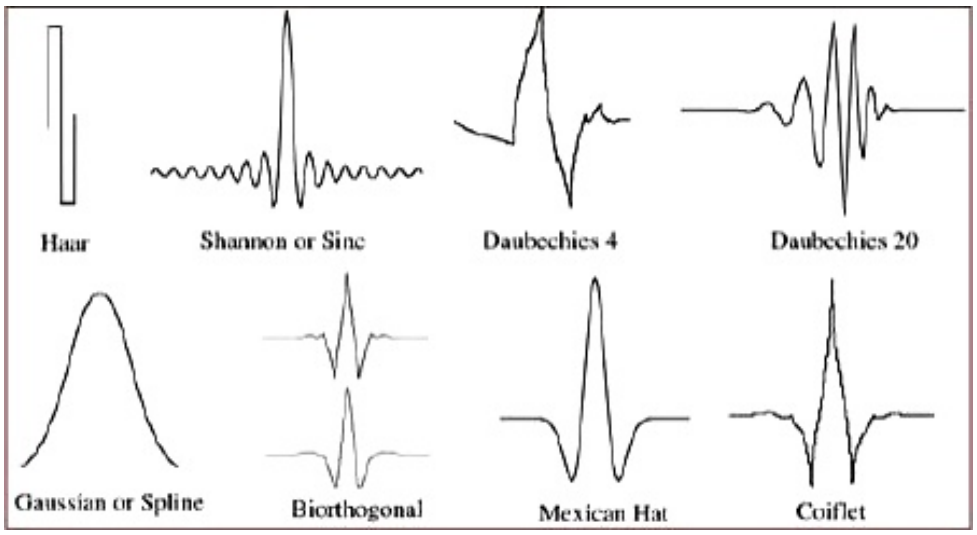


Figure C.1 – Illustration of different wavelet functions [326]

From a unique wavelet function, we can generate a family of functions by translation and dilation called atoms (also referred to as wavelets for conciseness) which are defined as follows :

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}} \psi \left(\frac{t-b}{a} \right), \ a \in \mathbb{R}^+, b \in \mathbb{R} \tag{C.11}$$

we can then define the wavelet transform of any function f with finite energy ($f \in L^2(\mathbb{R})$) as follows :

$$C_f(a,b) = \int_{\mathbb{R}} f(t)\overline{\psi_{a,b}(t)}dt \tag{C.12}$$

The wavelet coefficients C_f essentially measure the fluctuation of the function f on the scale a. Precisely, the value of the wavelet coefficient $C_f(a,b)$ depends on the values of f in a neighbourhood of b of size proportional to a. It is common to define a and b based on $(j,k) \in \mathbb{Z}^2$as follows :

$$\forall (j,k) \in \mathbb{Z}^2,\ a = 2^j,\ b = ka \tag{C.13}$$

When a family of wavelets is orthogonal, a scale function usually noted ϕ is often associated to ψ. In this case, ϕ is also called the father wavelet and ψ the mother wavelet. A family of wavelets can also be generated from the father wavelet ϕ in the same way as presented in equation (C.11). In this special case, the wavelet coefficients of a function f are given by the following :

$$\alpha_{j,k} = \int_{\mathbb{R}} f(t)\psi_{j,k}(t)dt \tag{C.14}$$

with $\psi_{j,k}(t) = 2^{-\frac{j}{2}}\psi\left(2^{-j}t - k\right),\ \forall (j,k) \in \mathbb{Z}^2$, which is equation C.11 where a and b have been replaced with their definition in equation C.13. From an algebric point of view, the scale family $\{\phi_{j,k}\}$ generates a vector space V_j called an approximation space while the family $\{\psi_{j,k}\}$ generates a vector space W_j called a detail space. Two general properties are particularly useful with regards to these vectorial spaces. The first one linking both spaces is that $V_{j-1} = V_j \oplus W_j,\ \forall j \in \mathbb{Z}$. We say that an element of the aprroximation space of level $j-1$ can be decomposed in an approximation of level j (which is more rough) and the detail of level j. The second property is that $L^2(R) = \bigoplus_{j\in\mathbb{Z}} W_j$. Therefore, any signal of $L^2(R)$ can be written as the sum of all its details. That is, the family $\{\psi_{j,k}\}_{(j,k)\in\mathbb{Z}^2}$ is an orthonormal wavelet basis of $L^2(R)$.

A famous example of this class of functions is the Haar father and mother wavelets. These functions have the rare property amongst wavelets to have an explicit analytical form and also benefiting from a compact support. The Haar mother wavelet is defined as follows :

$$\psi(x) = \begin{cases} -1 & \text{if } 0 \leq x < \frac{1}{2} \\ 1 & \text{if } \frac{1}{2} \leq x < 1 \\ 0 & \text{otherwise.} \end{cases} \tag{C.15}$$

and the the Haar father wavelet is defined as follows :

$$\phi(x) = \begin{cases} 1 & \text{if } 0 \leq x < 1 \\ 0 & \text{otherwise.} \end{cases} \tag{C.16}$$

This wavelet originated from Alfred Haar in 1909, who first proposed the "Haar transform". However, it is only in 1987 that Ingrid Daubechies demonstrated that wavelet transforms, of which the Haar transform is a particular case, were highly relevant in the field of signal processing (and beyond). The Haar wavelet has the particularity to be piecewise constant, a property which will be useful for the applications led in this manuscript. A thorough classification of wavelets accompanied with many examples is available in [327].

C.3.2 Wavelet Basis Construction

The way we construct our wavelet basis before estimating our model parameters is greatly inspired from the work produced in [248]. Essentially, the classical construction of wavelets on a grid of $2^J,\ J \in \mathbb{N}^*$ equidistant points on [0, 1] can be summarised as follows. Let us denote the $2^J - 1$ wavelets (details) $\{z_k(u),\ k \in \{1, \dots, 2^J - 1\}\}$, defined on [0, 1), to which we add the constant scale function equal to β_0 on [0, 1] (thus V_0 is the space of constant functions on [0,1]). Therefore the usual approximation used for any function f is given by

$$f(x) = \beta_0 + \sum_{k=1}^{K} u_k z_k(x) \tag{C.17}$$

For the sake of this example we write β_0 as part of the basis but in OBO and BAC the β_0 is considered to be part of the intercept of the model which was already estimated before entering the main BAC or OBO routine.

Let us then note $W = 2^{-\frac{J}{2}} \left(z_k(\frac{i-1}{2^J})\right)_{k\in\{1,\dots,2^J-1\},\ i\in\{1,\dots,2^J\}}$ where W is the orthonormal matrix of size $2^J \times 2^J$ known in the literature as the orthonormal basis of wavelets. If y is the 2^J dimensional vector composed of the values of f on the dyadic grid of the 2^J points of [0, 1] then it is represented by $y = \boldsymbol{W}\hat{\boldsymbol{\theta}}$ where, by orthogonality of W,

$$\hat{\boldsymbol{\theta}} = (\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{W}^T\boldsymbol{y} = \boldsymbol{W}^T\boldsymbol{y} \tag{C.18}$$

A fast O(R) algorithm, known as the discrete wavelet transform, exists for the determination of $\hat{\boldsymbol{\theta}}$. This is implemented in the `wavethresh` package in R and the wavelet basis construction as we described above is available trough the `WavD` function in the `gamwave` package [248].