

Location Intelligence: an integrated Decision Support System for improved infrastructure placement

Gonçalo Filipe Ferreira Rodrigues

Master's Dissertation

Supervisor: Prof. Flávia Barbosa



Mestrado em Engenharia e Gestão Industrial

2023-07-21

Abstract

The growing demand for electric charging poses challenges in terms of managing and scaling the infrastructure effectively. To address this, a generic methodology for infrastructure placement is proposed in this work. The methodology focuses on disaggregating demand to granular and precise locations, demand forecasting for projection in the future and identifying booming areas, and demand capture for prediction of sales and trade area estimation. Each step is accompanied by insightful visualizations, enabling more robust and supported location decision methodologies for key stakeholders and decision-makers. The methodology is applied in a case study of the public electric charging grid in Portugal, using real data from a charging point operator (CPO), as well as publicly available data.

The work starts by obtaining the data through web scraping, followed by disaggregating demand to a high-granularity level that incorporates sociodemographic variables and their interactions. The model is trained using aggregated sales data. Once current demand is disaggregated, future demand is forecasted with hierarchical time series, and then captured by the charging stations in the grid. The model takes into account distance as a factor in determining the capture of demand by each charging station.

To validate the methodology, the predicted sales per charging station are compared with the actual sales, with cross-validation in all steps. The results demonstrate the effectiveness of the methodology in accurately forecasting demand and capturing it within the charging grid. The insights gained from each step are highlighted, emphasizing its impact on the placement of charging infrastructure.

In addition to the forecasting and capture models, the work also emphasizes the importance of visualizations in supporting decision-making. Adequate visualizations are presented in all steps, providing a comprehensive view of the charging infrastructure and aiding in strategic planning.

The proposed methodology offers a general framework capable of effectively addressing a wide range of location problems involving demand forecast and capture. Not only can this general methodology be used in any other location problem with demand forecast and capture, but the specific implementation can serve the same purpose, as it was developed to be as robust as possible, dealing with data quality issues in hard problems. Its comprehensive guidelines ensure the robustness of each step while maximizing data insights. The developed implementation strikes a balance between robustness and performance, making it adaptable to other use cases with minor adjustments. The modular approach employed allows for easy substitution of techniques in individual steps to suit specific problem requirements. The implementation focuses on transitioning from the use of private data to public data, in an effort to make it user-agnostic, unbiased, and generalized. Although the implementation may initially appear intricate due to its sequential nature, its primary objective is to leverage all available data effectively. As a result, it yields a readily deployable global solution for location problems, featuring accurate models that go from actual sales to their forecast and new location recommendations, augmented by intuitive visualizations to facilitate decision-making.

Overall, this work offers a comprehensive methodology for infrastructure placement, applied to the electric charging sector, addressing the challenges of forecasting demand and capturing it effectively. The case study in Portugal demonstrates the applicability of the methodology and its potential to support charging point operators in optimizing their infrastructure deployment. The end result was a self-contained product for use by decision-makers in the electric vehicle charging stations (EVCSs) market.

Keywords: Location intelligence, Spatial decision support system, GIS, Electric vehicle charging stations, Location decision

Resumo

O crescimento da procura de carregamento elétrico vem trazer novos desafios em termos de gestão e dimensionamento da infraestrutura. De modo a contribuir para a resolução desta questão, propõe-se com o presente trabalho uma metodologia genérica para a localização de nova infraestrutura de carregamento. A metodologia proposta foca-se na desagregação da procura para localizações exatas e com elevada granularidade, bem como na previsão da procura, identificação das áreas com maior potencial de crescimento, e a captura da procura para previsão de vendas e estimação da área de influência. Cada etapa do processo é acompanhada por visualizações representativas, permitindo uma tomada de decisão mais robusta, clara e suportada por parte dos decisores. A metodologia é aplicada a um caso de estudo da rede de carregamento elétrico em Portugal, utilizando dados reais de um operador de postos de carregamento, bem como dados públicos.

Começa-se por obter os dados através de *web scraping*, seguido-se a desagregação da procura a um nível elevado de granularidade que incorpora variáveis sociodemográficas e as suas interações. O modelo é treinado com dados de vendas agregados. Uma vez desagregada a procura atual, a procura futura é prevista com séries temporais hierárquicas, sendo depois capturada pela rede de carregamento elétrico. O modelo tem em conta a distância como fator determinante para a captura da procura por cada posto de carregamento.

Para validar a metodologia, as vendas previstas por posto de carregamento são comparadas com as vendas reais, e o seu desempenho é avaliado com validação cruzada em todos os passos. Os resultados demonstram a eficácia da metodologia na previsão da procura e na sua captura pela rede de carregamento elétrico. Em cada etapa são retiradas conclusões, fundamentando o seu impacto na localização de novos postos de carregamento elétrico.

Para além dos modelos de previsão e de captura da procura, o trabalho torna evidente a importância das visualizações no apoio à tomada de decisões. Estas são apresentadas em todas as etapas, proporcionando uma visão abrangente da infraestrutura de carregamento e suportando o planeamento estratégico.

A metodologia proposta apresenta uma estrutura capaz de abordar uma vasta gama de problemas de localização que envolvem a previsão e a captura da procura. Esta metodologia geral pode ser utilizada em qualquer outro problema de localização com previsão e captura da procura. A implementação específica realizada pode também ser aplicada a outros problemas, sendo capaz de lidar com dados de má qualidade em problemas complexos. As directrizes definidas asseguram a robustez de cada passo, permitindo obter conclusões relevantes a partir da globalidade dos dados. A implementação desenvolvida estabelece um equilíbrio entre robustez e desempenho, sendo adaptável a outros casos de utilização com pequenos ajustes. A abordagem modular permite a substituição do método utilizado em cada etapa individual, indo ao encontro dos requisitos de novos problemas específicos. A implementação centra-se ainda na transição da utilização de dados privados para dados públicos, procurando tornar a ferramenta independente do utilizador, imparcial e genérica. Embora a natureza sequencial da implementação a torne complexa, o objetivo principal

é utilizar todos os dados disponíveis, com modelos adequados a cada tipo e natureza dos dados.

A solução desenvolvida representa uma ferramenta completa para aplicação em problemas de localização, permitindo chegar das vendas históricas a previsões de vendas futuras com elevado nível de granularidade e recomendação de novas localizações potenciadas por visualizações intuitivas, que suportam e facilitam a tomada de decisão.

Em suma, este trabalho propõe uma metodologia abrangente para a localização de infraestrutura, aplicada ao sector do carregamento elétrico, abordando os desafios da previsão da procura e da sua captura. O caso de estudo em Portugal demonstra a aplicabilidade da metodologia e o seu potencial para apoiar os fornecedores na otimização da localização dos seus postos de carregamento. O resultado final consiste num produto independente, para utilização por decisores no mercado dos postos de carregamento de veículos eléctricos.

Palavras-chave: *Location intelligence*, Sistemas de apoio à decisão espaciais, GIS, Postos de carregamento de veículos eléctricos, Decisão de localização

Acknowledgments

This thesis concludes my Masters in Industrial Engineering and Management at FEUP. It closes a five-year chapter of my life, for which I am most grateful.

I'd like to begin by thanking my academic supervisor Professor Flávia Barbosa, for the timely advice, availability, and support during this work. I would especially like to thank my company supervisor Eng. Pedro Schuller for the restless support and endless patience in guiding me through this arduous but rewarding challenge, as well as for his availability to revise not only this work but also the published article. I would also like to thank LTPlabs for the opportunity to take part in such interesting projects, showing that even the most academic knowledge and complex models have a place in industry, in this very special enterprise, where no problem is impossible to solve.

Throughout my passage in FEUP I have grown as a person, and most importantly as an engineer under guidance of the notable Professors Bernardo Almada-Lobo, Luís Guimarães and Pedro Amorim, who have continuously challenged and believed in me, not only in FEUP but also in this project. To my good friends (more like comrades) that I got to meet in FEUP, my thanks for the good times, as well as the hardships and long nights we shared in this journey.

All of this was possible because I had the unconditional support of my family, especially my parents and my brother, who made me who I am today. They inspired me in the ways of engineering, and taught me to go as high as I can. Resilience is something that you get by doing, and every problem becomes an opportunity with the right attitude.

To all, I extend my warmest gratitude.

*The only way to deal with an unfree world is to become so absolutely free
that your very existence is an act of rebellion.*

Albert Camus

Contents

1	Introduction	1
1.1	Company description	2
1.2	Motivation	2
1.3	Thesis outline	3
2	Literature review	5
2.1	Location intelligence	5
2.1.1	Origin: from the early days to the present	6
2.1.2	Motivation: economic location theory	8
2.1.3	Data management: geographical information systems	10
2.1.4	Application: spatial decision support systems	11
2.2	Spatial analysis techniques overview	13
2.3	Location intelligence in EVCS placement	15
3	Problem description	17
3.1	Problem statement	17
3.1.1	Available data	18
3.2	Technical characterization	19
3.3	Market characterization	19
3.4	Previous methodology	22
3.4.1	Market Dimensioning	23
3.4.2	Market Capture	23
3.4.3	Charging Equipment Profile	23
3.4.4	Occupancy Rate Simulator	24
4	Methodology	25
4.1	Formulation	25
4.2	Approach	26
4.2.1	Demand disaggregation	27
4.2.2	Demand forecasting	28
4.2.3	Mitigation of disaggregation issues	28
4.2.4	Estimation of sales	29
5	Results and Discussion	31
5.1	Data web scraping	32
5.2	Demand disaggregation	35
5.3	Demand forecast	42
5.4	Sales forecasting	46
5.5	Demand capture	50

6 Conclusion	53
Bibliography	55
A Mapping techniques	61
B Detailed variables	63
C Disaggregation elastic net model	67
D Demand forecasting coefficients	77
E Detailed GBM model results	79
F Gravity model parameters	83

Acronyms and Symbols

ABM	Agent-based modelling
AC	Alternating current
AHP	Analytic hierarchy process
AI	Artificial intelligence
BEV	Battery electric vehicle
CAGR	Compound annual growth rate
CPO	Charging Point Operator
DC	Direct current
DSS	Decision support system
DT	District (Distrito)
ERSE	Entidade Reguladora dos Serviços Energéticos
EVCS	Electric vehicle charging station
EV	Electric vehicle
FR	Parish (Freguesia)
GBRT	Gradient Boosted Regression Trees
GBM	Gradient boosting machine
GDP	Gross domestic product
GIS	Geographic information systems
GPS	Global Positioning System
INE	Instituto Nacional de Estatística
LI	Location intelligence
MAE	Mean absolute error
MAPE	Mean absolute percentage error
MCDM	Multi-criteria decision-making
MN	Municipality (Município)
MOBILE	National Electric Mobility Network
MUAP	Modifiable areal unit problem
MV	Medium voltage
NLV	Normal low voltage
PHEV	Plug-in hybrid electric vehicle
POI	Point-of-interest
PPP	Purchasing power parity
RMSE	Root mean squared error
ROI	Return on investment
RTIEBT	Regras Técnicas de Instalações Eléctricas de Baixa Tensão
SDSS	Spatial decision support system
SEC	Statistical section (Secção estatística)
SLV	Special low voltage
SS	Subsection (Subsecção estatística)
SSE	Sum of squared errors
TSR	Trip success ratio
UGCoP	Uncertain geographic context problem
XGBoost	eXtreme Gradient Boosting

List of Figures

2.1	Hexagonal arrangement of centers. Adapted from Christaller (1933).	7
2.2	Evolution of (a) coverage, (b) operating costs, and (c-d) capital costs. Adapted from Chan (2011).	9
2.3	Steps of the decision-making process. Adapted from Sugumaran and DeGroote (2010)	12
2.4	Components of spatial decision support systems	13
3.1	EV charging plug types. Source: Mostyn (2021)	19
3.2	Number of EVs in Portugal's (a) fleet and (b) new sales. Source: EAFO (2022) .	20
3.3	Public charging (a) sales & charges, and (b) 12-month moving average of sales in Portugal per month. Source: MOBI.E (2022a)	21
3.4	Previously developed methodology	22
4.1	Demand forecasting process	27
5.1	New methodology developed	31
5.2	CPO and MOBI.E data comparison and correlations per EVCS	33
5.3	EVCS locations and scaled sales for 2023 in Porto municipality	34
5.4	Utilization rate (a) distribution and (b) comparison with socket power	34
5.5	Assumed distance decay function for macro-region selection by (a) radius and (b) area	35
5.6	Elastic net r^2 (a) mean and (b) standard deviation with fixed feature selection parameter	39
5.7	Actual sales and predicted demand in Porto municipality for 2023	41
5.8	Forecast model r^2 after C1 and C2, before C3	43
5.9	Average growth rate from 2023 to 2024 in Porto municipality	46
5.10	Acceleration of instantaneous growth rate from 2023 to 2024 in Porto municipality	46
5.11	Average growth rate from 2023 to 2024 in Porto municipality	49
5.12	Predicted EVCS trade area in Porto municipality	52
A.1	Map types by purpose and data type	62
C.1	Elastic net μ and σ of error measures from cross-validation for 1 variable	69
C.2	Elastic net μ and σ of error measures for 10 variables	70
C.3	Elastic net μ and σ of error measures for 30 variables	71
C.4	Elastic net μ and σ of error measures for 100 variable	72
C.5	Elastic net μ and σ of error measures for 200 variable	73
C.6	Elastic net μ and σ of error measures for 330 variable	74
C.7	Predicted and actual sales values on train data	75

C.8	Residuals (a) plot and (b) histogram on train data	75
D.1	Forecast model final (a) second- and (b) first-order coefficients	77
E.1	Correlation between AutoML models predictions	79
E.2	Learning curve of the GBM model	80
E.3	Predicted vs actual plot of the GBM model	81
E.4	Residual plot of the GBM model	81
E.5	Relative variable importance of the GBM model	82

List of Tables

5.1	Comparison of retrieved ¹ and global ² data	32
5.2	Area of the granularity levels	36
5.3	Best elastic net models per n	39
5.4	Group and variables importance	40
5.5	Reconciliation results at DT level	40
5.6	Time series coefficients and r^2 after quality criteria selection	43
5.7	Total forecasted demand for 2024 per level	44
5.8	Model parameters and forecasts for 2024 in specific region	45
5.9	Total forecasted demand for 2024 per level	48
5.10	Most importance variables and groups of GBM model	49
5.11	Huff's gravity model best results for $s = power$	51
5.12	Huff's gravity model best results for $s = autoML$	51
B.1	EVCS variables	63
B.2	Demographic variables	64
B.3	Point-of-interest groups and types	65
C.1	Computational performance experiments of the elastic net model	67
C.2	Tested parameter values for elastic net	67
C.3	Best elastic net models per n	68
E.1	Importance of density variables per level of GBM model	80
E.2	Importance of distance variables per radius of GBM model	82
F.1	Tested parameter values for gravity model	83

Chapter 1

Introduction

In the era of big data, businesses face a mounting challenge to effectively manage and analyse vast quantities of data in order to make informed decisions. This challenge is compounded by the decreasing time available for investment deployment, demanding a transition from traditional decision-making approaches based on experience and intuition to more objective and systematic techniques.

While spatial data presents a wealth of valuable information, it has been underutilized by businesses. Addressing this gap is the emerging field of location intelligence (LI), which provides businesses with the tools necessary to analyse and visualize their data in a spatial context. By leveraging location intelligence, organizations can gain valuable insights that can inform decision-making and drive business growth through faster and more robust location decisions.

Location intelligence can be applied to any business problem with spatial-dependent decisions. In particular, this project focuses on location decision for electric vehicle charging stations (EVCSs) to be installed, whose demand is driven by the rapid growth of electric vehicles (EVs) in recent years. The placement of EVCSs is an important strategic decision for charging point operators (CPO) as it directly impacts their ability to meet the growing demand for EV charging services, as well as for electric utilities companies due to the need for expansion and adaptation of the electric grid. It is a complex problem that requires careful consideration of multiple factors such as increasing demand, rapidly changing supply infrastructure, and where demand is located as well as what attracts demand to certain locations. The process involves analysing various data sources to identify optimal locations that can provide the required charging services while maximizing profits of the infrastructure.

This thesis presents a new methodology for the placement of new EV charging stations for a large CPO in Portugal, which can be integrated in a spatial decision support system (SDSS). It was designed to help the company visualize and monitor operations, as well as make informed decisions about the optimal placement of new EVCSs.

1.1 Company description

This project was undertaken in a renowned utilities company, a large contender in the national public electric vehicle charging market. The company has stated as one of its main objectives to increase its market share in the Portuguese market, as well as to become an international reference in public charging by expanding to neighbouring countries and regions, to which this thesis aims to contribute.

1.2 Motivation

During the United Nations Climate Change Conference in 2016, which marked the first meeting of the Paris Agreement, Portugal pledged to achieve carbon neutrality by 2050. This commitment entails an 85% reduction in emissions compared to 2005 levels (Cristino, 2023; APA, 2019). As of 2020, Portugal has reduced its total emissions by 32.9% and transportation emissions by 25.7% relative to 2005, with transportation accounting for 25.8% of the total emissions (Pereira et al., 2022). Furthermore, in early 2023 the European Parliament formally approved a law mandating the phase-out of fossil fuel vehicles, including a ban on the sale of new diesel and petrol cars by 2035 (Abnett, 2023).

To meet those goals, as well due to the lower fuel cost, electric mobility has recently gained significant momentum, with charging demand in Portugal reaching 36 GWh in 2022, a 113% increase from the previous year, mainly driven by 21.7% of new car registrations being electric vehicles (EVs) (MOBILE, 2022b; EAFO, 2022). Nevertheless, achieving both goals will require a significant increase of electric mobility, driving demand for EV charging infrastructure in the country.

Different stakeholders have different goals in this market. While suppliers seek to maximize profits and establish themselves as relevant players by controlling a significant part of the market, consumers are concerned with assuring convenient servicing of their demand at a low price, and politicians and agencies push to meet ambitious environmental goals. Regardless, these goals converge in the need to scale up the public charging grid.

The profitability of an EVCS is heavily dependent on its location, with the goal being to maximize its sales and return on investment (ROI). To increase profitability the focus should be on locating the supply of EVCS near the demand, as well as other attraction points, known as points-of-interest (POIs), such as workplaces, offices, multi-story buildings, malls, supermarkets, among others.

Intertwined with the previous aspect is the market power that a company holds, which is determined by its market share as well as control over the most profitable locations. By identifying the most lucrative locations that lack coverage, the company can pinpoint profitable areas that have not yet been served, as well as areas that are served but have insufficient installed capacity. As expected, from the business perspective, both profitability and market power are simultaneously considered, either as a trade-off between maximizing profits vs acquiring market share at the

expense of lesser profits, or as an enhancer of one another, given that serving unsatisfied demand or even driving demand with new infrastructure can both minimize cannibalization and maximize profits.

As suppliers quickly react to demand and to their competitors' new EVCSs, the sales forecasts for given locations can be estimated for several years, but must be frequently updated, given that expected sales can change drastically in the face of new supply in that same location.

In order to locate new infrastructure in the best locations, it is paramount to predict where demand is and will be located, and how it is captured by EVCSs accounting for intra- and inter-suppliers cannibalization. Robust and accurate location decisions enable the maximization of profit and ROI, driving grid coverage and competitiveness in the market, ultimately resulting in superior customer service quality.

1.3 Thesis outline

The thesis is organized as follows. Chapter 2 presents a review of location intelligence as well as methods for placement of EV charging stations. Chapter 3 formalizes the problem statement, while characterizing the market, and a description of the previous methodology. Chapter 4 describes the conceptual methodology developed to solve general location problems. Chapter 5 presents the specific modules implemented, their results and respective discussion. Finally, Chapter 6 provides a conclusion and recommendations for future research.

Chapter 2

Literature review

While the body of knowledge on location intelligence is still undergoing significant development, it is the modern-day descendant of established and mature research areas, namely Location Theory, Spatial analysis, Geographical Information Systems, and more recently Spatial Decision Support Systems. Its primary focus lies in extracting valuable business and economical insights from spatiotemporal data, marking a significant shift from the historical perspective of urban planning, and environmental, physical, and ecological sciences, as pointed by Sugumaran and DeGroote (2010). Therefore, a review on the predecessors and components of location intelligence was performed, aiming to understand the need for this new research area, as well as its current shortcomings that this work aims to address.

This literature looks to answer the following questions: (1) How did location intelligence come to be? (2) What are the current general-purpose location intelligence techniques used? (3) What are the currently used location intelligence approaches to the specific problem of EVCS placement?

These questions directly map to the structure of the literature review, which focuses on the methods and results of recent and high-quality quantitative papers and books in the area. In the last paragraph of each question addressed, the main insights are summarized.

2.1 Location intelligence

Location intelligence (LI) is the practice of collecting, enriching, and analysing information, geo-referenced data, and geospatial information to discover contextual insights for location-based challenges or opportunities, including visualizing it on a map and identifying relationships between objects based on their relative placement (Wooler, 2020). An important part of LI is human geography, which explores the relationships between human behaviour and our environment, such as economic geography and urbanism (Steiner, 2023). The cornerstone of LI and spatial analysis lies in Tobler's first law of geography (Miller, 2004):

“Everything is related to everything else, but near things are more related than distant things”, Tobler (1970)

It serves as the main assumption for techniques based on the friction of distance and distance decay, meaning that the magnitude of any factor's influence is weighted by distance, namely its inverse or similar functions. If interrelation between entities increases with proximity in the real world, then representation in geographic space and assessment using spatial analysis techniques is appropriate (Tobler, 1965; Kemp, 2008; Marsh et al., 2008).

In order to understand what location intelligence is and how it came to be, this section starts with the historical origin of spatial analysis, followed by a systematization of the most relevant theoretical concepts and data management systems, and finally the tools used to translate information into business insights for decision-makers.

2.1.1 Origin: from the early days to the present

For as long as the human species has existed, the choice for the location of an activity has been relevant, from where to place a dwell, to hunting and natural resource extraction. From the location of cities regarding bodies of water, to planning aqueducts, roads and buildings, as well as map-based property register to levy taxes of the Roman Empire, these all show the importance of spatial data especially for governments and ruling entities (Church and Murray, 2009; Sugumaran and DeGroote, 2010).

Location intelligence emerged as a means to bring together different areas of knowledge that evolved independently and during distinct historical periods. Its objective is to merge their methodologies, strengths, and applications to build a fresh knowledge base that can effectively tackle specific location-related issues in a cohesive and unified manner. The main four fields that precede location intelligence are, in chronological order: cartography/geography, location science, operations research, and geographical information systems.

Although cartography and geography encompass much more than just maps, these can be regarded as the foundational element that enabled the visual depiction of spatial and location information. The use of maps as we know them can be traced back to as early as between the Bronze Age, with the Saint-Bélec slab from around 2000 BC, and the Iron Age, with the Bedolina Map from around 500 BC, with representations of respectively the Odet valley, France and cultivation fields in Bedolina, Italy (Nicolas et al., 2021; Bicknell, 1913). Geographers have long relied on maps to study the patterns of where objects are, how they have changed and how they came to be. Perhaps the most famous early use of maps from a descriptive approach and widely recognized as one of the first landmarks on spatial analysis, is the map drawn by Snow (1855), featuring the city of London with dots as the infection cases of the 1854 Broad Street cholera outbreak, which allowed him to identify the water pump responsible for the outbreak.

It is often considered that location science provides the theoretical location models, while operations research offers the practical methods for solving them. Location science initially dealt with simple models and geometric problems, such as the geometric median introduced by Fermat in the 1600s for three points, later broadened by Weber and Friedrich (1929) to include varying costs per unit of distance for each point in the single-source Weber problem. The multiple-source variant of the problem is particularly relevant for selecting multiple locations. In an attempt to

explain the intra- and inter- organization of cities in location science, several models have been developed over time, such as the central place theory proposed by geographer Christaller (1933). His theory affirms that the arrangement and size of centres, i.e. population agglomerates, depends on the type of goods and services that it provides, proposing the hexagonal arrangement of centres in Figure 2.1, where bigger centres offer all the goods available in lower-ordered centres. Operations research has evolved considerably since the 1950s, with application to business, industry, and society especially driven by technological advances. It has its focus on modelling real-world problems into mathematical problems and solving them as accurately and quickly as possible (Church and Murray, 2009).

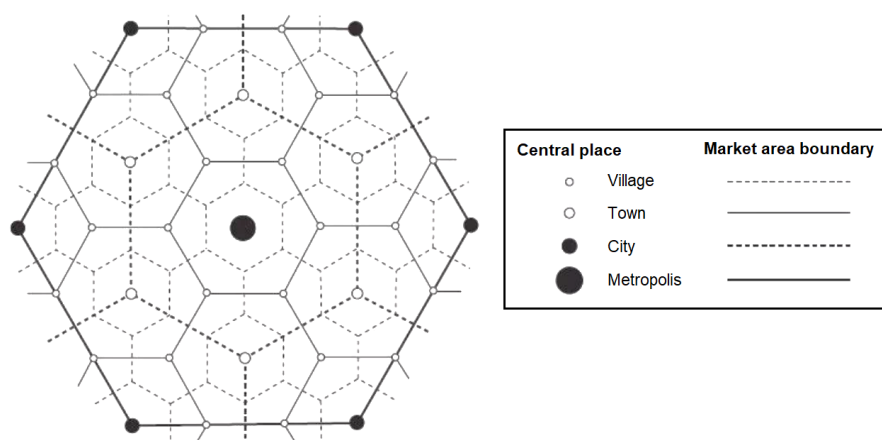


Figure 2.1: Hexagonal arrangement of centers. Adapted from Christaller (1933).

The last piece of the location intelligence puzzle is its technological and computational tool, geographic information systems (GIS). From the use of the first spatial statistics software programs such as SPSS (SPSS Inc., Chicago, USA) and SAS (SAS Institute, Cary, USA) in the 1980s, and the introduction of proper GIS and GPS that enabled widespread access to spacial data manipulation software in the 1990s, especially motivated by governments and local authorities to assist in mapping public resources, infrastructures, and urban planning. Since the 2000s, spatial analysis has gone through an era of big data, in most part fuelled by the internet. As stated by Gao (2022), the democratization of access to location data with OpenStreetMap (OpenStreetMap Foundation, London, UK), as well as the improvements in GIS software programs in terms of usability and providing insights to both researchers and businesses have driven the current trends in spatial analysis, translating in the combination of massive public datasets with business-specific data.

It is important to note that while spatial analysis concerns any formal technique that studies entities based on their topological, geometric, or geographic properties, location analysis is a particular case of spatial analysis as it encompasses techniques used to analyse and determine the location of existing and new infrastructure, respectively (Gao, 2022; Thrift and Kitchin, 2009).

2.1.2 Motivation: economic location theory

Location is a critical factor that affects the success of businesses and the well-being of individuals and communities. Through an economic perspective, decision-makers can strike a balance between efficiency, equity, and externalities associated with location decision. To do so, Chan (2011) identified the following determinants of location to be incorporated in decisions:

- **Technological factors** refer to the physical characteristics that make the functioning of the analysed infrastructure possible. This includes infrastructure (e.g. transportation network), environmental aspects (e.g. Lagrangian points of equilibrium for stable satellite placement), and external aspects such as technological characteristics and limitations of other technologies (e.g. service stations with regular distance to serve vehicles on the highway).
- **Economic and geographic factors** broadly refer to consumers' economic behaviour, in particular willingness to pay (shaping the organisation of cities as a function of accessibility relative to its financial centre and transportation systems), closeness to areas of higher economic activity (cities are usually located near trade-routes), and demand for products and services (leading infrastructure to locate near consumers).
- **Political factors** refer to the zoning of land (dividing land into portions for property delimitation), jurisdictions (e.g. areas with special economic laws), and political decisions (e.g. the location of public infrastructure).
- **Social factors** refer to sociological factors that determine the interactions between individuals and groups, and is observable by the congregation in communities and the reservation (explicit or not) of certain land to selected groups.

When designing an optimized system of supply points that can cater to geographically dispersed demand, it is important to understand the general dynamics of the cost-benefit trade-off (materialized in the size-coverage trade-off) in the sizing of the system for short- and long-term optimization. To illustrate, the model from Chan (2011) is used, and represented in Figure 2.2. It entails a zero-price homogeneous service, with transportation cost. The coverage V depends on the size of the system (p facilities of size P), assuming that the location of the facilities is optimized. The coverage is limited to a given population V^* , as well as the size of the system in the form of a cost $C_t = C_s + C_o$, composed of capital cost $C_s(P, p)$ and operating cost $C_o(V)$, globally limited to budget b_U .

In the short-run, coverage behaves like a sigmoid function of system size, tending asymptotically to V^* . Operating costs are assumed to behave as variable cost of service. Capital marginal cost are expected to decrease with scale at the start as economies of scale take action, but eventually increase due to diseconomies of scale. As for the increase in the number of facilities, the capital marginal cost decreases since the same facility design can be replicated, reducing per-unit fixed costs. In the long-run, however, the market changes and competitors also make decisions.

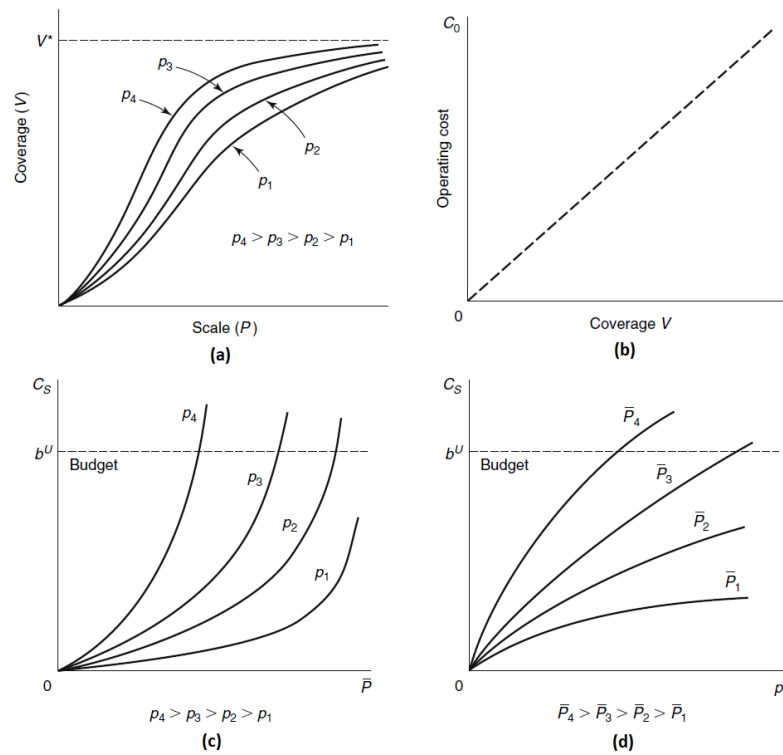


Figure 2.2: Evolution of (a) coverage, (b) operating costs, and (c-d) capital costs. Adapted from Chan (2011).

Therefore, optimizing in the short-run can lead to significant losses in the long-run. As can be seen, this and other challenges in the short- and long-term can hamper location analysis:

- **Pre-existing infrastructure** Both the decision-maker's and competitors' already existing infrastructure must be considered, in order to increase return on investments (past and new) in the short-term based on past information.
- **Market changes** What is a good decision today can be a bad solution in the future if market changes are not considered. This issue can be tackled with forecasts and more sophisticated planning techniques, capable of devising plans not only in space but also in time, so as to not compromise future investments with present decisions.
- **Competition** Most infrastructure location decisions are made in a competitive market. Therefore, it is important to understand how competitors act and react to predict their behaviour, since it means having the upper hand in the market.
- **Information and time** Decision-makers have to deal with imperfect information, which can either be missing or faulty, as well as balance assumptions and forecasts, given that a strategy is only as sound as its assumptions and forecasts. Another issue is the time from the decision to deployment. The mitigation of both issues enhances more solid strategic location decisions and quicker deployment.

To face these issues, a policy of suboptimizing may be less costly than an attempt at complete optimization, especially if this optimization is based only in past information. It is important to devise plans not only in space but also in time, based on trustworthy information and forecasts, accounting for competitors' future actions.

2.1.3 Data management: geographical information systems

A geographical information system (GIS) operates much like a relational database enhanced for spatial data. It consists of integrated procedures, hardware, and software to acquire, manage, manipulate, analyse, and display spatial data represented in multiple layers. The digital representation of the real world in GIS implies a set of digital representation approximations to improve data storage and computational efficiency: (1) the Earth's surface has a known regular shape (usually sphere); (2) spatial data is digitally represented in either the raster data model (grid-like), or in the vector data model (objects with attributes); (3) there is selective inclusion of information, meaning relevant layers or objects might be omitted, usually due to time, budget, or knowledge constraints. The components of GIS according to Church and Murray (2009) and Sugumaran and DeGroote (2010) are presented by order in the data management process.

Data acquisition If spatial data is not in a format compatible with GIS, it must be converted. Digital maps in image files can either be scanned to raster format pixel by pixel, or digitized to vector format by tracing the objects (points, lines, and polygons) manually or through automated image processing. Geocoding is employed to derive latitude and longitude from addresses.

Data management A coordinate system is a method of representing space in two dimensions by establishing an origin point and measuring relative positions in two directions from that point, and is used as the absolute positioning method. Given the added complexity of dealing with spatial data for database management systems, only two data models are fundamentally used in GIS.

The raster data model represents space as a set of grid cells that fill the entire space in a process known as tessellation. Each cell typically has a regular shape, usually but not limited to squares. Despite the large volume of data generated by an attribute value for each cell, the regularity of the grid allows for efficient storage by only storing the values and the grid configuration. Overall, the raster data model has a simple data structure that is easy to understand and is particularly useful for phenomena that vary continuously across space, especially in land cover mapping over a large area.

The vector data model is capable of representing various types of spatial objects, including points, lines, and area/polygons, that may not occupy the entire space. Since each object is exactly defined by its vertices, it is a relatively complex data structure and hard for computational processing. Vector data is usually low volume and its format is well suited for discrete features, but requires positional data for all objects. It can be extremely useful when a given

region has multiple attributes, as well as if there is an underlying geographic hierarchical structure, facilitating its aggregation.

Data manipulation Besides storing, GIS also allows for some manipulation operations, particularly coordinate system conversion, rubber-sheeting (translation and scale correction), and projecting data onto the Earth's surface. Moreover, GIS software can aggregate data by changing raster data cell size and merging vector data polygons. It also enables the overlay of different data layers, allowing for logical or arithmetic operations to be performed on raster data using map algebra, and topological overlay to be used on vector data (e.g. union, intersect). Interpolation stands out as the most diverse manipulation feature, given the range of methods with varying degrees of complexity employed to estimate attribute values in areas where information is not directly accessible.

Data analysis Despite being rather simple and limited, GIS is capable of performing some forms of analysis. It is possible to perform queries in SQL, with the typical SQL commands and added functions for spatial queries such as AREA, INTERSECTS, TOUCHES, WITHIN, CONTAIN, among others. It can also return proximity measures, from simply the distance (global or planar) to other topological relationships such as adjacency, contiguity, intersection, connectivity, and buffers. The centre of objects can also be determined, in particular the centroid of polygons and lines. The service zone of an object can be estimated by proximal area (similar to Voronoi diagram), catchment area, or trade area.

Data display Perhaps the most relevant part of GIS, the visualization of spatial data and spatial model outputs leverages the instinctive capability of people to read maps, given their regular use in daily affairs. It is a powerful tool, which is usually self-explanatory, and therefore crucial for insightful decision-making. There are a variety of map classifications, which were consolidated and systematized by purpose and data type in Figure A.1, as a review of Kraak and Ormeling (2010), Dent et al. (2009), and Longley (2015).

2.1.4 Application: spatial decision support systems

Inspired in decision theory, it is possible to address these problems in three ways, as a systematization of Church and Murray (2009) and Gao (2022). The first approach consists of simply mapping the spatiotemporal data by locational coordinates. This can be called the mapping approach, and it can be useful in navigation, visualization, and geographic calculations (distances and areas), with the possible representation systematized in Appendix A. It mainly relies on GIS to collect, store, retrieve, map and analyse spatial data, allowing for the management of multiple layers of data.

The second approach is to describe arrangements that have emerged and explain decisions that were made. This is the descriptive approach, as it aims at finding patterns that explain the location of a given phenomenon that occurred or a decision made. This is where location analysis, spatial analysis, and location theory play a major role.

The third approach is to identify the best location(s) for an activity. This is the prescriptive approach, where the goal is to guide and optimize the decision made in the present and the future.

In fact, spatial decision-making is a complex process due to the involvement of multiple stakeholders with differing goals, the multidimensionality of the problems and conflicting criteria, and the presence of uncertainty, conflicts, and errors. The structure of decision problems depends on the inputs required from the decision-maker as well as the level of automatization of the process. While structured problems are usually low level repetitive and routine problems where the decision-maker can follow a definite procedure for handling them efficiently based on hard data (reported facts, quantitative estimates, systematic surveys), unstructured problems often come as high level unique problems that require the decision-maker's judgement, evaluation, and insights for the problem definition. Unstructured problems are harder to address due to lack of knowledge about the specific problem, a larger search space, as well as non-quantifiable data in the form of soft data (opinions, priorities, preferences).

Most spatial decision problems are semi-structured, given that only a part of the problem has a clear-cut answer provided by an accepted procedure, and therefore require a multidisciplinary scientific approach to find the most appropriate solution. Sugumaran and DeGroote (2010) systematized the spatial decision-making process in a set of steps represented schematically in Figure 2.3. These steps usually overlap, especially the design and evaluation of scenarios which, through an iterative process, converge to a final selection.

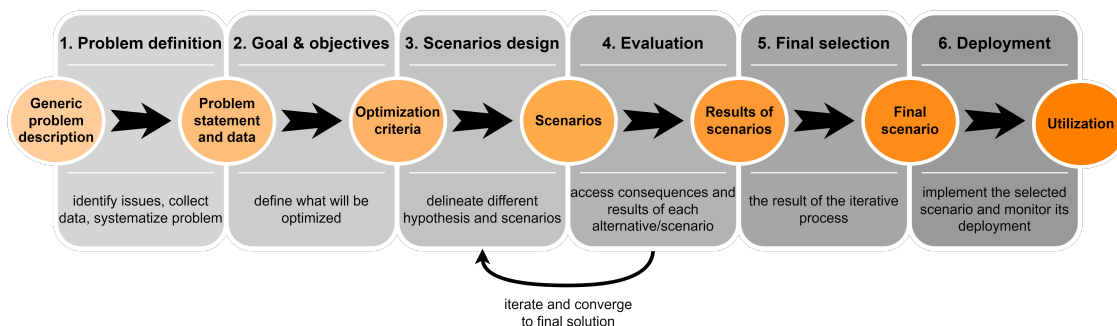


Figure 2.3: Steps of the decision-making process. Adapted from Sugumaran and DeGroote (2010)

A spatial decision support system (SDSS) is an interactive, computer-based system designed to assist decision-makers in solving semi-structured spatial problems, according to Sugumaran and DeGroote (2010). It combines the database management and spatial data manipulation features of GIS with the analytical tools and models of decision support systems (DSS) to create a flexible and customizable tool to accommodate for stakeholder's preferences and restrictions. As presented in Figure 2.4, it incorporates spatial and non-spatial database management and analysis, analytical modelling capabilities, decision support utilities, scenario analysis and evaluation, multi-criteria evaluation techniques, sensitivity analyses, visualization and data presentation utilities, and report generation capabilities in an interactive and recursive problem-solving environment, with graphical user interfaces.

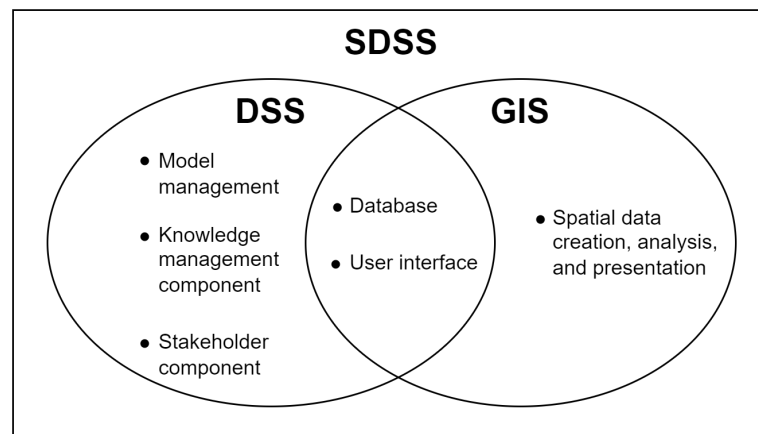


Figure 2.4: Components of spatial decision support systems

While GIS saw its development mainly due to advances in hardware and theoretical spatial sciences, DSS development has been driven by artificial intelligence (AI) applied to knowledge-based and intelligent DSS, as well as its migration to the web. According to Sugumaran and DeGroote (2010), the core drivers for the development of SDSS were the advancements in information and communication technology, the increase in spatial data availability, the development of spatially explicit modelling techniques, the broadening of applications, and wider user-/developer-base and expert domain knowledge. Despite the increase in the range of SDSS applications, the vast majority of research is still in visualization and monitoring for government agencies and the public sector in what can be broadly referred to as land use and urban planning, as stated in the review from Sugumaran and DeGroote (2010). Current applications range from natural resources management and environmental monitoring to agriculture and design of public services. Therefore, it can be said that the appearance of location intelligence as a separate field from GIS and SDSS is a reaction to the path these two areas have followed, as it looks to apply their knowledge-base to the private sector and business needs.

2.2 Spatial analysis techniques overview

The used spatial analysis techniques greatly depend on the problem at hands, given the very different possible problems, objectives, and data. Nonetheless, their focus on the common spatial component carries a series of possible generic techniques that, according to Sugumaran and DeGroote (2010), are used in most location intelligence applications, namely:

Map algebra Applies operators to input attribute layer(s) to produce a new output layer, such as arithmetic, relational, boolean, combinatorial, logical, accumulative, assignment, among other operators. These can be applied in four ways: for each individual cell using its values from multiple layers (local), for each cell using values from itself and neighbouring cells on a single layer (focal), for a group of cells with the same category with one or more

layers (zonal), and summarizing the grid with global operators (global). This also includes weighted linear combination, which combines multiple layers by assigning a weight to each.

Multi-criteria decision-making (MCDM) Aims at balancing conflicting criteria. A particular example is analytic hierarchy process (AHP), which uses pairwise comparison between options and influencing factors to yield the weight of each factor and the preference for each option. The quality of the evaluations of the decision-maker can be assessed with the consistency ratio.

Forecasting and prediction Enables the use of statistics and machine learning methods for making predictions or explain independent variables. For example, time-series analysis can be used for forecasting, while machine learning can be used for predictions on large datasets, with its major drawback being that the models are not explainable or understandable by humans.

Mathematical optimization Refers to a broad range of techniques to solve optimization problems. The algorithms can range from exact such as the simplex algorithm for small linear problems, to branch-and-bound for mixed-integer linear problems, and iterative and meta-heuristics for good solutions in useful time such as genetic algorithms and simulated annealing.

Simulation Applied to represent the evolution over time of a model based on a complex real-world process or system with nonlinear behaviour (Banks, 2001). Simulation models differ in the level of uncertainty, namely deterministic and stochastic simulation. There are three main simulation modelling approaches, which can range from a high level of abstraction in the form of system dynamics such as cellular automata, to a low level of abstraction in discrete event modelling aimed at standardized processes, with agent-based modelling (ABM) in the middle being the most widely used approach, as mentioned by Chan (2011). ABM involves providing agents with a set of individual action and interaction rules, along with the possibility to learn. The outcome of these individual rules is then analysed on a global or collective level.

Fuzzy modelling techniques Instead of an element belonging to a set being a binary attribute, fuzzy logic greys out this degree of membership to between 0 and 1. This is particularly useful in spatial analysis to avoid unjustified hard discontinuities in space, such as defining a transition zone for suitable-unsuitable land parcels.

2.3 Location intelligence in EVCS placement

The placement of EVCSs has garnered considerable attention from the scientific community and companies alike, owing to its contribution towards sustainable mobility as well as lucrative investment opportunities. Despite being a relatively new research topic, established models for demand forecasting and infrastructure placement have been repurposed for locating EV charging infrastructure, leading to considerable recent literature.

Flow and traffic The most prevalent road-centric approach corresponds to realistic traffic agent-based modelling and simulations. These methods integrate forecasting and capture of demand by modelling users' individual charging behaviour, in the context of their commuting activities or trips, as in the works of Pan et al. (2020) at Grigorev et al. (2021). Stochastic demand has also been modelled, as by Wu and Sioshansi (2017). Users incur a penalty for deviating from their intended path to charge their EV, therefore local search algorithms are typically employed to minimize total deviation. However, the goal of pinpointing users in space and time can fail under faulty modelling of user behaviour, such as unwillingness to stop mid-trip for charging or the routing algorithm used.

Demand forecasting In contrast to the realistic traffic models, spatial data and indirect predictors can also be used to forecast demand. To do so, machine learning methods such as Gradient Boosted Regression Trees (GBRT), eXtreme Gradient Boosting (XGBoost), and gradient boosting machines (GBM) can be trained on socio-demographic data, including working population density, mobility indexes, and points-of-interest, in order to predict demand and vehicle flows, as in the works of Chen et al. (2022a), Luo et al. (2021) and Dong et al. (2019).

Optimization models Classical optimization models rely on strong assumptions that simplify reality instead of replicating it, resulting in more understandable models that are faster to optimize. Numerous formulations and solving methods have been proposed, as in the reviews by Ahmad et al. (2022) and Kizhakkan et al. (2019).

Coverage models excel in simplicity and “fairness”, as it places hard constraints on the maximum distance between demand and supply. These ensure full coverage at minimum cost or maximum coverage for limited budget (or their mix), yielding visualizations similar to Voronoi diagrams, where each EVCS serves a specific area as in the works of Dong et al. (2019) and Niccolai et al. (2021). Pan et al. (2020) and Bi et al. (2017) also applied coverage models on traffic models, by capturing trips instead of demand points and polygons.

As demand (or arrival rate) in an individual EVCS increases, so does waiting time. Queuing theory models have been used by Grigorev et al. (2021) and Wang et al. (2022), seeking to model this relationship, in order to minimize average waiting time. User equilibrium models take the extra step of modelling user behaviour by considering the distance and waiting time at each EVCS for the user, leading the system to a state of equilibrium where no user is

willing to switch EVCS given that all provide the same benefit (Huang and Kockelman, 2020).

Social cost models seek to minimize economic (operation and investment) and environmental (electricity consumption and CO_2 emissions) costs, which Zhou et al. (2022) and Ren et al. (2019) solved with genetic algorithms.

Qualitative data Researchers have also explored the subjective nature of users' charging behaviour, as discussed in the review by Hardman et al. (2018). For instance, Liu et al. (2019) applied multi-criteria decision-making models on categorical/ordinal data converted to numerical preferences using grey theory and fuzzy logic.

Research on EV charging infrastructure planning has mainly focused on two distinct aspects: forecasting demand and optimally locating charging stations. However, there is a lack of research that integrates these two critical elements, a gap that has presented a significant challenge as both are inherently interconnected. While forecasting future demand enables infrastructure planning to accommodate growing demand, demand capture is crucial in ensuring that infrastructure is located optimally in the short term and accounting for existing EVCSs.

Chapter 3

Problem description

This chapter will begin by presenting the problem statement along with the available data. Next, a short technical and market contextualization is performed, already with some insights regarding the market forecast. Finally, the previous methodology is described as it was at the moment this work was initiated, followed by a discussion of its limitations that this work intends to address.

3.1 Problem statement

The problem at hand is to determine the optimal locations for EVCSs in order to maximize profitability. The goal is to identify areas where there is high demand for EV charging services but limited supply, and to prioritize these areas for the placement of EVCSs. This requires accurate forecasting of demand for EV charging with sufficient granularity, as well as considering where supply is located and understanding how EVCSs capture demand. Ultimately, the aim is to provide convenient EV charging options while ensuring the financial viability of the EVCSs.

The primary focus of the project team in the previous methodology was on predicting the sales and utilization rate of each existing EVCS, therefore requiring precise prediction of demand for small time intervals. Given the considerable differences in market settings of motorway and street/parking lot chargers, the team used different approaches for each. On the other hand, the new goal explored in this work is to identify new profitable areas, especially for new infrastructure rather than the existing one. Therefore, it is of higher interest to estimate demand, and not as much to predict hourly utilization and quit rates.

Street/parking lot EVCSs were considered, with motorway excluded. This is due to the straightforward relationship of socio-demographic variables with street/parking lot charging demand than with motorway charging, where demand can originate from more distant locations. These chargers account for about 3/4 of the CPO's sales.

The decision rules that competitors use to place their new EVCSs are unknown, and so are their future locations. Therefore, one can only account for the current EVCSs, and eventually try to predict, based on historical data, where competitors will place their new EVCSs.

As one can only know the current locations of EVCSs and not future, there are two main issues to address in order to solve this problem: (1) predict demand with sufficient granularity, and (2) model how EVCSs capture demand. By solving these two issues, it is possible to identify current and future best areas for locating new infrastructure, as well as simulate different scenarios with new EVCSs, both from the company itself and its competitors.

3.1.1 Available data

The data available originates both from public and private sources, with an effort to replace private with public data. All the data collection was done until end of June 2023. Portugal is divided in five administrative levels: districts (DT), municipalities (MN), parishes (FR), statistical sections (SEC), and subsections (SS). Tobler (1999, 2004) second law of geography states that “*The phenomenon external to a geographic area of interest affects what goes on inside*”, resulting in the loss of information by aggregation, artificial spatial discontinuities, and differing legislation due to the creation of arbitrary boundaries (Henley, 1981; Haining, 1990). These are known as the modifiable areal unit problem (MUAP) and the uncertain geographic context problem (UGCoP) and serve to remind that administrative divisions should not be considered as internally homogeneous and externally discontinuous (Kwan, 2012; Johnston and Semple, 1983; Chen et al., 2022b). The variables used are detailed in Appendix B, and are organized as follows:

- **Demographic** public census data from INE (2022a), which includes data on population (sex, age, education, occupation), buildings (residential, housing units, storeys, construction year), and households & housing units (familiar, primary and second home, accessibility, parking, owned by residents or rented, members per household, families with children).
- **Vehicles** public data from INE (2022b), namely the number of vehicles sold per thousand inhabitants, per municipality per year.
- **Economic** public data of purchasing power parity (PPP) every two years from INE (2019), relative to the national average by municipality.
- **EVCSs** public data available by MOBILE (2022a), and includes the **location** with latitude and longitude, and type of parking (street/parking lot vs motorway); their **specifications** including the CPO, format and type of the plugs, grid tension level, maximum voltage/ampere/power; the **status** of all EVCSs each minute from 2023/06/01 to 2023/06/14, which can in turn be used to estimate real sales in quantity and value; and the actual **charges** of one CPO (representing around 15% of global sales) from January 2019 to October 2022.
- **Points-of-interest** data from OpenStreetMap retrieved using Overpass turbo (OpenStreetMap contributors, 2023) for querying. It included nodes, ways (lines), and relations (polygons). The location used for each of these three corresponded to the first node point location of each element. Each POI has a type, and a group, as can be seen in Table B.3.

3.2 Technical characterization

In Europe, the standard plugs are the Mennekes (IEC 62196 T2) for AC charging and CCS2 (IEC 62196 T2 COMBO) for DC charging, also referred to just CCS (Bakker et al., 2015), presented in Figure 3.1. In Portugal, the CHAdeMO plug is also used for DC charging. Given the higher maximum power output of DC chargers, these usually come with cable (99%), while the lower power AC chargers usually have a socket (83%) that require the user to use his cable. AC chargers are either supplied by single-phase (23%) or three-phase (77%) AC power.

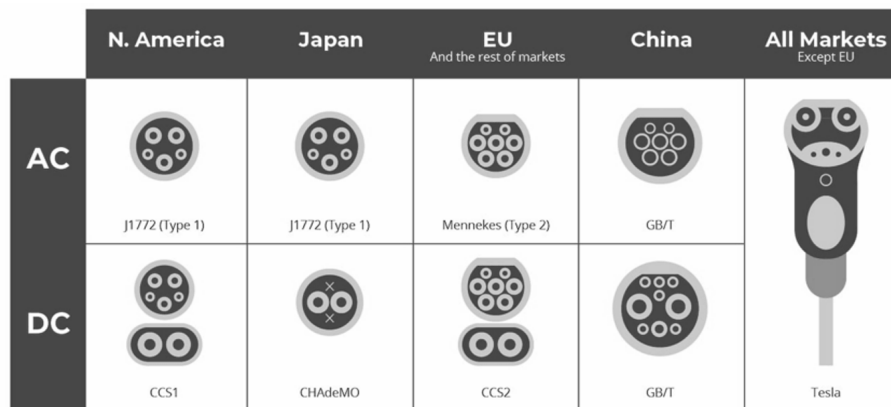


Figure 3.1: EV charging plug types. Source: Mostyn (2021)

Charging stations can be supplied by three tension levels, defined by the regulatory authority ERSE. Normal low voltage (NLV) supports up to 41.4kVA and does not require prior assessment of connection conditions. Special low voltage (SLV) supports power above 41.4kVA and requires prior evaluation of the low voltage grid connection conditions according to RTIEBT under Portuguese law. Medium voltage (MV) is destined for companies with power needs above 200 kVA and requires prior evaluation of grid connection conditions (E-REDES, 2020a,b). SLV and MV are equally used for AC (56%) and DC charging (44%), while NLV is mostly used for AC charging (98%) given its lower maximum power output.

Regarding the charging speed, when the EV is connected to a charger, the charger matches the battery voltage, and will maximize the amperage. Therefore, the actual charging power might be limited by the charger's maximum amperage or maximum power. Besides, the temperature can also highly impact charging speed, especially below 0°C (Yong et al., 2015; Ma et al., 2018; Denton, 2020).

3.3 Market characterization

As of the end of 2022, EVs account for 2.2% of the total national fleet of vehicles and 21.7% of new car sales, with the majority being BEVs as can be seen in Figure 3.2. This meant an increase in the number of EVs in Portugal from 2021 to 2022 of 37%, an increase that has remained fairly stable since 2020 with a CAGR of 40%.

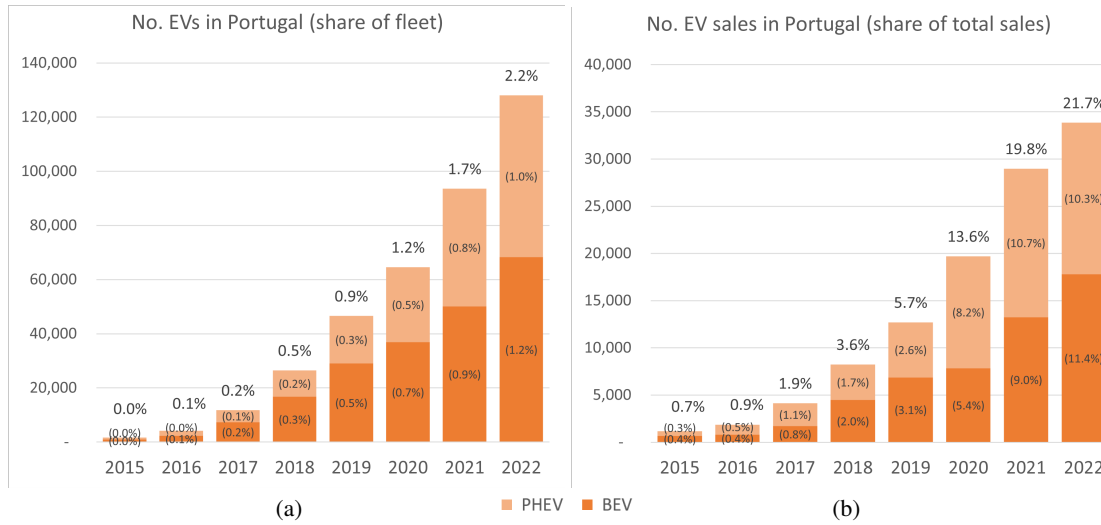


Figure 3.2: Number of EVs in Portugal's (a) fleet and (b) new sales. Source: EAFO (2022)

As for public charging sales, these have increased above the growth rate of EVs, as can be seen in Figure 3.3 (a). With an increase relative to the previous year of 78% in 2020, 172% in 2021, and 113% in 2022, total charging in 2022 has reached 35.9GWh. The number of charges closely follows, with a 0.995 correlation with sales in Figure 3.3 (a). It is also interesting to note that the centred 12-month moving average of total charging can be remarkably fitted by a second-degree polynomial regression with a very high coefficient of determination $r^2 = 0.9997$, represented in Figure 3.3 (b). In the same figure, it can also be seen that the monthly growth of the moving average (first derivative) can be modelled as a simple linear regression with an $r^2 = 0.9524$, projecting an increase in the annual growth rate (GWh/year).

These figures match the projections for Europe from Conzade et al. (2022), stating that public charging will have a CAGR of 103% between 2021-2025 and 38% between 2025-2030. As can be seen from both projections for Europe and actual sales in Portugal, the public charging is growing faster than the EV fleet, indicating an increase in importance of public charging. According to the same study, while only 10% of charging sales in 2021 were in public chargers, it is projected to be 34% by 2025, reaching 52% in 2030.

Not only does public charging demand drive supply, supply can also induce growth in demand, as pointed by Yong and Park (2017). If a country has adequate subsidies, the authors state that even if the GDP per capita of the country is low, the existence of charging infrastructure enables higher EV adoption.

The average station tariffs are 0.188€/charge, 0.021€/kWh, and 0.044€/min. Each station has its own prices and charge, minute, and kWh tariff combinations, some being applied in daily schedules according to the time of the day.

The public charging market in Portugal resembles an oligopoly, with 26% of market share belonging to the biggest CPO, and 94% concentrated in the seven biggest CPOs (Rodrigues, 2023). Given that electric charging is a utility, it can be considered a homogeneous product, where the main determinants of consumer preference are the location and power of the chargers.

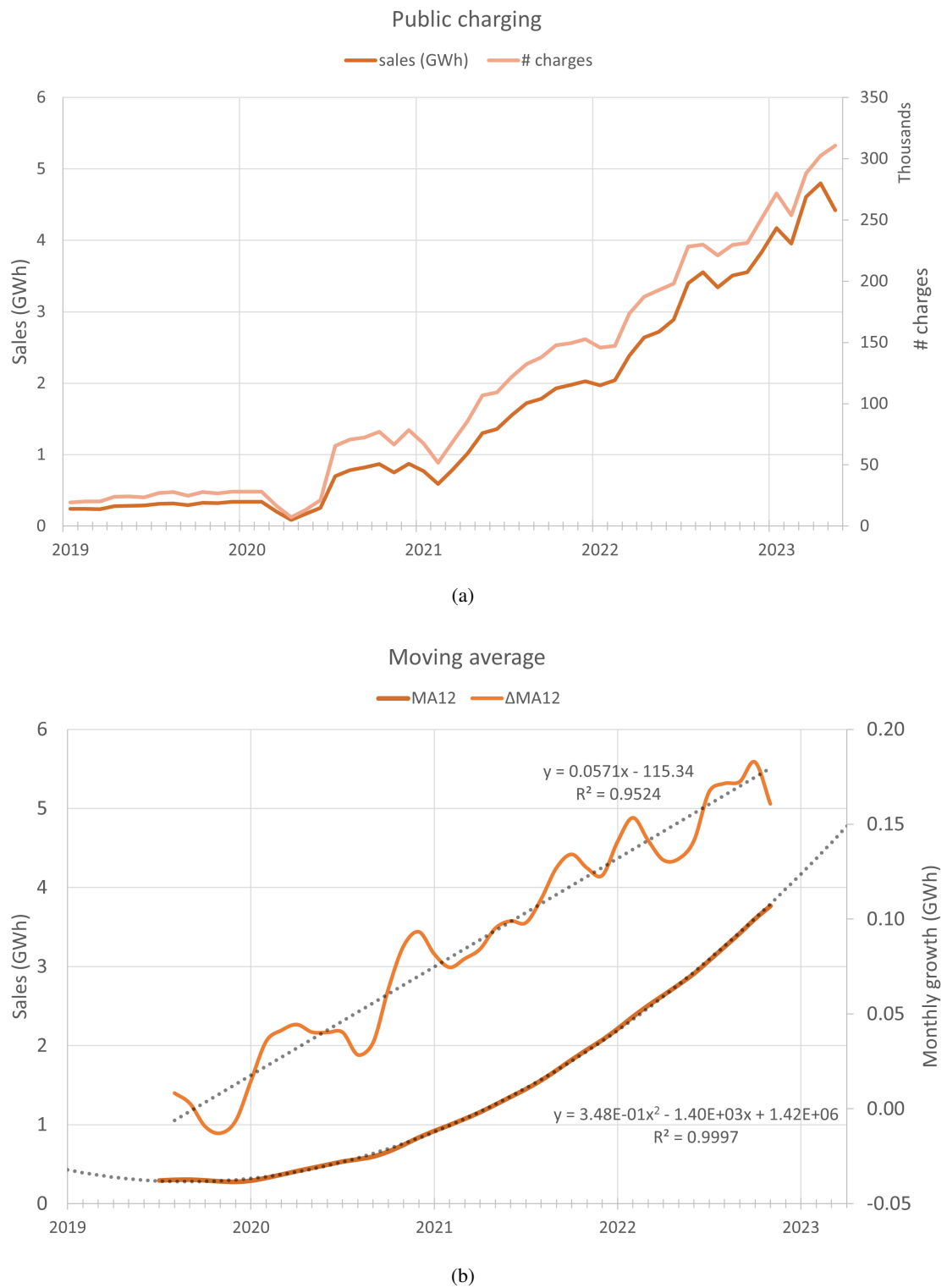


Figure 3.3: Public charging (a) sales & charges, and (b) 12-month moving average of sales in Portugal per month. Source: MOBI.E (2022a)

3.4 Previous methodology

The primal objective was to develop a model capable of aiding decision-makers in locating new EVCSs. For that purpose, the CPO's technological branch had previously developed a model that estimated the utilization rate of each location by using a random forest trained on data related to technical specifications, previous sales, demographic information, and various points of interest in the vicinity. The CPO's business representatives had difficulty in grasping and visualizing the inner working of the model, which was not transparent neither explainable.

In the initial phase of the project, before this work, the project team developed a preliminary methodology that also accounted for competition between EVCSs and used a gradient boosting machine (GBM) to improve accuracy, hereby referred to as the previous methodology. The methodology in this section was developed before this work, and therefore the author of this work did not take part in it. Having said that, it was considered relevant to systematize the previous methodology, in order to explain its shortcomings, as well as how the new methodology tackles them.

While the CPO's model predicted only the utilization rate, the preliminary methodology also predicted the sales. Both models could be employed to test different equipment specifications at discrete locations.

The previous methodology, represented in Figure 3.4, consisted of a four-step method to go from sales of one CPO, to the utilization rate and sales of all EVCSs. It allowed to manually test new EVCS locations and configurations, by predicting their sales and utilization with satisfactory performance.

Although this model outperformed the CPO's model and delivered satisfactory accuracy and precision, it was time-consuming to run, making it suitable to test a few possible locations and scenarios, but not for continuous space problems.

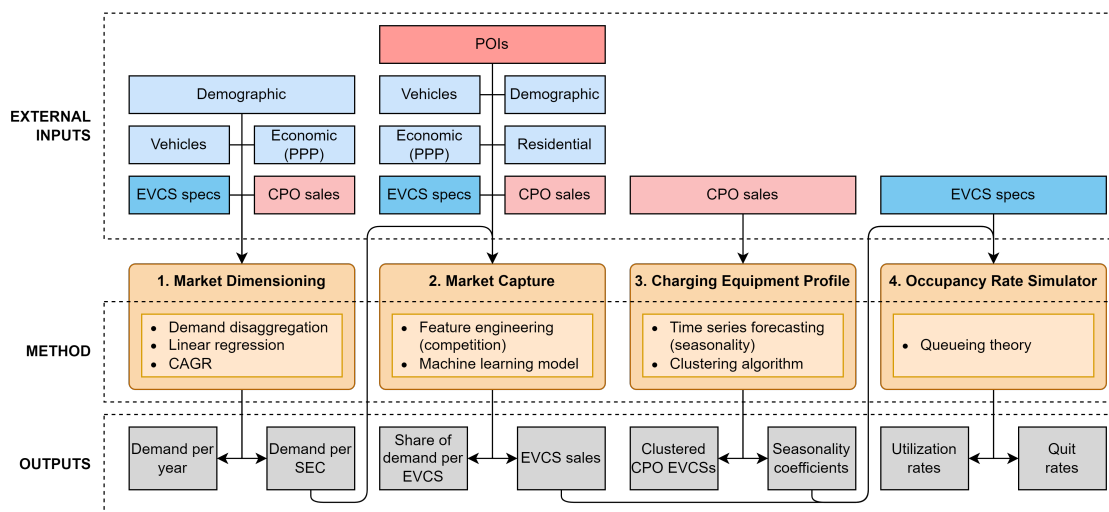


Figure 3.4: Previously developed methodology

3.4.1 Market Dimensioning

The objective of this step was to predict demand at a high-granularity level. To do so, a demand disaggregation model was trained on sociodemographic data at a low-granularity level, therefore accurately locating demand and its quantity. The CPO sales were used to compute average sales per EVCS charging type, therefore estimating the sales of the remaining EVCSs. These were aggregated at MN level, as it was considered safe to assume that at this level demand equals supply. This data was used to train a linear regression model using only three normalized variables: the number of new vehicles sold in 2022 with no importance (INE, 2022b), the PPP in 2019 with 80% importance (INE, 2019), and the active population in 2021 with 20% importance (INE, 2022a). It had a train r^2 of 0.773, and was used to estimate disaggregated demand at SEC level, while future demand forecasts were computed using a projected CAGR, estimated from previous years and from international agencies and reports on the subject.

This step suffers from three main drawbacks. First, it uses data from just one CPO. Second, it assumes the same sales for all EVCSs of the same charging speed. Finally, by assuming the same CAGR for all regions and years, it hindered the capability to differentiate between booming and stagnated regions, and therefore does not suit the new goal of identifying new investment areas.

3.4.2 Market Capture

Using the disaggregated demand as input, the model predicted, for each EVCS, the share of demand within a given radius that it captured. All data was assigned to points (the centroid of polygons and raster cell data), and their sum was computed within several radiuses around the EVCS, including EVCSs and points-of-interest in the vicinity. This data was fed to a gradient boosting machine. It had a cross-validated MAPE of 21.3%. The most important groups of variables corresponded to the equipment specifications, and the building and points-of-interest in its vicinity. The most important variable was the number of eating points or restaurants nearby.

While the feature engineering process was purely spatial, the model was not, given that it was a regression model after all. It also took significant time to make predictions (upwards of 10 minutes), and was only trained on the sales of one CPO.

3.4.3 Charging Equipment Profile

In order to account for seasonality, the seasonality coefficients of each CPO EVCS were computed, and clustered based on these coefficients. The number of clusters and bin width were chosen to maximize the Silhouette score, yielding cycles with length of one year (per month), week (per weekday), and day (in intervals of 6h), with two clusters for each cycle. The year cycle clusters obtained were *Full year* (82% of EVCSs) and *Summer peak*, for the week *Weekday* (72% of EVCSs) and *Weekend peak*, and for the day *Afternoon peak* (83%) and *Morning peak*.

This step presents as relatively robust, capturing the general utilization behaviour during the day by maximizing a clustering quality measure. On the other hand, there is no way to know the cluster to which an EVCS from another CPO belongs, neither for new EVCSs. Nevertheless, the

focus of the present work is not on the utilization or the seasonality patterns of demand, but rather on the overall trend of demand and its location.

3.4.4 Occupancy Rate Simulator

To compute the utilization rate, a queueing system M/M/S (GD/K) was used based on the seasonality coefficients and sales forecasts (from market capture), allowing to obtain utilization for small intervals of 6h. Using the average energy consumption and time per charge, it was possible to obtain the arrival rate and service rate, while the number of servers corresponds to the number of plugs, and the queue length was assumed for each charging type.

The estimation of the utilization rate is not the focus of this work. Despite this, it should be noted that the unmet demand from the quit rate does not impact the demand in surrounding EVCSs, while it is logic that if demand is not served in a given EVCS, it will be served in another, probably nearby. This indeed does not account for competition between the EVCSs.

Chapter 4

Methodology

It is crucial to model the problem at hands mathematically before proceeding to solve it. It is indeed what sets the goal (such as maximizing profit, return on investment, or revenue) and the constraints, highlighting which parameters and functions need to be defined. After that, a general methodology was developed to locate new supply infrastructure. The proposed methodology in this chapter and its validation in the following chapters through implementation and results analysis presents as the scientific contribution of the author in this work. It is therefore an original methodology, inspired by the literature and the previous methodology.

4.1 Formulation

As stated in Chapter 3, the goal is to determine optimal locations for supply points that capture demand, in order to maximize profit. There are essentially two factors in play that define this problem: supply and demand. While demand can be estimated and forecasted based on past demand, the same can not be said for supply.

In an optimal scenario, where the decision rules of competitors were known, the strategy to follow would be to plan the infrastructure to be installed across the years, with a set of locations for each year, that would account for competitors' new infrastructure.

The fast growth of supply infrastructure and the considerable impact that a new supply point has on the sales of its neighbours means that the issue of unknown competitors' decision rules can not be overlooked. In fact, predicting the location and specifications of new supply points from competitors is extremely hard, if not impossible. It indeed implies that a grand plan spanning multiple years would quickly turn obsolete, and therefore decisions must be made on the fly, with yearly (if not more regularly) infrastructure investments that account for the most recent supply network. As such, regarding supply, decisions must be made accounting only for the current infrastructure, and assuming that competitors will not place more supply points.

While predicting new supply points is not feasible, predicting demand is. This means that places where demand is growing faster are more attractive for investment, pointing for the need to account for growing demand. If competitors do not place new supply points in the vicinity of

the locations analysed, it indeed means that it is possible to obtain an estimate of sales across the years, with diminishing accuracy through the years due to new EVCSs placed nearby.

What this effectively means is that supply is static and demand changes with the years. The location of new supply points seeks to maximize profit in the long term, but given that as the years advance predictions are less precise, the weight of predicted sales should decrease with the years. The weight would account for the error in demand prediction but, most importantly, for the uncertainty of new competitor supply points. It can be interpreted as a gradual loss of validity of the assumption that competitors will not place further supply points, creating a gradual decrease in trustworthiness of predictions but at the same time considering growing demand in the next few years. For example, one can predict demand for 10 years, and considering the current network, maximize the sales of new supply points. These sales would be weighted according to the year with some decay function, like the inverse of the year $(year_{plan} - year_{current})^{-1}$.

The formal statement of the problem is a set of demand points $d \in D$ with their sole attribute being the demand amount v_d^y for each year y , and a set of supply points $s \in S$ each with their technical specifications. There is a function that can predict sales of each s as a function of its specifications, the specifications of other supply points, and demand in a given year as $P_s^y = f(s, S, w^y)$. With the decaying weights w^y for the year $y \in \{1, \dots, +\infty\}$, the goal is to maximize sales for a set of new supply points $s' \in S'$ as

$$\max T = \sum_{y=1}^{+\infty} w^y * \sum_{s' \in S'} P_{s'}^y \quad (4.1)$$

It is assumed that the user provides the new supply points (their number and specifications) as input. By experimenting different sets S' , it is possible to obtain different combinations of sales and costs/investment, allowing to adjust the value of new invested infrastructure against its return.

4.2 Approach

To tackle the identified shortcomings in the previous methodology, an extensive review was made on location intelligence, particularly on location models, in order to develop a methodology capable of deriving location decisions and insights from raw spatial data. In particular, forecasting and demand capture models were studied and implemented to solve the specific problem of this project. The final goal is therefore to develop a tool validated with real data that is capable of showing the most profitable areas for investment.

The followed approach to solve this problem focuses on the first and second steps of the previously developed methodology, namely forecasting and capture of demand. The proposed methodology divides the demand forecast in two steps, therefore leading to a new three-step methodology.

The first step, demand disaggregation, uses historical sales data to estimate demand at a high-granularity level, for a given set of years. In the second step, this estimated demand is projected into the future using time series forecasting. Both these steps are performed in hierarchical fashion, in order to reconcile the predicted demand. In the third and final step, demand is captured by the supply points, and their respective sales are predicted.

4.2.1 Demand disaggregation

The demand disaggregation module estimates demand at the lowest granularity level (called micro-region) with sales as the dependent variable, and socio-demographic data (available from census and statistics agencies) as independent variables. Therefore, the granularity of the socio-demographic must be much higher than the approximate area of influence of supply points (usually the case). The steps followed in this module are illustrated in Figure 4.1. Demand does not emanate from the location of the supply point. If only two supply points are considered, it is not known if demand emanates from each supply point, from a point in between, or any other number and location of points. Nevertheless, if a trade area A_{TA} around a supply point is considered so that it is big enough to ignore demand outside its perimeter, it can be assumed that, within that area, demand is equal to sales, since customers are not willing to travel further for the good or service.

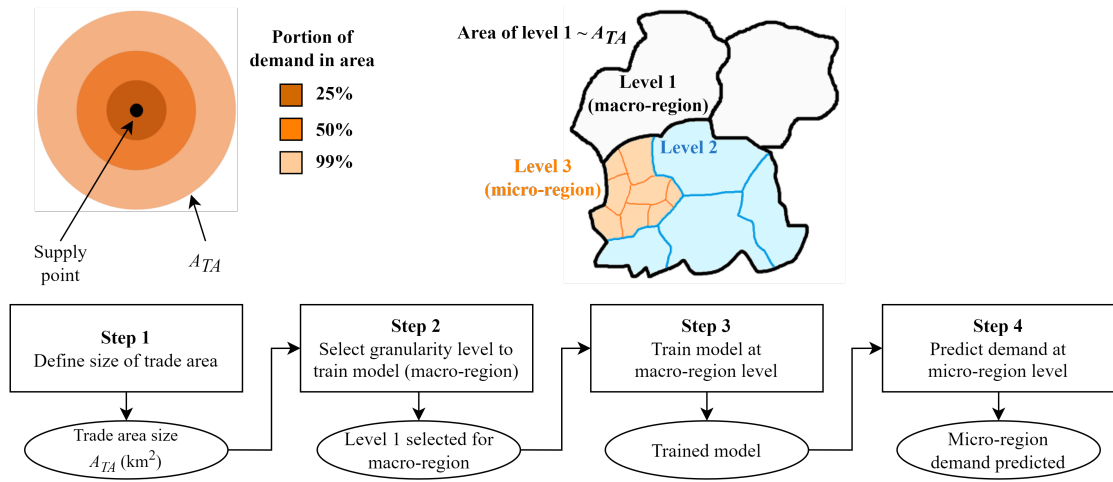


Figure 4.1: Demand forecasting process

In step 1 (Figure 4.1) the trade area can be determined by assuming that demand of a supply point decays with distance from its location. This step need not be exhaustive, as this value is used for magnitude comparison with the available granularity levels. In step 2 (Figure 4.1), the granularity level for model training is selected. It is assumed that granularity levels have approximately uniform areas. The average area size A_x is computed for all granularity levels $x \in X$, and the macro-region level is selected so that $A_{macro} \approx A_{TA}$. The model is trained and evaluated at the macro-region level in step 3 (Figure 4.1). Finally, demand at the micro-region level is estimated in step 4 (Figure 4.1). It should be robust to differences in the magnitude of the independent variables, given that it will be fitted at the macro-region level and used at the micro-region level ($A_{micro} \ll A_{macro}$), so predictions for low values of the independent variables should be verified. The model should also be as simple as possible, given that the demographic phenomena that impact demand at the micro-region level can be substantially different from the macro-region level. This is particularly evident in the fraction of demand from a region that is served outside it, which is ignored at the macro-region level but can be substantial at the micro-region level.

4.2.2 Demand forecasting

The goal is to forecast future demand at the lowest granularity level possible. Having said that, there is usually not enough time series data at that level for good forecasting models.

Each region in each level is considered as a time series, and its values are standardized (such as dividing by the most recent value). Then, a forecast model is fitted, and its optimized parameters are obtained, where these parameters can preferably translate (directly or not) how demand is growing in the region. This is standard procedure, but still does not handle the previously mentioned issue.

There are two key steps in the method for tackling this issue. The first is to set up quality criteria for a time series, that will allow it to be considered suitable. It can include the model fit quality (e.g. r^2), a minimum number of observations for model training, among other criteria. As this step will deem the time series of some regions as unsuitable, there is a need to obtain the parameters. This is where the second step comes in, where the parameters of the upper-level region that it belongs to (parent) are used. Naturally, if its direct parent also does not have parameters, it will search in the next level, and so on. This is the reason for standardizing all time series at all regions and levels, so that the parameters can be pulled from upper levels to lower levels.

This procedure is nothing more than using the parameters of its closest parent if its own time series is unsuitable. The main advantage of this procedure is that, in the worst case, every region at every level will have the same growth rate as the global time series. By expanding parameters to lower levels, there is no risk of forecast instability, only the risk of not capturing rapid changes at this level.

4.2.3 Mitigation of disaggregation issues

There are two main issues that arise from demand disaggregation: the first is that lower independent variable values and a lesser degree of aggregation at the micro-region level, summarized as higher input variability, will lead to instability and high variability of the estimated/forecasted demand (in both modules); the second concerns the mismatch between the global predicted demand and global actual demand, leading to over- or under-estimation of the future market size, ultimately resulting in either profitability losses from oversized infrastructure or lost opportunities from undersized infrastructure.

Regarding the inconsistency in predicting demand at high-granularity levels, as mentioned in step 4 (Figure 4.1) of the demand disaggregation module, data at a micro-region level changes faster than at macro-region level, leading to extremely variable predictions, much like the consequence of complex models with low bias and high variance (Geman et al., 1992). In the case of the demand disaggregation module, the procedure followed to mitigate this issue consists in training the model at the macro-region and predict at micro-region level. On the other hand, time series forecasting models do this inherently when trained and tested on the same data by adjusting the trend smoothing intensity to the variability in the train data, therefore the levels for training and testing should be the same.

As for the mismatch between global predicted and actual demand, this issue affects both the disaggregation and forecasting modules. In particular, take Figure 4.1 as an example, assuming demand is disaggregated to level 3 with a model trained at Level 2. In a perfect demand disaggregation model, actual and predicted total demand at level 1 would be equal. In reality, these differ, so to tackle this issue, reconciliation of aggregated demand is needed. Consider a level 1 region R that contains a set of level 3 regions $r \in R$, disaggregated demand $\hat{d}_r$ in region r , and actual demand D in level 3 region R . Defining global predicted demand as $\hat{D} = \sum_{r \in R} \hat{d}_r$, the predicted disaggregated demand can be corrected by $\hat{d}'_k = \hat{d}_k * (D/\hat{D})$. As total actual demand corresponds to a constant value independently of the chosen reconciliation level, it means that all levels above the chosen level will also be automatically reconciled. As the portion of estimated demand in each r within R is kept constant, this is the reason why choosing different reconciliation levels results in different estimated demand.

A similar process can be applied to the forecasting of future demand, by replacing $\hat{d}_r$ with the demand forecasted at the level 3 region r , and replacing D with the demand forecasted based at the level 1 region R . In contrast with disaggregating past demand, time series applied at different levels will result in different values of total forecasted demand. Therefore, reconciliation can be done at multiple levels, in what can be named as hierarchical reconciliation of disaggregated demand forecast. As an example, consider the levels in Figure 4.1. Demand at level 3 can initially be determined by applying individual time series, or by simply assuming the values of the last disaggregated past demand. Next, forecasts at level 2 are performed, and demand at level 3 is scaled so that, within the level 2 regions, it equals the level 2 forecast. This procedure is applied to level 1, by forecasting demand at that level and scaling predictions at the levels below to match the new forecasts. The main advantage of this procedure is that forecasts are done at the highest granularity level possible, but at the same time these are smoothed by taking advantage of the well-behaved nature of aggregated time series forecasts.

It should be noted that the choice of the reconciliation level should also consider the approximate scale of the area of influence of a supply point, much like as in choosing the macro-region level for demand disaggregation. On the other hand, while choosing the macro-region had a soft impact on influence area delimitation, it did not set a cap on estimated demand at those levels, while reconciliation does, effectively stating that the all demand is contained within the borders of regions in that level. It is therefore recommended to reconcile at levels above macro-region, and to use the same reconciliation levels for past demand disaggregation and future demand forecast.

4.2.4 Estimation of sales

This module aims to estimate EVCS sales, either by assigning a fraction or market share of each demand to a supply point, or by directly computing the sales. Keep in mind that, in this context, the market share should be interpreted as the fraction of demand captured by each supply point. For example, say there are two customers and two supply points. If demand of customers A and B is 10 and 30 respectively, and they serve 4 and 20 of that demand in supply point X respectively, then the market share of A captured by X is 40% and the market share of B captured by X is 66.7%. This

market share allows to compute the influence or trade area of each supply point for visualization, as well as estimate supply point sales by multiplying the market share of each demand by the demand value. In this case, supply point X sales will be $10 * 40\% + 30 * 66.7\% = 24$.

When aiming to predict market shares, it is possible that the location of supply and demand is known, as well as sales at each supply point, but the flow of demand from its location to the supply points is not known, therefore not allowing for direct computation of market share to train the model. Therefore, it can be assumed that, for example, a supply point captures market share from demand within a given radius, and use those approximated market shares to train, for instance, a machine learning model.

This can be an optimization model, such as a user equilibrium or a location-allocation model, or an attraction model, such as gravity models. Agent-based simulation can also be employed, modelling individual agent decision rules instead of global optimization goals, in what can be considered a more realistic simulation of charging behaviour.

Chapter 5

Results and Discussion

In this chapter, the implementation of the methodology defined in the previous chapter is shown, and the decisions on the parameters and specific techniques used are explained. Given that the implementation consists of several coupled models, the models will be presented sequentially according to the data flow, and for each model will be presented its theoretical background, parameterization, and respective results, by this order. The new methodology developed is summarized schematically in Figure 5.1. All the developed methods are in python, mainly using the packages *pandas* for dataframe organization, *geopandas* for geodata manipulation, and *numpy* for computational performance. All the computational experiments were done in the same machine, with the CPU 12th Gen Intel® Core™ i7-1265U, and RAM 16.0 GB DDR4.

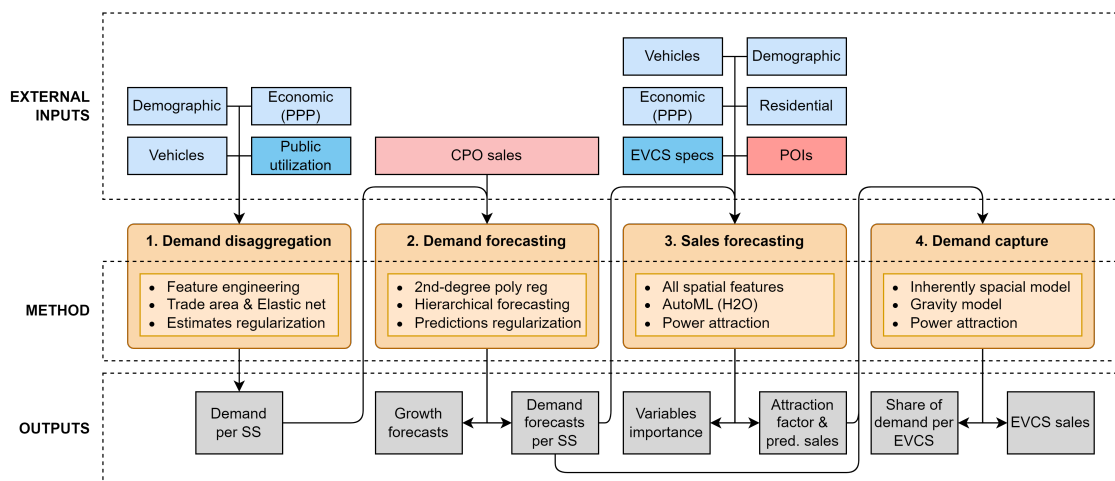


Figure 5.1: New methodology developed

5.1 Data web scraping

A very relevant improvement in the new methodology corresponds to the retrieval of live utilization of all EVCSs. This data was collected each minute for the mentioned two-weeks period. The station specifications and the tariffs were recovered daily. The possible statuses are shown in Table B.1 in variable *equipment_status*. To obtain the intervals of each status for each EVCS, timestamps of the observations were ordered and consecutive timestamps with the same status were replaced by a single observation, with the status, the start and end timestamp, with duration computed as their difference, therefore obtaining the duration and number of charges. The charge amount in kWh was obtained by multiplying the intervals with status *charging* by the maximum power of the EVCS. The cost of the charge is then computed based on the station tariffs, which can apply a fixed cost per charge, a variable cost per kWh, and a variable cost per minute.

This process inevitably stands on a set of impactful assumptions. It is assumed that different charges can be detected by a change in status retrieved every minute, therefore impacting the number of charges and the associated fixed costs. It is also assumed that if an EVCS is plugged (status *charging*) it directly translates to kWh charged, which might not be true if the user leaves his vehicle plugged while already at maximum battery capacity, impacting the kWh charged and respective variable cost.

The market size of the obtained data can be seen in Table 5.1. The equivalent yearly values are obtained by accounting for the global demand monthly seasonality index of June (8.06% of yearly demand in 2022) from MOBI.E (2022b), as well as for the two weeks in that month (14/30). It is clearly visible that, comparing with the trend from previous years, the increase in the number of charges by 45% is consistent with the previous trend. On the other hand, a kWh sales increase of 350% does not seem reasonable when compared to an increase of 113% in 2022 relative to the previous year.

Table 5.1: Comparison of retrieved¹ and global² data

	sales (GWh)	# charges	duration (min)	cost (€)
<i>Estimation</i> ¹				
Two-week period	6.1	132 514	13 044 419	895 921
2023 (approx)	161.4 +350%	3 522 618 +45%	346 759 594	23 816 262
<i>Actual</i> ²				
2022	35.9 +113%	2 435 971 +83%	—	—
2021	16.8 +172%	1 331 456 +144%	—	—
2020	6.2 +78%	545 003 +86%	—	—
2019	3.5	293 680	—	—

¹Data retrieved from MOBI.E (2022a)

²Data from MOBI.E (2022b)

It is not possible to validate the retrieved data with CPO data from the same period, given the mismatch of the periods for both. However, it is possible to use data from the CPO for the previous year. To do so, the retrieved data was scaled as previously mentioned, while CPO data was scaled using the seasonality coefficients from global sales (MOBI.E, 2022b) of November

and December (total 20.3%) given that these months are not available in CPO data. The CPO EVCSs with valid values were selected, plotted and the correlations computed, as can be seen in Figure 5.2. The correlations of CPO and MOBI.E kWh, minutes and charges are 0.810, 0.774 and 0.814 respectively, therefore pointing to the validity of the retrieved data, even when comparing CPO data from 2022 and MOBI.E data from two weeks in 2023. The retrieved data from MOBI.E (2022a) will therefore be used for demand disaggregation due to its valuable higher-granularity level, with the previously described scaling. The representation of these scaled sales can be seen for the municipality of Porto in Figure 5.3.

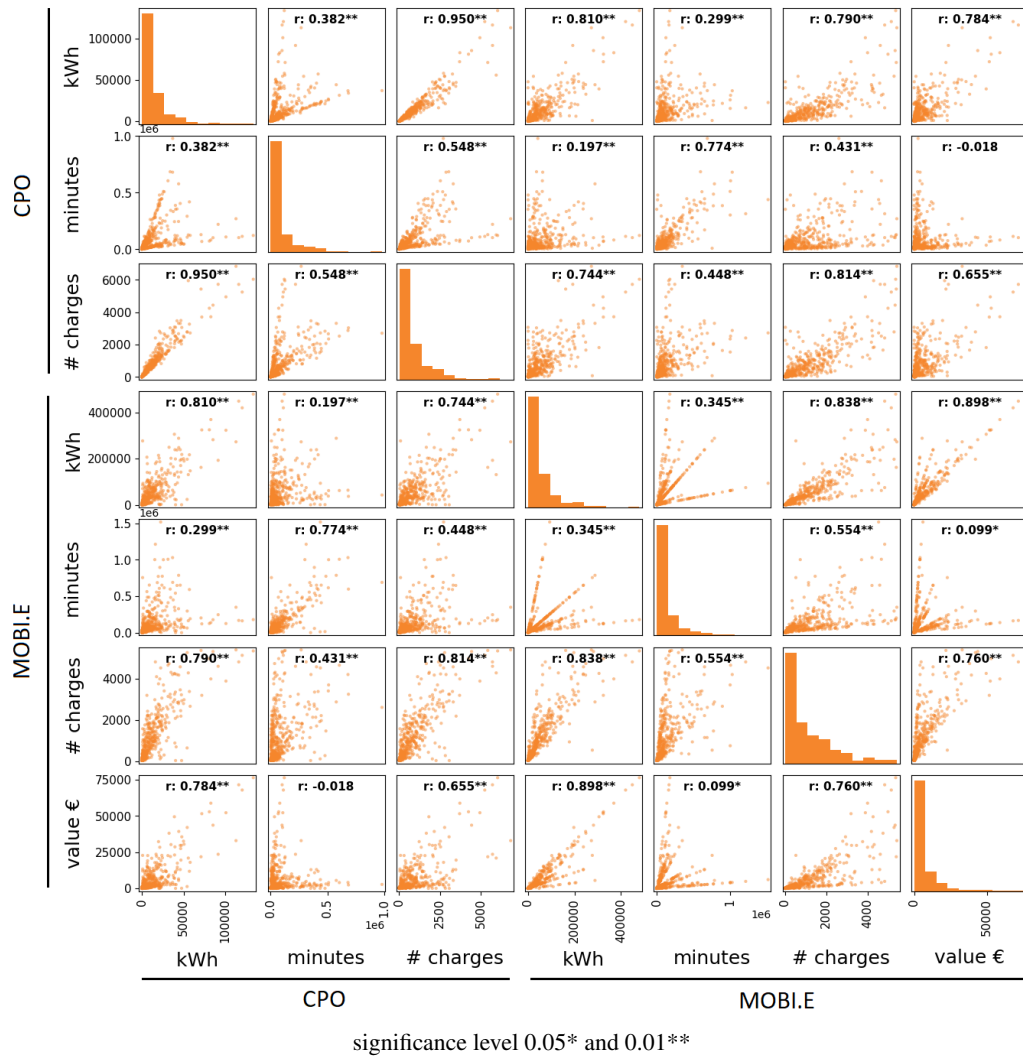


Figure 5.2: CPO and MOBI.E data comparison and correlations per EVCS

Globally, for the two-week period, the utilization rate is of 8.6%, with the sockets being *out-of-order* 3.6% of the time, *unknown* for 8.6%, and the remainder *available*. Having said that, the utilization rates of the sockets are not centred around its mean, given that it is heavily right skewed as can be seen in Figure 5.4 (a), with a median utilization rate of 5.4%. In fact, 85% of the sockets have a utilization rate below 20%. The utilization rate of the sockets does have an inverse

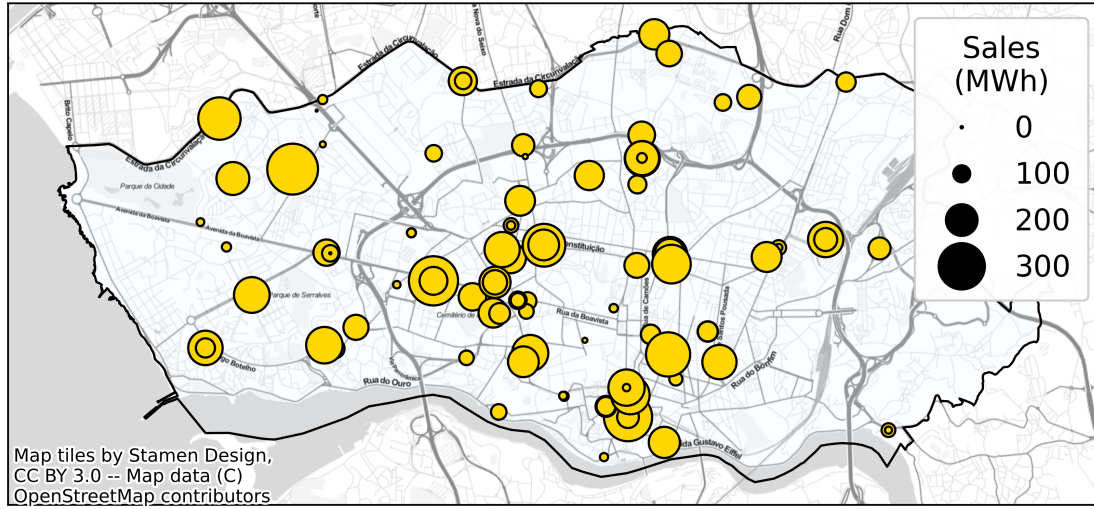


Figure 5.3: EVCS locations and scaled sales for 2023 in Porto municipality

relationship with its maximum power, where higher power sockets tend to have a lower utilization rate, given that these charge faster, as can be seen in Figure 5.4 (b). On the other hand, the fact that some lower power sockets have utilization rates above 50% might be due to users leaving their vehicles plugged overnight without actually charging, in part explaining the overestimation of sales in Table 5.1.

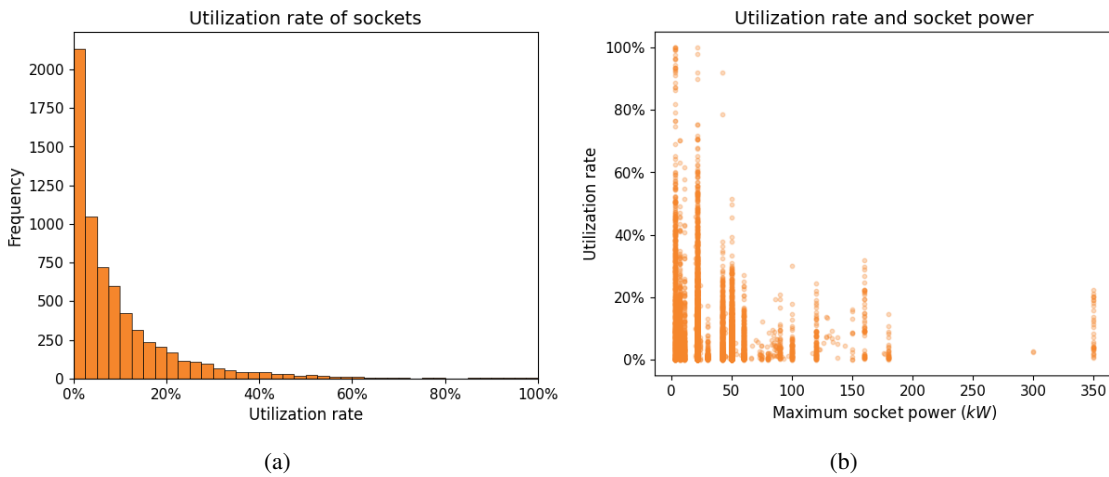


Figure 5.4: Utilization rate (a) distribution and (b) comparison with socket power

The collection and usage of this data enabled the use of more granular data from all CPOs. Despite only using data from two weeks, data collection is still being done. When the period of data recovery is significant to extract yearly trends (at least one year, preferably two or three years), the CPO data will only be needed to validate the recovered data, and the methodology will become somewhat independent of any CPO.

5.2 Demand disaggregation

The described process maps directly to Figure 4.1. It was applied to 2023 with utilization data for two weeks (scaled to the whole year using seasonality coefficients from 2022).

1. Define trade area size: Assume that demand decays with distance from the EVCS as $f(d) = (1 + d^2)^{-1}$, with d in kilometres. The definite integral of $f(d)$ in polar coordinates over a circular area of radius r is given by $F(r) = \pi \cdot \ln(r^2 + 1)$ and can be used to define the fraction of demand in that area as $D(r) = \frac{F(r)}{F(r_{max})} = \frac{\ln(r^2 + 1)}{\ln(r_{max}^2 + 1)}$. Since $F(+\infty) \rightarrow +\infty$, the denominator is necessary to limit demand, assuming that beyond r_{max} demand is null.

To set a value for r_{max} , the trip success ratio (TSR) method can be used. Given the distribution of battery capacities, the probability of each state-of-charge, and EV user trips/regular commuting activities, it outputs the fraction of trips that can be made within r_{max} without stopping for recharging. According to Alhazmi et al. (2017), an r_{max} of 10 km covers 99% of trips. Assuming null demand $D(r)$ past this radius, it can be seen in Figure 5.5 (a) that 50% of demand is located within 3 km, 80% within 6 km, and 95% within 9 km. These insights allow to choose the granularity level for model training (i.e. macro-region).

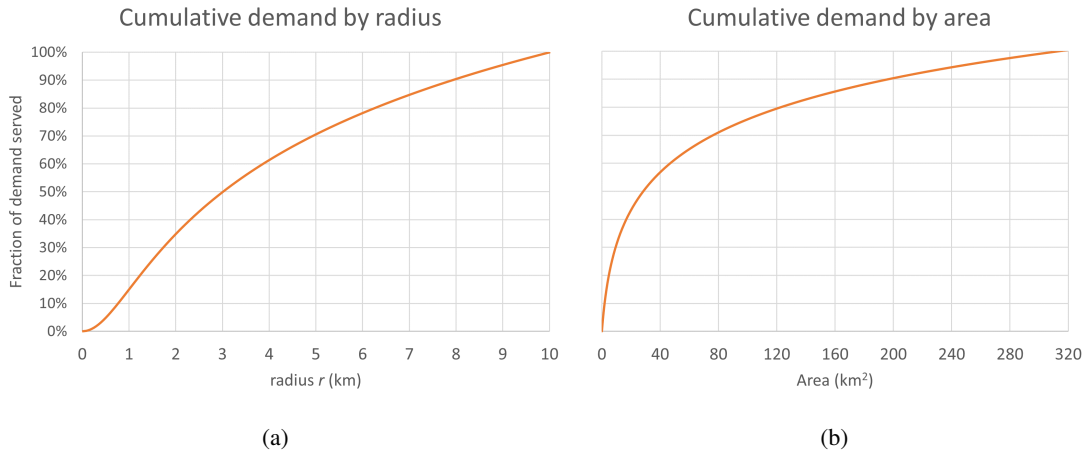


Figure 5.5: Assumed distance decay function for macro-region selection by (a) radius and (b) area

2. Select training granularity level: The subsection level SS is defined as the micro-region. Assuming an EVCS centred in a circular region with an equivalent area to the average area of each granularity level in Table 5.2, it can be seen in Figure 5.5 (b) that an SS would contain 3% of demand, SEC 29%, FR 52%, and MN 99.9%. The chosen granularity level should contain most of the demand, therefore a granularity level between MN and FR was chosen, with the model being trained simultaneously on both levels. The main advantage of training with two levels is the possibility to capture phenomena from both, at the cost of non-independence between observations, given that each MN contains many FRs.

It should be noted that, when sales per EVCS were aggregated by levels, there are some residual differences in global demand between levels. Therefore, the global actual demand is 158.18

GWh at COUNTRY and DT levels, 157.81 GWh at MN, 157.89 GWh at FR, 157.86 at SEC, and 153.40 at SS, representing a maximum deviation of 3.0% between these values. These are due to representation precision issues of region polygons in the same level, where some intersect each other or do not cover all the places where there are EVCSs.

Table 5.2: Area of the granularity levels

	DT	MN	FR	SEC	SS
#	18	278	2 882	9 929	203 264
μ	4950	321	30.9	8.97	0.438
σ	2122	284	47.1	21.50	1.547
min	2219	8	0.2	0.01	0.000
med	5003	229	16.9	1.43	0.062
max	10263	1721	888.4	325.31	135.435

3. Train model at macro-region level:

This step is divided in three substeps: feature engineering, computation of interactions, and the regression model. As the starting point, the applied feature engineering aims at obtaining relative variables from absolute variables and vice versa. The independent variables are composed by the demographic, vehicles, and economic (PPP) data, while the dependent variable corresponded to the sales in kWh. All the raw variables were considered, and new features computed. Given that vehicles consists of the sales of new vehicles per thousand inhabitants and PPP is per capita, both were multiplied by the number of inhabitants to obtain the number of new vehicles sold and an absolute proxy of wealth, respectively. The active and dependent population were obtained based on the demographic data. These calculations are done at the highest-granularity level SS, and compose the base dataset. This data is then aggregated to other granularity levels according to the need, computing the sum of count variables (raw demographic, number of vehicles sold, absolute proxy of wealth, sales in kWh) and the weighted average of normalized features (raw density of vehicles sold per thousand inhabitants, raw PPP) weighted by the number of inhabitants. After aggregation at the intended granularity level, the demographic count variables are normalized, dividing by the number of inhabitants, and by the area in km^2 . Let these new features be called the original variables. The observations (regions) with no inhabitants were excluded.

Next, the granularity levels of the training data are selected. In this case, as explained in step 2, these are MN and FR. The observations of these levels are concatenated in a single dataset. Next, interactions are computed. It aims to uncover non-additive relationships among two or more variables, inspired by two-way ANOVA, where a linear relationship is established between the dependent variable and the independent variables as well as their interactions. To accelerate this process, the interactions are computed for the best variables. Feature selection of the best variables is always done with Pearson correlation coefficient relative to sales. Therefore, for orders $l \in L$, the fraction of the best $f_l |_{0 \leq f_l \leq 1}$ variables are used to compute the l^{th} degree interactions and power of the variables. For example, say there are $k = 100$ transformed variables, and that $L = 1, 2, 3$

and $f_1, f_2, f_3 = (1, 0.5, 0.2)$. This means that all 100 original variables would be used as first-order features ($l = 1$) so $C_1^{1k+1-1} = 100$ variables, for the first-order interactions ($l = 2$) it would mean using the best 50% original variables so $C_2^{0.5k+2-1} = 5050$ new features in the forms $a * b$ and a^2 , and for the second-order interactions ($l = 3$) using the 20% best variables would result in $C_3^{0.2k+3-1} = 255$ new features in the forms $a * b * c$, $a^2 * b$ and a^3 , totalling 5405 features to be fed to the model. Let these features be called the transformed features. This procedure is especially important if little feature engineering is performed. To divide the variables by each other, one can also feed the transformed variables and their inverses to this interactions step. Having said that, it is important to avoid the computation of many overly correlated variables, as it will negatively impact the model if these are simultaneously considered. Finally, of these transformed variables, the best n features are fed to the model, thereby called the selected features.

The number of original variables and their correlations with the dependent variable were considered high enough to justify not computing any interactions, so $L = \{1\}$ and $f_L = \{1\}$. The number of best features tested included $n \in N = \{1, 10, 30, 100, 200, 330\}$. For the same reason, an elastic net was fitted. The chosen prediction method is one of high bias and low variance, where a simpler model is fed with engineered features, rather than a complex non-linear model trained with the original variables. It requires careful parameterization and slower deployment but results in a more explainable model.

The elastic net is a reconciled linear regression that deals with correlated variables and limits overfitting by minimizing not only the sum squared of errors (SSE) but also a penalty for the coefficients. This penalty is a sum of the penalties from ridge regression and lasso, l^1 and l^2 norms of the coefficients, respectively (Zou and Hastie, 2005). The model is given by

$$\min_{\beta} \left\{ \frac{1}{2N} \|y - X\beta\|^2 + \alpha \ell_1 \|\beta\|^1 + \frac{\alpha}{2} (1 - \ell_1) \|\beta\|^2 \right\} \quad (5.1)$$

The ridge part (l_1 penalty) tends to remove variables from the regression by equalling its coefficient to zero, while the lasso part (l_2 penalty) tends to decrease the weights of unimportant variables but does not remove them. This difference in behaviour results from the different norms used, and is a part of what makes elastic net more versatile than any of the two separately, by combining the strengths of both as the cost of having more parameters. The relative strength of the two norms is given by the ℓ_1 ratio, which ranges from 0 to 1, while the intensity of the coefficient penalties is given by α . Given that the absolute values (squared or not) of all coefficients are summed, the independent variables must be normalized. In this particular case, because most variables used are counts that scale with the dependent variable, the independent variables are scaled to unit standard deviation, but the mean of the independent variables is not removed, so that the linear regression is fitted with null intercept. This guarantees positive predictions, and predicts zero for values of zero in all independent variables. While α depends on the behaviour of the phenomenon at hands, namely the model fit quality and the predictive power of the variables used, its magnitude is mainly driven by the magnitude of the response variable, given that it is not scaled.

The model is cross-validated with k-fold $k = 10$ and $r = 10$ repeats, on a fixed random seed for comparability among the different combinations of parameters and reproducibility. The cross-validation is used to select the best (ℓ_1, α) combination of the input parameters. This is done based on the coefficient of determination r^2 , given that it is the optimized parameter of linear regression, it can handle differences in magnitude across multiple years (unlike MAE and RMSE) making the models somewhat comparable, while also being stable for estimations close to zero (unlike MAPE). Having said that, MAE and RMSE will also be shown.

The tested combinations of (ℓ_1, α) can be seen in C.2. As the method to obtain the regression parameters is not deterministic (due to the l_1 and l_2 penalties), these are obtained iteratively. The model tolerance for stopping was set at 0.0001 and the maximum number of iterations at 1000, to keep computation time within reasonable limits.

These combinations of (ℓ_1, α) for each $n \in N$ result in $\# \ell_1 * \# \alpha * \# N = 8 * 10 * 6 = 480$ parameters combinations, each originating $k * r = 10 * 10 = 100$ cross-validation models, totalling 48 000 models. Therefore, it is important to do a previous test on the computational performance of these models. For $n = 10$, the 16 combinations of $\ell_1 \in \{0, 0.001, 0.1, 10\}$ and $\alpha \in \{0, 0.25, 0.75, 0.99\}$ were tested with the respective cross-validation, and the average time per combination was computed. The time was registered for different values of n_jobs in two python functions from the *sklearn* package: *ElasticNetCV* and *cross_val_score*. The results are shown in Table C.1, and led to choosing 1 and 16 n_jobs for *ElasticNetCV* and *cross_val_score*, respectively.

Given that there are three parameters at play, two for model calibration and one for feature selection, the analysis of the results needs to fix some parameters to vary others. As such, first the best models of each n will be selected and analysed, with the best of these being considered the reference. Second, using this model as reference, the feature selection parameter n is fixed, and the calibration parameters (ℓ_1, α) are changed. The best model of each n was selected according to maximum r^2 , assuring that there was continuity in the r^2 among neighbouring ℓ_1 and α parameters. This meant that despite $(\ell_1, \alpha) = (1, 1000)$ being the best combination for $n = 100$, it was not considered given its disparity of r^2 from neighbouring parameter combinations, as well as the very different (ℓ_1, α) parameters from the best at the other n , as can be seen in Figure C.4.

It can be seen in Table 5.3 that the improvement of the elastic net relative to the linear regression ($\alpha = 0$, Table C.3) is relatively low, with the improvement of r^2 always below 0.02. As such, the n with the best model (0.722) is indeed the one where the r^2 of the linear regression is highest (0.709), namely $n = 100$, where less or more selected features worsens the model fit. The penalty of the best models is always 0.1 except at $n = 10$, and as the number of features used increases, the model goes from lasso (l_1 penalty) to ridge (l_2 penalty). This implies an increasing focus on decreasing the coefficients instead of directly removing the variables, except for $n = 1$ where reconciliation has little meaning given the use of a single feature. The best models of each n have $\alpha = 0.1$ and ℓ_1 between 0.75 and 0.99, therefore one can choose as reference the model $(\ell_1, \alpha, n) = (0.75, 0.1, 100)$. Of these models, it is actually the best in all three metrics MAE, RMSE, and r^2 . As computation time increases significantly with the number of features, it is important to choose the minimum number of features that assures sufficient performance.

Table 5.3: Best elastic net models per n

n	ℓ_1	α	Cross-validation ($\mu \pm \sigma$)				time (s)
			MAE	RMSE	r^2		
1	0.25	0.1	68 433 $\pm$ 12 733	227 312 $\pm$ 101 531	0.710 $\pm$ 0.175		1.2
10	0.9	0.01	59 614 $\pm$ 12 389	228 332 $\pm$ 90 263	0.683 $\pm$ 0.193		3.7
30	0.99	0.1	58 743 $\pm$ 12 364	223 150 $\pm$ 92 258	0.703 $\pm$ 0.182		8.0
100	0.75	0.1	57 264 $\pm$ 11 646	208 268 $\pm$ 81 472	0.722 $\pm$ 0.177		32.7
200	0.75	0.1	64 031 $\pm$ 12 251	222 282 $\pm$ 96 503	0.719 $\pm$ 0.171		90.4
330	0.1	0.1	67 982 $\pm$ 12 672	223 330 $\pm$ 98 223	0.714 $\pm$ 0.174		210.6

To analyse the effect of the calibration parameters ℓ_1 and α , the feature selection parameter is fixed at $n = 100$. Model accuracy reaches its maximum at intermediate values of ℓ_1 and α , whereas precision improves for lower ℓ_1 and higher α , as can be seen in Figure 5.6. Relative to a linear regression with the same variables, the best model presents an r^2 improvement from 0.709 to 0.722 (excluding the singularity at 0.731), and a reduction in variability from 0.190 to 0.177.

μ_{r^2}		α									
		0	0.0001	0.001	0.01	0.1	1	10	100	1000	10000
ℓ_1	0.00	0.709	0.710	0.713	0.721	0.711	0.678	0.648	0.523	0.115	0.005
	0.10	0.709	0.710	0.713	0.719	0.713	0.680	0.649	0.540	0.127	0.007
	0.25	0.709	0.710	0.713	0.718	0.719	0.683	0.650	0.566	0.150	0.010
	0.50	0.709	0.709	0.712	0.715	0.715	0.690	0.654	0.608	0.210	0.018
	0.75	0.709	0.709	0.711	0.713	0.722	0.699	0.664	0.640	0.342	0.044
	0.90	0.709	0.709	0.710	0.713	0.721	0.711	0.678	0.648	0.522	0.113
	0.99	0.709	0.709	0.709	0.710	0.713	0.721	0.711	0.678	0.648	0.516
	1.00	0.709	0.709	0.709	0.709	0.709	0.709	0.709	0.713	0.731	0.674

(a)

σ_{r^2}		α									
		0	0.0001	0.001	0.01	0.1	1	10	100	1000	10000
ℓ_1	0.00	0.190	0.189	0.187	0.182	0.176	0.176	0.169	0.142	0.043	0.019
	0.10	0.190	0.189	0.187	0.182	0.175	0.176	0.170	0.143	0.047	0.019
	0.25	0.190	0.189	0.188	0.183	0.177	0.176	0.172	0.145	0.054	0.019
	0.50	0.190	0.189	0.188	0.184	0.177	0.176	0.176	0.145	0.072	0.020
	0.75	0.190	0.190	0.189	0.186	0.177	0.176	0.177	0.151	0.108	0.024
	0.90	0.190	0.190	0.189	0.187	0.182	0.176	0.176	0.169	0.142	0.043
	0.99	0.190	0.190	0.190	0.189	0.187	0.182	0.176	0.176	0.168	0.142
	1.00	0.190	0.190	0.190	0.190	0.190	0.190	0.190	0.188	0.178	0.171

(b)

Figure 5.6: Elastic net r^2 (a) mean and (b) standard deviation with fixed feature selection parameter

Having explored the parameterization of the model, the best (and chosen) model will be analysed, as well as its predictions. As previously stated, it has parameters $(\ell_1, \alpha, n) = (0.75, 0.1, 100)$ and cross-validation performance $\mu_{r^2} \pm \sigma_{r^2}$ of 0.722 ± 0.177 , which took 34.1s to compute the 10-fold cross validation with 10 repeats. Its has a train fit of $r^2 = 0.908$. Further analysis of the model fit and residuals analysis can be found at the end of Appendix C

The variables' importance were obtained based on the absolute coefficients of the features.

The groups of variables and their respective most important variables are shown in Table 5.4. The *Vehicles sold vars* is the most important group, with 56.2%. The model gives high importance to heavy and light goods vehicles, which is unexpected given that light and light passenger vehicles have higher correlation with sales. On the other hand, *Sociodemographic vars* weight 32.1%, with its variable *# w/ college degree* being the most important variable, with 5.8%, when its correlation with sales is the 25th highest. These results are due to *# w/ college degree* being a very unique variable in that it is not highly correlated with others, while the number of vehicles sold as well as PPP variables are highly correlated within their respective groups.

Table 5.4: Group and variables importance

Grouped		Most important variable per group	
Group	Imp.	Variable	Imp.
Vehicles sold vars	56.2%	2017 Heavy goods vehicles	3.8%
Sociodemographic vars	32.1%	# w/ college degree	*5.8%
# inhabitants * PPP vars	11.7%	# inhabitants * PPP 2005	2.8%

*global most important variable

4. Predicting demand at micro-region level: The trained model was applied to predict demand in the 203 264 SS. This resulted in a total predicted demand $\hat{D}_{total}$ of 163.7 GWh compared to the total actual demand D_{total} of 158.2 GWh from the train data, a 3.5% gap. Given that the macro-region corresponds to MN and FR, conciliation was performed at DT and COUNTRY levels in order to assess model quality, whose results can be seen by DT in Table 5.5.

Table 5.5: Reconciliation results at DT level

Level		Sales (GWh)			Predicted/Actual	Relative error
id	Name	Actual	Predicted	Error		
PT01	Aveiro	6.62	8.15	1.53	123%	23%
PT02	Beja	4.40	2.48	-1.92	56%	-44%
PT03	Braga	7.61	8.45	0.84	111%	11%
PT04	Bragança	0.76	1.69	0.93	223%	123%
PT05	Castelo Branco	1.73	2.22	0.49	129%	29%
PT06	Coimbra	4.88	5.26	0.38	108%	8%
PT07	Évora	2.50	2.48	-0.02	99%	-1%
PT08	Faro	9.70	9.98	0.28	103%	3%
PT09	Guarda	0.63	1.81	1.18	288%	188%
PT10	Leiria	11.38	9.39	-1.99	83%	-17%
PT11	Lisboa	58.16	58.73	0.56	101%	1%
PT12	Portalegre	0.76	1.61	0.86	214%	114%
PT13	Porto	25.46	26.68	1.21	105%	5%
PT14	Santarém	5.37	4.42	-0.96	82%	-18%
PT15	Setúbal	13.08	11.15	-1.92	85%	-15%
PT16	Viana do Castelo	1.64	1.98	0.33	120%	20%
PT17	Vila real	1.06	2.16	1.10	203%	103%
PT18	Viseu	2.44	3.46	1.02	142%	42%

In the comparison of actual sales and disaggregated estimated demand at the reconciliation level DT, it is clearly visible that the regions with lower actual sales have higher relative error, with a tendency for overestimating demand, namely PT04, PT09, PT12, PT17. This is due to three factors: estimations for regions with less demand (and less population) have lower accuracy due to a lesser degree of aggregation; these regions are less densely populated, therefore the longer travel distances and more dispersed charging infrastructure corresponds to very different demand patterns from urban areas; given that the model predicts absolute demand, relative errors of low sales regions have little importance for the model, therefore low actual sales values in denominator leads relatively small absolute errors to present as big relative errors. On the other hand, regions with higher sales tend to present lower relative errors, in particular PT11 and PT13. The reconciliation at DT level corrected these errors, by setting the total estimated demand equal to total actual sales at that level. Given that the total actual sales at DT and COUNTRY level are the same, the reconciliation at COUNTRY level was not necessary afterwards.

The new demand disaggregation model is a major contribution of the new methodology, automatizing the process of demand disaggregation (previously a manual process), improving its performance with a more sophisticated and parameterized model, and using it to predict at a higher granularity level (SS) than the previous model (SEC). It proved capable of going from actual sales to estimated disaggregated demand, as can be seen in Figure 5.7.

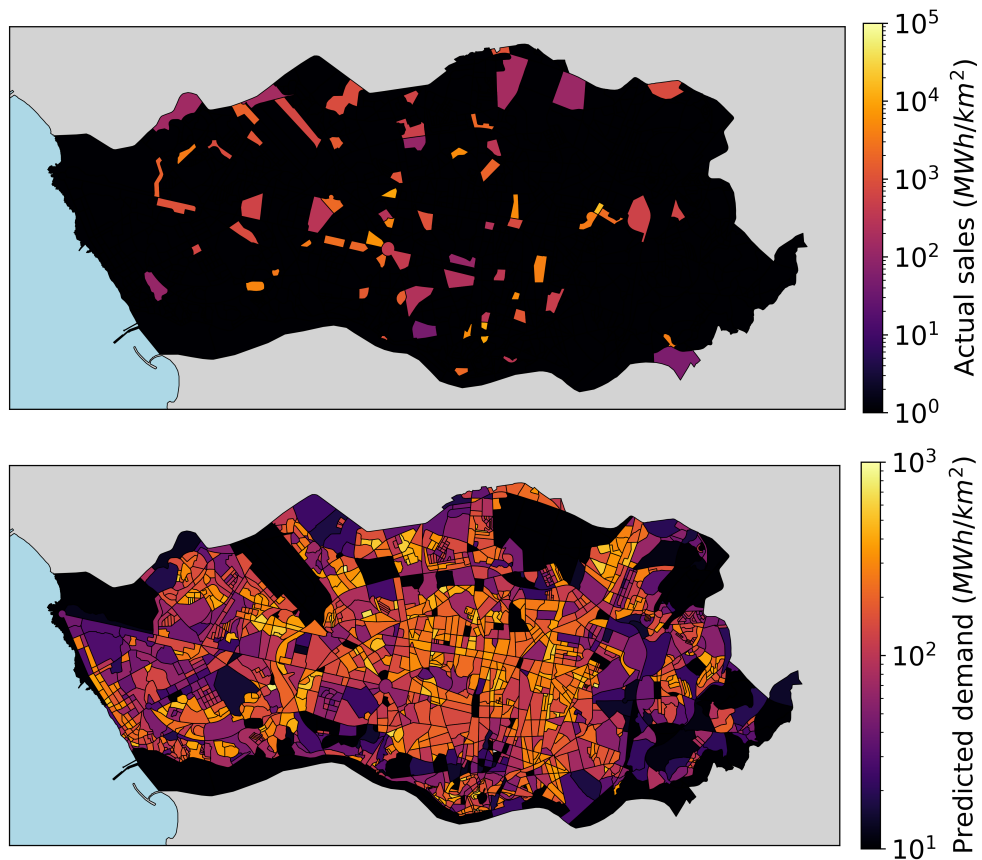


Figure 5.7: Actual sales and predicted demand in Porto municipality for 2023

5.3 Demand forecast

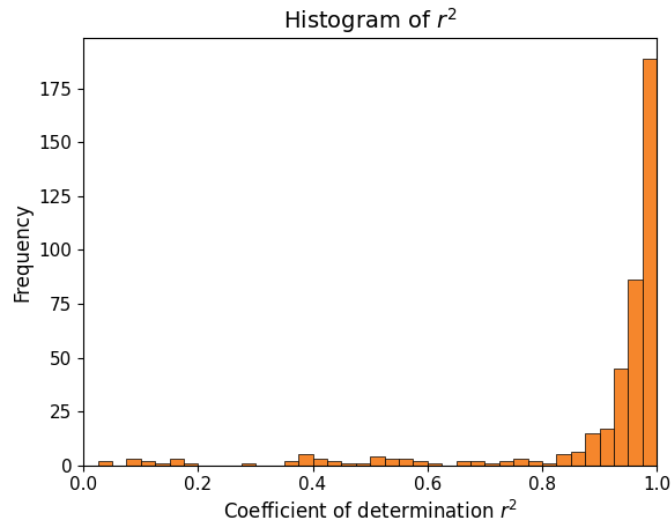
It is important to forecast demand at the highest granularity level, but doing so can present issues with data instability, as previously discussed. Given the good fit on Figure 3.3 (b) of a second-degree polynomial regression on the global demand, it was the chosen model for time series forecasting. It can incorporate both a fixed growth in the first-degree part, and if this growth is accelerating or decelerating with its second-degree part in the form $\hat{y} = a * x^2 + b * x + c$, where $\hat{y}$ is the predicted demand and x is time. The moving average of width 12 periods (12 months or a year) was computed for each time series, in order to smooth the series as well as remove seasonality, given that the focus is on medium-term forecasts (1 to 5 years).

As mentioned in the methodology of section 4.2.2, each time series is standardized by dividing its values by the most recent observation. Besides, the time of observations is converted to a number, with January 2019 as $2019 + 0/12$, February 2019 as $2019 + 1/12$, and so on. The most recent observation time is subtracted from all the observation times, so the most recent observation is at time 0, past observations are negative time, and forecasts are positive time. Therefore, both time x and demand y are standardized, allowing to use the model parameters at different levels. This means that the last observations is located at $(x, y) = (0, 1)$.

Imagine that instead of a second-degree polynomial regression, a simple linear regression was applied on the global demand. It would predict a demand for the immediate next period (in this case June 2023) that is much lower than the last observed value, given that it would fit to minimize error from all past data equally. This issue can be attenuated by forcing an intercept of the model at the most recent observation. Therefore, by standardizing the most recent observation as value 1 at time 0, it is possible to set the intercept c of the second-degree polynomial regression to 1. The model will intercept the most recent observation, and optimize the parameters to minimize error.

The time series inclusion criteria are: (C1) minimum of 6 periods (months); (C2) the value of the last observed month must be at least 3000 KWh; (C3) model fit criteria, with minimum r^2 of 0.8. The first (C1) aims at having sufficient data for computing a time series, namely avoiding a very good model fitted on little data that predicts extreme growths or shrinkage of demand. Given that the aim is to forecast in a period of around 1–5 years, it was deemed that 6 months moving average (18 months of actual observations) was the minimum necessary data to extrapolate for that horizon. This length also proves adequate given that the 12-period moving average is being used, corresponding to roughly half of it. As for the second (C2), the threshold was estimated by computing the average sales per EVCS for the last month, and stating that the sales in the last month must be at least the equivalent to three times that value, in order to ensure that the parameters are representative of a trend. Regarding the third criteria (C3), selecting based on the r^2 aims to include most of computed time series while providing sufficient fit quality, as can be seen by its distribution in Figure 5.8.

The first derivative of the model corresponds to $2a * x + b$, while the second derivative is $2a$. This means that the first-degree coefficient b can be interpreted as the current growth rate, while $2a$ is interpreted as the variation in the growth rate. For positive b , positive a means the growth

Figure 5.8: Forecast model r^2 after C1 and C2, before C3

rate is positive and accelerating, and negative a means the growth rate is positive and decelerating. From Table 5.6 and Figure D.1 it can be seen that, globally, the current relative growth rate is 68.4% per year, and is accelerating 26.2pp per year. Keep in mind that growth rate is in energy per year, and that the standardized coefficients a and b are relative to the last observed value (moving average of November 2022), not relative to the previous observation or year. The growth rate b of the selected regions is mostly positive, and while the same is true for its variation a , there are a significant number of regions where growth is decelerating.

Table 5.6: Time series coefficients and r^2 after quality criteria selection

level (#)	COUNTRY (1)			DT (15)			MN (68)		
	a	b	r^2	a	b	r^2	a	b	r^2
μ	0.131	0.684	0.999	0.117	0.685	0.970	0.203	0.850	0.960
σ	—	—	—	0.143	0.322	0.027	0.783	0.970	0.040
min	—	—	—	-0.268	-0.153	0.888	-1.204	-2.511	0.837
med	—	—	—	0.132	0.715	0.973	0.143	0.768	0.976
max	—	—	—	0.409	1.314	0.998	4.673	4.293	0.999

level (#)	FR (113)			SEC (167)		
	a	b	r^2	a	b	r^2
μ	0.082	0.823	0.965	-0.049	0.649	0.964
σ	0.767	0.970	0.036	0.608	0.918	0.034
min	-2.563	-2.511	0.837	-2.563	-1.568	0.825
med	0.058	0.734	0.979	-0.063	0.547	0.976
max	4.673	4.293	1.000	2.822	3.677	1.000

In the end, 364 valid time series were obtained, each for a given region in a given level. This means that the regions whose time series were not suitable for parameter estimation do not

currently have parameters. To obtain those coefficients, the procedure in section 4.2.2 is applied, namely for a given region in a given level, use the parameters of its closest parent, i.e. from the smallest region in which it is contained. Say that a given SEC has no parameters, nor does its FR and MN, but its DT and COUNTRY do have parameters. This means that the SEC itself, its FR and MN will get their parameters from their corresponding DT.

As stated in sections 5.1 and 5.2, the disaggregated demand for 2023 needs to be corrected. The forecast in x years relative to the last known period for a given region k is computed as

$$\hat{D}_k(x) = D_k * (a_k * x^2 + b_k * x + 1) \quad (5.2)$$

where D_k is the demand of the last known period. As such, using the sales from the previous year 2022 of 35.9 GWh, and the global coefficients $a = 0.131$ and $b = 0.684$ from Table 5.6, it is possible to predict global sales in 2023 as $D(1) = 35.9 * (0.131 * 1^2 + 0.684 * 1 + 1) = 65.2 \text{ GWh}$. The 2023 disaggregated demand was reconciled at COUNTRY level so that it matches this amount.

At this point, the time series for all regions at all levels have been obtained, as well as the corrected disaggregated demand. Having the corrected disaggregated demand for 2023 and the time series parameters, the 2023 demand will be projected into the future. In particular, these were computed for next year, in order to show the procedure and results. First, equation 5.2 is used to forecast demand at all levels, yielding the total demand per level in Table 5.7. Now, these gaps to COUNTRY are brought to zero by using the hierarchical forecasting procedure. It is similar to reconciliation, in that it is applied from levels COUNTRY to SEC, so that total demand at all levels is the same by assuming. When reconciling a given upper level region, it assumes that the proportion of demand of each lower level region within it remains constant, and re-scales the lower level demand so that its sum matches demand of the upper level region.

Table 5.7: Total forecasted demand for 2024 per level

Level	Forecasted demand (GWh)	Gap to COUNTRY
COUNTRY	118.1	—
DT	112.9	-4.4%
MN	117.5	-0.5%
FR	116.8	-1.1%
SEC	116.4	-1.4%
SS	117.1	-0.9%

To grasp how the forecast model works at different granularity level, the higher-level regions of the specific SS PT13121000802, shown in the left most side of Table 5.8, are analysed. The model parameters for the regions of levels COUNTRY to FR belong to the region itself, while SEC and SS parameters came from FR, as can be seen in column *origin* from Table 5.8. As stated previously, the 2024 demand predictions are obtained based on the 2023 demand and the model parameters, therefore the percentage growth from year 2023 to 2024 is the same for levels FR, SEC and SS given that the model parameters are the same. It can be seen that the DT is growing faster

(102.6%) than the national average (81.5%), with the MN growing considerably slower (47.0%). The national demand will surpass 118 GWh in 2024.

The instantaneous growth rate can be computed based on the predicted demand for that year and the model parameters, while the average growth rate from 2023 to 2024 corresponds to the difference between the demand predictions of those years. The relative instantaneous growth rates are similar, but divided by the demand predicted in that year. It can therefore be seen that, as of 2023, the market is growing 68.4%, while next year it will grow at 52.1%, resulting in a deceleration of the growth rate of -16.3pp. While the MN presented a slower percentage average growth from 2023 to 2024 (47.0%), this relative growth is expected to take longer to slow down (decelerating at only -6.2pp) in contrast with its DT and the national average.

Table 5.8: Model parameters and forecasts for 2024 in specific region

Level		Model parameters				Demand predictions (kWh)		
Group	Id	a	b	r2	origin	2023	2024	growth
COUNTRY	PT	0.131	0.684	0.999	—	65 080 949	118 128 553	81.5%
DT	13	0.156	0.780	0.994	—	10 709 743	21 698 543	102.6%
MN	12	0.120	0.707	0.963	—	3 234 336	4 753 883	47.0%
FR	10	0.132	0.898	0.996	—	648 324	1 136 328	75.3%
SEC	008	0.132	0.898	0.996	FR	11 483	20 126	75.3%
SS	02	0.132	0.898	0.996	FR	3 463	6 069	75.3%

Level		Growth rate (kWh/year)			Relative growth rate		
Group		*2023	2023-24	*2024	*2023	*2024	acceleration
COUNTRY		44 541 052	53 047 604	61 554 157	68.4%	52.1%	-16.3pp
DT		8 358 354	10 988 800	11 706 814	78.0%	54.0%	-24.1pp
MN		2 287 114	1 519 547	3 065 837	70.7%	64.5%	-6.2pp
FR		582 189	488 004	752 855	89.8%	66.3%	-23.5pp
SEC		10 311	8 643	13 334	89.8%	66.3%	-23.5pp
SS		3 110	2 607	4 021	89.8%	66.3%	-23.5pp

*instantaneous rates, the remaining are average rates from 2023 to 2024

Analysing the outputs visually at SS level, the areas of highest demand growth can be seen in Figure 5.9. Indeed, the coastal area seems to have stagnated, while the Northwest area is shrinking. Other than the coastal and Northwest areas, the demand in the municipality is growing at a significant growth rate. These insights are complemented by Figure 5.10, where it can be seen that indeed the coastal and Northwest areas are rapidly lowering their growth rate, meaning that the coastal area demand will soon start to shrink, and the demand in the Northwest area will ever-rapidly decrease. As for the remainder of the municipality, the growth rate remains fairly constant (with null acceleration), highlighting that the demand in the East area of the municipality is to grow ever-rapidly. Overall, analysing only the predicted demand and not accounting for the current EVCSs, all areas of the municipality except the coastal and Northwest areas are suitable for installing new infrastructure, particularly the East area, which is projected to not only grow fast, but this growth rate will increase with the years.

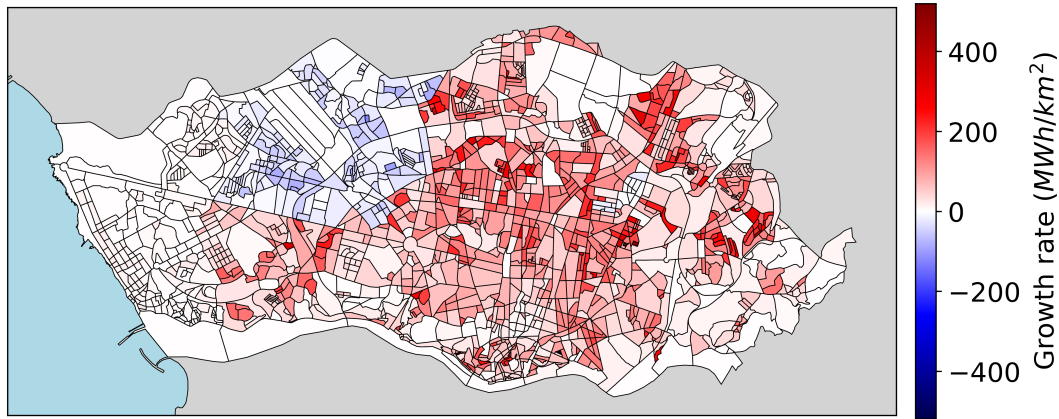


Figure 5.9: Average growth rate from 2023 to 2024 in Porto municipality

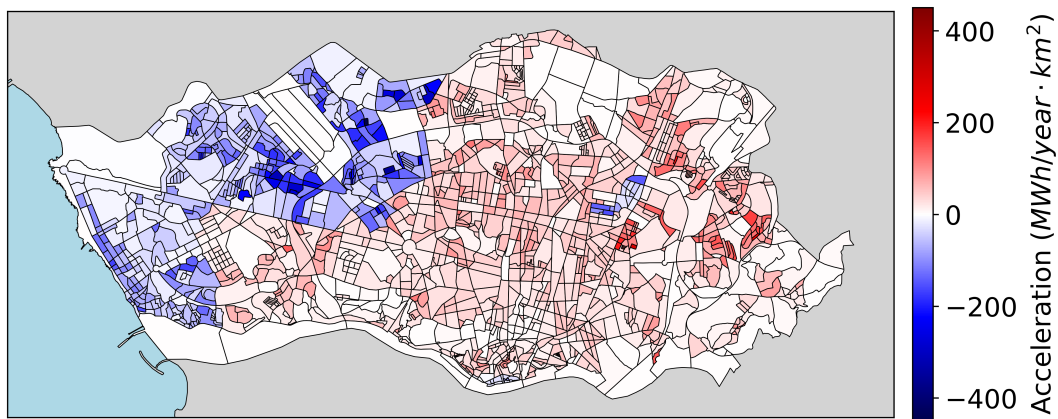


Figure 5.10: Acceleration of instantaneous growth rate from 2023 to 2024 in Porto municipality

The new demand forecast model constitutes a relevant novelty from the previous methodology, given that it used the same CAGR from market reports and projections for all regions and years. With this new methodology, the growth rate is much more granular, and is extracted from sales time series instead of using external projections of growth. Having said that, it is good practice to validate the obtained growth rates using those external sources. It also provides insightful visualizations, critical for pinpointing demand hot-spots.

5.4 Sales forecasting

In this step, a machine learning algorithm is trained to estimate an attractivity coefficient of demand to the EVCSs. The process of feature engineering is fully spatial resulting in 1727 features, which are used to predict sales. The sales amount used correspond to the raw estimates of sales from the two-week period, about 5 GWh. It does not need to be scaled, as it would only entail multiplying by a constant, which has no effect on the model training and in the gravity model.

The model is trained and predictions are made for the 5813 connectors (or sockets). The features created are of the following types:

- **Density features** are computed as the sum of original features in a given region divided by *# inhabitants* and by *# Area (km²)*. These are applied to all levels from SS to DT, and to the sociodemographic, vehicles sold, PPP variables and the predicted demand.
- **Proximity features** are computed as the sum of original features within a given radius from the EVCS. These include the previous features per SS (sociodemographic, vehicles sold, PPP, predicted demand) located at a point in the centroid of their respective SS, as well as the points-of-interest, therefore both are located in points. The proximity features were computed for radius of 1, 3, 6, 10 kilometres. In the case of the SS features, the sum of the variables was computed, yielding for instance the *# inhabitants* in each of those radiuses. As for points-of-interest, the count per amenity type and amenity grouping was computed.
- **EVCS features** correspond to the technical specifications of the EVCS, namely socket type, power, CPO, and tariffs. Using the charges from MOBI.E, a linear regression was fitted to each variable duration, sales (kWh), and cost of charges as a function of power, therefore obtaining an “average” duration, sales (kWh) and cost of charge per EVCS power, with this “average” duration and sales (kWh) being used to compute an average cost based on the tariffs. Keep in mind that the linear regression to compute cost returns the average cost per charge, while the tariffs-based cost is the cost of the average charge.

The method used is the AutoML from the package H2O, which includes gradient boosting machine (GBM), distributed random forest (DRF), extensible R-Trees (XRT), deep learning (DL), and generalized linear model (GLM). It is also possible to search parameters (grid), and combine these models in stacked ensembles (all models, or the best per family). Ten models were obtained and validated using 10-fold cross-validation, with results in Table 5.9 sorted by RMSE. It can be seen that the DL and GLM models present the worst results, with GBM, XRT, DRF presenting the best single-model results, while both stacked models present the best results overall. Given that the stacked models have lower explainability and their improvement relative to the best individual model (GBM2) is low, from 0.546 to 0.555, GBM2 was chosen as the model to use. The predictions of the best eight models are highly correlated, as can be seen in Figure E.1.

The model was stopped at 103 trees, when there is a spike in the cross-validation RMSE, and the residuals tend to increase with the predicted value, further visualizations in Appendix E. The distribution of the actual sales values is $(\mu \pm \sigma) = (847.8 \pm 1185.2)$, and the train metrics are RMSE of 503.7 and MAE of 311.1. Comparing those to the model’s cross-validated metrics RMSE (774.3 ± 101.9) , MAE (461.5 ± 29.6) , and r^2 (0.546 ± 0.109) , these are considerably different from the train metrics but with similar order of magnitude. The quality of the model is suitable (despite not extraordinary), highlighting the consistency of the cross-validation performance, with low standard deviation of the three quality metrics when compared to their expected

Table 5.9: Total forecasted demand for 2024 per level

	Model	Cross-validation ($\mu \pm \sigma$)			Time (s)	
		RMSE	MAE	r^2	train	predict all
#0	Stacked Family	759.8 $\pm$ 117.0	445.5 $\pm$ 28.7	0.555 $\pm$ 0.097	1.2	0.125
#1	Stacked All	767.5 $\pm$ 104.6	444.5 $\pm$ 32.1	0.559 $\pm$ 0.088	1.2	0.294
#2	GBM2	774.3 $\pm$ 101.9	461.5 $\pm$ 29.6	0.546 $\pm$ 0.109	15.3	0.045
#3	GBM4	775.8 $\pm$ 109.1	449.7 $\pm$ 29.0	0.547 $\pm$ 0.101	26.1	0.051
#4	GBM3	778.3 $\pm$ 100.0	458.1 $\pm$ 28.9	0.542 $\pm$ 0.107	16.2	0.043
#5	DRF1	786.6 $\pm$ 100.2	444.8 $\pm$ 28.1	0.537 $\pm$ 0.089	195.3	0.042
#6	GBM5	791.9 $\pm$ 111.4	460.1 $\pm$ 30.3	0.523 $\pm$ 0.125	14.1	0.037
#7	GBM grid	814.2 $\pm$ 106.6	481.6 $\pm$ 35.4	0.491 $\pm$ 0.144	12.3	0.034
#8	XRT1	815.8 $\pm$ 107.2	492.9 $\pm$ 20.7	0.503 $\pm$ 0.089	179.3	0.044
#9	GBM1	837.1 $\pm$ 155.6	473.5 $\pm$ 28.4	0.490 $\pm$ 0.052	25.4	0.048
#10	DL1	999.5 $\pm$ 139.1	627.7 $\pm$ 34.9	0.265 $\pm$ 0.059	2.9	0.048
#11	GLM1	1 067.2 $\pm$ 172.8	680.8 $\pm$ 32.9	0.169 $\pm$ 0.019	12.6	0.026

values. Each instance of the model takes 15.3s to train, striking a balance between computational time and performance, especially relevant when cross-validation is used.

In terms of variable importance, the most important feature types are the EVCS features which account for 65.9% of importance, namely its specs, charges & tariffs, and region as can be seen in Table 5.10. Proximity variables follow, particularly the POIs and sociodemographic variables. These latter features account for practically all the density feature's importance. It is interesting to note how little importance the model gave to vehicles sold, PPP, and predicted demand variables, which together account for less than 2%. Proximity variables are more important at lower radius as can be seen in Table E.1, with a similar phenomenon for the density variables where importance is higher at lower granularity levels, as in Table E.2. The top five variables account for 50% of the importance, as can be seen in Table 5.10. It can be seen that the most important variables belong to the EVCS, especially its technical specifications. This is not only because they directly refer to the EVCS whose sales are being predicted, but also because they have little correlation with other variables, which is the same for most of the proximity and density variables. The visual representation of the relative importance of the variables can be seen in Figure E.5.

The variables' importance is supplemented by the SHAP, in order to analyse how the different levels of the variables map to the response, and if the model is over-fitting or not. As can be seen from Figure 5.11, there is a clear mapping from most variables to the response, especially for the *connector standard* and the *connector max power*. The mapping of *DT region CPO* (*station_party_id* in the chart) is not as clear, because despite having a significant impact on the response, these have many category levels (there are 18 DTs and 7 CPOs with market share above 4%) which might indirectly affect the sales. It is expected that an EVCS has higher sales due to its surrounding context, not the *DT* and *CPO* assuming there are no loyalty programs. However, it is possible that given CPOs choose better spots and tariffs than others, and that a categorical DT might model sales more accurately than the combination of its demographic variables according to a predefined function or mathematical operator.

Table 5.10: Most importance variables and groups of GBM model

Variables group	Variable types			Σ
	Density	Proximity	EVCS	
sociodemographic	12.4%	8.4%	—	20.8%
vehicles	0.2%	1.2%	—	1.4%
PPP	0.0%	0.0%	—	0.0%
predicted demand	0.4%	0.0%	—	0.4%
POIs	—	11.4%	—	11.4%
tech specs	—	—	32.0%	32.0%
charges & tariffs	—	—	15.0%	15.0%
regions	—	—	18.9%	19.0%
Σ	13.0%	21.1%	65.9%	100.0%

Feature	Imp.	Type	Source
<i>DT</i>	18.8%	EVCS	regions
<i>connector standard</i>	14.3%	EVCS	tech specs
<i>CPO</i>	9.0%	EVCS	tech specs
<i>linear reg. kWh w/ max power</i>	4.2%	EVCS	tariffs
<i>connector max power</i>	3.8%	EVCS	tech specs

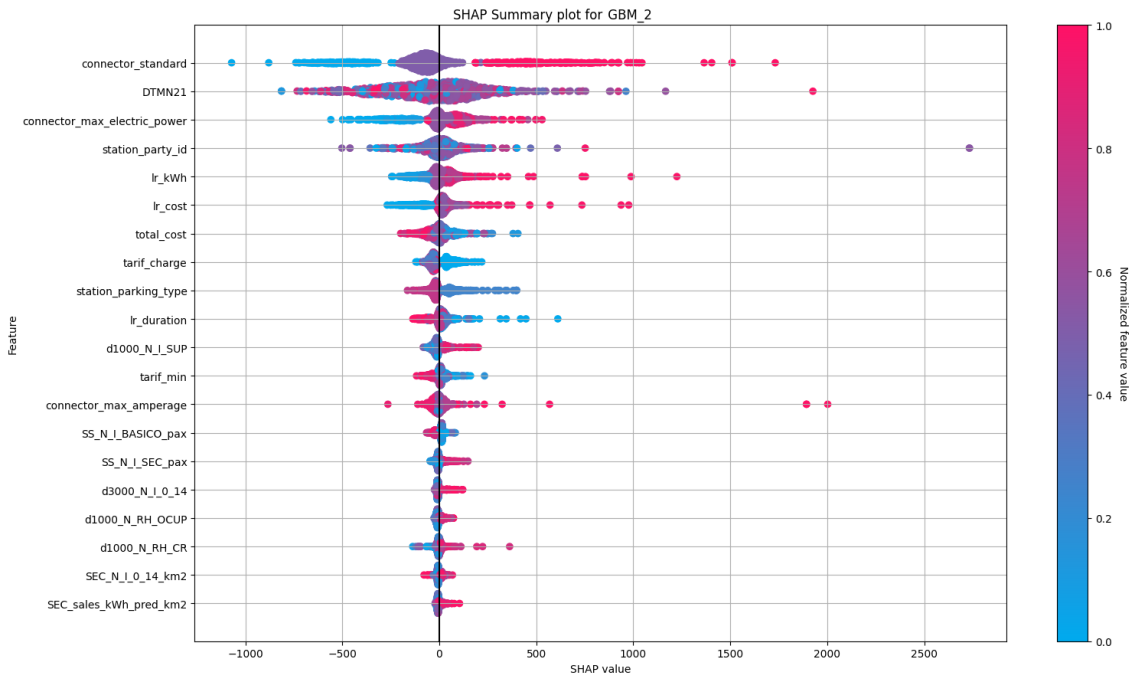


Figure 5.11: Average growth rate from 2023 to 2024 in Porto municipality

5.5 Demand capture

In this step, an attraction size of each station/connector along with the predicted demand per SS are used to capture demand to the EVCSs. Inspired by the law of universal gravitation $F = G \frac{m_1 m_2}{d^2}$, gravity models draw an analogy where, according to Reilly (1931) law of retail gravitation, the attraction force corresponds to sales/ share/ probability of a customer at a given supply point, while mass corresponds to the activity/size (e.g. sales area or revenue), effectively stating that market share increases with supply size and decreases with distance. It can be used to determine the ratio of attractiveness between two supply points and one demand point. Consider a household located at I , somewhere in the line between stores at A and B , it stands $\frac{T_A}{T_B} = \frac{W_A}{W_B} \left(\frac{d_B}{d_A} \right)^2$ where W_A and W_B are the sizes of A and B , T_A and T_B are the proportions of trade (market share for example) from I to A and B respectively, and d_A and d_B are the distances from A and B to I respectively. The point of indifference (or break point) can be obtained by setting $T_A = T_B$ and solving the equation for d_B , with known $d_{AB} = d_A + d_B$. This model was further generalized to more than two locations and other distance exponents by Lakshmanan and Hansen (1965), and Huff (1962). The fraction M_{ij} of the household i income spent in the supply point j is defined as

$$M_{ij} = \frac{W_j / \tau_{ij}^\beta}{\sum_k W_k / \tau_{ik}^\beta} \quad (5.3)$$

where τ_{ij} is the time or distance between i and j , and β is the distance exponent (generally 2, can be fine-tuned), according to Chan (2011). There are numerous variants of gravity models according to their application as well as the theoretical principles used.

With the goal of obtaining a simple and explainable model, gravity models were used due to their clear assumptions and computational performance. In particular, Huff's gravity model was used, with modifications to the distance function. A module was developed with inputs:

- **Size** as the variable used for supply point mass
- **Distance decay function** is equivalent to $1/\tau_{ik}^\beta$ in the original model. Its type corresponds to how attraction strength decays with distance, and can be inverse to distance raised to a power ("pow"), or inverse to the exponential of distance ("exp"). Both functions have three parameters, k_1 , k_2 , and k_3 . If using the "pow" function the distance factor is given by $(k_1 + (d + k_3)^{k_2})^{-1}$, while for the "exp" function the distance factor is given by $(k_1 + e^{k_2 \cdot (d^{k_3})})^{-1}$.

The model's results consist in experimenting with the input factors and parameter fine-tuning, namely size s , the two designated types of distance functions t , and the three parameters of the function k . The parameters tested, that can be found in Appendix F, were defined depending on the distance function. The best parameters were found for sizes $s = \{power, autoML\}$ and distance function $t \in T = \{pow, exp\}$. Firstly three default parameters were defined, and each parameter was changed individually. Then, the best parameters were selected based on the metrics (i) RMSE, (ii) MAPE, (iii) MAE, (iv) correlation of predicted with actual sales, and all combinations of

the selected parameters were tested. The model was run for the charging stations using their total maximum power as size $s = power$, and also for the connectors using their autoML cross-validation sales predictions as size $s = autoML$, with results in Tables 5.11 and 5.12 respectively.

Table 5.11: Huff's gravity model best results for $s = power$

t	Parameters			Quality measures			r^2	Correl. of predicted		time (s)
	k_1	k_2	k_3	RMSE	MAPE	MAE		w/ sales	w/ power	
pow	0	2	0	**24 721	5.09	15 004	**0.149	0.530	0.739	40.0
	0	2	1	25 309	**4.71	15 197	0.108	0.527	0.816	31.5
	0	1.5	0	24 907	4.86	**14 933	0.136	0.533	0.808	39.4
exp	0	0.1	0.5	*25 266	5.00	*15 800	*0.111	0.482	0.989	38.5
	0	0.1	1	26 214	*4.75	15 866	0.043	0.505	0.851	39.0
reference model				25 700	5.07	16 249	0.081	0.459	1.000	16.1

*best value for t per metric

**best value of metric

Table 5.12: Huff's gravity model best results for $s = autoML$

t	Parameters			Quality measures			r^2	Correl. of predicted		time (s)
	k_1	k_2	k_3	RMSE	MAPE	MAE		w/ sales	w/ power	
pow	0	0.5	100	10 351	4.403	**5 929	0.564	0.751	0.998	33.9
	0	2	100	10 670	*4.317	6 060	0.537	0.734	0.978	35.1
exp	0	0.1	0.1	10 308	4.435	*5 933	0.568	0.754	1.000	52.1
	1000	0.1	0.1	*10 305	4.439	5 936	*0.568	0.754	1.000	53.7
	1000	0.1	1	10 500	**4.301	6 058	0.552	0.743	0.983	47.4
reference model				**10 305	4.439	5 936	**0.568	0.754	1.000	52.6

*best value for t per metric

**best value of metric

Using the station power as size, the r^2 of the reference model (using only size as predictor) is 0.081 which, when compared to the obtained r^2 of up to 0.149, proves the importance of considering the distance between supply and demand, as can be seen in Table 5.11. The model with best RMSE and r^2 uses the inverse of the squared distance for its decay, which is usually the default exponent assumed in gravity models, while the "pow" function mostly outperforms the "exp" function. The results of the model $s = power$, $t = pow$ and $k = \{0, 2, 0\}$ are represented in Figure 5.12, with a bubble per EVCSs and a different colour for each EVCS trading area.

When using the autoML cross-validation sales predictions as size, the gravity model can marginally improve relative to the reference model in MAPE using the "exp" function, and in MAE using the "pow" function, as can be seen in Table 5.12. In the case of MAPE, it should be noted that the high values are due to low sales volume of some stations, which inflate the

score. Note that the reference model metrics are the same as the cross-validation from autoML, but computed for all observations simultaneously instead of the average per cross-validation fold.

Comparing the two sizes, it can be seen that incorporating distance from the gravity model when using power as the size of the stations improves the prediction of the sales, while using the sales predictions as the size of the connectors does not present improvements in the r^2 . This is because the autoML predictions already incorporated numerous distance features, as well as the reference model already explaining a significant portion of the sales with $r^2 = 0.568$. As such, the gravity model is especially useful when the previous autoML step is not performed, and the user wishes for faster testing of given features as size. It points that the prediction of sales with autoML (step 3) and the use of gravity model (step 4) should be alternatives, rather than sequential.

It should be noted that the gravity model outputs the market share of each demand captured by each station/connector, while the autoML model does not (but could, if what was predicted was the market share within a given radius instead of the actual sales. Therefore, if the focus is to maximize performance, the autoML model should be used, given that it outputs overall better predictions than the gravity model. On the other hand, if the focus is on insightful visualizations, the gravity model can be applied to raw EVCS technical characteristics, in this case the maximum power, to improve predictions by using distance besides only power, as well as to estimate the trade area of each EVCS and visualize it.

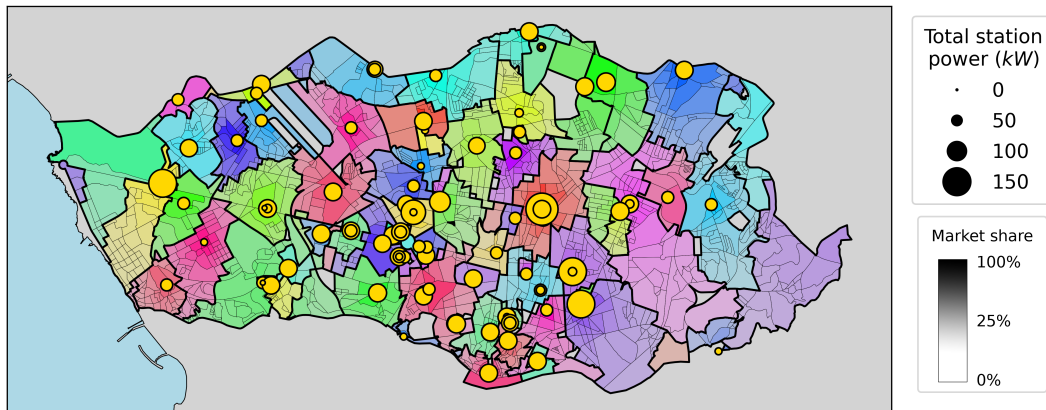


Figure 5.12: Predicted EVCS trade area in Porto municipality

Although the model's absolute performance may seem underwhelming, it proved capable of incorporating distance in continuous space, yielding better results than if it had not, especially when the initial predictor of sales is weak. The main advantage of this model is that it is much more transparent than a machine learning model, despite its lower performance. From a computational perspective, running the model again in the vicinity of the grid changes instead of running it for the whole grid allows significant computation time reduction, which in turn improves the user experience, as well as the automatic or manual test of scenarios.

Chapter 6

Conclusion

The objective of this work was to determine optimal locations for EVCSs to maximize profitability. To do so, there was the need for accurate demand pinpointing in space and forecasting in time, as well as understanding how street/parking lot EVCSs capture demand in order to prioritize areas with high demand and limited supply. This is in the context of unknown competitor decision rules.

The general developed methodology proved capable of tackling a particular location problem, for which a particular implementation was put to practice. Demand was disaggregated and forecasted with good accuracy, pointing for the potential of the new vehicles sold as well as the number of inhabitants with college degree to model and disaggregated demand, as well as for the importance of considering not only an average growth rate (like a CAGR), but also how fast this growth rate is changing, in order to understand where demand is booming and shrinking. The sales forecasting and demand capture models proved as good alternatives, each with different strong points. The first revealed suitable performance in predicting charging station sales, while the second estimates the trade area of each EVCS. Throughout the data flow process in the implementation developed, a set of insightful visualizations allow for not only the visualization of the data, but for decisions to be made, given the focus on simplicity of representation and straightforward indicators.

Relative to the methodology developed prior to this work, the proposed methodology represents an improvement in its formalization, as well as additional guidelines that are fundamental in improving its robustness, as well as the additional steps added, namely demand forecasting and demand capture. As for the implementation, it successfully increased the independency of the tool from private data, replacing CPO sales with publicly available utilization data on all EVCS, and replacing the private POIs dataset with OpenStreetMap POIs. The CPO data was still relevant in estimating the time series parameters. The utilization data collection is still underway, which will enable to compute the time series once a sufficient period of data is collected (at least one year), therefore dismissing the need for CPO data. Nonetheless, it is important to validate the retrieved utilization data with sales from a particular CPO, which was not possible in this work given that the time interval of CPO sales and the time interval of utilization data do not intersect.

Given the modularity of the developed methodology, it is possible to add new steps to it.

The current model allows for demand disaggregation, forecast, and capture, therefore a location recommendation step could be further formalized in a continuation of this work. This step could locate the actual new EVCSs to be installed based on their number and specifications. This could be done using a metaheuristic or other iterative methods, that would use the sales predictions to locate the new stations.

As for the implementation, in the demand disaggregation module ML methods can be used instead of the elastic net to improve performance, considering that the method should not predict negative demand, as stated in the general methodology guidelines. In demand forecasting, other time series techniques such as exponential smoothing or ARIMA/ARIMAX can be used, while verifying if the chosen technique suits the hierarchical time series forecasting, enabling the correct forecast of sales at the different granularity levels. In sales forecasts, the autoML model could be further improved by using the competition variables other than only the number of EVCSs in the vicinity (such as number of higher power EVCSs or from the same CPO), and predict market share instead of straight out sales. For demand capture, alternative models to the gravity model can be used to account for capacity, such as location-allocation models with hard constraints, or user-equilibrium models accounting for the average waiting time in queue to charge during peak hours.

This work had significant and real impact in the actual market, given its deployment in a CPO to support decision-making, with potential to develop a new location intelligence product by adding new techniques to each step. Relative to the rather limited performance of the most common methods, the proposed general methodology expands the range of tools available for tackling spatial location problems, with a robust and sequential set of interchangeable steps.

Bibliography

- Abnett, K. (2023). EU lawmakers approve effective 2035 ban on new fossil fuel cars. *Reuters*.
- Ahmad, F., Iqbal, A., Ashraf, I., Marzband, M., and Khan, I. (2022). Optimal location of EVCS and its impact on distribution network: A review. *Energy Reports*, 8:2314–2333.
- Alhazmi, Y. A., Mostafa, H. A., and Salama, M. M. (2017). Optimal allocation for electric vehicle charging stations using Trip Success Ratio. *IJEP&ES*, 91:101–116.
- APA (2019). Roadmap for Carbon Neutrality 2050. Technical report, APA.
- Bakker, S., Leguijt, P., and Van Lente, H. (2015). Niche accumulation and standardization – the case of electric vehicle recharging plugs. *Journal of Cleaner Production*, 94:155–164.
- Banks, J., editor (2001). *Discrete-event system simulation*. Prentice-Hall international series in industrial and systems engineering. Prentice Hall, Upper Saddle River, NJ, 3rd ed edition.
- Bi, R., Xiao, J., Pelzer, D., Ciechanowicz, D., Eckhoff, D., and Knoll, A. (2017). A simulation-based heuristic for city-scale EVCS placement. In *2017 ITSC*, pages 1–7, Yokohama. IEEE.
- Bicknell, C. (1913). *A guide to the prehistoric rock engravings in the Italian Maritime Alps / C. Bicknell*. Printed by G. Bessone, Bordighera.
- Chan, Y. (2011). *Location Theory and Decision Analysis: Analytics of Spatial Information Technology*. Springer, Heidelberg, 2nd edition.
- Chen, X., Lei, Z., and Ukkusuri, S. V. (2022a). Prediction of Road-level Energy Consumption of Battery Electric Vehicles. In *2022 ITSC*, pages 2550–2555, Macau, China. IEEE.
- Chen, X., Ye, X., Widener, M. J., Delmelle, E., Kwan, M.-P., Shannon, J., Racine, E. F., Adams, A., Liang, L., and Jia, P. (2022b). A systematic review of the modifiable areal unit problem (MAUP) in community food environmental research. *Urban Informatics*, 1(1):22.
- Christaller, W. (1933). *Die zentralen Orte in Süddeutschland*. Wissenschaftliche Buchgesellschaft, Darmstadt.
- Church, R. L. and Murray, A. T. (2009). *Business site selection, location analysis, and GIS*. John Wiley & Sons, Hoboken, N.J.
- Conzade, J., Nägele, F., Ramanathan, S., and Schaufuss, P. (2022). Europe’s EV opportunity—and the charging infrastructure needed to meet it.
- Cristino, J. (2023). Carbono: o princípio da transição. *Visão*.

- Dent, B. D., Torguson, J., and Hodler, T. W. (2009). *Cartography: thematic map design*. McGraw-Hill Higher Education, New York, 6th ed edition.
- Denton, T. (2020). *Electric and hybrid vehicles*. Routledge, Abingdon, Oxon ; New York, NY, 2nd edition edition.
- Dong, G., Ma, J., Wei, R., and Haycox, J. (2019). Electric vehicle charging point placement optimisation by exploiting spatial statistics and maximal coverage location models. *TR_D*, 67:77–88.
- E-REDES (2020a). Definitive premises - Low voltage | E-REDES. URL: <https://www.e-redes.pt/en/definitive-premises-low-voltage>.
- E-REDES (2020b). Medium Voltage | E-REDES. URL: <https://www.e-redes.pt/en/companies/business/average-voltage>.
- EAFO (2022). Vehicles and fleet | Portugal | European Alternative Fuels Observatory. URL: <https://alternative-fuels-observatory.ec.europa.eu/transport-mode/road/portugal/vehicles-and-fleet>. Source: EV Volumes (ev-volumes.com).
- Gao, J. (2022). *Fundamentals of spatial analysis and modelling*. CRC Press, Boca Raton, 1st edition.
- Geman, S., Bienenstock, E., and Doursat, R. (1992). Neural Networks and the Bias/Variance Dilemma. *Neural Computation*, 4(1):1–58.
- Grigorev, A., Mao, T., Berry, A., Tan, J., Purushothaman, L., and Mihaita, A.-S. (2021). How will electric vehicles affect traffic congestion and energy consumption: an integrated modelling approach. In *2021 ITSC*, pages 1635–1642, Indianapolis, IN, USA. IEEE.
- Haining, R. (1990). *Spatial Data Analysis in the Social and Environmental Sciences*. Cambridge University Press, 1 edition.
- Hardman, S., Jenn, A., Tal, G., Axsen, J., Beard, G., Daina, N., Figenbaum, E., Jakobsson, N., Jochem, P., Kinnear, N., Plötz, P., Pontes, J., Refa, N., Sprei, F., Turrentine, T., and Witkamp, B. (2018). A review of consumer preferences of and interactions with electric vehicle charging infrastructure. *TR_D*, 62:508–523.
- Henley, S. (1981). *Nonparametric Geostatistics*. Springer Netherlands, Dordrecht.
- Huang, Y. and Kockelman, K. M. (2020). Electric vehicle charging station locations: Elastic demand, station congestion, and network equilibrium. *TR_D*, 78:102179.
- Huff, D. L. (1962). *Determination of intra-urban retail trade areas*. University of California.
- INE (2019). Poder de compra per capita. URL: https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_indicadores&indOcorrCod=0008614.
- INE (2022a). BGRI 2021. URL: <https://mapas.ine.pt/download/index2021.phtml>.
- INE (2022b). Veículos novos vendidos. URL: https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_indicadores&indOcorrCod=0008464.
- Johnston, R. J. and Semple, R. K. (1983). *Classification using information statistics*. Number 37 in Concepts and techniques in modern geography. Geo Books, Norwich [Norfolk].

- Kemp, K. K., editor (2008). *Encyclopedia of geographic information science*. SAGE Publications, Los Angeles.
- Kizhakkan, A. R., Rathore, A. K., and Awasthi, A. (2019). Review of Electric Vehicle Charging Station Location Planning. In *2019 ITEC-India*, pages 1–5, Bengaluru, India. IEEE.
- Kraak, M. J. and Ormeling, F. (2010). *Cartography: visualization of geospatial data*. Prentice Hall, Harlow ; New York, 3rd ed edition.
- Kwan, M.-P. (2012). The Uncertain Geographic Context Problem. *Annals of the Association of American Geographers*, 102(5):958–968.
- Lakshmanan, J. R. and Hansen, W. G. (1965). A retail market potential model. *JAPA*, 31(2):134–143.
- Liu, H.-C., Yang, M., Zhou, M., and Tian, G. (2019). An Integrated Multi-Criteria Decision Making Approach to Location Planning of Electric Vehicle Charging Stations. *T-ITS*, 20(1):362–373.
- Longley, P. (2015). *Geographic information science & systems*. Wiley, Hoboken, NJ, fourth edition edition.
- Luo, R., Zhang, Y., Zhou, Y., Chen, H., Yang, L., Yang, J., and Su, R. (2021). Deep Learning Approach for Long- Term Prediction of Electric Vehicle (EV) Charging Station Availability. In *2021 ITSC*, pages 3334–3339, Indianapolis, IN, USA. IEEE.
- Ma, S., Jiang, M., Tao, P., Song, C., Wu, J., Wang, J., Deng, T., and Shang, W. (2018). Temperature effect and thermal impact in lithium-ion batteries: A review. *Progress in Natural Science: Materials International*, 28(6):653–666.
- Marsh, M., Alagona, P. S., and Alagona, P. S. (2008). *AP human geography*. Barron's Educational Series, Hauppauge, N. Y, 2nd ed edition.
- Miller, H. J. (2004). Tobler's First Law and Spatial Analysis. *Annals of the Association of American Geographers*, 94(2):284–289.
- MOBIE (2022a). Finding charging points. URL: <https://mobie.pt/redemobie/encontrar-posto>.
- MOBIE (2022b). MOBIDATA - Mobility, Portugal. URL: <https://mobie.pt/en/mobidata/data>.
- Mostyn, M. (2021). EV Charging Connectors: Plugs Types and Speeds Explained | OVO Energy.
- Niccolai, A., Bettini, L., and Zich, R. (2021). Optimization of electric vehicles charging station deployment by means of evolutionary algorithms. *Int. Jour. Intel. Syst.*, 36(9):5359–5383.
- Nicolas, C., Paillet, Y., Stéphan, P., Pierson, J., Aubry, L., Gall, B. L., Lacombe, V., and Rollet, J. (2021). La carte et le territoire : la dalle gravée du Bronze ancien de Saint-Bélec (Leuhan, Finistère). *Bulletin de la Société préhistorique française*, 118(1):99.
- OpenStreetMap contributors (2023). Amenities retrieved from Overpass turbo. URL: <https://overpass-turbo.eu/>.
- Pan, L., Yao, E., Yang, Y., and Zhang, R. (2020). A location model for electric vehicle (EV) public charging stations based on drivers' existing activities. *SCS*, 59:102192.

- Pereira, T. C., Amaro, A., Borges, M., Silva, R., Seabra, T., and Canaveira, P. (2022). Portuguese National Inventory Report on Greenhouse Gases, 1990-2020. Technical report, APA, Amadora, Portugal.
- Reilly, W. J. (1931). *The law of retail gravitation*. Knickerbocker Press, New York.
- Ren, X., Zhang, H., Hu, R., and Qiu, Y. (2019). Location of electric vehicle charging stations: A perspective using the grey decision-making model. *Energy*, 173:548–553.
- Rodrigues, J. V. (2023). Novos postos de carregamento rápido e ultrarrápido mais do que duplicaram em 2022. URL: <https://www.dinheirovivo.pt/empresas/novos-postos-de-carregamento-rapido-e-ultrarrapido-mais-do-que-duplicaram-em-2022-15703646.html>.
- Snow, J. (1855). On the mode of communication of cholera.
- Steiner, D. (2023). 1.4 Location Intelligence | GEOG 850: Location Intelligence for Business. Course notebook.
- Sugumaran, R. and DeGroote, J. (2010). *Spatial Decision Support Systems*. CRC Press, Boca Raton.
- Thrift, N. J. and Kitchin, R. (2009). *International encyclopedia of human geography*. Elsevier, Amsterdam Boston.
- Tobler, W. (1999). Linear pycnophylactic reallocation comment on a paper by D. Martin. *International Journal of Geographical Information Science*, 13(1):85–90.
- Tobler, W. (2004). On the First Law of Geography: A Reply. *Annals of the Association of American Geographers*, 94(2):304–310.
- Tobler, W. R. (1965). Computation of the correspondence of geographical patterns. *Papers of the Regional Science Association*, 15(1):131–139.
- Tobler, W. R. (1970). A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography*, 46:234.
- Wang, Q., Zhang, D., and Du, B. (2022). A Queue Balancing Approach for Electric Vehicle Charging Allocation. In 2022 ITSC, pages 2750–2755, Macau, China. IEEE.
- Weber, A. and Friedrich, C. J. (1929). *Alfred Weber's theory of the location of industries*. The University of Chicago Press, Chicago.
- Wooler, A. (2020). How Location Intelligence Is Powering Today's Most Disruptive Products. Section: Innovation.
- Wu, F. and Sioshansi, R. (2017). A stochastic flow-capturing model to optimize the location of fast-charging stations with uncertain electric vehicle flows. *TR_D*, 53:354–376.
- Yong, J. Y., Ramachandramurthy, V. K., Tan, K. M., and Mithulananthan, N. (2015). A review on the state-of-the-art technologies of electric vehicle, its impacts and prospects. *Renewable and Sustainable Energy Reviews*, 49:365–385.
- Yong, T. and Park, C. (2017). A qualitative comparative analysis on factors affecting the deployment of electric vehicles. *Energy Procedia*, 128:497–503.

- Zhou, G., Zhu, Z., and Luo, S. (2022). Location optimization of electric vehicle charging stations: Based on cost model and genetic algorithm. *Energy*, 247:123437.
- Zou, H. and Hastie, T. (2005). Regularization and Variable Selection Via the Elastic Net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320.

Appendix A

Mapping techniques

The quality of insights generated from data and models is only as good as the quality of the chart used to visualize them. A bad chart can undermine the quality of even the best data and models, causing confusion and misinterpretation. In location intelligence, this is even more important, given the added complexity of representing spatial data on a map. Choosing the right map type and following design best-practices principles are critical for accurately conveying insights and realizing the full potential of the data and results.

There are a variety of map classifications introduced by many authors. The most used map types were systematized by purpose followed by data type in Figure A.1, as a review of Kraak and Ormeling (2010), Dent et al. (2009), and Longley (2015). Thematic maps are the most suitable for this work, considering their flexibility and well as their ability to provide valuable insights in a business context. The objective of this subsection is to present a non-exhaustive overview of all possible map types to consider when representing spatial data, therefore further discussion on the details of each map type and best practices is available in the aforementioned works.

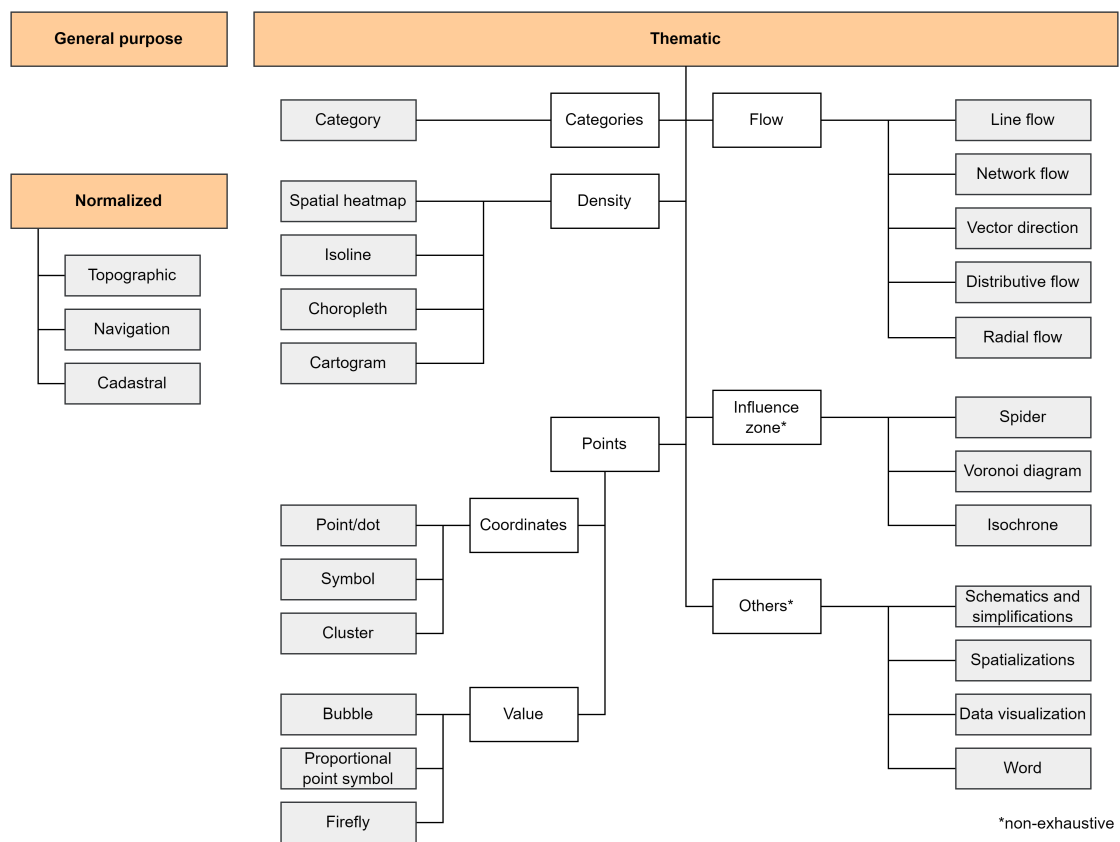


Figure A.1: Map types by purpose and data type

Appendix B

Detailed variables

The demographic variables are all in counts per sub-section (SS), and come from census each 10 years. Therefore only the latest (from 2021) were used, and can be seen in B.2. The vehicles data, consisting of the new car sales per municipality (MN), is available every year, with values for 2019 shown in B.2.

The PPP data consists of the PPP per capita per municipality (MN) relative to the national per capita average. The data used is from years 2004, and from 2005-2019 every two years. For the EVCSs, the location and specifications data is collected daily, and encompasses the variables in Table B.1. The tariffs table is also collected daily to calculate the value of the chargers. Every minute, the status data is collected, and at the end of the day it is condensed, and the tariffs of the day are used to estimate the value of the charges.

As for the points-of-interest, the property of interest is the *amenity*. The types and groups are shown in Table B.3. The type is the property itself, and is usually standardized, despite existing some that are not (those in the *Unknown* group). The groups are defined loosely by OpenStreetMap, and are not standardized. The percentage corresponds to the fraction of POIs within that group that have a given type.

Table B.1: EVCS variables

	Variable	Metric	Value
Location	station_address	mode	Av. Dr. Óscar Lopes
	station_city	mode	Lisboa
	station_postal_code	mode	2614-520
	station_parking_type	options	parking lot (54%), street (27%), motorway (9%), ...
	station_latitude	—	—
	station_longitude	—	—
Specifications	station_party_id	options	KLC (18%), EDP (17%), HRZ (15%), GLP (15%), ...
	station_mobie_voltage_level	options	MV (46%), SLV (31%), LV (23%)
	connector_standard	options	Mennekes (66%), CCS (19%), CHAdeMO (16%)
	connector_format	options	SOCKET (55%), CABLE (45%)
	connector_power_type	options	AC 1 ϕ (15%), AC 3 ϕ (50%), DC (34%)
	connector_max_voltage	$\mu \pm \sigma$	441 $\pm$ 163
	connector_max_amperage	$\mu \pm \sigma$	72 $\pm$ 69
	connector_max_electric_power	$\mu \pm \sigma$	35294 $\pm$ 34721
Status	connector_id	—	—
	equipment_status	options	available, charging, unknown, out-of-order
	timestamp	range	2023/06/01 to 2023/06/14

Table B.2: Demographic variables

	Variable	Sum
Population	N_INDIVIDUOS	9 855 909
	N_INDIVIDUOS_H	4 687 055
	N_INDIVIDUOS_M	5 168 854
	N_INDIVIDUOS_0_14	1 264 697
	N_INDIVIDUOS_15_24	1 031 659
	N_INDIVIDUOS_25_64	5 225 083
	N_INDIVIDUOS_65_OU MAIS	2 334 470
	N_IND_ENS_BASICOM_COMPLETO	4 733 882
	N_IND_ENS_SEC_OU_POS_COMPLETO	2 127 241
	N_IND_ENS_SUPERIOR_COMPLETO	1 717 145
	N_IND_EMPREGADOS	4 220 423
	N_IND_REFORMADOS	2 297 887
Buildings	N_EDIF_CLASS	3 381 968
	N_EDIF_CLASS_CONST_1_OU_2_ALOJ	3 089 482
	N_EDIF_CLASS_CONST_3_OU MAIS_ALOJ	286 077
	N_EDIF_EXCLUSIV_RESID	3 356 711
	N_EDIF_1_OU_2_PISOS	2 822 094
	N_EDIF_3_OU MAIS_PISOS	559 874
	N_EDIF_CONSTR_ANTES_1945	426 208
	N_EDIF_CONSTR_1946_1980	1 282 585
	N_EDIF_CONSTR_1981_2000	1 075 366
	N_EDIF_CONSTR_2001_2010	494 277
	N_EDIF_CONSTR_2011_2021	103 532
	N_EDIF_NECESS_REPARACAO	1 204 971
Households & Housing units	N_ALOJ_TOTAL	5 736 759
	N_ALOJ_FAMILIARES	5 730 434
	N_ALOJ_FAM_CLASS_RHABITUAL	3 962 715
	N_ALOJ_FAM_CLASS_VAGOS_OU_RESID_SEC	1 763 766
	N_RHABITUAL_ACESS_CADEIRAS_RODAS	1 343 452
	N_RHABITUAL_COM_ESTACIONAMENTO	2 276 444
	N_RHABITUAL_PROP_OCUP	2 766 166
	N_RHABITUAL_ARRENDADOS	891 434
	N_AGREGADOS_DOMESTICOS_PRIVADOS	3 968 951
	N_ADP_1_OU_2_PESSOAS	2 318 739
	N_ADP_3_OU MAIS_PESSOAS	1 650 212
	N_NUCLEOS_FAM	2 983 280
	N_NUCLEOS_FAM_FILHOS_MENORES_25	1 415 039
Vehicles*	Total	166 373
	Ligeiros	155 516
	Ligeiros de passageiros	129 115
	Ligeiros de mercadorias	26 402
	Pesados	4 052
	Pesados de passageiros	369
	Pesados de mercadorias	1 204
	Tratores de espécie diversa	2 483
	Tratores agrícolas	6 790

*year 2019

Table B.3: Point-of-interest groups and types

Group (type)	# (%)		Group (type)	# (%)
<i>Education</i>	9 084		<i>Sustenance</i>	35 267
school	67.8%		restaurant	45.8%
kindergarten	12.5%		cafe	35.4%
library	7.8%		bar	10.0%
university	3.7%		fast_food	4.3%
driving_school	3.5%		pub	3.3%
<i>Entertain., Art & Cult.</i>	6 185		<i>Transportation</i>	66 183
fountain	56.7%		parking	68.6%
community_centre	15.5%		parking_space	11.1%
theatre	8.4%		fuel	5.5%
nightclub	4.7%		charging_station	5.0%
arts_centre	4.0%		<i>Waste Management</i>	17 315
social_centre	3.1%		recycling	52.4%
<i>Facilities</i>	21 532		waste_disposal	27.9%
bench	51.1%		waste_basket	18.8%
drinking_water	22.1%		<i>Others</i>	18 101
toilets	11.1%		place_of_worship	74.0%
shelter	6.9%		vending_machine	9.6%
telephone	4.0%		public_building	4.9%
<i>Financial</i>	5 621		marketplace	4.4%
bank	60.2%		<i>Unknown</i>	1 539
atm	39.2%		lavoir	41.4%
<i>Healthcare</i>	8 499		ticket_validator	10.0%
pharmacy	37.7%		prep_school	5.0%
clinic	25.7%		vacuum_cleaner	4.2%
social_facility	13.7%		letter_box	4.1%
dentist	7.9%			
veterinary	5.8%			
hospital	4.6%			
doctors	3.9%			
<i>Public Service</i>	6 093			
townhall	36.2%			
police	19.6%			
post_office	16.6%			
post_box	11.4%			
fire_station	10.0%			
courthouse	4.5%			

only the most frequent types of each group are shown

Appendix C

Disaggregation elastic net model

The n_jobs was fixed at 1 and changed for each, followed by looking to combine the results of the previous analysis, as can be seen in Table C.1.

Table C.1: Computational performance experiments of the elastic net model

n_jobs		CPU usage	Average time per combination (s)	Improvement*
ENCV	CVscore			
1	1	20%	12.8	0%
1	2	50%	10.2	20%
1	4	70%	6.9	46%
1	8	85%	4.6	64%
1	16	95%	3.6	72%
2	1	50%	12.0	6%
4	1	60%	10.4	19%
8	1	70%	12.2	5%
16	1	70%	12.2	5%
4	16	95%	4.0	69%
2	16	95%	3.7	71%
4	4	80%	6.7	48%

*relative to the first combinations of n_jobs

The parameters tested for the elastic net can be seen in Table C.2. These were chosen in order to explore the whole range of effects that the parameters could have on the distance function.

Table C.2: Tested parameter values for elastic net

Parameter	Values
ℓ_1	{0, 0.1, 0.25, 0.5, 0.75, 0.9, 0.99, 1}
α	{0, 0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000, 10000}
n	{1, 10, 30, 100, 200, 330}

In Table C.3, one can find the linear regression model performance, for comparison with the best of each n .

These are the complete results of the elastic net model for the disaggregation of 2023 demand. It presents the mean (μ) and standard deviation (σ) of the mean absolute deviation (MAD), root

Table C.3: Best elastic net models per n

n	Linear Regression r^2 ($\mu \pm \sigma$)	
1	$0.678 \pm$	0.197
10	$0.677 \pm$	0.199
30	$0.695 \pm$	0.183
100	$0.709 \pm$	0.190
200	$0.699 \pm$	0.187
330	$0.699 \pm$	0.190
for comparison		

mean squared error (RMSE) and coefficient of determination (r^2), as well as the time to cross-validate each combination with 10-fold and 10 repetitions Figures C.1 to C.6.

The model shows a good fit, as can be seen in Figure C.7. However, it should be noted that the existence of very high values of the dependent variable (20 observations of 3160 not shown in the charts) can lead r^2 to a good value, but with bad predictions given that these high values have higher importance for the model. The first 278 observations are MN while the remaining are FR, with their split being visible in Figure C.8 (a). Besides, FRs of the same MN are next to each other in the dataset, so clusters of FRs with differing residuals' variance can be seen, probably sharing the same MN. When the size of an MN increases, so does the size of its FRs, so FRs in bigger MNs will have higher values, and therefore higher residuals. The facts conflict with the assumption of independence between observations from linear regression. Despite some extreme residuals, the histogram in Figure C.8 (b) shows that the distribution of the residuals is centered in zero and is not skewed.

Δt (s)		α									
		0	0.0001	0.001	0.01	0.1	1	10	100	1000	10000
e_1	0.00	0.128	0.040	0.036	0.038	0.037	0.036	0.032	0.033	0.033	0.032
	0.10	0.033	0.019	0.020	0.022	0.023	0.021	0.020	0.021	0.020	0.019
	0.25	0.034	0.020	0.022	0.021	0.020	0.022	0.020	0.019	0.019	0.020
	0.50	0.033	0.019	0.020	0.019	0.019	0.019	0.022	0.020	0.019	0.019
	0.75	0.033	0.019	0.019	0.019	0.019	0.019	0.019	0.021	0.019	0.019
	0.90	0.033	0.020	0.019	0.019	0.019	0.020	0.019	0.019	0.022	0.019
	0.99	0.032	0.019	0.019	0.019	0.019	0.019	0.019	0.019	0.022	0.019
	1.00	0.033	0.019	0.019	0.019	0.019	0.019	0.020	0.019	0.022	0.019
μ_{MAD}		α									
		0	0.0001	0.001	0.01	0.1	1	10	100	1000	10000
e_1	0.00	60,376	60,376	60,374	60,382	63,729	101,898	153,407	164,585	165,829	165,954
	0.10	60,376	60,376	60,374	60,378	63,125	98,986	152,165	164,433	165,813	165,953
	0.25	60,376	60,376	60,374	60,375	62,248	94,085	149,775	164,131	165,782	165,950
	0.50	60,376	60,376	60,375	60,372	61,109	84,439	143,276	163,231	165,689	165,940
	0.75	60,376	60,376	60,376	60,372	60,525	72,302	127,832	160,609	165,412	165,913
	0.90	60,376	60,376	60,376	60,374	60,382	63,729	101,899	153,409	164,587	165,831
	0.99	60,376	60,376	60,376	60,376	60,374	60,382	63,730	101,908	153,429	164,609
	1.00	60,376	60,376	60,376	60,376	60,376	60,376	60,376	60,376	60,377	60,494
σ_{MAD}		α									
		0	0.0001	0.001	0.01	0.1	1	10	100	1000	10000
e_1	0.00	12,995	12,997	13,015	13,240	16,565	31,437	32,366	31,648	31,559	31,549
	0.10	12,995	12,996	13,013	13,213	16,156	30,955	32,436	31,659	31,560	31,550
	0.25	12,995	12,996	13,010	13,173	15,565	29,993	32,557	31,680	31,562	31,550
	0.50	12,995	12,996	13,005	13,109	14,634	27,208	32,822	31,743	31,569	31,551
	0.75	12,995	12,995	13,000	13,049	13,713	21,964	33,062	31,922	31,589	31,553
	0.90	12,995	12,995	12,997	13,015	13,240	16,565	31,437	32,366	31,648	31,559
	0.99	12,995	12,995	12,995	12,997	13,015	13,240	16,566	31,442	32,370	31,552
	1.00	12,995	12,995	12,995	12,995	12,995	12,995	12,995	12,998	13,035	13,534
μ_{RMSE}		α									
		0	0.0001	0.001	0.01	0.1	1	10	100	1000	10000
e_1	0.00	229,751	229,767	229,911	231,419	248,691	355,180	483,051	509,912	512,887	513,187
	0.10	229,751	229,765	229,894	231,246	246,768	347,590	480,062	509,549	512,850	513,184
	0.25	229,751	229,763	229,870	230,988	243,852	334,819	474,273	508,825	512,776	513,176
	0.50	229,751	229,759	229,831	230,566	238,953	308,895	458,300	506,674	512,554	513,155
	0.75	229,751	229,755	229,791	230,153	234,139	274,859	420,238	500,400	511,891	513,089
	0.90	229,751	229,753	229,767	229,911	231,419	248,692	355,183	483,057	509,918	512,893
	0.99	229,751	229,751	229,753	229,767	229,911	231,419	248,696	355,212	483,114	509,980
	1.00	229,751	229,751	229,751	229,751	229,751	229,751	229,754	229,781	230,059	233,195
σ_{RMSE}		α									
		0	0.0001	0.001	0.01	0.1	1	10	100	1000	10000
e_1	0.00	90,877	90,893	91,046	92,839	126,085	294,280	365,474	373,572	374,380	374,461
	0.10	90,877	90,892	91,028	92,616	121,829	287,138	364,475	373,472	374,370	374,460
	0.25	90,877	90,889	91,003	92,293	115,498	273,874	362,482	373,273	374,350	374,458
	0.50	90,877	90,885	90,960	91,783	105,435	241,117	356,536	372,674	374,291	374,452
	0.75	90,877	90,881	90,918	91,311	96,806	182,788	339,240	370,876	374,112	374,435
	0.90	90,877	90,878	90,893	91,046	92,839	126,086	294,284	365,479	373,577	374,385
	0.99	90,877	90,877	90,878	90,893	91,046	92,839	126,095	294,326	365,533	373,632
	1.00	90,877	90,877	90,877	90,877	90,877	90,877	90,880	90,908	91,197	95,203
μ_2		α									
		0	0.0001	0.001	0.01	0.1	1	10	100	1000	10000
e_1	0.00	0.678	0.678	0.678	0.679	0.682	0.534	0.122	0.006	0.007	0.009
	0.10	0.678	0.678	0.678	0.679	0.682	0.552	0.135	0.008	0.007	0.009
	0.25	0.678	0.678	0.678	0.679	0.682	0.578	0.158	0.011	0.007	0.009
	0.50	0.678	0.678	0.678	0.678	0.682	0.625	0.220	0.021	0.006	0.009
	0.75	0.678	0.678	0.678	0.678	0.680	0.668	0.354	0.048	0.003	0.008
	0.90	0.678	0.678	0.678	0.678	0.679	0.682	0.534	0.122	0.006	0.007
	0.99	0.678	0.678	0.678	0.678	0.679	0.679	0.682	0.534	0.122	0.006
	1.00	0.678	0.678	0.678	0.678	0.678	0.678	0.678	0.678	0.678	0.680
σ_2		α									
		0	0.0001	0.001	0.01	0.1	1	10	100	1000	10000
e_1	0.00	0.197	0.197	0.197	0.195	0.178	0.140	0.044	0.019	0.019	0.019
	0.10	0.197	0.197	0.197	0.195	0.179	0.142	0.048	0.019	0.019	0.019
	0.25	0.197	0.197	0.197	0.196	0.182	0.145	0.054	0.019	0.019	0.019
	0.50	0.197	0.197	0.197	0.196	0.187	0.150	0.072	0.020	0.019	0.019
	0.75	0.197	0.197	0.197	0.197	0.192	0.160	0.107	0.024	0.019	0.019
	0.90	0.197	0.197	0.197	0.197	0.195	0.178	0.140	0.044	0.019	0.019
	0.99	0.197	0.197	0.197	0.197	0.197	0.195	0.178	0.140	0.044	0.019
	1.00	0.197	0.197	0.197	0.197	0.197	0.197	0.197	0.197	0.197	0.193

Figure C.1: Elastic net μ and σ of error measures from cross-validation for 1 variable

Δt (s)		α									
		0	0.0001	0.001	0.01	0.1	1	10	100	1000	10000
e_1	0.00	0.061	0.059	0.059	0.064	0.065	0.070	0.068	0.070	0.078	0.069
	0.10	0.065	0.060	0.061	0.062	0.046	0.032	0.033	0.031	0.031	0.031
	0.25	0.066	0.060	0.061	0.062	0.048	0.034	0.032	0.032	0.031	0.031
	0.50	0.066	0.063	0.062	0.061	0.063	0.034	0.031	0.032	0.031	0.031
	0.75	0.070	0.064	0.067	0.067	0.069	0.041	0.034	0.034	0.033	0.032
	0.90	0.068	0.063	0.061	0.062	0.063	0.043	0.032	0.031	0.031	0.031
	0.99	0.066	0.067	0.062	0.063	0.061	0.062	0.037	0.033	0.034	0.031
	1.00	0.068	0.062	0.062	0.061	0.062	0.067	0.062	0.061	0.055	0.034
μ_{MAD}		α									
		0	0.0001	0.001	0.01	0.1	1	10	100	1000	10000
e_1	0.00	60,253	60,081	59,614	59,829	61,457	64,201	102,501	153,441	164,589	165,829
	0.10	60,253	60,097	59,613	59,768	61,453	63,604	99,654	152,205	164,437	165,814
	0.25	60,253	60,121	59,622	59,672	61,327	62,832	94,788	149,821	164,136	165,783
	0.50	60,253	60,162	59,692	59,484	61,089	62,033	85,018	143,329	163,240	165,692
	0.75	60,253	60,205	59,889	59,616	60,560	61,710	72,845	127,973	160,630	165,420
	0.90	60,253	60,233	60,081	59,614	59,829	61,457	64,202	102,510	153,461	164,611
	0.99	60,253	60,233	60,233	60,081	59,614	59,829	61,457	64,214	102,602	153,661
	1.00	60,253	60,253	60,253	60,252	60,252	60,249	60,221	60,000	60,280	60,849
σ_{MAD}		α									
		0	0.0001	0.001	0.01	0.1	1	10	100	1000	10000
e_1	0.00	12,561	12,512	12,389	12,593	12,804	16,488	31,478	32,388	31,650	31,559
	0.10	12,561	12,516	12,397	12,577	12,785	16,054	31,002	32,460	31,661	31,560
	0.25	12,561	12,523	12,409	12,551	12,761	15,407	30,064	32,588	31,684	31,562
	0.50	12,561	12,536	12,428	12,467	12,737	14,272	27,385	32,869	31,749	31,570
	0.75	12,561	12,548	12,461	12,317	12,721	13,246	22,142	33,125	31,933	31,591
	0.90	12,561	12,556	12,512	12,389	12,593	12,804	16,489	31,482	32,392	31,655
	0.99	12,561	12,560	12,556	12,512	12,389	12,593	12,804	16,498	31,527	32,436
	1.00	12,561	12,561	12,561	12,561	12,561	12,561	12,562	12,600	12,787	13,433
μ_{RMSE}		α									
		0	0.0001	0.001	0.01	0.1	1	10	100	1000	10000
e_1	0.00	229,353	228,786	228,332	234,381	237,151	253,395	357,989	483,641	509,978	512,893
	0.10	229,353	228,835	228,179	234,139	237,051	251,487	350,531	480,709	509,623	512,858
	0.25	229,353	228,912	227,991	233,842	236,912	248,597	337,988	475,030	508,915	512,787
	0.50	229,353	229,048	227,881	233,081	236,673	243,779	312,565	459,353	506,811	512,574
	0.75	229,353	229,195	228,230	230,512	236,126	239,248	279,226	421,968	500,673	511,936
	0.90	229,353	229,288	228,786	228,332	234,381	237,152	253,399	358,019	483,698	510,040
	0.99	229,353	229,347	229,288	228,786	228,332	234,384	237,154	358,312	484,264	510,040
	1.00	229,353	229,353	229,353	229,353	229,353	229,348	229,310	229,449	238,853	244,293
σ_{RMSE}		α									
		0	0.0001	0.001	0.01	0.1	1	10	100	1000	10000
e_1	0.00	90,076	89,929	90,263	91,934	90,371	123,748	295,558	365,754	373,603	374,383
	0.10	90,076	89,940	90,182	91,954	90,330	119,277	288,431	364,784	373,507	374,374
	0.25	90,076	89,958	90,057	91,989	90,342	112,620	275,157	362,845	373,315	374,356
	0.50	90,076	89,993	89,880	91,795	90,633	102,058	242,163	357,051	372,740	374,302
	0.75	90,076	90,032	89,833	90,753	91,436	93,316	182,653	340,106	371,011	374,141
	0.90	90,076	90,058	89,929	90,263	91,934	90,371	123,756	295,600	365,808	373,657
	0.99	90,076	90,074	90,058	89,929	90,263	91,934	90,375	123,853	296,025	366,343
	1.00	90,076	90,076	90,076	90,076	90,077	90,081	90,129	91,069	96,252	103,824
μ_2		α									
		0	0.0001	0.001	0.01	0.1	1	10	100	1000	10000
e_1	0.00	0.677	0.678	0.683	0.675	0.670	0.673	0.528	0.120	0.006	0.008
	0.10	0.677	0.678	0.683	0.675	0.670	0.674	0.545	0.132	0.007	0.007
	0.25	0.677	0.678	0.683	0.676	0.670	0.674	0.572	0.156	0.011	0.007
	0.50	0.677	0.677	0.682	0.676	0.670	0.673	0.618	0.217	0.020	0.006
	0.75	0.677	0.677	0.680	0.681	0.671	0.671	0.660	0.349	0.047	0.003
	0.90	0.677	0.677	0.678	0.683	0.675	0.670	0.673	0.528	0.120	0.006
	0.99	0.677	0.677	0.677	0.678	0.683	0.675	0.670	0.673	0.527	0.118
	1.00	0.677	0.677	0.677	0.677	0.677	0.677	0.677	0.680	0.672	0.671
σ_2		α									
		0	0.0001	0.001	0.01	0.1	1	10	100	1000	10000
e_1	0.00	0.199	0.198	0.193	0.195	0.190	0.171	0.136	0.043	0.019	0.019
	0.10	0.199	0.198	0.194	0.195	0.191	0.173	0.137	0.046	0.019	0.019
	0.25	0.199	0.198	0.194	0.196	0.191	0.176	0.140	0.053	0.019	0.019
	0.50	0.199	0.199	0.195	0.196	0.193	0.181	0.144	0.070	0.019	0.019
	0.75	0.199	0.199	0.197	0.193	0.194	0.186	0.154	0.104	0.024	0.019
	0.90	0.199	0.199	0.198	0.193	0.195	0.190	0.171	0.136	0.043	0.019
	0.99	0.199	0.199	0.199	0.198	0.193	0.195	0.190	0.171	0.136	0.042
	1.00	0.199	0.199	0.199	0.199	0.199	0.199	0.199	0.196	0.194	0.190

Figure C.2: Elastic net μ and σ of error measures for 10 variables

Δt (s)		α									
		0	0.0001	0.001	0.01	0.1	1	10	100	1000	10000
e_1	0.00	0.123	0.133	0.163	0.152	0.144	0.148	0.149	0.149	0.149	0.149
	0.10	0.141	0.130	0.131	0.129	0.132	0.077	0.061	0.060	0.060	0.062
	0.25	0.142	0.132	0.134	0.131	0.133	0.083	0.060	0.060	0.060	0.062
	0.50	0.140	0.136	0.133	0.134	0.133	0.095	0.060	0.059	0.060	0.060
	0.75	0.144	0.136	0.133	0.133	0.134	0.130	0.063	0.061	0.060	0.061
	0.90	0.142	0.134	0.133	0.137	0.132	0.132	0.073	0.061	0.061	0.061
	0.99	0.145	0.134	0.136	0.132	0.133	0.137	0.140	0.073	0.064	0.063
	1.00	0.152	0.135	0.135	0.133	0.133	0.134	0.133	0.129	0.107	0.070
μ_{MAD}		α									
		0	0.0001	0.001	0.01	0.1	1	10	100	1000	10000
e_1	0.00	59,907	59,655	58,743	59,454	61,828	63,686	75,067	133,936	161,782	165,538
	0.10	59,907	59,679	58,763	59,364	61,660	63,679	73,432	131,396	161,334	165,491
	0.25	59,907	59,715	58,812	59,178	61,511	63,651	70,883	126,720	160,448	165,397
	0.50	59,907	59,776	58,993	58,904	61,080	63,492	66,366	115,430	157,873	165,117
	0.75	59,907	59,841	59,349	58,850	60,302	62,925	63,705	95,484	150,769	164,285
	0.90	59,907	59,880	59,655	58,743	59,454	61,829	63,686	75,078	133,986	161,849
	0.99	59,907	59,904	59,655	58,743	59,454	59,455	61,833	63,684	75,189	134,489
	1.00	59,907	59,907	59,907	59,907	59,907	59,903	59,868	59,506	59,815	62,683
σ_{MAD}		α									
		0	0.0001	0.001	0.01	0.1	1	10	100	1000	10000
e_1	0.00	12,462	12,425	12,364	13,507	13,197	13,969	24,570	32,989	31,830	31,578
	0.10	12,462	12,428	12,352	13,479	13,202	13,829	23,814	33,005	31,859	31,582
	0.25	12,462	12,434	12,343	13,418	13,236	13,636	22,522	32,995	31,916	31,589
	0.50	12,462	12,443	12,361	13,291	13,297	13,362	19,893	32,672	32,074	31,609
	0.75	12,462	12,453	12,389	12,789	13,395	13,176	16,360	30,519	32,469	31,670
	0.90	12,462	12,458	12,425	12,364	13,507	13,197	13,969	24,578	33,002	31,842
	0.99	12,462	12,462	12,458	12,425	12,364	13,508	13,198	13,979	24,656	33,139
	1.00	12,462	12,462	12,462	12,462	12,462	12,459	12,434	12,201	13,406	14,986
μ_{RMSE}		α									
		0	0.0001	0.001	0.01	0.1	1	10	100	1000	10000
e_1	0.00	224,617	223,936	223,149	243,493	245,980	247,915	290,218	436,645	503,380	512,208
	0.10	224,617	223,999	222,819	242,786	245,968	247,459	285,983	430,442	502,327	512,098
	0.25	224,617	224,096	222,481	241,350	246,229	246,807	279,403	419,015	500,242	511,879
	0.50	224,617	224,263	222,430	238,468	246,462	245,869	267,791	391,080	494,156	511,223
	0.75	224,617	224,437	223,133	231,076	246,219	245,410	255,381	341,250	477,301	509,272
	0.90	224,617	224,544	223,936	223,149	243,494	245,981	247,918	290,253	436,789	503,565
	0.99	224,617	224,610	224,544	223,936	223,150	243,502	245,987	247,955	290,607	438,229
	1.00	224,617	224,617	224,617	224,617	224,615	224,598	224,421	221,780	237,380	260,725
σ_{RMSE}		α									
		0	0.0001	0.001	0.01	0.1	1	10	100	1000	10000
e_1	0.00	90,540	90,442	92,257	110,906	104,804	108,660	214,471	348,711	371,847	374,207
	0.10	90,540	90,450	91,826	110,024	105,224	107,412	206,081	345,963	371,558	374,180
	0.25	90,540	90,462	91,269	108,620	105,990	105,653	192,023	340,559	370,979	374,126
	0.50	90,540	90,486	90,626	106,750	107,669	103,218	164,000	325,120	369,244	373,963
	0.75	90,540	90,512	90,390	100,002	110,100	102,334	129,892	285,838	364,056	373,473
	0.90	90,540	90,529	90,442	92,257	110,907	104,805	108,667	214,547	348,855	372,003
	0.99	90,540	90,539	90,529	90,442	92,258	110,918	104,813	108,760	215,313	350,295
	1.00	90,540	90,540	90,540	90,540	90,539	90,526	90,419	89,215	108,052	138,019
μ_{r2}		α									
		0	0.0001	0.001	0.01	0.1	1	10	100	1000	10000
e_1	0.00	0.695	0.697	0.703	0.682	0.665	0.664	0.652	0.300	0.035	0.004
	0.10	0.695	0.697	0.703	0.683	0.665	0.664	0.657	0.321	0.040	0.004
	0.25	0.695	0.697	0.702	0.685	0.666	0.664	0.662	0.359	0.049	0.003
	0.50	0.695	0.696	0.701	0.690	0.668	0.663	0.669	0.444	0.076	0.000
	0.75	0.695	0.696	0.699	0.698	0.673	0.663	0.669	0.568	0.146	0.009
	0.90	0.695	0.695	0.697	0.703	0.682	0.665	0.664	0.652	0.300	0.035
	0.99	0.695	0.695	0.695	0.697	0.703	0.682	0.665	0.664	0.652	0.295
	1.00	0.695	0.695	0.695	0.695	0.695	0.695	0.696	0.699	0.690	0.662
σ_{r2}		α									
		0	0.0001	0.001	0.01	0.1	1	10	100	1000	10000
e_1	0.00	0.183	0.183	0.182	0.184	0.196	0.191	0.152	0.094	0.022	0.019
	0.10	0.183	0.183	0.182	0.183	0.195	0.192	0.154	0.099	0.023	0.019
	0.25	0.183	0.183	0.182	0.183	0.195	0.193	0.157	0.109	0.025	0.019
	0.50	0.183	0.183	0.182	0.181	0.194	0.195	0.165	0.126	0.031	0.019
	0.75	0.183	0.183	0.182	0.182	0.190	0.197	0.179	0.144	0.051	0.019
	0.90	0.183	0.183	0.183	0.182	0.184	0.196	0.191	0.152	0.094	0.022
	0.99	0.183	0.183	0.183	0.183	0.182	0.184	0.196	0.191	0.152	0.093
	1.00	0.183	0.183	0.183	0.183	0.183	0.183	0.183	0.183	0.181	0.182

Figure C.3: Elastic net μ and σ of error measures for 30 variables

Δt (s)		α									
		0	0.0001	0.001	0.01	0.1	1	10	100	1000	10000
e_1	0.00	0.464	0.504	0.537	0.553	0.572	0.570	0.601	0.617	0.630	0.617
	0.10	0.571	0.555	0.546	0.544	0.546	0.541	0.258	0.247	0.243	0.246
	0.25	0.565	0.544	0.549	0.541	0.544	0.546	0.257	0.246	0.246	0.251
	0.50	0.573	0.570	0.545	0.545	0.543	0.571	0.296	0.247	0.245	0.245
	0.75	0.569	0.553	0.548	0.554	0.544	0.559	0.343	0.252	0.251	0.250
	0.90	0.581	0.562	0.567	0.574	0.545	0.545	0.521	0.254	0.248	0.256
	0.99	0.579	0.547	0.561	0.555	0.549	0.549	0.549	0.523	0.252	0.261
	1.00	0.587	0.555	0.554	0.562	0.577	0.556	0.549	0.511	0.405	0.271
μ_{MAD}		α									
		0	0.0001	0.001	0.01	0.1	1	10	100	1000	10000
e_1	0.00	59,480	59,295	58,479	58,115	58,638	62,689	67,938	101,353	153,603	164,614
	0.10	59,480	59,312	58,511	58,260	58,411	62,430	67,823	98,318	152,376	164,468
	0.25	59,480	59,338	58,582	58,424	57,629	61,978	67,574	93,133	150,000	164,177
	0.50	59,480	59,383	58,776	58,523	58,153	61,044	66,834	82,857	143,485	163,312
	0.75	59,480	59,430	59,062	58,465	57,264	60,038	65,078	71,090	127,906	160,786
	0.90	59,480	59,460	59,295	58,478	58,115	58,638	62,691	67,937	101,453	153,817
	0.99	59,480	59,478	59,295	58,478	58,115	58,638	62,691	67,937	101,453	153,817
	1.00	59,480	59,480	59,480	59,480	59,479	59,469	59,376	58,584	56,484	62,601
σ_{MAD}		α									
		0	0.0001	0.001	0.01	0.1	1	10	100	1000	10000
e_1	0.00	11,158	11,164	11,240	11,536	12,144	14,166	19,199	31,551	32,216	31,632
	0.10	11,158	11,163	11,236	11,551	12,522	14,043	18,726	31,221	32,273	31,641
	0.25	11,158	11,162	11,230	11,549	12,160	13,844	17,997	30,562	32,378	31,660
	0.50	11,158	11,160	11,209	11,488	12,121	13,488	16,763	28,608	32,606	31,716
	0.75	11,158	11,159	11,182	11,318	11,646	12,846	15,404	24,660	32,816	31,870
	0.90	11,158	11,158	11,164	11,240	11,536	12,143	14,167	19,211	31,589	32,256
	0.99	11,158	11,158	11,158	11,164	11,536	12,133	12,133	14,181	19,328	31,967
	1.00	11,158	11,158	11,158	11,158	11,159	11,160	11,176	11,240	11,862	15,202
μ_{RMSE}		α									
		0	0.0001	0.001	0.01	0.1	1	10	100	1000	10000
e_1	0.00	209,720	209,111	206,824	208,088	216,636	245,513	278,785	363,824	485,245	510,179
	0.10	209,720	209,167	206,889	207,938	216,742	244,011	277,088	356,867	482,451	509,852
	0.25	209,720	209,254	207,046	207,462	212,660	241,405	274,259	345,401	477,029	509,200
	0.50	209,720	209,403	207,524	206,949	213,970	235,665	268,283	323,375	461,982	507,261
	0.75	209,720	209,558	208,371	207,024	208,268	227,322	258,494	297,793	425,741	501,591
	0.90	209,720	209,655	209,111	206,824	208,088	216,636	245,528	278,819	364,097	485,801
	0.99	209,720	209,714	209,655	209,111	206,823	208,085	216,636	245,680	279,165	366,843
	1.00	209,720	209,720	209,720	209,720	209,718	209,702	209,520	207,595	205,637	254,488
σ_{RMSE}		α									
		0	0.0001	0.001	0.01	0.1	1	10	100	1000	10000
e_1	0.00	77,134	77,130	77,491	79,594	86,021	111,996	177,977	314,393	367,591	373,777
	0.10	77,134	77,129	77,458	79,362	88,141	110,098	173,429	308,994	366,840	373,706
	0.25	77,134	77,128	77,400	78,893	85,254	107,018	166,084	298,912	365,343	373,563
	0.50	77,134	77,128	77,283	78,051	84,727	101,096	151,800	273,760	360,902	373,134
	0.75	77,134	77,130	77,168	77,675	81,472	93,574	132,058	227,937	348,096	371,849
	0.90	77,134	77,133	77,130	77,491	79,594	86,016	112,017	178,087	314,734	368,015
	0.99	77,134	77,134	77,133	77,130	77,492	79,599	85,969	112,230	179,192	318,137
	1.00	77,134	77,134	77,134	77,135	77,136	77,149	77,263	77,878	83,551	128,823
μ_2		α									
		0	0.0001	0.001	0.01	0.1	1	10	100	1000	10000
e_1	0.00	0.709	0.710	0.713	0.721	0.711	0.678	0.648	0.523	0.115	0.005
	0.10	0.709	0.710	0.713	0.719	0.713	0.680	0.649	0.540	0.127	0.007
	0.25	0.709	0.710	0.713	0.718	0.719	0.683	0.650	0.566	0.150	0.010
	0.50	0.709	0.709	0.712	0.715	0.715	0.690	0.654	0.608	0.210	0.018
	0.75	0.709	0.709	0.711	0.713	0.722	0.699	0.664	0.640	0.342	0.044
	0.90	0.709	0.709	0.710	0.713	0.721	0.711	0.678	0.648	0.522	0.113
	0.99	0.709	0.709	0.709	0.710	0.713	0.721	0.711	0.678	0.648	0.516
	1.00	0.709	0.709	0.709	0.709	0.709	0.709	0.709	0.713	0.731	0.674
σ_2		α									
		0	0.0001	0.001	0.01	0.1	1	10	100	1000	10000
e_1	0.00	0.190	0.189	0.187	0.182	0.176	0.176	0.169	0.142	0.043	0.019
	0.10	0.190	0.189	0.187	0.182	0.175	0.176	0.170	0.143	0.047	0.019
	0.25	0.190	0.189	0.188	0.183	0.177	0.176	0.172	0.145	0.054	0.019
	0.50	0.190	0.189	0.188	0.184	0.177	0.176	0.176	0.145	0.072	0.020
	0.75	0.190	0.190	0.189	0.186	0.177	0.176	0.177	0.151	0.108	0.024
	0.90	0.190	0.190	0.189	0.187	0.182	0.176	0.176	0.169	0.142	0.043
	0.99	0.190	0.190	0.190	0.189	0.187	0.182	0.176	0.176	0.168	0.142
	1.00	0.190	0.190	0.190	0.190	0.190	0.190	0.190	0.188	0.178	0.171

Figure C.4: Elastic net μ and σ of error measures for 100 variable

Δt (s)		α									
		0	0.0001	0.001	0.01	0.1	1	10	100	1000	10000
e_1	0.00	1.533	1.548	1.509	1.500	1.503	1.504	1.650	1.630	1.629	1.562
	0.10	1.511	1.413	1.491	1.519	1.501	1.473	0.601	0.537	0.529	0.566
	0.25	1.533	1.519	1.513	1.566	1.495	1.490	0.634	0.546	0.548	0.557
	0.50	1.591	1.553	1.494	1.520	1.514	1.546	0.726	0.547	0.540	0.585
	0.75	1.531	1.418	1.530	1.489	1.506	1.470	1.087	0.550	0.536	0.535
	0.90	1.558	1.481	1.473	1.492	1.598	1.520	1.498	0.575	0.537	0.535
	0.99	1.486	1.465	1.485	1.385	1.547	1.514	1.498	1.419	0.565	0.534
	1.00	1.510	1.521	1.510	1.531	1.520	1.512	1.511	1.363	1.001	0.652
μ_{MAD}		α									
		0	0.0001	0.001	0.01	0.1	1	10	100	1000	10000
e_1	0.00	67,706	67,436	65,963	64,048	63,414	65,553	74,592	88,033	148,237	163,956
	0.10	67,706	67,462	66,063	64,161	62,991	65,386	74,039	85,329	146,559	163,741
	0.25	67,706	67,501	66,239	64,369	63,149	65,138	73,029	81,529	143,352	163,314
	0.50	67,706	67,567	66,604	64,736	63,101	64,613	70,765	77,663	134,761	162,049
	0.75	67,706	67,636	67,081	65,241	64,031	63,673	67,626	77,204	115,698	158,426
	0.90	67,706	67,678	67,436	65,963	64,047	63,414	65,553	74,589	88,187	148,654
	0.99	67,706	67,703	67,435	65,962	64,042	63,421	65,550	74,563	89,770	148,654
	1.00	67,706	67,706	67,706	67,706	67,703	67,673	67,376	65,243	60,673	64,640
σ_{MAD}		α									
		0	0.0001	0.001	0.01	0.1	1	10	100	1000	10000
e_1	0.00	11,675	11,691	11,718	11,783	12,599	15,246	19,525	30,457	32,338	31,660
	0.10	11,675	11,689	11,726	11,720	12,491	15,055	19,381	30,010	32,392	31,673
	0.25	11,675	11,687	11,734	11,629	12,497	14,768	19,124	28,997	32,485	31,697
	0.50	11,675	11,682	11,739	11,495	12,409	14,187	18,476	25,692	32,650	31,769
	0.75	11,675	11,679	11,717	11,546	12,251	13,439	17,112	21,535	32,475	31,955
	0.90	11,675	11,677	11,691	11,718	11,783	12,600	15,249	19,527	30,503	32,396
	0.99	11,675	11,675	11,677	11,691	11,718	11,784	12,615	15,283	19,551	30,956
	1.00	11,675	11,675	11,675	11,675	11,675	11,677	11,714	11,454	12,432	17,041
μ_{RMSE}		α									
		0	0.0001	0.001	0.01	0.1	1	10	100	1000	10000
e_1	0.00	220,556	220,535	218,474	218,961	221,941	247,346	285,937	349,095	476,812	509,145
	0.10	220,556	220,543	218,694	218,471	218,979	245,726	284,156	343,482	473,312	508,711
	0.25	220,556	220,552	219,053	217,557	221,770	243,065	281,067	334,660	466,587	507,847
	0.50	220,556	220,561	219,709	216,123	219,733	237,467	274,165	319,087	448,412	505,283
	0.75	220,556	220,562	220,318	216,459	222,282	228,574	262,352	302,021	407,680	497,852
	0.90	220,556	220,559	220,535	218,474	218,961	221,962	247,370	285,973	349,404	477,685
	0.99	220,556	220,556	220,559	220,535	218,472	218,959	222,178	247,608	286,359	352,552
	1.00	220,556	220,556	220,556	220,555	220,554	220,541	220,576	216,377	218,298	263,872
σ_{RMSE}		α									
		0	0.0001	0.001	0.01	0.1	1	10	100	1000	10000
e_1	0.00	88,420	88,781	86,695	88,508	94,339	124,670	199,281	313,760	367,348	373,756
	0.10	88,420	88,756	87,036	87,660	92,896	122,075	195,229	308,652	366,568	373,683
	0.25	88,420	88,715	87,559	86,065	94,971	117,874	188,431	299,325	365,013	373,537
	0.50	88,420	88,633	88,391	83,410	92,134	109,975	174,061	277,203	360,402	373,098
	0.75	88,420	88,535	88,893	83,204	96,503	100,650	151,149	239,936	347,202	371,780
	0.90	88,420	88,468	88,781	86,695	88,508	94,356	124,704	199,375	314,125	367,827
	0.99	88,420	88,425	88,468	88,781	86,691	88,509	94,525	125,055	200,360	317,765
	1.00	88,420	88,420	88,420	88,420	88,421	88,429	88,853	83,618	92,462	146,425
μ_2		α									
		0	0.0001	0.001	0.01	0.1	1	10	100	1000	10000
e_1	0.00	0.706	0.707	0.712	0.715	0.715	0.688	0.646	0.570	0.155	0.010
	0.10	0.706	0.707	0.711	0.715	0.718	0.690	0.648	0.581	0.170	0.012
	0.25	0.706	0.707	0.711	0.714	0.716	0.693	0.651	0.599	0.197	0.016
	0.50	0.706	0.707	0.710	0.713	0.715	0.698	0.658	0.622	0.269	0.028
	0.75	0.706	0.706	0.708	0.712	0.719	0.708	0.672	0.636	0.411	0.063
	0.90	0.706	0.706	0.707	0.712	0.715	0.715	0.688	0.646	0.569	0.152
	0.99	0.706	0.706	0.706	0.707	0.712	0.715	0.688	0.688	0.646	0.562
	1.00	0.706	0.706	0.706	0.706	0.706	0.706	0.707	0.711	0.719	0.667
σ_2		α									
		0	0.0001	0.001	0.01	0.1	1	10	100	1000	10000
e_1	0.00	0.187	0.186	0.182	0.176	0.173	0.168	0.160	0.145	0.053	0.018
	0.10	0.187	0.186	0.182	0.176	0.174	0.168	0.161	0.145	0.057	0.018
	0.25	0.187	0.186	0.182	0.177	0.174	0.169	0.162	0.145	0.066	0.018
	0.50	0.187	0.186	0.183	0.178	0.173	0.170	0.164	0.144	0.087	0.019
	0.75	0.187	0.186	0.185	0.180	0.171	0.172	0.166	0.148	0.123	0.026
	0.90	0.187	0.186	0.186	0.182	0.176	0.173	0.168	0.160	0.145	0.052
	0.99	0.187	0.187	0.186	0.186	0.182	0.176	0.173	0.168	0.159	0.146
	1.00	0.187	0.187	0.187	0.187	0.187	0.187	0.186	0.183	0.174	0.165

Figure C.5: Elastic net μ and σ of error measures for 200 variable

Δt (s)		α									
		0	0.0001	0.001	0.01	0.1	1	10	100	1000	10000
e_1	0.00	3.514	3.428	3.396	3.403	3.431	3.380	3.700	3.731	3.554	3.596
	0.10	3.302	3.260	3.249	3.359	3.511	3.352	1.189	1.021	1.038	1.037
	0.25	3.505	3.398	3.402	3.358	3.444	3.373	1.252	1.037	1.048	1.064
	0.50	3.426	3.318	3.375	3.383	3.339	3.325	1.455	1.079	1.054	1.044
	0.75	3.377	3.222	3.243	3.272	3.334	3.335	2.290	1.031	1.027	1.057
	0.90	3.418	3.344	3.359	3.390	3.313	3.460	3.334	1.087	1.031	1.031
	0.99	3.375	3.414	3.445	3.338	3.364	3.326	3.345	3.073	1.054	0.998
	1.00	3.264	3.288	3.267	3.232	3.284	3.225	3.247	2.753	1.807	1.188

μ_{MAD}		α									
		0	0.0001	0.001	0.01	0.1	1	10	100	1000	10000
e_1	0.00	73,466	73,194	71,841	70,046	68,153	68,165	72,326	86,962	147,034	163,797
	0.10	73,466	73,220	71,926	70,181	67,982	68,173	71,862	84,892	145,269	163,568
	0.25	73,466	73,259	72,077	70,407	68,433	68,170	71,099	82,213	141,901	163,110
	0.50	73,466	73,327	72,404	70,867	68,529	68,255	69,668	78,863	132,953	161,757
	0.75	73,466	73,396	72,857	71,314	69,664	67,905	68,332	76,253	113,499	157,892
	0.90	73,466	73,438	73,194	71,841	70,045	68,150	68,157	72,321	87,082	147,556
	0.99	73,466	73,464	73,194	73,194	71,839	70,035	68,131	68,073	72,306	88,541
	1.00	73,466	73,466	73,466	73,466	73,463	73,432	73,123	70,801	63,472	64,538

σ_{MAD}		α									
		0	0.0001	0.001	0.01	0.1	1	10	100	1000	10000
e_1	0.00	12,437	12,476	12,569	12,607	12,653	14,947	19,719	30,114	32,345	31,665
	0.10	12,437	12,472	12,558	12,552	12,672	14,775	19,482	29,512	32,396	31,678
	0.25	12,437	12,467	12,547	12,448	12,733	14,541	19,089	28,257	32,485	31,704
	0.50	12,437	12,458	12,532	12,392	12,741	14,021	18,218	25,423	32,627	31,777
	0.75	12,437	12,448	12,509	12,559	12,908	13,280	16,691	22,257	32,407	31,970
	0.90	12,437	12,442	12,476	12,569	12,607	12,654	14,951	19,727	30,177	32,409
	0.99	12,437	12,438	12,442	12,476	12,569	12,608	12,663	14,991	19,801	30,743
	1.00	12,437	12,437	12,437	12,437	12,437	12,441	12,486	12,371	12,984	16,996

μ_{RMSE}		α									
		0	0.0001	0.001	0.01	0.1	1	10	100	1000	10000
e_1	0.00	228,922	228,944	227,062	227,134	224,779	248,373	285,493	348,566	475,390	508,950
	0.10	228,922	228,949	227,228	226,560	223,330	246,882	283,691	343,161	471,792	508,497
	0.25	228,922	228,953	227,527	225,619	227,312	244,401	280,589	334,651	464,899	507,596
	0.50	228,922	228,953	228,138	224,760	224,840	239,413	273,764	319,455	446,399	504,924
	0.75	228,922	228,943	228,745	225,887	228,847	231,036	262,398	302,132	405,631	497,200
	0.90	228,922	228,932	228,944	227,062	227,134	224,795	248,395	285,534	348,873	476,384
	0.99	228,922	228,923	228,932	228,944	227,060	227,133	224,955	248,613	285,955	352,019
	1.00	228,922	228,922	228,922	228,922	228,920	228,910	228,899	225,451	223,525	263,792

σ_{RMSE}		α									
		0	0.0001	0.001	0.01	0.1	1	10	100	1000	10000
e_1	0.00	103,009	103,528	101,327	100,934	98,208	127,578	200,237	314,526	367,529	373,780
	0.10	103,009	103,493	101,659	99,601	98,223	125,118	196,130	309,492	366,764	373,709
	0.25	103,009	103,433	102,196	97,458	101,531	121,176	189,270	300,310	365,235	373,567
	0.50	103,009	103,316	103,119	95,383	97,517	113,647	174,946	278,528	360,689	373,142
	0.75	103,009	103,175	103,707	97,726	106,164	104,220	152,692	241,459	347,617	371,862
	0.90	103,009	103,079	103,528	101,327	100,935	98,218	127,611	200,331	314,883	367,999
	0.99	103,009	103,016	103,079	103,528	101,322	100,941	98,328	127,937	201,304	318,437
	1.00	103,009	103,009	103,009	103,009	103,010	103,037	103,598	96,264	97,824	146,465

μ_2		α									
		0	0.0001	0.001	0.01	0.1	1	10	100	1000	10000
e_1	0.00	0.699	0.700	0.705	0.708	0.712	0.688	0.648	0.572	0.162	0.011
	0.10	0.699	0.700	0.705	0.708	0.714	0.690	0.650	0.583	0.177	0.014
	0.25	0.699	0.700	0.704	0.707	0.710	0.692	0.653	0.599	0.206	0.018
	0.50	0.699	0.700	0.703	0.706	0.711	0.697	0.660	0.622	0.278	0.030
	0.75	0.699	0.700	0.702	0.705	0.711	0.706	0.673	0.637	0.420	0.066
	0.90	0.699	0.699	0.700	0.705	0.708	0.712	0.688	0.648	0.571	0.158
	0.99	0.699	0.699	0.699	0.700	0.705	0.708	0.712	0.688	0.648	0.565
	1.00	0.699	0.699	0.699	0.699	0.699	0.699	0.700	0.703	0.713	0.667

σ_2		α									
		0	0.0001	0.001	0.01	0.1	1	10	100	1000	10000
e_1	0.00	0.190	0.189	0.185	0.179	0.173	0.167	0.160	0.145	0.055	0.018
	0.10	0.190	0.189	0.186	0.180	0.174	0.168	0.161	0.146	0.059	0.018
	0.25	0.190	0.189	0.186	0.180	0.175	0.168	0.162	0.145	0.068	0.018
	0.50	0.190	0.189	0.187	0.182	0.174	0.169	0.164	0.144	0.089	0.019
	0.75	0.190	0.190	0.188	0.184	0.175	0.172	0.166	0.148	0.125	0.026
	0.90	0.190	0.190	0.189	0.185	0.179	0.173	0.167	0.160	0.146	0.054
	0.99	0.190	0.190	0.190	0.189	0.185	0.179	0.173	0.167	0.159	0.146
	1.00	0.190	0.190	0.190	0.190	0.190	0.190	0.189	0.187	0.174	0.165

Figure C.6: Elastic net μ and σ of error measures for 330 variable

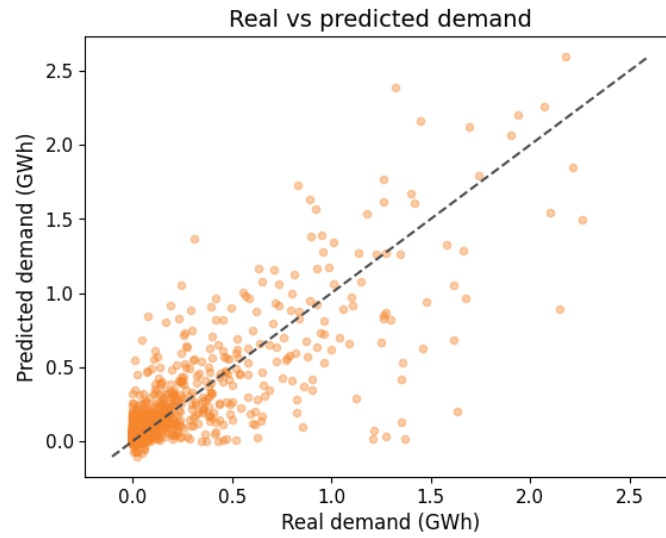


Figure C.7: Predicted and actual sales values on train data

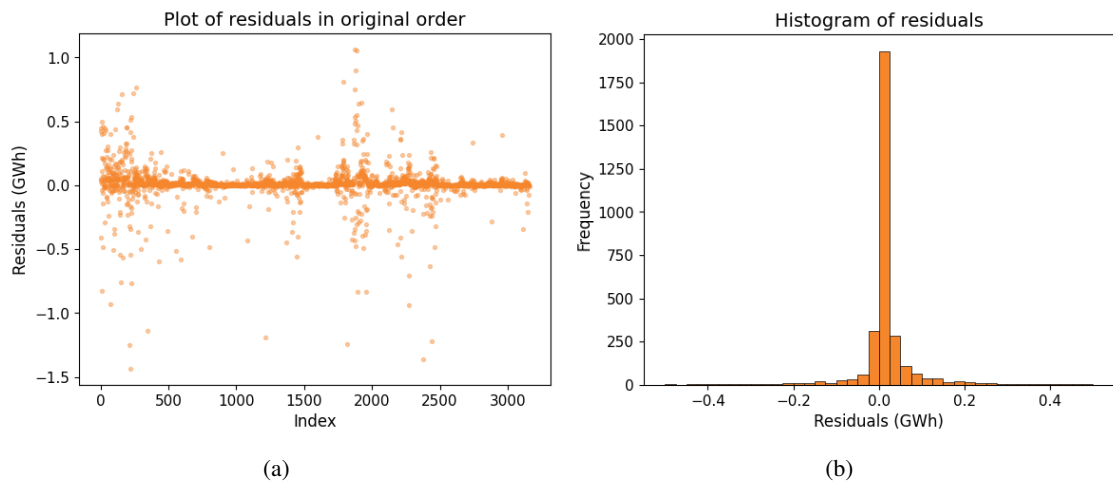


Figure C.8: Residuals (a) plot and (b) histogram on train data

Appendix D

Demand forecasting coefficients

While most of the growth coefficients b are positive, the acceleration coefficients a are split, so there are a significant number of regions whose growth is decelerating.

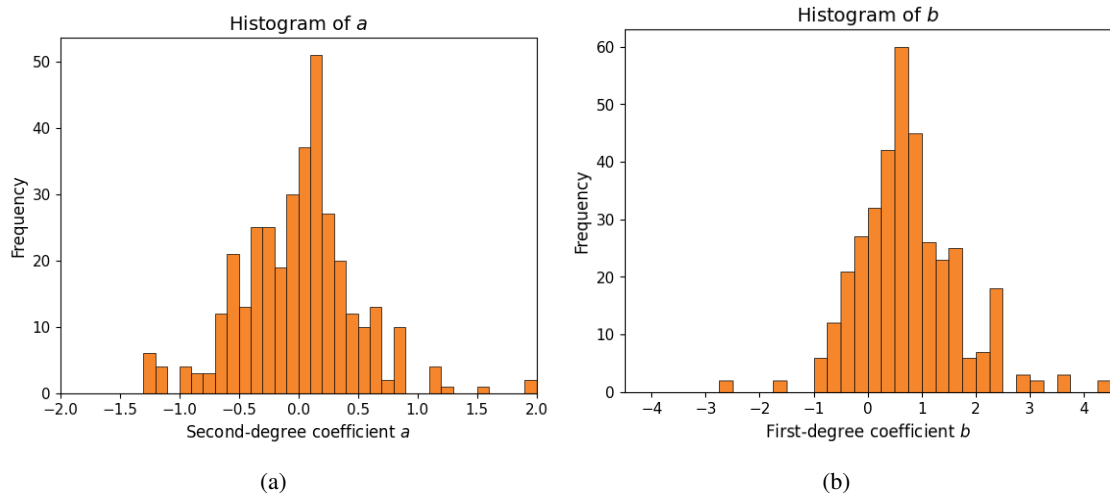


Figure D.1: Forecast model final (a) second- and (b) first-order coefficients

Appendix E

Detailed GBM model results

The outputs of the best eight models (all except GLM and DL) are highly correlated, above 0.9 as can be seen from Figure E.1.

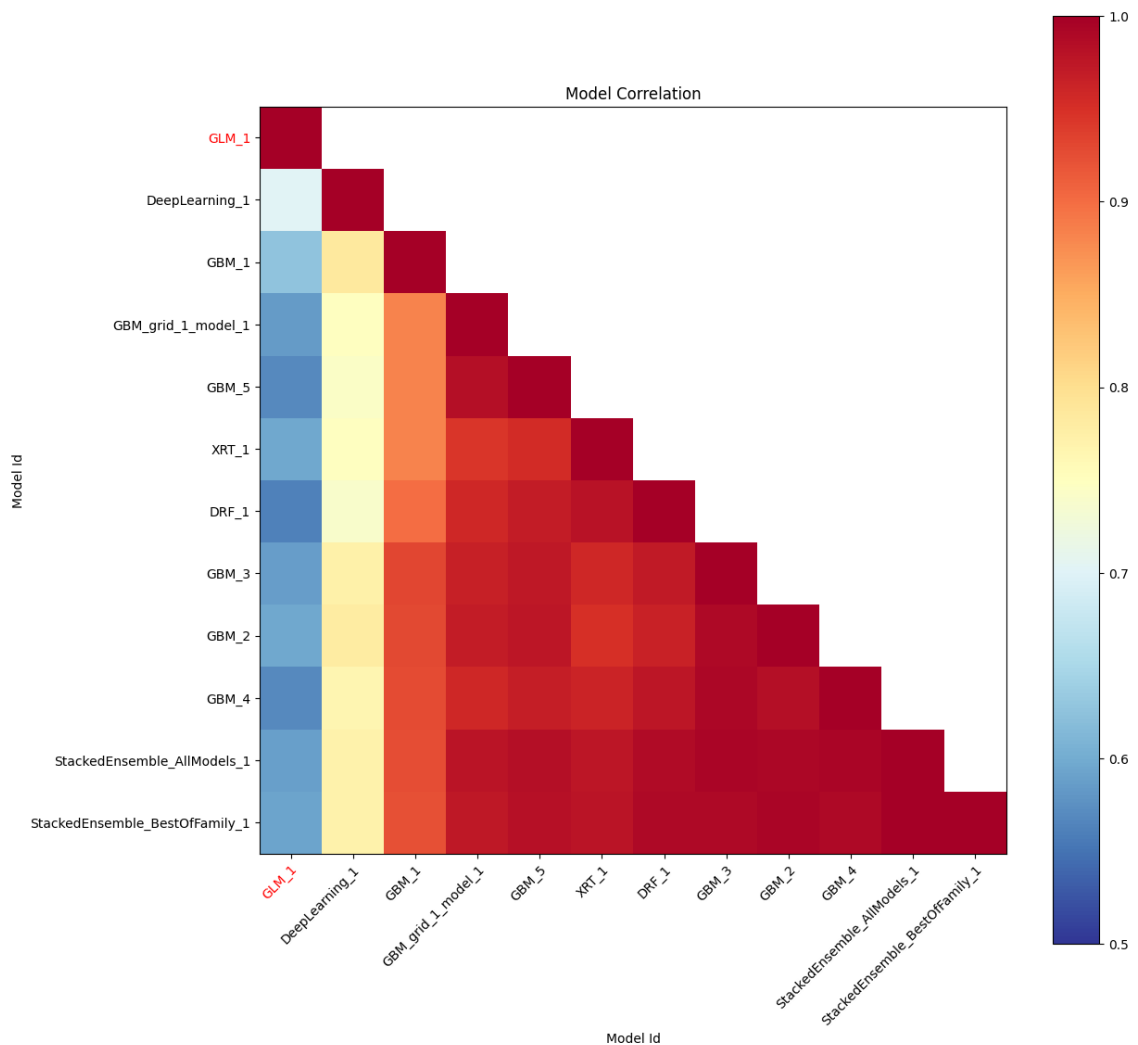


Figure E.1: Correlation between AutoML models predictions

As can be seen from Figure E.2, the GBM2 model has significant difference between the train and cross-validated model fit, finishing at a train RMSE of about 500, with a cross-validation RMSE of around 800. This is for 103 trees, when the cross-validation RMSE suddenly starts increasing. From the residual plot in Figure E.3 and the predicted vs actual plot in Figure E.4, it can be seen that residuals increase significantly with the predicted value. The relative variable importance can be visualized in Figure E.5, with the same variables as in Table 5.10 and some extra ones for relative visual comparison.

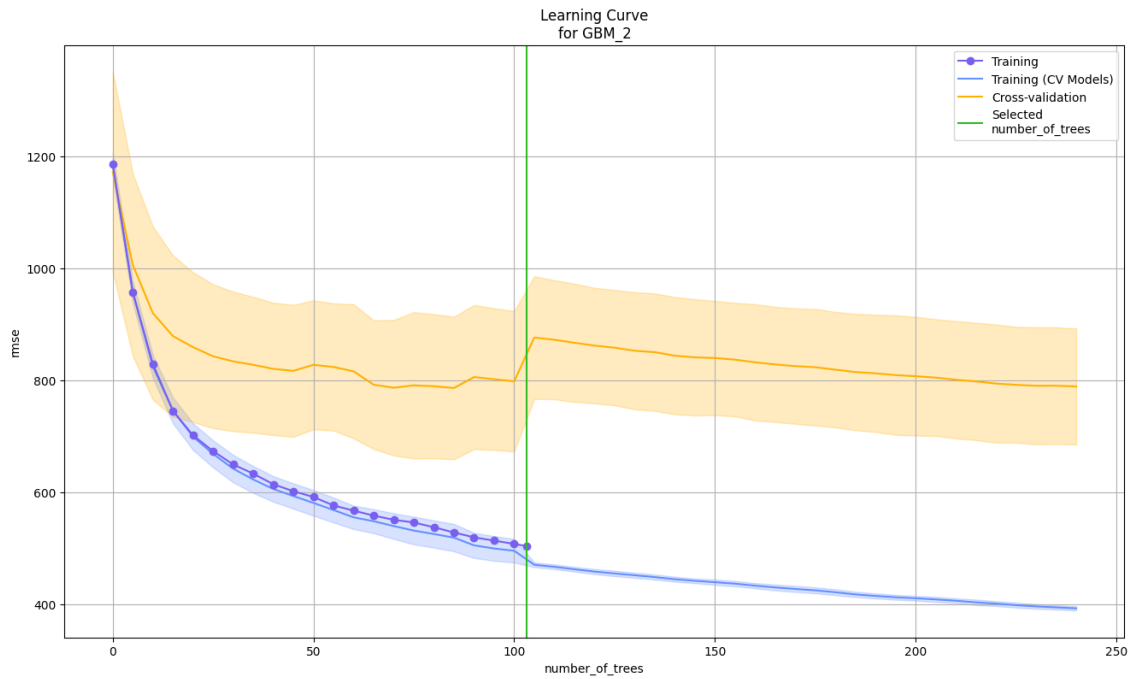


Figure E.2: Learning curve of the GBM model

Table E.1: Importance of density variables per level of GBM model

Variables group	Density levels					Σ
	DT	MN	FR	SEC	SS	
sociodemographic	0.0%	0.3%	2.9%	4.8%	4.5%	*12.4%
vehicles	0.0%	0.0%	0.0%	0.1%	0.1%	0.2%
PPP	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
predicted demand	0.0%	0.0%	0.1%	0.2%	0.0%	0.4%
POIs	—	—	—	—	—	—
tech specs	—	—	—	—	—	—
charges & tariffs	—	—	—	—	—	—
regions	—	—	—	—	—	—
Σ	0.0%	0.3%	*3.0%	*5.1%	*4.7%	*13.0%

*most relevant variables groups, mentioned in the results analysis

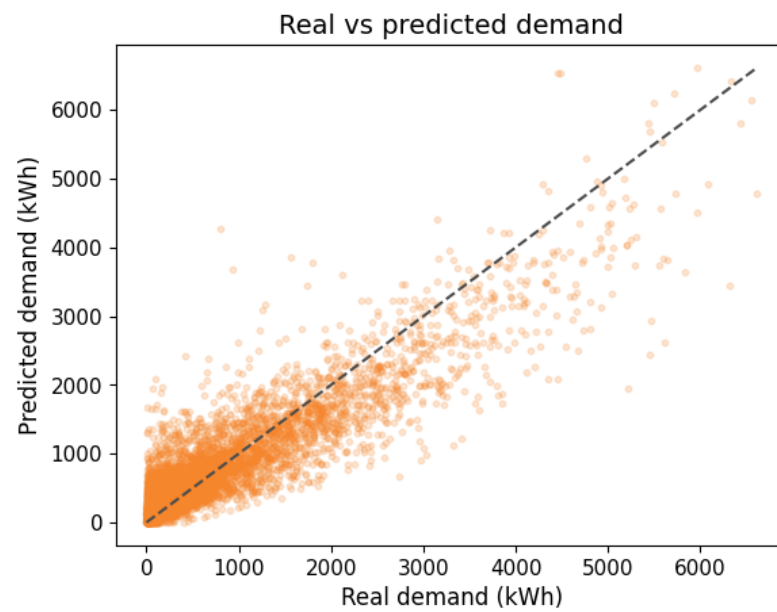


Figure E.3: Predicted vs actual plot of the GBM model

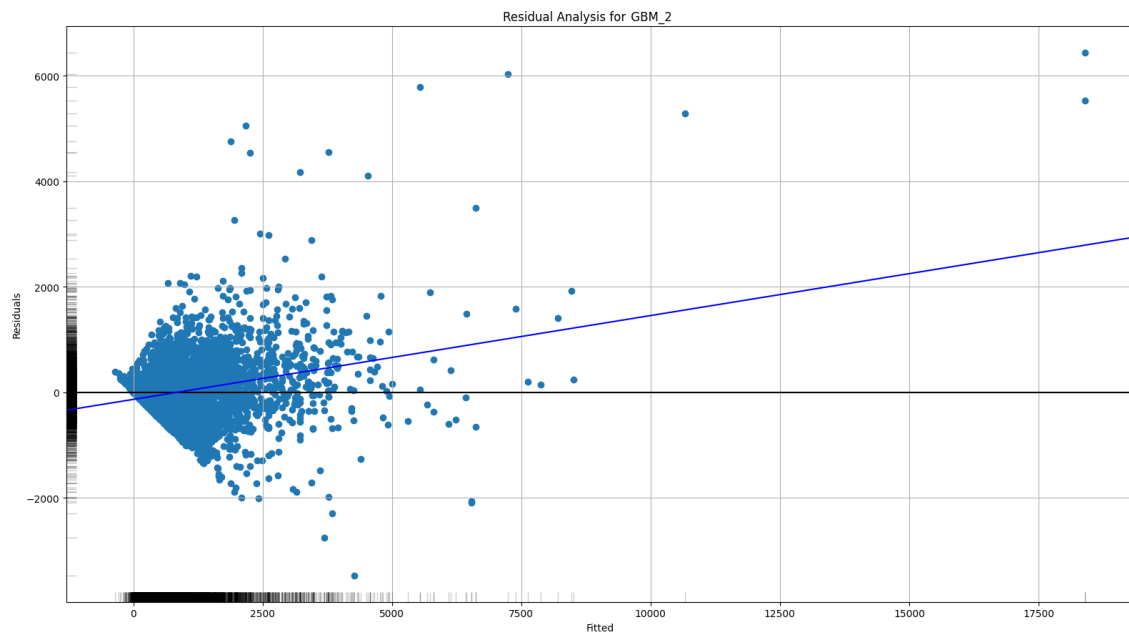


Figure E.4: Residual plot of the GBM model

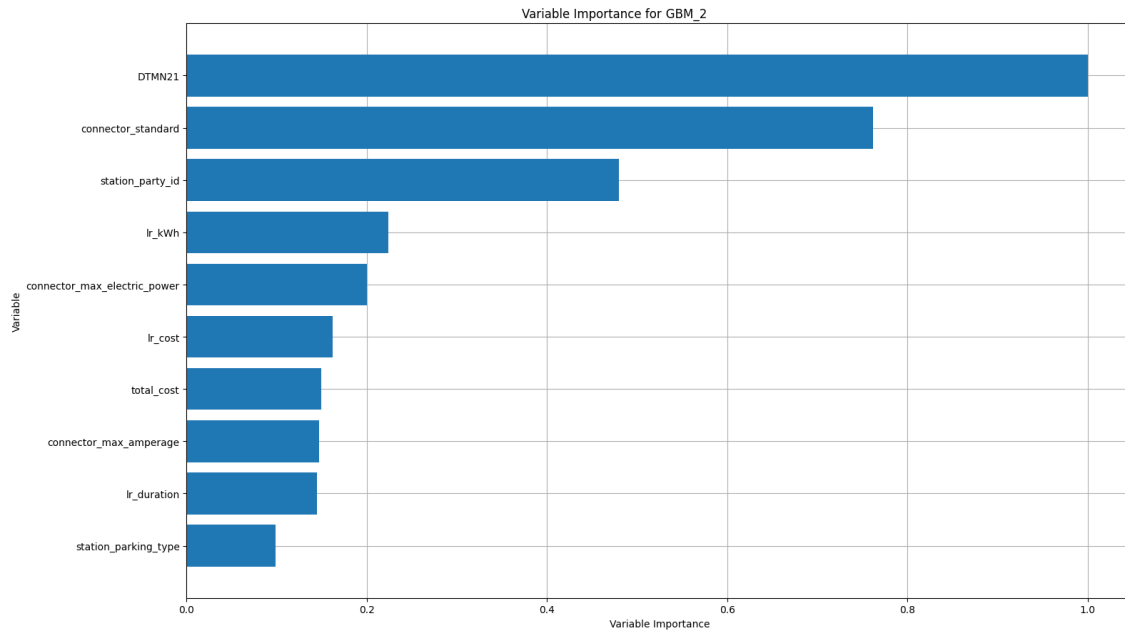


Figure E.5: Relative variable importance of the GBM model

Table E.2: Importance of distance variables per radius of GBM model

Variables group	Distance levels - radius (m)				Σ
	1000	3000	6000	10000	
sociodemographic	4.7%	2.5%	0.7%	0.5%	*8.4%
vehicles	0.2%	0.6%	0.1%	0.4%	1.2%
PPP	0.0%	0.0%	0.0%	0.0%	0.0%
predicted demand	0.0%	0.0%	0.0%	0.0%	0.0%
POIs	4.0%	1.7%	3.3%	2.4%	*11.4%
tech specs	—	—	—	—	—
charges & tariffs	—	—	—	—	—
regions	—	—	—	—	—
Σ	*8.9%	*4.8%	4.1%	3.2%	21.0%

*most relevant variables groups, mentioned in the results analysis

Appendix F

Gravity model parameters

Table F.1: Tested parameter values for gravity model

t ; param.	Values
<i>pow</i>	
k_1	[0*, 0.01, 0.1, 1, 10, 100, 1000]
k_2	[4.0, 3.5, 3.0, 2.5, 2.0*, 1.5, 1.0, 0.5, 0.0]
k_3	[0*, 0.1, 1, 10, 100, 1000]
<i>exp</i>	
k_1	[0*, 0.01, 0.1, 1, 10, 100, 1000]
k_2	[0.1*, 0.5, 1.0]
k_3	[0.1, 0.5, 1.0*]
*default values	