Description of an Electric Vehicle Charge Station Network using Knockoff Filters

DANIEL ANTONIO MARTINEZ-MUNOZ

Abstract

In this work we analyse the utilisation of electric vehicle (EV) public charging stations in the Netherlands to understand and describe their usage as a function of surrounding premises (such as hospitals, casinos and schools, among others) and population. Also, we analyse the charging performance of such charging stations taking into account temporal values and charging measures taken from transactions registered within the years 2012 and 2016. In order to identify the (potentially) explanatory variables that are meaningful, we will use a False Discovery Rate (FDR) control approach known as *Knockoffs filter*. The results reveal that charging stations located close to Kindergartens, Fuel stations and Car sharing points are more likely to be used more frequently and for the longest time; whereas those users who charge their vehicles either on a weekend or in July between 12 AM and 6 AM are expected to charge their vehicles faster than in other configurations.

Description of an Electric Vehicle Charge Station Network using Knockoff Filters

By

DANIEL ANTONIO MARTINEZ MUNOZ

Supervisor
DR RUI CARVALHO

A dissertation submitted to the Durham University
Department of Engineering in accordance with the
requirements to obtain the degree of MASTER IN SCIENCE.

DECEMBER 2021

Table of Contents

Page

List of Tables

List of Figures

Statement of Copyright

Acknowledgements

I can only thank God for his goodness and faithfulness throughout this whole process. He will always receive my gratitude. But if I am allowed to acknowledge and thank all the people who stood by me and helped me to finish this work, I must start by naming my family: my wife Abigail, who always encouraged me to complete this dissertation; my daughter Sophia, who was born in the middle of this process and has been our greatest gift from God so far, and my parents who mean everything to me.

I must afterwards thank my supervisor Rui Carvalho, who believed in me and helped me from the very beginning of this journey.

Later, my gratitude goes to all the new friends we made while in the UK: Jordy and Lis, Yohan and his lovely family, David Edgar, Kay Snow and Daniel Edgar, and to all our friends at Elim's Church.

I would also like to extend my special thanks to my alma mater, Universidad Autonoma de Ciudad Juarez (in Mexico), Diana Galarza and to Lupita Almeida, which financial help was key to start off this project.

This project was supported by the National Council of Science and Technology (Consejo Nacional de Ciencia y Tecnología - CONACyT) of Mexico, under the programme Fondo Sectorial CONACyT-Secretaría de Energía-Sustentabilidad Energética 2017 - Sexto Periodo (consortium CONACyT-SENER); CVU 763843.

Dedication

To my wife Abigail, my daughter Sophia and my parents. Anabel, you will be missed and loved forever.

Chapter 1

Introduction

In the last years, private and governmental institutions have invested in promoting Electric Vehicles (EVs) as an environment-friendly and energy efficient alternative [1, 2], resulting in numerous studies about EVs to explore strategies of implementation and thus reduce the environmental impact caused by transportation in metropolitan areas.

In this work we will make an effort to contribute with some formal analysis on the charging stations network in the Netherlands, one of the countries with the highest rates of EV usage [3], their geographical distribution and performance variation. This in hopes that we can develop further analysis techniques for this type of settings.

1.1 Literature review

Research on Electric Vehicles is broad, focusing in a variety of aspects; for instance, on the study of performance of EV energy consumption rates [4], and even on the economic profiles of society for switching into EV usage rather than conventional transport [5].

A study conducted by Helmus et al. [6] analyses two roll-out strategies for the placement of charging stations in the Netherlands: a demand-driven placement (i.e., the charging station is requested by EV drivers, generally near to home) and a strategic placement (the charging stations are placed by decision of the Government near to public facilities such as hospitals and schools). This study concluded that the effectiveness of these strategies depend on the market of EVs: in an immature market, demand-driven placement of charging stations is more effective as there is a better performance on energy transfer per charging point, whereas in a mature market, a strategic placement of charging stations is more beneficial.

On the other hand, another study that deals with the strategic location of public EV charging stations was published by Xi, Sioshansi and Marano [7]. In their investigation they determine the most viable charging technology to find an optimal distribution of locations for such charging points. In their model, they used the same approach as Curtin et al. [8], considering a linear model to examine EV adoption probabilities (i.e., the likelihood of drivers to move from conventional transportation to EV), with clusters of demographic variables (such as income, age, number of vehicles owned, average monthly gasoline used, among others) and macroeconomic variables (such as gasoline and electricity prices and the price premium of acquiring a conventional EV). Their approach comprises a simulation-optimisation model (linear integer programming) to determine the optimal location of charging stations based on previous simulations they made.

Researchers have also analysed EV charging demand as an important factor to a successful implementation of EV policies. Robinson et al. [9] explored the behaviour of EV drivers in the north-east of England. They suggest strategies such as a pay-as-you-go recharging to be implemented at all public charging points as well as smart meters in order to reduce peak demand on local power grids and also reduce the carbon emissions associated with EV charging. Moreover, Sadeghianpourhamami et al. [10] conducted a quantitative analysis on the EV flexibility (i.e., the extent to which the charging load can be controlled) to characterise the peak demand on a network of charging stations in the Netherlands and thus help to develop strategies to stimulate more flexibility.

Further details on the theoretical background and major literature available on each topic and technique implemented on this study will be described throughout each one of the Chapters 2, 3 and 4.

1.2 Objective and Motivation

This project aims to provide a different perspective on the study of an EV charging network, identifying the main factors (temporal and in relation to surrounding premises) that affect the charging rates (performance) of charge points; all through a combination and comparison of variable selection methods (described in Chapter 3) and the implementation of a fairly novel False Discovery Rate (FDR) control method: the Knockoffs filter [11].

Previous research have implemented data-driven techniques to describe the demand of EV and energy consumption (for instance, [12] and [13]); however, their statistical modelling methodology do not perform variable selection or account for Type I errors (i.e. false positives) when it comes to decide what features truly affect the behaviour of their data.

To the best of our knowledge, no FDR control technique has been applied to an EV setting before, so we pretend in this work to explore the analysis of this type of data sets, adjusting the available parameters of each method, to be able to draw valid and meaningful conclusions on the explanatory variables affecting the dynamics of an EV network. This work also aims to contribute with geographical approaches that ease the study of proximity and for a better understanding of the data set we were provided with (Chapter 5).

1.2.1 Delimitation

Up until 2016, the Netherlands registered a total of 26,088 public and semi-public charging stations around the country, with 115,223 electric vehicles (EVs) including plug-in hybrid electric vehicles and full electric vehicles [3]. In 2018, the Netherlands was the second country with the largest number of EV users in Europe [3] with 36,049 public charging points and 128,612 electric cars [14], which indicates a notorious growth on the purchase and usage of such type of vehicles.

The ElaadNL is a leading organisation that manages technology for EV in the Netherlands [15]. They provided records from 1,060,763 transactions over 1,747 charging stations (distributed in 1,725 different geographical points around the country) with attributes such as `Start Time`, `Stop Time`, `Connected Time`, `Idle Time`, `Latitude`, `Longitude`, `Total Energy`, among others (17 attributes in total), dating from the 1st of January, 2012 and until the 30th of March, 2016 (see Section 5.1 for more details). This work will be centred on the study of these records.

1.2.2 Research questions

We expect to be able to couple hypothesis testing with fitting; more precisely, we aim to answer the following **research questions**:

- **Spatial and temporal behaviour of users:** *Are there any spatial and temporal patterns in how users utilise the network of charging stations?* We will address this question by performing a geographical analysis on the location of charging stations (performing linear regression methods), varying temporal features and taking into account nearby amenities and proximity between charging points.

- **Performance of charging stations:** *What are the factors influencing the performance of charging stations?* To answer this question, we will use variable selection techniques such as the Least Absolute Shrinkage and Selection Operator (LASSO) [16] and the knockoffs filter, so that we can characterise and identify the temporal parameters that have a direct influence on the energy supply.

Chapter 2

Linear regression

One of the most helpful and interpretative tools one can use to make predictions from a specific collection of points (representing data) is *Linear Regression*. Nowadays, linear regression stands as one of the most employed techniques among scientific researchers given its simplicity, interpretation and its versatility as it has been used to achieve newer approaches.

In this chapter we will explore some fundamentals of linear regression and will describe how it works for a better comprehension of this method.

2.1 Ordinary Least Squares Regression

Suppose we require to obtain a quantitative response vector Y from a vector X consisting of predictors (we call X to be a *predictor variable* or the *input vector*). The linear regression method assumes that there exists a linear relationship between such variables. We express such relationship as

$$Y = \beta_0 + \beta_1 X + \epsilon, \tag{2.1}$$

where β_0 is the *intercept*, β_1 is the *slope* and ϵ is the *expected error*, which is assumed to be independent of X, with $E(\epsilon) = 0$; this means that either Y is an affine function, or it is a reasonable approximation to the actual phenomenon given the input vector X. Both, β_0 and β_1, receive the name of *parameters* or *coefficients* of the model. We will now discuss how these parameters are estimated.

2.1.1 Estimating the Parameters

In order to make predictions from (2.1), we require to estimate first β_0 and β_1. Let

$$(x_1, y_1), \dots, (x_n, y_n)$$

be n observations of the experiment. Our aim is, therefore, to find estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ such that $y_i \approx \hat{\beta}_0 + \hat{\beta}_1 x_i$ for $i = 1, \dots, n$. For that, we will use the *Ordinary Least Squares* (OLS) approach, which measures the average lack of fit.

Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ be the prediction for Y at the ith component of X. Let $e_i = y_i - \hat{y}_i$ be defined as the ith-*residual* (the difference between the actual response value and the prediction based on our linear model). We define the *Residual Sum of Squares* (RSS) as

$$RSS(\hat{\beta}_0, \hat{\beta}_1) = e_1^2 + e_2^2 + \cdots + e_i^2$$

that is

$$RSS(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2. \tag{2.2}$$

OLS aims to find $\hat{\beta}_0$ and $\hat{\beta}_1$ such that $RSS(\hat{\beta}_0, \hat{\beta}_1)$ is minimised.

Differentiating with respect to $\hat{\beta}_0$ we obtain:

$$\frac{\partial RSS(\hat{\beta}_0, \hat{\beta}_1)}{\partial \hat{\beta}_0} = 2n\hat{\beta}_0 - 2\sum_{i=1}^{n} (y_i - \hat{\beta}_1 x_i)$$

We set this derivative to zero to obtain

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{n} y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i}{n} = \bar{y} - \hat{\beta}_1 \bar{x}. \tag{2.3}$$

where $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$, $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$ are the arithmetic means.

Now let us differentiate (2.2) with respect to $\hat{\beta}_1$:

$$\frac{\partial RSS(\hat{\beta}_0, \hat{\beta}_1)}{\partial \hat{\beta}_1} = 2\sum_{i=1}^{n} (\hat{\beta}_1 x_i - x_i y_i + \hat{\beta}_0 x_i) \tag{2.4}$$

We once more set this derivative to zero, distribute the terms, and substitute $\hat{\beta}_0$ in (2.4) by $\hat{\beta}_0$ obtained from (2.3) to get

$$0 = \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i y_i + \sum_{i=1}^{n} \left(\bar{y} - \hat{\beta}_1 \bar{x}\right)(\bar{x})$$

Thus,

$$\sum_{i=1}^{n} x_i y_i = \bar{y} \sum_{i=1}^{n} x_i - \hat{\beta}_1 \bar{x} \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2. \tag{2.5}$$

Before continuing, it is convenient to note that the following identities hold true (recall that $\sum_{i=1}^{n} x_i = n\bar{x}$ and that $\sum_{i=1}^{n} a = na$ for all a not indexed):

$$\begin{aligned}
\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) &= \sum_{i=1}^{n}(x_i y_i - x_i \bar{y} - \bar{x} y_i + \bar{x}\bar{y}) \\
&= \sum_{i=1}^{n} x_i y_i - \bar{y} \sum_{i=1}^{n} x_i - \bar{x} \sum_{i=1}^{n} y_i + \sum_{i=1}^{n} \bar{x}\bar{y} \\
&= \sum_{i=1}^{n} x_i y_i - \bar{y} \sum_{i=1}^{n} x_i - \bar{x} \sum_{i=1}^{n} y_i + n\bar{x}\bar{y} \\
&= \sum_{i=1}^{n} x_i y_i - \bar{y} \sum_{i=1}^{n} x_i - \bar{x} n \bar{y} + n\bar{x}\bar{y} \\
&= \sum_{i=1}^{n} x_i y_i - \bar{y} \sum_{i=1}^{n} x_i \qquad (2.6)
\end{aligned}$$

$$\begin{aligned}
\sum_{i=1}^{n} (x_i - \bar{x})^2 &= \sum_{i=1}^{n} \left(x_i^2 - 2x_i\bar{x} + \bar{x}^2\right) \\
&= \sum_{i=1}^{n} x_i^2 - 2\bar{x} \sum_{i=1}^{n} x_i + n\bar{x}^2 \\
&= \sum_{i=1}^{n} x_i^2 - 2\bar{x} n \bar{x} + n\bar{x}^2 \\
&= \sum_{i=1}^{n} x_i^2 - 2n\bar{x}^2 + n\bar{x}^2 \\
&= \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \\
&= \sum_{i=1}^{n} x_i^2 - \bar{x} \sum_{i=1}^{n} x_i \qquad (2.7)
\end{aligned}$$

Now, from (2.5) we have

$$\hat{\beta}_1 = \frac{\sum\limits_{i=1}^{n} x_i y_i - \bar{y} \sum\limits_{i=1}^{n} x_i}{\sum\limits_{i=1}^{n} x_i^2 - \bar{x} \sum\limits_{i=1}^{n} x_i}.$$

Substituting with the identities obtained in (2.6) and in (2.7) we obtain:

$$\hat{\beta}_1 = \frac{\sum\limits_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum\limits_{i=1}^{n} (x_i - \bar{x})^2} \tag{2.8}$$

2.1.2 Assessing the Accuracy of the Parameters

Now that we have an estimation of the parameters for our model given the n first observations, we require to know how accurate these coefficients are with respect to the actual model $Y \approx \beta_0 + \beta_1 X$. To answer that question we require to compute the *standard error* of $\hat{y}$, written as $SE(\hat{y})$:

$$SE(\hat{y})^2 = \frac{\sigma^2}{n} \tag{2.9}$$

where σ is the standard deviation of each observation y_i (assuming that the n observations are uncorrelated).

The identity (2.9) indicates that the standard error of $\hat{y}$ decreases as n increases, i.e. the greater amount of observations (n) we have, the more accurate our estimation is.

Furthermore, we can also estimate how close the parameters $\hat{\beta}_0$ and $\hat{\beta}_1$ are from β_0 and β_1 by computing their respective standard errors:

$$SE\left(\hat{\beta}_0\right)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum\limits_{i=1}^{n} (x_i - \bar{x})^2} \right], \tag{2.10}$$

$$SE\left(\hat{\beta}_1\right)^2 = \frac{\sigma^2}{\sum\limits_{i=1}^{n} (x_i - \bar{x})^2} \tag{2.11}$$

where σ^2 =Var(ϵ). From equation (2.10) we note that $SE\left(\hat{\beta}_0\right) = SE(\hat{y})$ when $\bar{x} = 0$ and then $\hat{\beta}_0 = \hat{y}$.

The estimate of σ is called the *Residual Standard Error (RSE)* and we can compute it from data as indicated:

$$RSE = \sqrt{\frac{RSS}{n-2}} \tag{2.12}$$

We will use the standard errors to run *hypothesis tests* on the coefficients, taking the *null hypothesis*:

H_0 : X and Y are not related.

versus the *alternative hypothesis*:

H_a : X and Y are related.

Taking into consideration that if $\beta_1 = 0$, then (2.1) gets reduced to $Y = \beta_0 + \epsilon$ and X and Y are not related; therefore, we aim to test

$$H_0 : \beta_1 = 0 \text{ versus } H_a : \beta_1 \neq 0,$$

In order to test the null hypothesis, we proceed to estimate *how far* from zero is our estimate $\hat{\beta}_1$ so that we can infer that β_1 is non-zero. This distance depends on the accuracy of $\hat{\beta}_1$, i.e., we must compute (2.11); if this value is small, then we can deduce that our estimate is considerably accurate. The smaller it is, the more accurate is our estimate. If $SE\left(\hat{\beta}_1\right)$ is small, then even for small values of our estimate $\hat{\beta}_1$ we may show that $\beta_1 \neq 0$ in which case the alternative hypothesis H_a holds. However, if $SE\left(\hat{\beta}_1\right)$ is large, then $|\hat{\beta}_1|$ must be large as well in order for us to reject the null hypothesis.

Typically, we calculate a *t-statistic*, given by

$$t = \frac{\hat{\beta}_1 - 0}{SE\left(\hat{\beta}_1\right)}, \tag{2.13}$$

which measures the number of standard deviations that $\hat{\beta}_1$ is away from zero; so, if H_0 happens to be true, then we expect that (2.13) will have a *t-distribution* with $n-2$ degrees of freedom.

2.1.3 Assessing the Accuracy of the Model

Once we have assessed the accuracy of our coefficients, we now require to know to what extent our model actually does fit the data. We will proceed taking into account two quantities that are about to be detailed: the Residual Standard Error (RSE) and the R^2 statistic.

2.1.3.1 Residual Standard Error RSE

As stated before, RSE is an estimate of the standard deviation of ϵ. In general terms, it is the average amount that the prediction is deviated from the actual regression line (recall from (2.1) that, due to the presence of ϵ, which does not depend on X, we will not be able to perfectly predict Y from X under a linear approach). We compute the RSE using

$$RSE = \sqrt{\frac{1}{n-2}RSS} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2.14}$$

We must consider RSE as a measure of lack of fit of our predictive model to the data. A quite large RSE indicates a greater lack of fit, which means that our model does not predict the data very well. On the other hand, if RSE is considerably small, then we may expect our model to be a good predictor for the true output data of the phenomenon.

2.1.3.2 R^2 Statistic

We will consider the R^2 statistic as an alternative measure of fit, apart from RSE, which indicates an estimation of the accuracy of the model from the units of Y. The R^2 statistic, unlike the RSE, expresses this accuracy in terms of a proportion which takes on values from 0 to 1 in $\mathbb{R}$. We will use the following formula to calculate R^2:

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS} \tag{2.15}$$

where $TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$ is the *Total Sum of Squares* and can be interpreted as the amount of variability in the response *before* the regression is executed while RSS measures the amount of variability left *after* the regression is carried out.

According to [17], an R^2 statistic closer to 1 indicates that a large proportion of the variability in the response is explained by the regression meanwhile as R^2 approaches 0 means that the regression did not explain much of the variability in the response. This can be due to either a wrong linear model or σ^2 is considerably high (or both).

So far, we may notice that using the R^2 statistic approach represents an advantage over RSE in terms of interpretation, since it takes on values only within the interval $[0,1] \in \mathbb{R}$.

2.2 Multiple Linear Regression

So far we have learnt how to perform a linear regression for a response vector using a single predictor. In practice, it turns out that we usually have more than one predictor that might affect the response vector.

In order for us to obtain a model considering those predictors, we might perform the simple linear regression method separately for each one of the predictors and then proceed to analyse each result trying to compare them between each other. However, this approach might result insufficient since every model is not taking into account the participation of the other predictors over the response.

Instead of performing numerous simple linear regressions for the predictors, we can extend the known method so that it takes into account multiple predictors (as many as required). We express the multiple linear regression model as follows, assigning an independent slope coefficient to each predictor in the same model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon, \tag{2.16}$$

where β_i is the slope coefficient corresponding to the predictor X_i, being p the total amount of predictors considered in the model.

2.2.1 Multiple Linear Regression Coefficients

Just like it occurred in Simple Linear Regression, the coefficients β_i are unknown and must be computed. Using our data set, consisting of a finite number n of observations, we can once more obtain predictions for such identities of the coefficients using the formula

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p, \tag{2.17}$$

and as in Simple Linear Regression (where $\hat{y} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n)^\top$ and $x_i = (x_{i1}, x_{i2}, \ldots, x_{in})^\top$, with $i = 1, \ldots, p$), we make use of the RSS defined in this case as

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip} \right)^2. \tag{2.18}$$

Thus, the values $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ that minimise (2.18) are obtained by applying the least squares method. In this case, it is particularly convenient to express such identities using matrix algebra [18].

Let $\hat{y} = \mathbf{X}\hat{\beta}$ be the matrix representation of (2.17), where $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)^\top$ and $\mathbf{X}$ is the *data matrix* which columns are $x_0, x_1, \ldots, x_p$, where x_0 is the vector of n ones.

In other words, we can express (2.17) as

$$\hat{y} = \begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{pmatrix} = \begin{pmatrix} \hat{\beta}_0 + \hat{\beta}_1 x_{11} + \hat{\beta}_2 x_{21} + \cdots + \hat{\beta}_p x_{p1} \\ \hat{\beta}_0 + \hat{\beta}_1 x_{12} + \hat{\beta}_2 x_{22} + \cdots + \hat{\beta}_p x_{p2} \\ \vdots \\ \hat{\beta}_0 + \hat{\beta}_1 x_{1n} + \hat{\beta}_2 x_{2n} + \cdots + \hat{\beta}_p x_{pn} \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{21} & \ldots & x_{p1} \\ 1 & x_{12} & x_{22} & \ldots & x_{p2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \ldots & x_{pn} \end{pmatrix} \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{pmatrix} = \mathbf{X}\hat{\beta} \qquad (2.19)$$

This representation allows us to obtain the next known result [18]:

$$\hat{\beta} = \left(X^\top X\right)^{-1} X^\top \hat{y} \qquad (2.20)$$

2.2.2 Assessing the accuracy of the model

We can assess how accurate our Multiple Linear Regression model is by proceeding similarly as in OLS, where we only required to verify whether $\beta_1 = 0$ so that we can accept or refuse the null hypothesis. In Multiple Linear Regression (MLR), we require to determine whether $\beta_i = 0$, with $i = 1, \ldots, p$. We then test the null hypothesis:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

versus the alternative hypothesis:

$$H_a : \texttt{at least one } \beta_j \texttt{ is non-zero}$$

We perform this hypothesis test by using the *F-Statistic*:

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)}, \qquad (2.21)$$

where TSS is defined as in OLS. Therefore, when the F-statistic is expected to take a value close to 1, we accept the null hypothesis. Otherwise, if H_a is true, then we expect F to be greater than 1.

2.3 The Bias-Variance Trade-off

In practice, we usually experience issues in our models regarding to their complexity and their accuracy for real-world predictions. For example, we can obtain complex models *trained* by a specific set of observations, which might be quite sensitive to changes in the training set, preventing us from getting good predictions of real phenomena with future input data. On the other hand, with simpler models, we might have results far from reality. We will now introduce two concepts that will help to adjust our models to a more convenient fit.

We call *variance* to the fluctuations caused as we change the training data set. Ideally, the estimated coefficients should approximately be the same regardless of the training data set (given that the model is describing a real-world phenomenon). We say that a model has *high variance* (or that it is *overfitted*) if it is too sensitive to such changes in the training set, yielding to models with some behaviour that occur only in some particular circumstances, rather than in general scenarios.

Additionally, we call *bias* to the error caused by assumptions that are different from what happens in reality. For example, linear regression assumes that there exists a linear relationship between Y and X, however, real life events are very unlikely to follow a linear behaviour, meaning that our linear approach might not be explaining very well our case of study (an *underfitted* model).

In general terms, a simple fit is most likely to have *high bias*, whilst a complex model would have high variance. Our goal, then, is to achieve an estimation low in both, bias and variance, so that it is a not-so-complex model (understandable and computationally economic) that is also a true-explanatory model (describes real behaviour without overfitting). To meet these objectives, we can measure and balance the model through the expected value of its respective RSS, which can be decomposed in terms of variance and bias of the model itself.

Let f be a function describing the phenomenon y, so that $y = f + \epsilon$, with ϵ being the error term, which is assumed to be irreducible. Also, let $\hat{f}$ and $\mathbb{E}\left[(y-\hat{f})^2\right]$ be an estimated fit for f and the expected value of the RSS of $\hat{f}$, respectively. Then,

$$\begin{aligned}
\mathbb{E}\left[(y-\hat{f})^2\right] &= \mathbb{E}\left[y^2+\hat{f}^2-2y\hat{f}\right] \\
&= \mathbb{E}\left[y^2\right]+\mathbb{E}\left[\hat{f}^2\right]-2\mathbb{E}\left[y\hat{f}\right] \\
&= Var(y)+\mathbb{E}[y]^2+Var(\hat{f})+\mathbb{E}\left[\hat{f}\right]^2-2\mathbb{E}\left[y\hat{f}\right] \\
&= Var(y)+\mathbb{E}[y]^2+Var(\hat{f})+\mathbb{E}\left[\hat{f}\right]^2-2\mathbb{E}[y]\mathbb{E}\left[\hat{f}\right].
\end{aligned}$$

Now, recall that $\mathbb{E}[y] = \mathbb{E}[f+\epsilon] = \mathbb{E}[f] + \mathbb{E}[\epsilon]$, and given that ϵ is the error term, which is irreducible, its expected value is 0. Therefore, $\mathbb{E}[y] = \mathbb{E}[f] = f$. Hence,

$$\begin{aligned}
\mathbb{E}\left[(y-\hat{f})^2\right] &= Var(y) + f^2 + Var(\hat{f}) + \mathbb{E}\left[\hat{f}\right]^2 - 2f\mathbb{E}\left[\hat{f}\right] \\
&= Var(y) + Var(\hat{f}) + \left(f^2 - 2f\mathbb{E}\left[\hat{f}\right] + \mathbb{E}\left[\hat{f}\right]^2\right) \\
&= Var(y) + Var(\hat{f}) + \left(f - \mathbb{E}\left[\hat{f}\right]\right)^2 \\
&= Var(y) + Var(\hat{f}) + Bias(\hat{f})^2.
\end{aligned}$$

2.4 Summary of chapter

We will now highlight some remarkable results from this chapter:

- Least Squares approaches aim to model a certain phenomenon Y from observations X, by finding coefficients $\beta \in \mathbb{R}^N$ such that the model $Y = \beta X + \epsilon$ is the best linear approximation to Y (assuming that Y follows a linear behaviour depending on X).

- We call *variance* to the fluctuations caused as the training data set varies. A model with a high variance is said to be *overfitted*.

- We call *bias* to the error caused by assumptions that are different from what happens in real life. A model with a high bias is said to be *underfitted*.

- Typically, a simple model tends to have high bias and low variance, whilst a complex model has lower bias but high variance.

Chapter 3

Variable Selection

Initially, when researchers intended to describe a specific phenomenon through a fitting model given a data set, techniques such as ordinary least squares (OLS) and logistic regression models (LRM) gained popularity among the scientific community due to their interpretability and simple implementation. However, this task has become a challenge over the last five decades [19], since investigators have been facing data settings describing even more complex phenomena with an increasing amount of explanatory variables (over millions of factors) where least squares and logistic regression fail to provide an accurate, unique solution (when features outnumber observations), not to mention their high variance, which increases with the number of variables.

Along with the development of data analysis through computer-based algorithms, the ability of scientists to obtain larger data sets also increased. In many instances, it turns out that it is even more likely to identify more features p than observations N, that is, $p >> N$. For example, Sesia, Sabatti and Candès [20] investigated genome-wide association studies (GWAS) in Crohn's disease, described by Ogura et al. [21] as "a chronic inflammatory disorder of the gastrointestinal tract, which is thought to result from the effect of environmental factors in a genetically predisposed host", working with a data set consisting of 4,913 binary-type samples with 377,749 single-nucleotide polymorphisms (SNPs) as variables. The researchers aimed to identify the SNPs that are related to Crohn's disease (in such large settings, it is often of enormous importance to identify those variables that are true signals, i.e., those features that have an actual effect over the response).

3.1 Variable selection methods

Variable selection methods are techniques designed to discard the variables that do not influence the response vector. Many reasons support the necessity of performing variable selection when we want to model a high-dimensional phenomenon; for instance:

- **We can improve the prediction accuracy of our model considerably**. Least squares and logistic regression frequently have low bias but high levels of variance. We can sacrifice some bias to reduce variance by shrinking variables or set them to zero (promoting a sparse setting) and by doing so, improving the accuracy of our model [18].
- **A better interpretation of the data**. Having knowledge of those relevant variables allow us to have a better understanding of the nature of the prediction problem we are working on [22]. We are interested in defining the most influential variables that affect the response; therefore, we can dismiss some features that have little influence on the model. As Friedman, Hastie and Tibshirani declare [18]: “In order to get the “big picture," we are willing to sacrifice some of the small details".

Nowadays, researchers use many well-developed variable selection methods in their investigations. We will now briefly introduce some of these techniques: Best-subset selection, backwards-stepwise regression and forward-stepwise regression.

3.1.1 Best-subset Selection

Best-subset selection consists of selecting the best subset of size $k \in \{1,\dots,p\}$ of variables, where p is the number of available variables. In other words, this method finds, for each k, the subset of k variables with the smallest mean squared error [18]. It is essential to keep in mind that, if $j \in \{1,\dots,p\}$ (with $p \leq k$), the best $j-$sized subset is not necessarily contained in the best $k-$sized subset. Unsurprisingly, an exhaustive evaluation of all the possible subsets is computationally demanding since it is a combinatorial problem. For a set of p variables, the amount of possible $k-$sized subsets is

$$\binom{p}{k} = \frac{p!}{k!(p-k)!}, \tag{3.1}$$

whereas the total amount of all the possible combinations for all the possible values of k is

$$\sum_{i=1}^{p} \binom{p}{i} = 2^p. \tag{3.2}$$

As a consequence of the costs of time and resources required to invest in a best-subset selection approach, researchers developed various techniques to make the process computationally more affordable. The studies published by Narendra and Fukunaga [23] and Furnival and Wilson

[24] propose *branch and bound* algorithms that perform this task efficiently for data sets from 24 to 35 variables by pivoting variables and establishing an enumeration system over the features.

Selecting the appropriate k, however, is again a matter of balance between bias and variance in the model. Later in this chapter, we will provide more details on *resampling methods*, which are techniques designed to rearrange the samples so that it is more feasible to find an optimal parameter for variable selection.

3.1.2 Forward- and Backward-stepwise Regression

Forward-stepwise Regression (FSR) is a method that works gradually, step by step, over the predictors. Considered a "simple modification of best subset" [25], FSR, in the first step, takes the intercept (first predictor) and estimates the best candidate to be added to the *active set* (i.e., the set of selected variables) which will be the variable that minimises the residual sum of squares at stage *k* (RSS_k). The process is an iteration of the same principle, adding to the active set those variables that decrease the most the respective RSS_k, so that the test statistic

$$R_k = \frac{1}{\sigma^2}(RSS_{k-1} - RSS_k) \tag{3.3}$$

(where σ^2 is assumed to be known) is compared to a χ_1^2 distribution [26].

Hastie, Tibshirani and Friedman [18] suggest two main advantages as reasons why FSR could be preferred over best-subset selection:

- **Computational**. In contrast with best-subset selection, FSR can perform variable selection for large values of p (even if $p >> N$, being N the number of observations), providing a nested series of models (given that the sequence of models updates at every step of the procedure; however this is a direct consequence of the linear design as well as of squared error loss [25]).

- **Statistical**. Unlike best-subset selection, FSR is a more constrained method that reduces the variance of our model. However, it might also mean that we have more bias.

Backwards-stepwise regression (BSR), on the other hand, does the same procedure as FSR but it starts with the full set of predictors, and in each step, BSR discards that variable that has the least impact on the response. Nevertheless, BSR can only be performed when $N > p$, whereas FSR can be used always, independently of the relation between p and N [18].

One of the variable selection methods we will use in this project is the Least Absolute Shrinkage Selection Operator (LASSO) which, apart from providing a reliable fit to the data, also promotes a sparse active set. More details on the LASSO appear in this chapter.

3.2 Regularisation

Regularisation is a common way to control error in a flexible manner. Sometimes we may have multiple curves fitting the data in space, but then we need to know which one we must choose, depending on the parameter values of β.

The main idea behind regularisation lies in the bias-variance trade-off by adding a *penalty term* to the classic MLR optimisation problem, encouraging a selecting measure among all possible solutions:

$$\min \|y - X\beta\| + \lambda f(\beta) \tag{3.4}$$

By doing so, we now balance one term (the data reconstruction error) with another term (the regularisation penalty, $\lambda f(\beta)$, for some function f).

3.2.1 Resampling Methods

Before continuing directly with the description of the regularised methods, we will introduce the concept and usefulness of resampling methods. In practice, whenever we build a model to describe a certain phenomenon, we need to assess how well our model describes response points from the actual experiment. Usually, we will not have new data available to compare and determine if the model is actually working properly. Therefore, we can only use part of the same data we were given to assess and even improve our approach.

Resampling methods are a nowadays, a useful tool in statistics [17], moreover, these methods have come to be one of the basis for new Machine Learning techniques. *Resampling* our data, means that we are selecting a subset of the population of our original sample to *train* our model. In this work we will perform the *k-fold cross-validation*, which is described in the following subsection, although there are other techniques used by researchers such as the *Jackknife* [25] and *the Bootstrap* [18, 25] (methods that assess the accuracy of the model).

3.2.1.1 k-Fold Cross-validation

In general, *cross-validation* (CV) is a simple, intuitive method that allows us to assess how accurately our model predicts the behaviour in practice, estimating prediction errors as well.

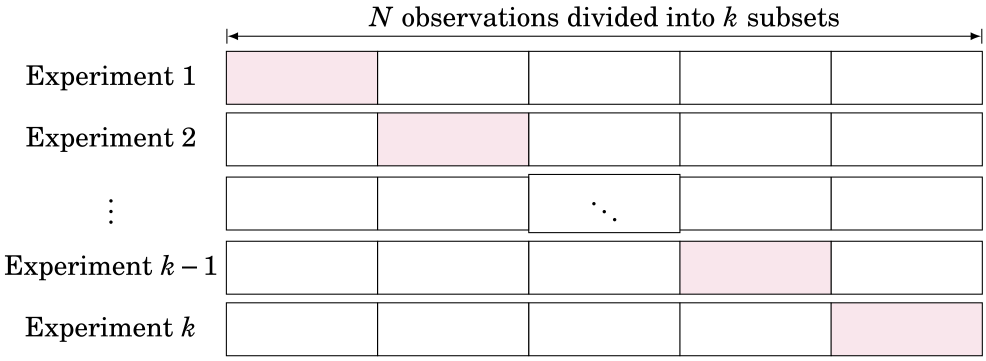

Figure 3.1: k-fold Cross-Validation. A resampling method that runs a model over $k-1$ training sets to finally test it over the remaining subset (coloured cells).

This method splits the n samples or observations into k different subsets of roughly equal size. The elements of each subset must be selected randomly. Then we *'take apart'* the first subset and perform the corresponding fitting technique over the $k-1$ remaining subsets (called the *training set*) with a range of different values for λ (tuning parameter of fitting technique). Finally, we record the squared error of these applications of the fitting technique for each value of λ, by testing the results over the subset that was taken apart at the beginning (known as *testing set*). This process is repeated k times so that each subset is taken as the testing set, by performing the fitting technique with different values for λ over the remaining $k-1$ subsets (Figure 3.1).

As a consequence, we have a total of k different squared errors for each value of λ, which are averaged for each λ in order to create a *CV error curve*, $CV(\lambda)$. We will select the value of λ for which $CV(\lambda)$ reaches its minimum as the tuning parameter that optimises the fit. In practice, we use 5 and 10 as typical values for k [17].

3.2.2 Ridge Regression

In 1970, Hoerl and Kennard [27] proposed a regularised method: *Ridge regression*, which takes into consideration a shrinkage effect as the penalty term. Such penalty resulted in a model with less variability than the usual best-subset selection approach.

Ridge regression minimises the residual sum of squares penalising with an ℓ_2 norm the vector of coefficients:

$$\hat{\beta}_{Ridge} = \arg\min_{\beta\in\mathbb{R}^p}\left\{\frac{1}{2N}\|y-X\beta\|_2^2+\lambda\|\beta\|_2^2\right\}, \tag{3.5}$$

where, when $\lambda = 0$ we have a typical MLR approach.

It is important to note that Ridge regression allows us to control, in a way, the bias-variance trade-off since, as λ increases, the regression fit flexibility decreases, yielding a model with decreased variance but increased bias. In this case, the tuning parameter serves as shrinkage factor on the penalty term, minimising the estimated association of all predictors (variables) with the response vector.

Since we will obtain a different fit each time we vary the tuning parameter λ in Ridge regression, we will make use of CV as a resampling method. This will let us evaluate multiple values of λ in order to obtain a fit with the minimum error.

3.2.3 Least Absolute Shrinkage and Selection Operator (LASSO)

Whenever we work on a linear model that fits a large data set, considering several potentially explanatory variables, it is convenient for us to try to minimise issues such as lack of interpretability, overfitting, among others. This section explores the LASSO method, a penalised approach introduced in 1996 by Robert Tibshirani [16]. We will also use cross-validation (as resampling method) in order to deal with those problems.

Robert Tibshirani [16] first introduced the Least Absolute Shrinkage and Selection Operator (LASSO) in 1994. It is a method that takes the same inspiration as the Ridge Regression proposed by Hoerl and Kennard [27], adopting an optimisation approach by implementing penalisations to the design of the problem: while Ridge regression penalises the parameters with an ℓ_2 norm (i.e. the Euclidean norm of vectors; see equation (3.5)), the LASSO uses an ℓ_1 norm (the sum of absolute values of the components of a vector; see equation (3.7)).

The idea behind an ℓ_q-penalised method is that it promotes shrinkage of the values of coefficients in our model. Ridge regression shrinks the coefficients but does not perform variable selection, while the LASSO takes advantage of both, subset-selection and shrinkage so that it provides an interpretable model with the advantage that an ℓ_1-penalisation induces a sparse model by shrinking or even setting some coefficients to zero.

We will consider a linear regression approach to our data. Let $y = \beta_0\mathbf{1} + X\beta$ be the matrix representation of our setting where $y \in \mathbb{R}^N$ is a vector of responses, $\beta_0 \in \mathbb{R}$ is the intercept coefficient, $\mathbf{1}$ is the vector of N ones, $\beta = (\beta_1, \dots, \beta_p)^\top$ is the vector of coefficients that best fit the model to the collected observations and $X = [X_1 \dots X_p]$ is the *design matrix* which columns are $X_1, \dots, X_p \in \mathbb{R}^N$, i.e., the variables of the model. We aim to describe a quantitative response vector

y from the matrix X consisting of predictors. The LASSO problem can be written as the following convex optimisation problem:

$$\begin{aligned} &\underset{\beta_0,\beta}{\text{minimise}} && \left\{\frac{1}{2N}\|y-\beta_0\mathbf{1}-X\beta\|_2^2\right\} \\ &\text{subject to} && \|\beta\|_1 \le t, \end{aligned} \tag{3.6}$$

where $\|\cdot\|_2$ is the Euclidean norm of vectors and $\|\beta\|_1 = \sum_{j=1}^{p}|\beta_j|$.

Typically, when the features are expressed in different units, we standardise the columns (i.e. $\bar{X}_i = 0$ and with unit variance $\frac{1}{N}\|X_i\|_2^2 = 1$). We also center the values y_i (i.e. $\frac{1}{N}y_i = 0$) so that we can omit the intercept term β_0 since, once we get an optimal solution for $\hat{\beta}$ on the centralised data, we can recover those for the uncentralised setting by maintaining $\hat{\beta}$ the same, and computing

$$\hat{\beta}_0 = \bar{y} - \sum_{j=1}^{p}\bar{X}_j\hat{\beta}_j,$$

where $\bar{y}$ and $\{\bar{X}_j\}_1^p$ are the arithmetic means.

For convenience purposes, we will rewrite the LASSO problem (3.6) in its Lagrangian form [18, 26]:

$$\hat{\beta}_{LASSO} = \arg\min_{\beta\in\mathbb{R}^p}\left\{\frac{1}{2N}\|y-X\beta\|_2^2 + \lambda\|\beta\|_1\right\}. \tag{3.7}$$

Lagrangian duality [28] guarantees a bijection between the constraint t in (3.6) and λ in (3.7). We will call $\lambda \ge 0$ the *tuning parameter* of the model. When $\lambda = 0$, we have an ordinary least squares fit; and for a λ sufficiently large, this method yields a model where all coefficients are null.

In order to select an ideal tuning parameter which does not fall into any extreme case (where the estimation error might be inflated), we will use Cross-validation (CV) as the *resampling method*.

3.2.4 Elastic net

One of the biggest drawbacks of the LASSO is that it does not perform very well when there are highly correlated variables in the design matrix. Even though it still promotes shrinkage and performs variable selection, the LASSO tends to disregard correlated variables indistinctly. The *Elastic net* [29] arises as a method that takes the best out of the Ridge and LASSO penalties by finding solutions to the next expression:

$$\underset{\beta\in\mathbb{R}^p}{\arg\min}\left\{\frac{1}{2}\|y-X\beta\|_2^2+\lambda\left[\frac{1}{2N}(1-\alpha)\|\beta\|_2^2+\alpha\|\beta\|_1\right]\right\}, \tag{3.8}$$

where $\alpha \in [0,1]$ is a varying parameter. When $\alpha = 1$, we have an ℓ_1-norm penalty or LASSO approach, whereas with $\alpha = 0$, we obtain an ℓ_2-norm penalty corresponding to Ridge regression.

Once again, we will use CV to evaluate multiple values for λ as well as for α.

3.3 Summary of chapter

We will now highlight some remarkable results from this chapter:

- $k-$fold Cross-validation (CV) is a resampling method that splits N samples into k different randomly selected subsets (of roughly equal size). All but one of these subsets are used to *train* the fitting model and the remaining subset is used to *test* the accuracy of the model. The same process is repeated k times, until all subsets have been used as testing sets. Common values for k are 5 and 10.

- Ridge regression is a regularised method that promotes shrinkage of the values of the coefficients of β by adding an ℓ_2-norm penalty to the least squares approach (which reduces variance by adding some bias). Ridge regression seeks to solve the next optimisation problem:

$$\hat{\beta}_{Ridge} = \arg\min_{\beta\in\mathbb{R}^p}\left\{\frac{1}{2N}\|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2\right\},$$

  where $\|\cdot\|_2^2$ is the Euclidean norm.

- Least Absolute Shrinkage and Selection Operator (LASSO) is a regularised method that promotes both, shrinkage and variable selection, by adding an ℓ_1-norm penalty term to the least squares approach. The LASSO approach solves the following optimisation problem:

$$\hat{\beta}_{LASSO} = \arg\min_{\beta\in\mathbb{R}^p}\left\{\frac{1}{2N}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1\right\},$$

  where $\|\beta\|_1 = \sum_{j=1}^{p}|\beta_j|$.

- The Elastic net is a regularised method that decides which penalisation is best for fitting by adding another parameter $\alpha \in [0,1]$ to the penalty function:

$$\arg\min_{\beta\in\mathbb{R}^p}\left\{\frac{1}{2}\|y - X\beta\|_2^2 + \lambda\left[\frac{1}{2N}(1-\alpha)\|\beta\|_2^2 + \alpha\|\beta\|_1\right]\right\},$$

  such parameter can be tuned by CV as well. When $\alpha = 1$, we have the LASSO approach, whereas with $\alpha = 0$, we have Ridge regression.

Chapter 4

False Discovery Rate (FDR) control methods

Another approach to perform variable selection is through a statistical inference point of view. Given that we can estimate the respective $p-$value associated to the hypothesis of having a null variable (i.e., a variable that has no relevant effect on the response), one of the further methods followed by researchers was to execute *tests of significance* for each one of the variables.

In the procedure proposed by Holm [30], the $p-$values corresponding to their respective testing hypotheses are ordered and compared to a critical value $f(\alpha)$ (where $0 < \alpha < 1$ is a fixed bound), to reject one hypothesis at a time until it is impossible to do further rejections. Hochberg [31] presented a modified method that rejects all hypotheses with $p-$value lower or equal than a certain critical value and introduced the *family-wise error rate* (FWER), which is the probability that at least one rejection was made by error among p (number of variables) simultaneous hypothesis tests [25].

4.1 Introduction to the False Discovery Rate (FDR)

In 1995, Benjamini and Hochberg [32] proposed a different point of view for the same problem: they thought that it was also important to take into account the number of hypotheses falsely rejected among all the rejected hypotheses. The authors considered the generalised event illustrated in Table 4.1 [32], where m null hypotheses were simultaneously tested, of which m_0 represents the amount of true null hypotheses, and R is the number of null hypotheses rejected.

	Declared non-significant	**Declared significant**	**Total**
True null hypothesis	U	V	m_0
Non-true null hypothesis	T	S	$m - m_0$
	$m - R$	R	m

Table 4.1: Number of errors committed when testing m null hypotheses.

We define the False Discovery Proportion (FDP) of a testing rule $\mathscr{D}$ as the proportion of hypothesis declared non-null (or significant) by error (*false discoveries*) among all the hypotheses declared significant. In other terms,

$$FDP(\mathscr{D}) = \frac{V}{R} \tag{4.1}$$

(we define the FDP to be 0 if $R = 0$).

Although we cannot determine which null hypotheses are true or not (and therefore we cannot compute the FDP), what we can do is to control its expectation. We define the *False Discovery Rate* (FDR) as the expectation of the FDP:

$$FDR(\mathscr{D}) = \mathbb{E}[FDP(\mathscr{D})] \tag{4.2}$$

The testing rule proposed by Benjamini and Hochbergh [32] (BH_q) yields the following theorem, which indicates that $FDR(BH_q)$ is controlled [25] at a level $0 < q < 1$ (i.e. $FDR(BH_q) \leq q$):

Theorem 1 (Benjamini-Hochberg)**.** *For independent test statistics and for any configuration of false null hypotheses, the FDR is controlled at level q under the following procedure:*

1. *Let $P_{(1)} \leq P_{(2)} \leq \cdots \leq P_{(m)}$ be the ordered p−values and let $H_{(i)}$ denote their respective null hypothesis.*
2. *Let k be the largest i for which $P_{(i)} \leq \frac{i}{m}q$.*
3. *Reject all $H_{(i)}$, with $i \leq k$.*

The connection between significance test and variable selection is that we can obtain a set of variables from a linear setting that represents a good fit by controlling the FDR of the variables. Under this perspective, we define the FDR of a variable selection procedure that returns a subset $\hat{S} \subset \{1, \ldots, p\}$ as

$$FDR = \mathbb{E}\left[\frac{\#\{j : \beta_j = 0 \text{ and } j \in \hat{S}\}}{\#\{j : j \in \hat{S}\} \vee 1}\right], \tag{4.3}$$

where $a \vee b$ denotes $\max\{a, b\}$ [11] and $\beta = (\beta_1, \ldots, \beta_p)$ is the vector of estimated coefficients as indicated before.

Further main studies on the control of the FDR were published by Storey [33], Verhoeven, Simonsen and McIntyre [34] and by Su, Bogdan and Candès [35].

For this work, we have chosen to use an FDR control method given that we require to assess the certainty of having selected the most relevant features. Particularly, the technique we will use in this project is called the *knockoffs filter*.

4.2 Knockoffs filter

We will now give details about the **Knockoffs filter** approach, which tackles the problem of choosing true explanatory features on a model, through techniques such as the LASSO regression and hypothesis-testing.

Barber and Candès [11] introduced this method, which controls the FDR at a desired level q, appearing to perform better than the method proposed by Benjamini and Hochberg [32], and showing more statistical power (i.e., the probability of correctly rejecting a null hypothesis given that the alternative hypothesis). In other words, it is a method that selects the features that best fit the data (true signals) and omit those variables that are not meaningful in the model.

Suppose a setting $\mathbf{y} = \mathbf{X}\beta + \epsilon$ as before, where $\mathbf{X} = \left[X_1 \dots X_p\right]$ is the design matrix with p columns $X_i \in \mathbb{R}^N$. The idea behind this approach is to try to imitate the same correlation behaviour of the p original features, by creating *knockoff* copies $\tilde{X}_j$ for each variable X_j and extending the design matrix $\mathbf{X}$ by adding the columns $\tilde{X}_j$, so that, through a variable selection method one can select those variables that perform better than their respective knockoff copy.

Different variations have been made to the original proposal from Barber and Candès [11], depending on the restrictions of the design and the nature of the data set. Weinstein, Barber and Candès [36] provide a version of knockoffs filter using proper LASSO statistics and considering an independent and identically distributed Gaussian design. Furthermore, Dan and Barber [37] published a study that utilises the group LASSO [38], a technique that respects grouped sets of variables. In their study, they propose a version of knockoffs filter that selects relevant groups of variables that have a null influence on the response, obtaining satisfactory results on the same data set used by Barber and Candès [11]. On the other hand, Barber and Candès [39] present a study of the case $p > N$ (i.e. we have more variables than observations) for knockoffs, applying their analysis to a genome-wide association study (GWAS). More interesting applications of the knockoffs filter to GWAS can be found in the studies by Katsevich and Sabatti [40] and Sesia, Sabatti and Candès [20].

Finally, a new interpretation of knockoffs also worth to mention is the *Model-X Knockoffs* [41], which implements a conditional randomisation test to the knockoffs model that allows the knockoffs filter to be applied to settings where $p >> N$. However, one of the biggest limitations of the randomisation is that it is computationally quite expensive as it must compute numerous test statistics, while in the original knockoffs filter only one test statistic is computed.

The procedure we will follow in this project, as described by Dai and Barber [37], consists of two main steps: construction of knockoffs and filtration of the results.

4.2.1 Constructing the knockoffs

In the initial setting $X = [X_1 \dots X_p]$ (supposing each feature X_i is centered and normalised), we look for knockoff copies $\tilde{X}_i$ for each X_i, such that the matrix $\tilde{X} = \left[\tilde{X}_1 \dots \tilde{X}_p\right]$ holds the next properties:

$$\begin{aligned} \tilde{X}^\top \tilde{X} &= \Sigma && (4.4) \\ \tilde{X}^\top X &= \Sigma - \text{diag}\{s\}, && (4.5) \end{aligned}$$

for some $0 \preccurlyeq s \in \mathbb{R}^p$, where $\Sigma = X^\top X$ is the Gram matrix of the original columns, after normalising each one of original features (i.e., $\Sigma_{ii} = \|X_i\|_2^2 = 1$).

We will now perform the LASSO on the augmented matrix $[X\tilde{X}] = [X_1, \dots, X_p, \tilde{X}_1, \dots, \tilde{X}_p]$, with response vector y, varying the parameter λ from a large value and towards 0:

$$\widehat{\beta}(\lambda) = \arg\min_{b \in \mathbb{R}^{2p}} \{\frac{1}{2N} \|y - [X\tilde{X}]b\|_2^2 + \lambda \|\beta\|_1\}. \qquad (4.6)$$

If X_j is a true signal, then X_j should enter the LASSO earlier than its knockoff copy $\tilde{X}_j$, for a sufficiently large λ. It is important to note that, if X_j is null (i.e., X_j has no real effect on the response vector y), then it is equally likely to enter the model after or before its knockoff copy $\tilde{X}_j$.

4.2.2 Filtering the results

We will construct statistics Z_j and $\tilde{Z}_j$ (for each $j = 1, \dots, p$) that measure the relevance of each feature X_j and $\tilde{X}_j$ respectively, in the response vector:

$$\begin{aligned} Z_j &= |\hat{\beta}_j(\lambda)|, \\ \tilde{Z}_j &= |\hat{\beta}_{j+p}(\lambda)|. \end{aligned}$$

The knockoff filter compares the Z_j's to the $\tilde{Z}_j$'s and selects only those variables that are better than their knockoff copy. This is possible because of the construction of the knockoff copies: the copies of null variables hold the *pairwise exchangeability property*, that is, it is possible to interchange the position of the statistics corresponding to null variables without producing any change in the joint distribution $(Z_1,\dots,Z_p,\tilde{Z}_1,\dots,\tilde{Z}_p)$.

In order to compare the statistics, we require an anti-symmetric function h to compute the *symmetrised knockoff statistics*

$$W_j = h(Z_j,\tilde{Z}_j) = -h(\tilde{Z}_j,Z_j), \forall j \tag{4.7}$$

such that $W_j > 0$ means that X_j seems to be better (i.e. more important) than its knockoff copy. We will take $W_j = Z_j - \tilde{Z}_j$.

The final step is to select only those features with large, positive values of W_j, considering an adaptive threshold. In this case, we will use a threshold defined as

$$T = \min\left\{t \in \mathcal{W} : \frac{1 + \#\{j : W_j \leq -t\}}{\#\{j : W_j > t\} \vee 1} \leq q\right\}, \tag{4.8}$$

where q is the desired FDR bound and $\mathcal{W} = \{|W_j| : j = 1,\dots,p\}\backslash\{0\}$ is the set of unique nonzero values attained by the $|W_j|$'s. In their paper, Barber and Candès [11] propose this threshold demonstrating that it is a stopping time for selection, assuming, without loss of generality, that $|W_1| \geq \cdots \geq |W_p|$. Then, the process tests from $t = |W_p|$ onward, until a value of t is found such that it is the smallest possible value of T satisfying $\widehat{FDP}(t) := \frac{\#\{j:W_j \leq -t\}}{\#\{j:W_j > t\} \vee 1} \leq q$.

The resulting theorem from this procedure is stated below:

Theorem 2 (Knockoffs$_+$ filter [11])**.** *Let $\hat{S} = \{j : W_j \geq T\}$ be selection model for indexed parameters from $1,\dots,p$. For any $q \in [0,1]$, the knockoff method satisfies*

$$FDR = \mathbb{E}\left[\frac{\#\{j : \beta_j = 0 \; and \; j \in \hat{S}\}}{\#\{j : j \in \hat{S}\} \vee 1}\right] \leq q,$$

where the expectation is taken over the Gaussian noise ϵ, while treating X and $\tilde{X}$ as fixed.

This theorem guarantees that, under the constructions described above, the false discovery rate can be controlled significantly, resulting in a good alternative for variable selection.

It is important to note that, although the knockoff filter proposed by Barber and Candès [11] uses the LASSO, this method is not restricted to it. In particular settings, it is only required that the method from which the statistics Z_j and $\tilde{Z}_j$ satisfy the *antisymmetry* and *sufficiency* requirements; that is, swapping X_j with $\tilde{X}_j$ has the same effect of swapping Z_j with $\tilde{Z}_j$ (antisymmetry) and that W_j depends only on the Gram matrix of the augmented setting $[X\tilde{X}]$ and on feature-response inner products $[X\tilde{X}]^\top y$ (sufficiency). In the LASSO coefficients, computed with a fixed λ, we satisfy both, fairness and sufficiency requirements, however, computing the LASSO coefficients through cross-validation fails to meet the sufficiency requirement since it does not consider a fixed X. This is why in the knockoffs filter we cannot use cross-validation to tune an optimal λ.

4.3 Summary of chapter

We will now highlight some remarkable results from this chapter:

- The Knockoffs filter is a False Discovery Rate (FDR) control technique that performs hypothesis testing to select meaningful variables in a model. This method creates a matrix $\tilde{X}$ which columns are knockoff copies of the columns of X, satisfying $\tilde{X}^\top \tilde{X} = \Sigma$ and $\tilde{X}^\top X = \Sigma - \text{diag}(s)$, for some $0 \preccurlyeq s \in \mathbb{R}^p$, where $\Sigma = X^\top X$ is the Gram matrix of the original columns, after normalising each one of original features (i.e., $\Sigma_{ii} = \|X_i\|_2^2 = 1$). Then LASSO is performed on the augmented matrix $[X\tilde{X}] = [X_1,\ldots,X_p,\tilde{X}_1,\ldots,\tilde{X}_p]$ to solve

$$\widehat{\beta}(\lambda) = \arg\min_{b\in\mathbb{R}^{2p}}\{\frac{1}{2N}\|y-[X\tilde{X}]b\|_2^2 + \lambda\|\beta\|_1\}.$$

  - All the entry times of every variable into the LASSO path are registered and compared, yieding a statistic W_j, which indicates whether a column X_j is 'better' or not that its knockoff copy $\tilde{X}_j$. If $W_j > 0$, it means that X_j seems to be a better candidate than $\tilde{X}_j$ to enter into the model.
  - In this work we will use the adaptive threshold T proposed by Barber and Candès [11]:

$$T = \min\left\{t : \frac{1+\#\{j : W_j \le -t\}}{\#\{j : W_j > t\} \vee 1} \le q\right\},$$

    where q is the desired FDR bound.
  - This threshold yields the following theorem [11] that guarantees FDR control: Let $\hat{S} = \{j : W_j \ge T\}$ be selection model for indexed parameters from $1,\ldots,p$. For any $q \in [0,1]$, the knockoff method satisfies

$$FDR = \mathbb{E}\left[\frac{\#\{j : \beta_j = 0 \text{ and } j \in \hat{S}\}}{\#\{j : j \in \hat{S}\} \vee 1}\right] \le q,$$

    where the expectation is taken over the Gaussian noise ϵ, while treating X and $\tilde{X}$ as fixed.

Chapter 5

Parsing and description of data sets

Having a clear structure of the data we were working with was a key factor for the success of this project. Therefore, we required to manipulate the main data set, so that we could dispose of those transactions with unclear or incomplete information or that were not meaningful for the purpose of the investigation.

Throughout this chapter, we will describe all the data sets used to develop this research project. We will start by describing the main data set, which contains information about EV charging transactions in public and semi-public charging stations in the Netherlands, during the period 2012-2016. Additionally, we will construct new data settings by adding information from other two public sources containing information on the demographics of the Netherlands as well as the number of places of interest (schools, hospitals, etc.) surrounding each charging station.

5.1 EV Charging Transactions data set

In this section we will give details about the data set `Transactions.Rda` provided by the ElaadNL. This data set is available upon request for non-commercial purposes and therefore, cannot be publicly shared. It consists of 1,060,763 transactions comprising 17 features which are described below. We used version 3.4.3 of `R` [42] on `RStudio` [43] to parse this data set.

The following list is a description of the variables in the `Transactions.Rda` data set:

1. `TransactionId`. Transaction identifiers (since these IDs were not unique, the column `Index` was created. Only 960,438 unique IDs).
2. `Index`. Transaction IDs. 1'060,763 unique Transaction IDs.
3. `ChargePoint`. ID of the charging station. 1,747 unique IDs.
4. `Connector`. ID of the outlet were the vehicle was connected. Values: 1, 2, 3, 13, 14.
5. `UTCTransactionStart`. Local time when the transaction started.
6. `UTCTransactionStop`. Local time when the transaction stopped.
7. `MeterStart`. Readings of energy of the vehicle at the beginning of the transaction (Wh).
8. `MeterStop`. Readings of energy of the vehicle at the end of the transaction (Wh).
9. `StartCard`. Radio Frequency Identification (RFID) card used at the beginning of the transaction. 53,850 unique cards.
10. `StopCard`. RFID card used at the end of the transaction. 53,256 unique cards.
11. `ConnectedTime`. Amount of time that the vehicle was connected to the charging station.
12. `ChargeTime`. Amount of time on which the vehicle was charging (an effective energy transfer took place).
13. `IdleTime`. Difference between `ConnectedTime` and `ChargeTime`.
14. `TotalEnergy`. Total energy acquired during the charging transaction.
15. `MaxPower`. No description provided.
16. `lat`. Latitude of the charging station. 1,725 unique coordinates (along with `lon`).
17. `lon`. Longitude of the charging station. 1,725 unique coordinates (along with `lat`).

5.1.1 Data set parsing

In order for us to make progress with this project, we needed to filter the `Transactions.Rda` data set, discarding some transactions and variables that were not specific or meaningful to the purpose of this investigation. We will now describe the filtering process that led us to a cleaner configuration.

Initially we had 1,747 many different charging stations, however, only 1,725 of them had unique coordinates. This happens since duplicated coordinates (for two or more different IDs in the variable `ChargePoint`) correspond to multiple charging stations at the same premise (Figure 5.1 [44]). To facilitate the geographical analysis, we renamed these charging stations with shared coordinates, choosing the name of the first element from an alphabetically sorted list of their names for each coordinate (see Table 5.1).

Table 5.1: Charge points with shared coordinates (the names in **bold** letters were chosen to represent all the charging stations with their same coordinate).

Coordinate (lat, lon)	ChargePoint
(50.894937, 6.059926)	**1. AL194**
	2. AL462
(51.431825, 5.428635)	**3. AL268**
	4. AL269
	5. AL602
	6. AL603
	7. AL604
(52.016245, 5.040206)	**8. AL427**
	9. AL547
(52.439989, 4.81851)	**10. AL91**
	11. AL92
(52.158933, 4.478506)	**12. DB0351**
	13. DB0353
(52.550007, 4.661949)	**14. EV0034**
	15. EV0037
(52.544901, 4.660249)	**16. EV0035**
	17. EV0038
(52.645705, 4.749776)	**18. EV0121**
	19. EV0132
(51.807006, 5.729469)	**20. EV0178**
	21. EV0182

Coordinate (lat, lon)	ChargePoint
(52.300911, 4.47821)	**22. REE520**
	23. REE521
(51.813498, 4.657108)	**24. REE651**
	25. REE652
(51.484048, 3.962978)	**26. RWE003**
	27. RWE004
(51.486397, 3.955032)	**28. RWE005**
	29. RWE006
(51.263731, 3.907549)	**30. RWE007**
	31. RWE008
(51.515495, 3.995568)	**32. RWE011**
	33. RWE012
(51.335689, 3.838085)	**34. RWE013**
	35. RWE014
(51.228606, 3.801977)	**36. RWE015**
	37. RWE016
(51.334748, 3.821503)	**38. RWE017**
	39. RWE018
(51.400989, 3.534943)	**40. RWE019**
	41. RWE020

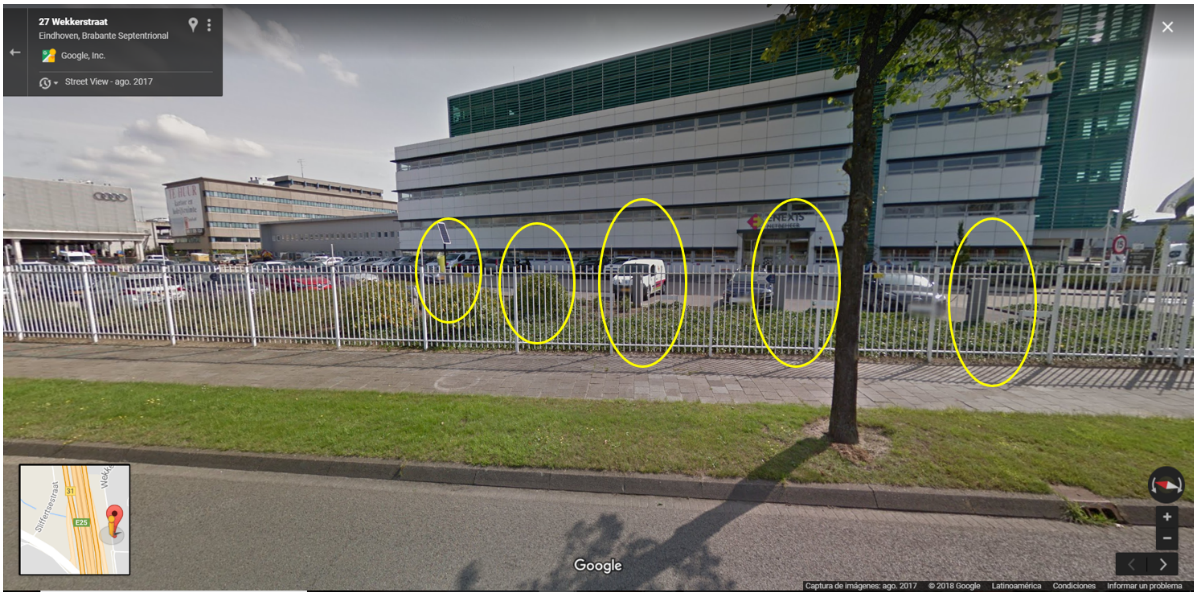

Figure 5.1: Charging stations located at the coordinate (51.431825, 5.428635). Here, 5 charging stations are located at the same premise (yellow ovals).

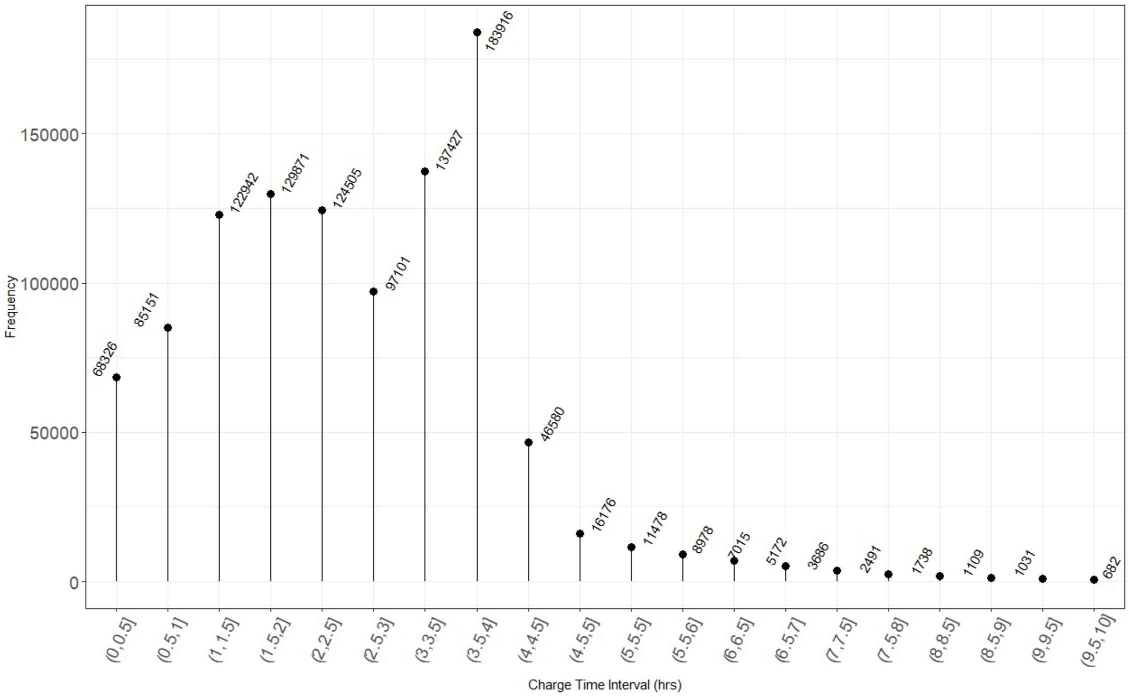

Figure 5.2: Frequency distribution with intervals of 30 min of transactions with charge time shorter than 10 hours.

The number of transactions with a charging time longer than 10 hours is 5,388, which is less than 0.51% of the total amount of transactions. The frequency distribution of the transactions with a charge time shorter than 10 hours is illustrated in Figure 5.2.

As to the cards used in the data set, 481 records had null values in either `StartCard` or `StopCard`, whereas the number of transactions with a `StartCard` different from the `StopCard` is 35,244, including the transactions with null values in their cards. This leaves a total of 53,245 unique cards in the data set where `StartCard` is identical to its `StopCard`. Within this remaining subset, 796 transactions have their `StartCard` and `StopCard` equal to “00000000". Additionally, there are 494 rows such that the difference between `ConnectedTime` and the sum of `IdleTime` plus `ChargeTime` is greater than 0.0002777778 hrs (i.e. 1 second).

Finally, if we consider a data set filtered under the next conditions:

1. `ChargeTime` is shorter than 10 hours;
2. `StopCard` is identical to its `StartCard` and both different from “00000000";
3. The difference between `ConnectedTime` and the sum of `IdleTime` plus `ChargeTime` is less than 0.0002777778 hrs (1 second);
4. Consolidate the names of duplicated coordinates per location (according to Table 5.1).

and disregard the following variables:

- `index` (not meaningful for regression),
- `TransactionId` (substituted by `index`),
- `UTCTransactionStop` (to avoid correlated variables, keeping `UTCTransactionStart` and `ChargeTime`),
- `MeterStart` (we will keep `TotalEnergy`),
- `MeterStop` (we will keep `TotalEnergy`),
- `StopCard` (we are keeping `StartCard`),
- `IdleTime` (we are keeping `ConnectedTime` and `ChargeTime`),
- `MaxPower` (no description provided),

we obtain a cleaner data set consisting of $1,019,091$ transactions and 9 variables. We will call this resulting data set `Transactions_filtered.Rda`, which will serve as the main source of information to describe the behaviour of users and the performance of charging stations.

Due to the nature of the variable `UTCTransactionStart`, we will also create a subset of the `Transactions_filtered.Rda` data set called `df.Rda`, which consists of the same 1,019,091 transactions, but with the next 9 variables:

- `ChargePoint`.
- `ConnectedTime`.
- `ChargeTime`.
- `TotalEnergy`.
- `Weekday`. Categorical variable with 7 character-class values, from Monday to Sunday.
- `Month`. Categorical variable with 12 character-class values, from January to December.
- `Time`. Categorical variable with 4 character-class values: After-midnight (from 12am to 6am), Morning (from 6am to 12pm), Afternoon (from 12pm to 6pm) and Evening (from 6pm to 12am).
- `Season`. Categorical variabe with 4 character-class values: Winter (from January to March), Spring (from April to June), Summer (from July to September) and Autumn (from October to December).
- `Year`. Categorical variable with 5 character-class values, from 2012 to 2016.

5.2 Netherlands demographic and geographic data

Along with the `Transactions_filtered.Rda` data set, we will consider the information provided by the Central Bureau of Statistics of Netherlands [45] through the shapefile contained in [46]. In this file, the Netherlands map is sectioned into squares of 100 meters by 100 meters each (Figure 5.3), containing information on population, housing, social security, among others, on which at least 5 inhabitants or 5 dwellings were counted (for the year 2016).

In this project, we created a Voronoi diagram generated by the 1,725 different coordinates, in order to facilitate the geographical analysis of the network as well as the spatial coverage of each charging point (Figure 5.4). We have extracted the information on the population contained in each $100m \times 100m$ square and summed up all the population values of the squares which centroid is contained inside each Voronoi cell (e.g. see Figure 5.5).

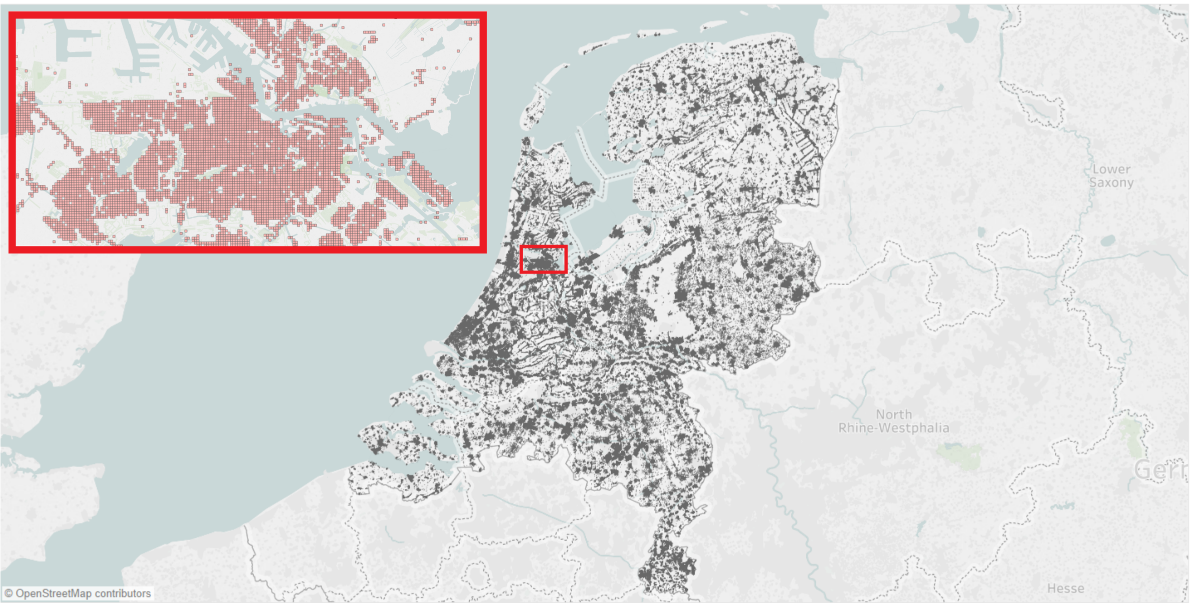

Figure 5.3: Visualisation of the Netherlands map sectioned into $100m \times 100m$ squares. The amplified area indicates the Amsterdam city centre.

Finally, we obtained the number of places of interest (schools, hospitals, restaurants, etc.) that are located within the area covered by each Voronoi polygon by extracting the information from the plugin `Quick OSM` included in QGIS [47] version 2.8.4. The resulting data set was called `chrg_spatial.Rda`, consisting of 1,725 observations (one per Vovonoi polygon) and 61 variables: charge point ID, population, and 59 types of places of interest (see Appendix A.1 for full list of variables).

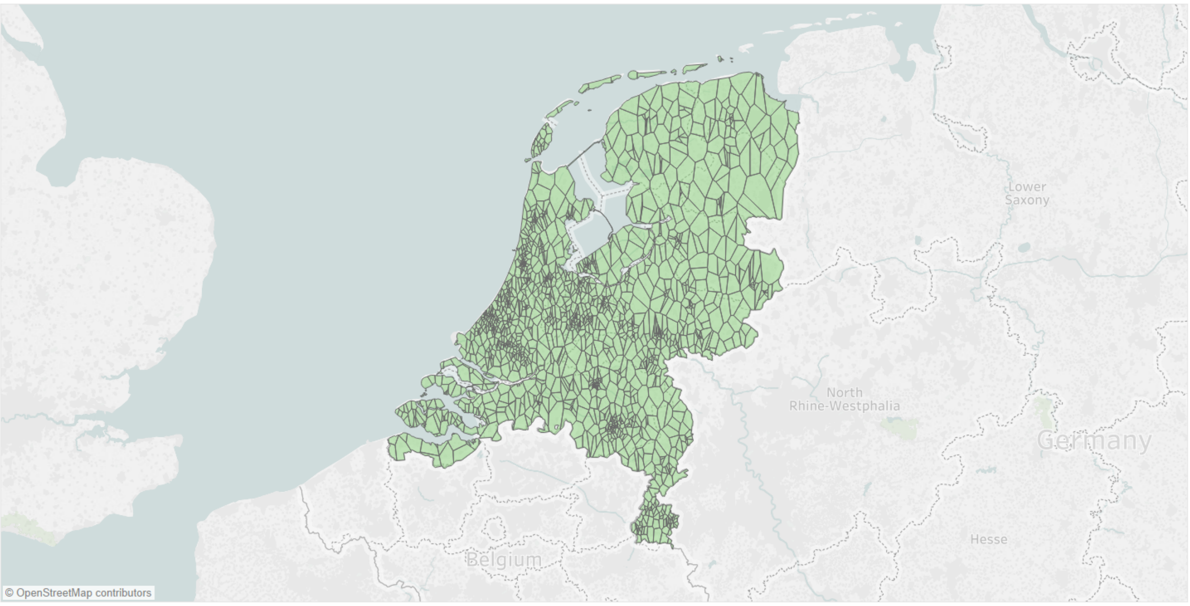

Figure 5.4: Voronoi diagram generated by the 1,725 different coordinates.

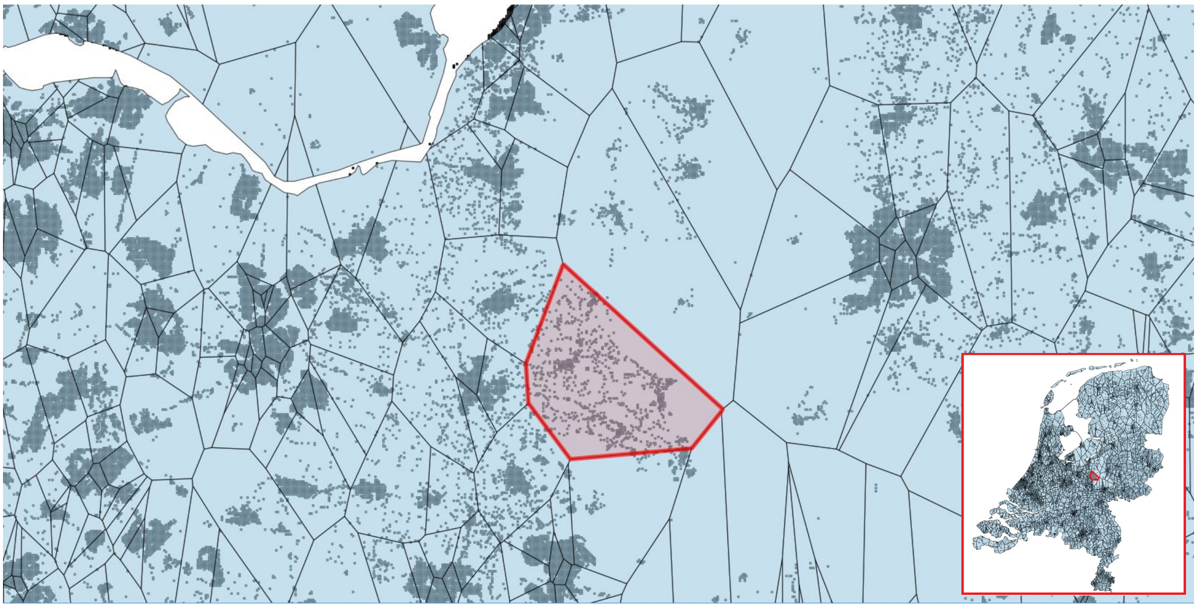

Figure 5.5: Amplified visualisation of the Voronoi diagram overlapping the sectioned map of the Netherlands. The red polygon corresponds to the surrounding area of the charging point REE065.

5.3 Summary of chapter

We will now provide a brief summary of the data sets obtained in this chapter:

- `Transactions_filtered.Rda`. Data set consisting of 1,019,091 observations (each one containing information of a single charging transaction) and 9 variables:

  - `ChargePoint`.
  - `Connector`.
  - `UTCTransactionStart`.
  - `StartCard`.
  - `ConnectedTime`.
  - `ChargeTime`.
  - `TotalEnergy`.
  - `lat`.
  - `lon`.

- `df.Rda`. Data set consisting of 1,019,091 rows (each row describes a single charging transaction) and 9 variables:

  - `ChargePoint`.
  - `ConnectedTime`.
  - `ChargeTime`.
  - `TotalEnergy`.
  - `Weekday`.
  - `Month`.
  - `Time`.
  - `Season`.
  - `Year`.

- `chrg_spatial.Rda`. Data set consisting of 1,725 rows (each row describes the population covered by each charging station as well as the number of places of interest (schools, hospitals, etc.) that are nearest to every charge point. A full list of the variables of this data set can be consulted in Appendix A.1.

Chapter 6

Analysis

We will show now how the methodology was applied onto the data sets described in Chapter 5. All the analyses were made using `Tableau` [48] (to visualise the data sets) and `R` [42, 43] mainly making use of the built-in functions from the packages `dplyr` [49] (for data manipulation), `glmnet` [50] (for linear regression methods) and `knockoff` [51] (for knockoffs filter).

6.1 Data sets overview

Throughout this section, we will explore the structure of `df.Rda` by analysing some results that arise from the data set only.

In Figure 6.1 we can see the behaviour of `TotalEnergy` with respect of `ChargeTime` for all transactions in the `df.Rda` data set. We have identified two clusters of points that seem to obey certain linear rule. In Figure 6.2, a yellow line was drawn, dividing these two clusters.

(6.1)

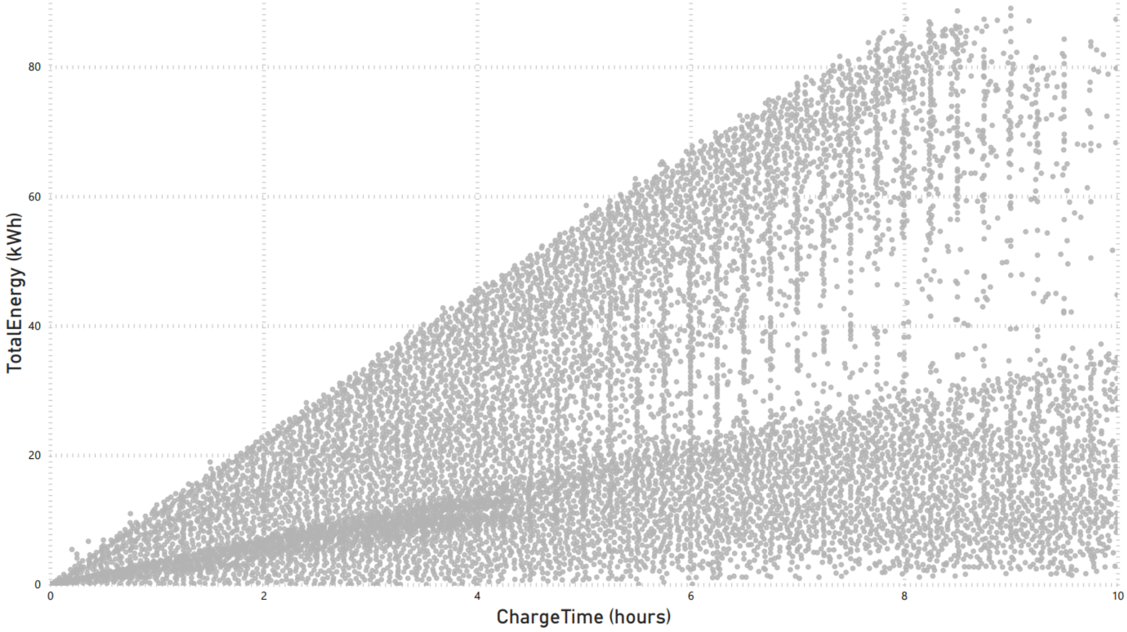

Figure 6.1: Comparison between variables ChargeTime and TotalEnergy.

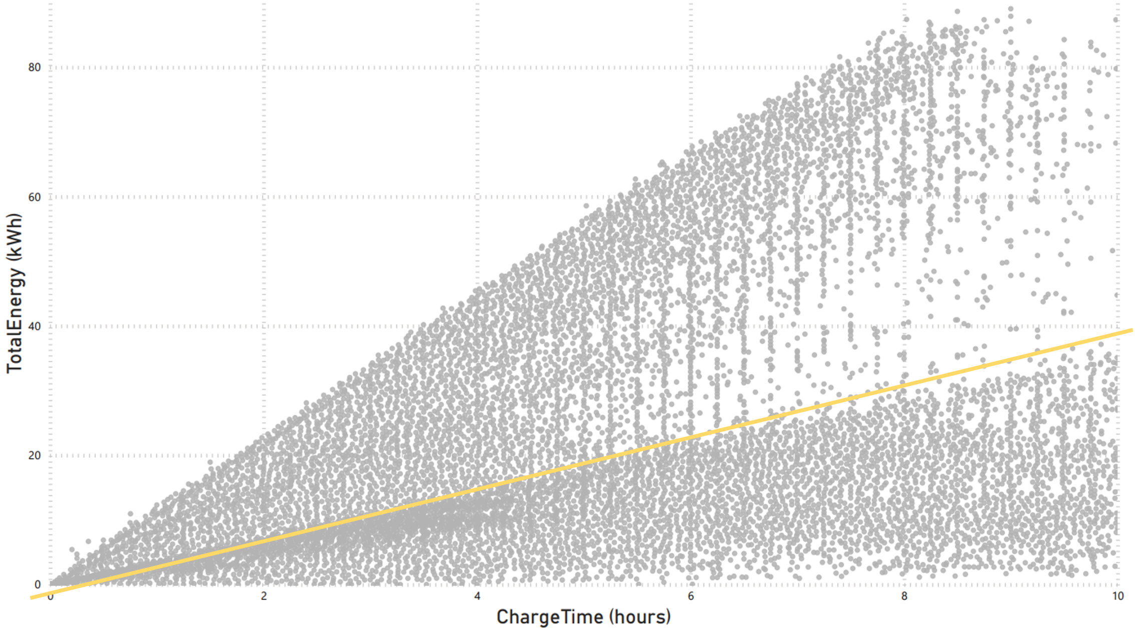

Figure 6.2: Dividing line (in yellow) of the two main clusters of points of Figure 6.1

Even though we had no further information from the data sets on what could be the reason for this effect, we can assume that it reflects the charging capacity and needs of PHEVs (Plug-in

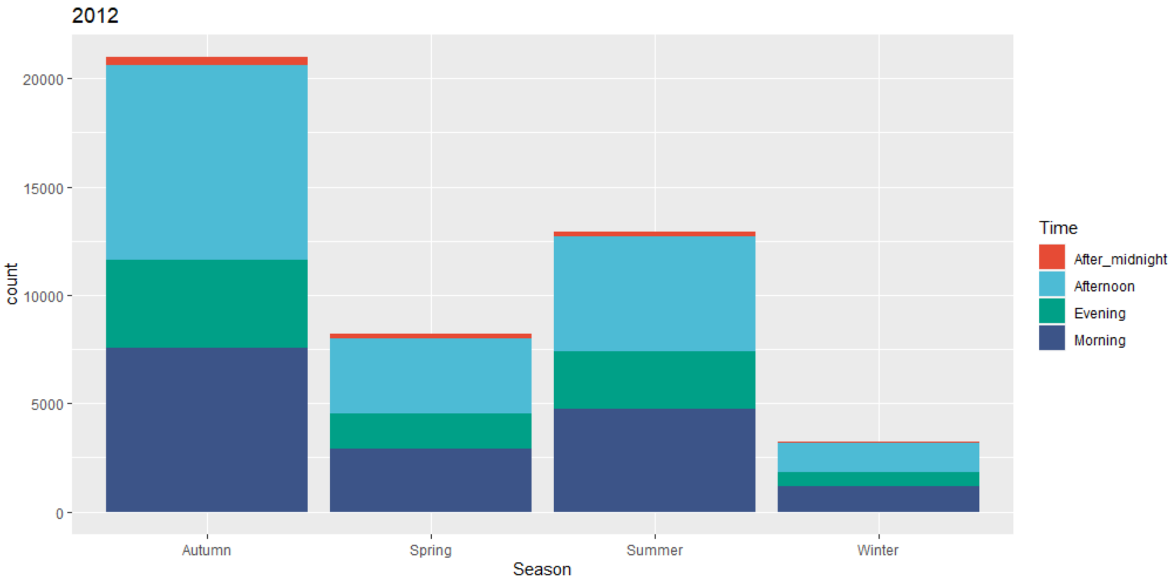

Figure 6.3: Amount of transactions made in 2012, every season, and distinguishing time of the day.

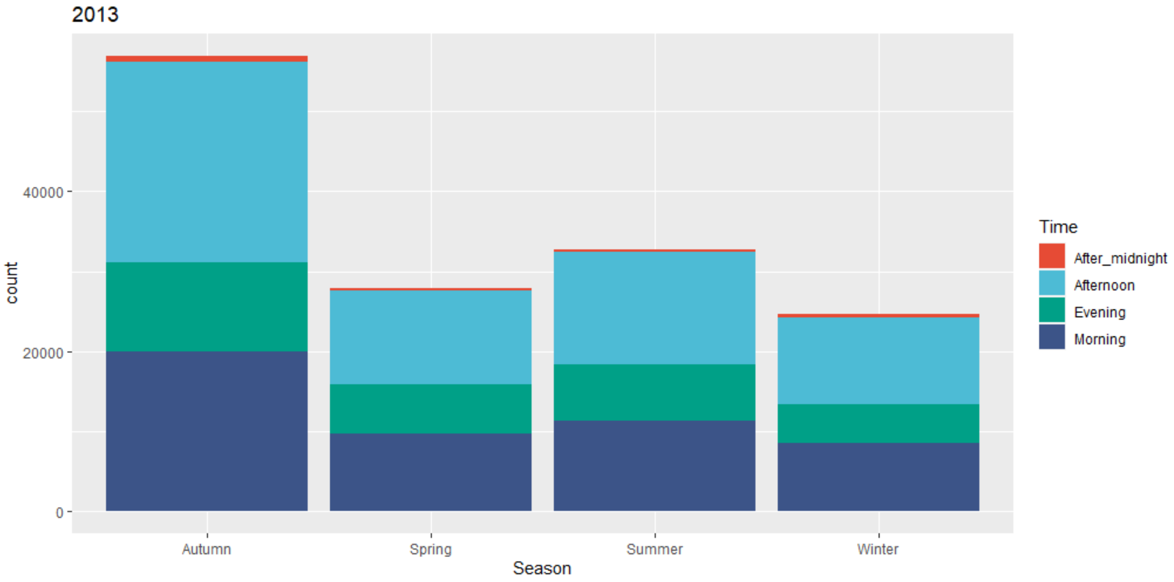

Figure 6.4: Amount of transactions made in 2013, every season, and distinguishing time of the day.

Hybrid Electric Vehicles) vs. BEVs (Battery electric vehicles) [52].

We can also notice that most of the transactions were made during the afternoons throughout the years, and that the day less utilised by the users to charge their vehicles is on Sunday, as shown in Figures from 6.3 to 6.7 and from 6.8 to 6.12.

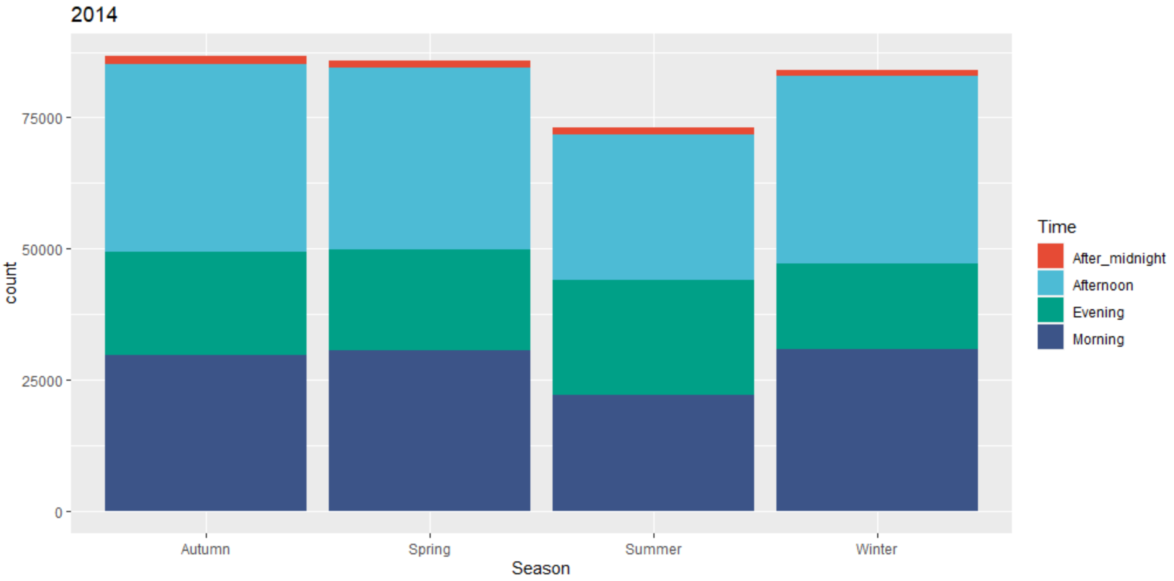

Figure 6.5: Amount of transactions made in 2014, every season, and distinguishing time of the day.

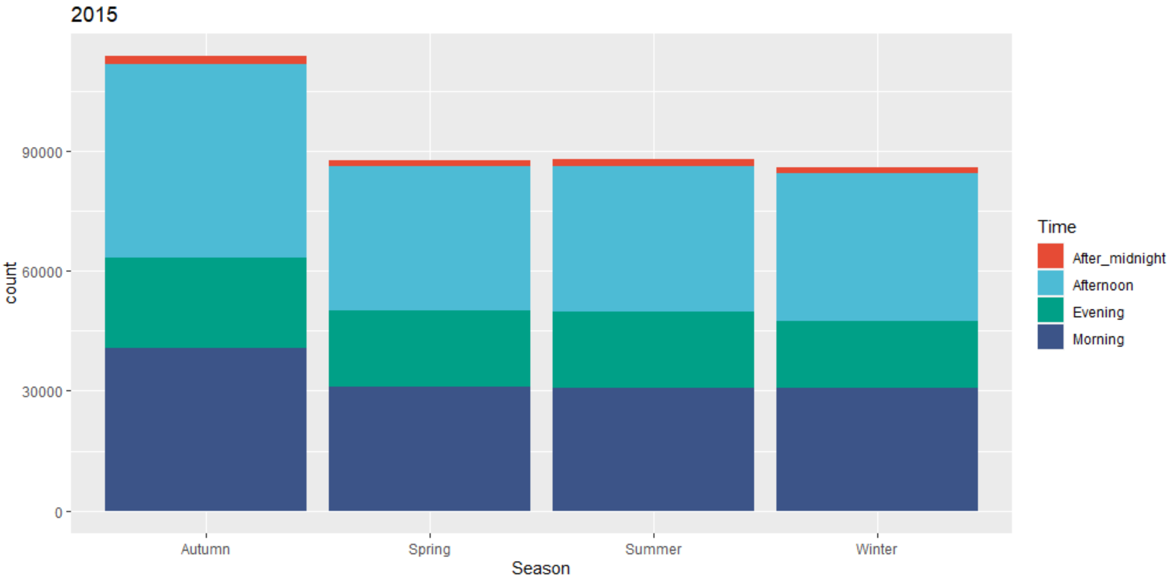

Figure 6.6: Amount of transactions made in 2015, every season, and distinguishing time of the day.

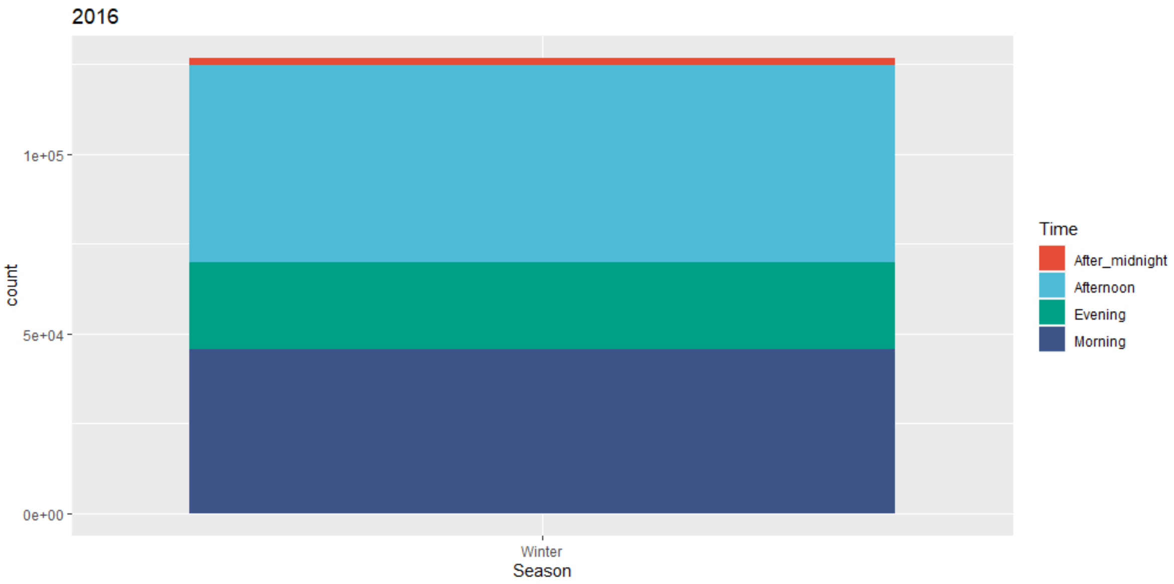

Figure 6.7: Amount of transactions made in 2016, every season, and distinguishing time of the day.

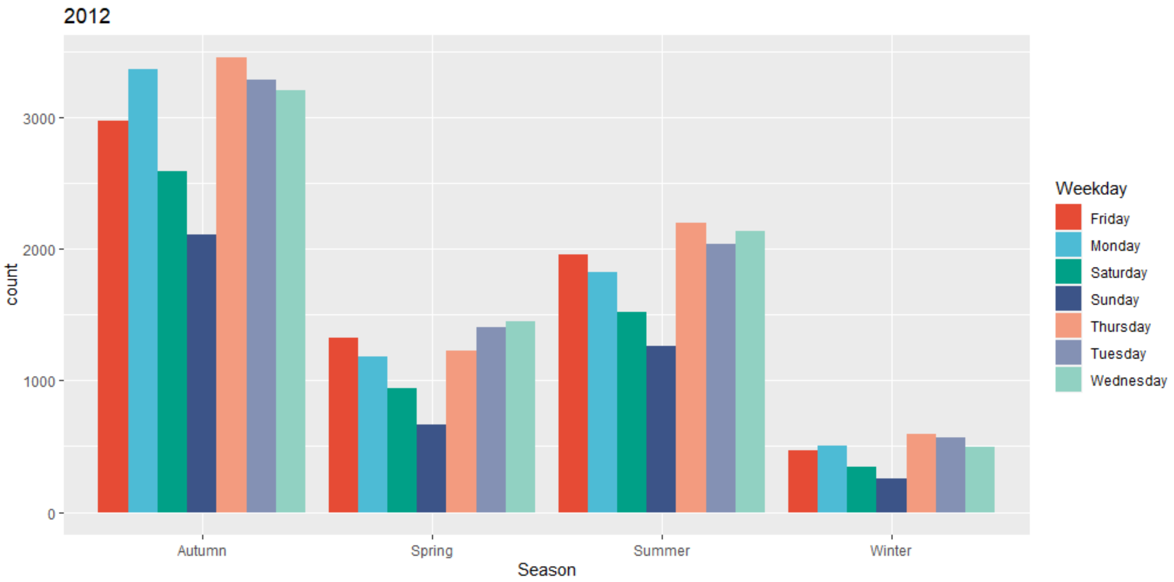

Figure 6.8: Amount of transactions made in 2012, every season, and separating by weekday.

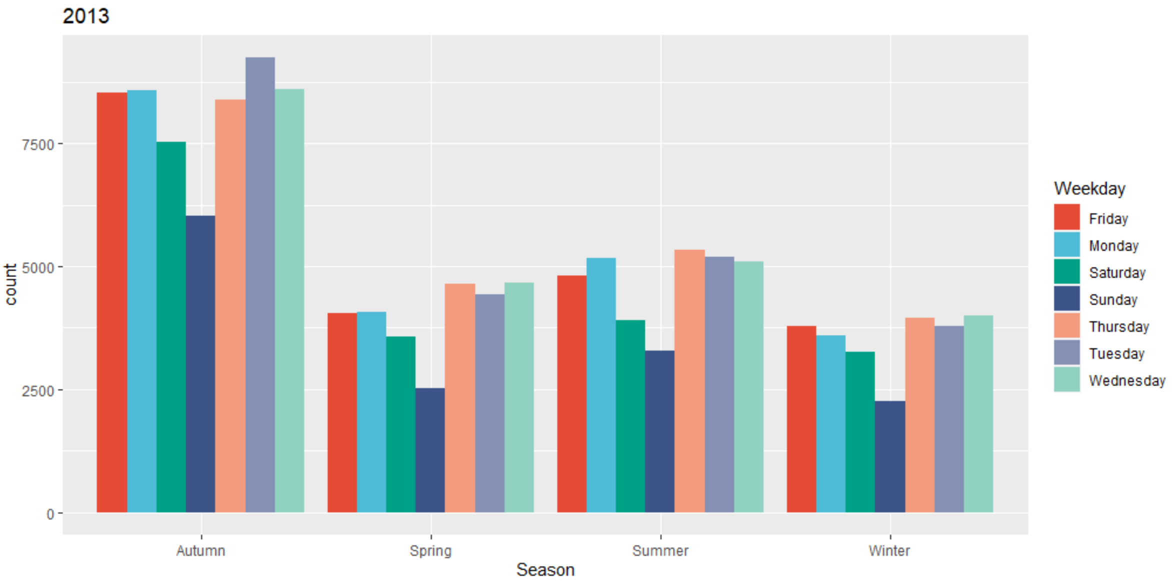

Figure 6.9: Amount of transactions made in 2013, every season, and separating by weekday.

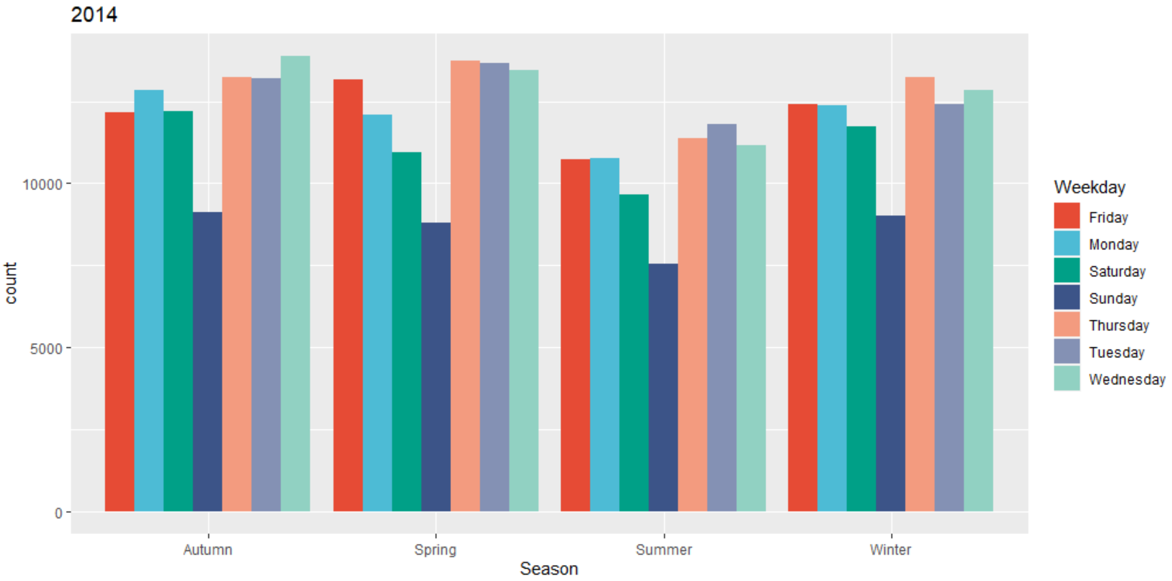

Figure 6.10: Amount of transactions made in 2014, every season, and separating by weekday.

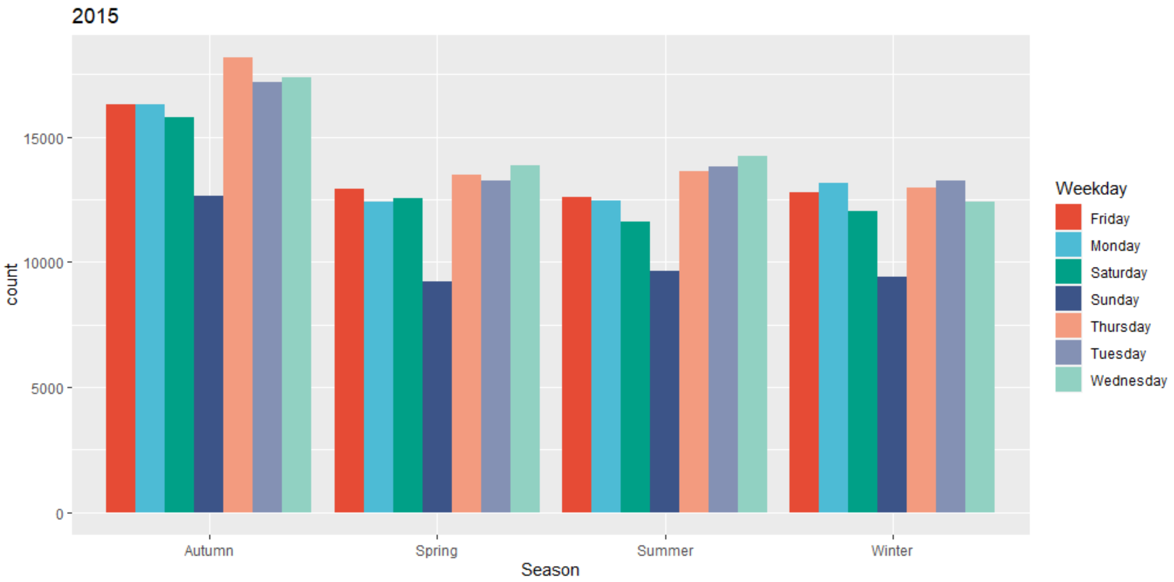

Figure 6.11: Amount of transactions made in 2015, every season, and separating by weekday.

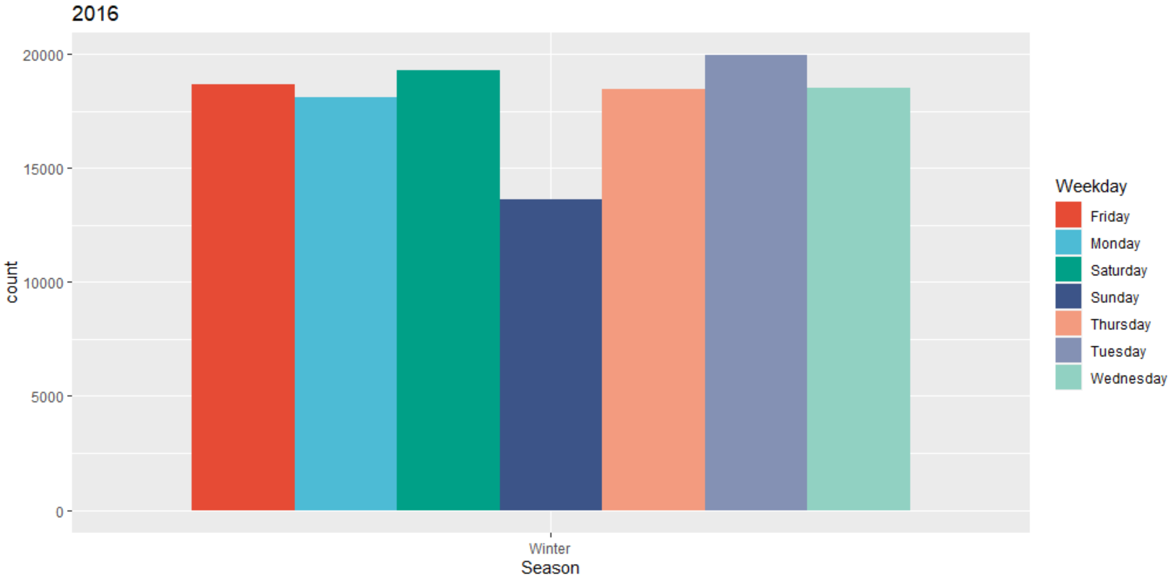

Figure 6.12: Amount of transactions made in 2016, every season, and separating by weekday.

6.2 Application of methodology

Now with the data sets cleaned and parsed, we can address each one of the research questions posed in Chapter 1.

6.2.1 Spatial and temporal behaviour of users

Research question 1: *Are there any spatial and temporal patterns in how users utilise the network of charging stations?*

In order for us to address the first research question of this project, we ran Ridge regression, LASSO, Elastic Net and the Knockoffs filter over the `chrg_spatial.Rda` data set, taking as the dependent variable the total sum of the time in which every charging station had a plug connected (i.e. `ConnectedTime`, from the `df.Rda` data set), varying only the year.

We used an $\alpha = 40/99$ for the Elastic Net, which was chosen via 5-fold Cross Validation after 100 iterations. This procedure led to the selected features shown in Table 6.1, which were the variables selected by the Elastic Net and LASSO models (the results of these two approaches coincided completely in all years, but in 2015, where the features BBQ, Bicycle Rental, Bus station and Social Centre were selected by the Elastic Net, but not by the LASSO). Full disclosure of the results of each model ran over the `df.Rda` data set filtered by year is shown in Figures 6.13 to 6.17. We can notice the disparity between the results from the Ridge and the LASSO-Elastic Net (the former with more variables selected than the later).

On the other hand, after performing the Knockoffs filter over the `df.Rda` data set (filtered by year), at the standard $FDR = 0.1$, we only obtained results for the 2016 setting (Table 6.2). This is why we continued increasing the FDR in steps of 0.05, obtaining the results shown in Tables 6.3 and 6.4, meaning that we are allowing up to a maximum of 20% of selected variables to be false discoveries (more variables were selected with higher False Discovery Rates, however we considered that such values were significantly high for us to be able to draw any kind of valid conclusion).

The results of this approach showed an almost perfect match between the coefficients of the selected variables by the Elastic Net and the LASSO, being the gambling places the most significant from 2013 through 2015; in 2012, planetariums were selected as the most significant amenity whereas, in 2016, 4 type of amenities were barely selected (positive coefficients: car sharing, dentist, and kindergarten; negative coefficients: fuel).

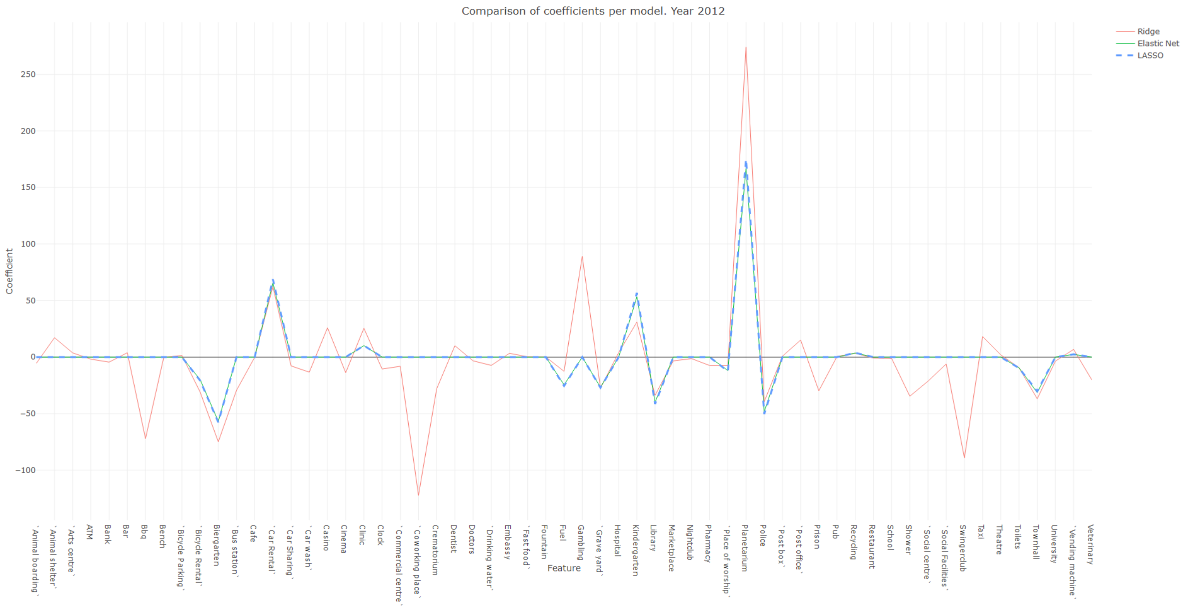

Figure 6.13: Amenities (features) coefficient comparison between the models Ridge, Elastic Net and LASSO. Year 2012.

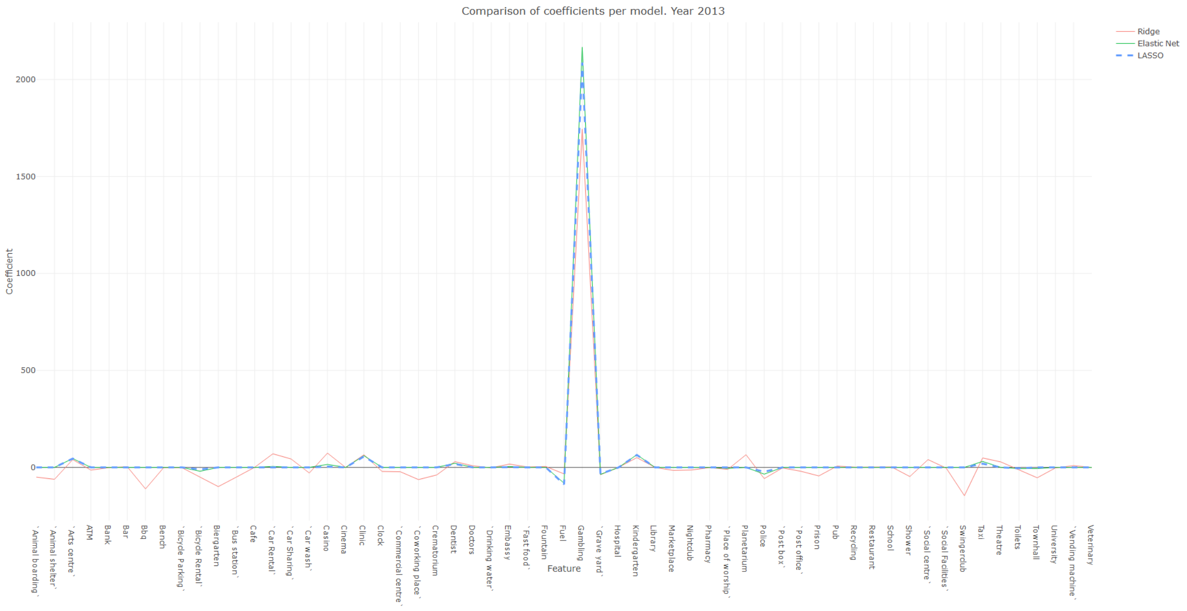

Figure 6.14: Amenities (features) coefficient comparison between the models Ridge, Elastic Net and LASSO. Year 2013.

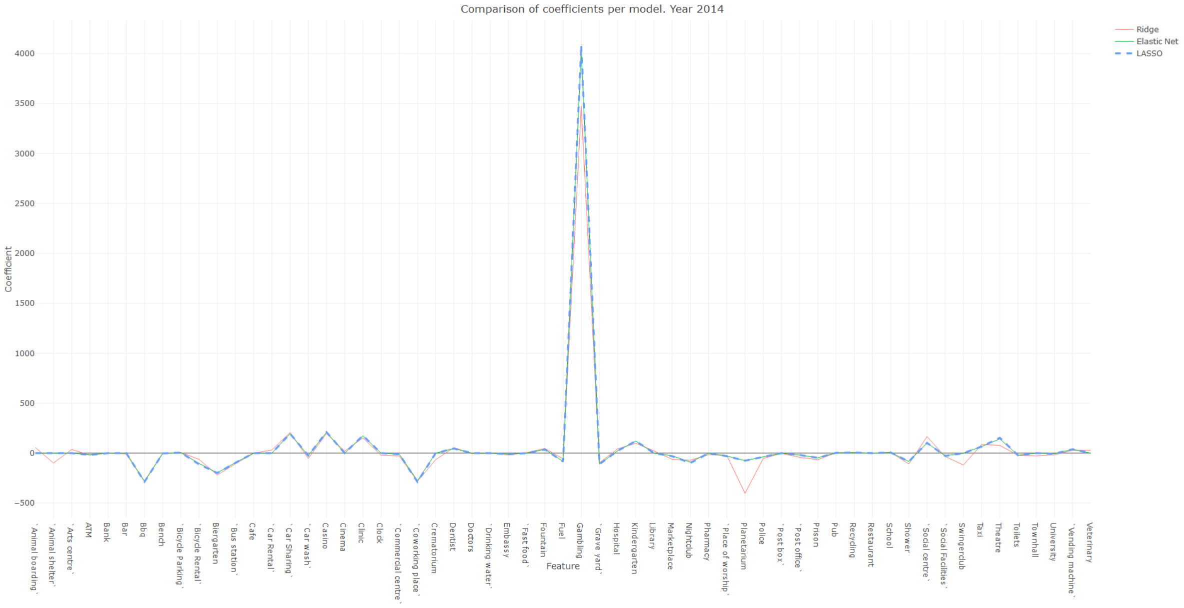

Figure 6.15: Amenities (features) coefficient comparison between the models Ridge, Elastic Net and LASSO. Year 2014.

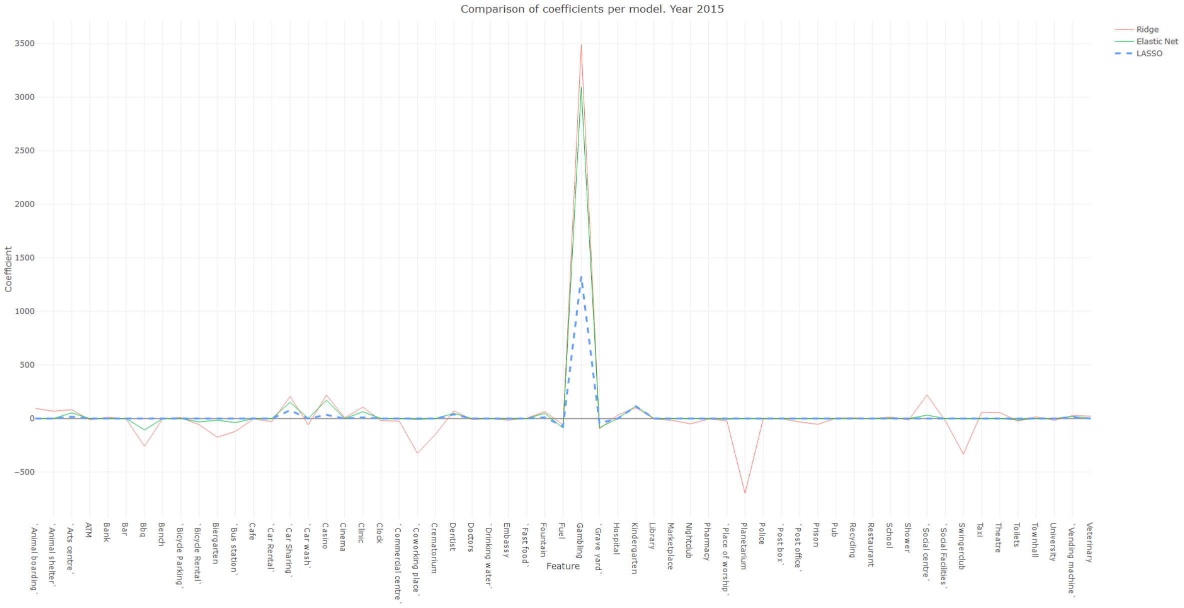

Figure 6.16: Amenities (features) coefficient comparison between the models Ridge, Elastic Net and LASSO. Year 2015.

2012	2013	2014	2015	2016
Bicycle Rental	Arts Centre	ATM	Arts Centre	Car Sharing
Biergarten	Bicycle Rental	BBQ	Car Sharing	Dentist
Car Rental	Casino	Bicycle Rental	Casino	Fuel
Clinic	Clinic	Biergarten	Clinic	Kindergarten
Fuel	Dentist	Bus Station	Dentist	
Grave yard	Fuel	Car Sharing	Fountain	
Kindergarten	Gambling	Casino	Fuel	
Library	Grave Yard	Clinic	Gambling	
Place of worship	Kindergarten	Coworking place	Grave yard	
Planetarium	Police	Dentist	Kindergarten	
Police	Taxi	Fountain		
Recycling		Fuel		
Toilets		Gambling		
Townhall		Grave yard		
Vending machine		Hospital		
		Kindergarten		
		Marketplace		
		Nightclub		
		Place of worship		
		Planetarium		
		Police		
		Prison		
		Shower		
		Social centre		
		Taxi		
		Theatre		
		Vending machine		

Table 6.1: Selected features per year, through models LASSO and Elastic Net.

6.2.2 Performance of charging stations

Research question 2: *Can we identify some factors influencing the performance of charging stations?*

We have addressed this question by first creating and adding another column to the data set `df.Rda`, called `efficiency`, and which entries are obtained by dividing the values of the features `TotalEnergy` over `ChargeTime` for each transaction. Of course, the higher the value, the higher the charging efficiency level as it measures the charging speed. By doing so, we can now analyse the charging performance throughout the time having the `efficiency` feature as the dependent variable.

We ran the LASSO built-in method, included in the package `glmnet` in `RStudio` over the

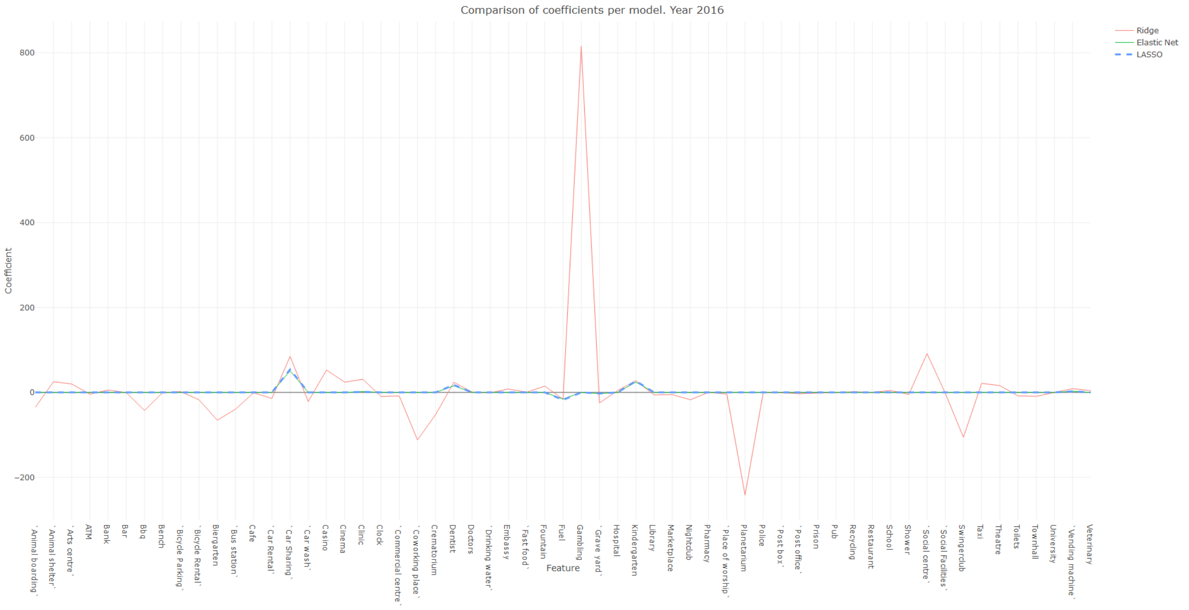

Figure 6.17: Amenities (features) coefficient comparison between the models Ridge, Elastic Net and LASSO. Year 2016.

2012	2013	2014	2015	2016
				Bicycle Rental
				Bus station
				Car Sharing
				Coworking place
				Dentist
				Fountain
				Gambling
				Kindergarten
				School
				Theatre
				Vending machine

Table 6.2: Selected features per year. Model: Knockoffs filter (FDR=0.1).

aforementioned configuration, filtering by each temporal type of feature (weekdays, months, time and season). The results of these compilations are shown in Table 6.5.

On the other hand, in order to compare the above results with an FDR method, we also applied the Knockoffs filter to the same configuration (i.e., with the `efficiency` feature added to the `df.Rda` data set). We set the FDR at 0.15 as lower rates would not select any variables at all, and higher rates would be too high to draw valid conclusions. Results of the selected variables after running the Knockoffs filter are also shown in Table 6.5.

2012	2013	2014	2015	2016
			Arts centre	Bus station
			Car Sharing	Car Sharing
			Casino	Dentist
			Clinic	Fuel
			Dentist	Kindergarten
			Fountain	School
			Fuel	Theatre
			Gambling	Toilets
			Grave yard	Vending machine
			Kindergarten	
			Recycling	
			Vending machine	

Table 6.3: Selected features per year. Model: Knockoffs filter (FDR=0.15).

2012	2013	2014	2015	2016
		Car Sharing	Arts centre	Car Sharing
		Clinic	Bench	Dentist
		Fuel	Bicycle Rental	Fuel
		Gambling	Car Sharing	Gambling
		Grave yard	Casino	Kindergarten
		Kindergarten	Clinic	Vending machine
		Vending machine	Dentist	
			Fountain	
			Fuel	
			Gambling	
			Grave yard	
			Kindergarten	
			Recycling	
			Toilets	
			Vending machine	

Table 6.4: Selected features per year. Model: Knockoffs filter (FDR=0.2).

We found that every variable not selected by the LASSO was not chosen by the Knockoffs filter either (in contraposition, every chosen variable by the Knockoffs filter was also selected by the LASSO); this reflects the correspondence between these two methods and brings certain reliability to the Knockoffs filter. Among the selected variables by the Knockoffs filter, we find that the sign of their respective LASSO coefficient is positive, which indicates a direct relationship between the variable and the efficiency. Additionally, every selected variable is related to non-busy temporal variables (say, Summer; weekends: Friday, Saturday and Sunday; school vacation months: July and December and non-operational hours: After midnight).

Weekdays		
Variable	LASSO	Knockoffs Filter
Monday	-0.01916	Not selected
Tuesday	-0.01465	Not selected
Wednesday	Not selected	Not selected
Thursday	-0.00800	Not selected
Friday	0.04095	✓
Saturday	0.12343	✓
Sunday	0.13866	✓
Months		
January	-0.00595	Not selected
February	Not selected	Not selected
March	-0.01713	Not selected
April	-0.01630	Not selected
May	-0.00721	Not selected
June	Not selected	Not selected
July	0.02857	✓
August	0.03510	Not selected
September	Not selected	Not selected
October	0.01574	Not selected
November	Not selected	Not selected
December	0.04235	✓
Time		
Morning	-0.11353	Not selected
Afternoon	Not selected	Not selected
Evening	0.10818	Not selected
After midnight	0.19800	✓
Season		
Spring	Not selected	Not selected
Summer	0.02083	✓
Autumn	0.03124	Not selected
Winter	-0.00953	Not selected

Table 6.5: Describing efficiency. List of coefficients of the selected values per group of temporal variables by the LASSO approach and selected values by the Knockoffs filter (FDR=0.15).

6.3 Interpretation of results

After comparing the features selected by the LASSO-Elastic Net and the Knockoffs filter, we noticed that most of the variables from the Knockoffs filter are included in the ones obtained from the LASSO and Elastic Net. Additionally, the amenities Kindergarten and Fuel were selected in all years by the LASSO, Elastic net and the Knockoffs filter (all but in 2016, where Fuel was not selected with FDR=0.1, as shown in Table 6.2). These amenities are followed by Car sharing (selected in all but two scenarios), Vending machine, Dentist and Gambling (selected in all but three scenarios).

The above findings suggest that charge points located close to facilities such as Fuel stations, Kindergartens and Car sharing points are more likely to be utilised by EV users than other type of public amenities.

Furthermore, after running the LASSO and Knockoffs Filter approaches to decide what temporal variables influence the performance of the charging stations, we could notice that the values of the variables that were not selected by the LASSO were not chosen by the Knockoffs filter either. Also, those values selected by the Knockoffs filter for each variable include the weekends (Friday to Sunday), July and December (which coincide with the Summer and Winter holidays), After midnight and the Summer; this makes, for instance, a transaction made on a July weekend in the after midnight, more likely to be an *efficient* transaction (probably because of a less-congested charging network caused by a lower demand).

Chapter 7

Conclusions

7.1 Conclusions

Electric Vehicles have gained popularity among the general society as they represent an innovative mean of transport that aims to decrease the CO_2 emissions. It is, therefore, of high importance, to recognise the necessity of the development and generation of knowledge on EV infrastructures and their different settings. In this work we have made an effort to contribute with the application of some statistical and geographic tools and analysis that can be used to understand and describe usage trends in EV charging networks (in this case, from the Netherlands).

The main results obtained from this study are listed below:

- We have determined, through the implementation of variable selection and false discovery rate control linear methods that charging stations located close to kindergartens, car sharing spots and fuel stations are more likely to be used more frequently and for the longest time.
- Likewise, we found that those charging transactions made during the weekends, the months of July and December, in the summer or between 12AM and 6AM, had the most efficient performance (meaning that they charged the highest amount of energy in the shortest time).

These findings can be taken into account when deciding where to install new charging stations and provide information to private and public organisations regarding the usage and performance of such charging points based on the moment they are intended to be used (e. g. to design and implement incentives and strategies of usage at certain hours, to avoid network congestion and promote charging efficiency).

Overall, we found that FDR techniques such as the Knockoffs filter here implemented become helpful to identify the variables that truly affect the dynamics of a certain network, avoiding the appearance of false positives (Type I errors) as much as possible. This certainly provides a new approach to the modelling and description of EV networks.

Moreover, even though this study was conducted for a specific type of EV network (the Netherlands), we believe the methodology can also be helpful and applicable to other scenarios and places. The approaches here implemented provide a new combination of tools (Voronoi diagrams, regularisation and a fairly novel FDR control method such as the Knockoffs filter) that can be used to study the impact and description of other geographical phenomena.

7.2 Limitations and further research

Throughout the development of this study, we encountered various limitations that represent important points to take into account. Some of them were derived from the deployment of the LASSO as a variable selection method, which is known to operate indistinctly among highly correlated variables (such as some of the variables we used in this work); however, the LASSO resulted to be the approach that provided the best fit for this study so we had to run various scenarios with and without highly correlated variables. Another important limitation was found at the time of applying the Knockoffs filter at standard levels (FDR, alphas, CV folds, etc.) as we were not obtaining any results at all with them. In order to obtain a significant outcome from their implementation and be able to draw valid conclusions, we had to perform a remarkable amount of different scenarios in order to find the best possible configuration without compromising the reliability of the results. Further research could be focused on the characteristics the data set feeding the Knockoffs filter should have in order to ease their study.

We also suggest a time series analysis in order to find more significant and interesting patterns in how users utilise the EV network (e. g., tracking how often and in which charging stations a particular card was used during the time frame of the data set).

APPENDIX A

APPENDIX A

A.1 `chrg_spatial`'s features names

	Feature name		Feature name		Feature name
1	Charge Point	22	Clock	43	Police
2	Population	23	Commercial centre	44	Post box
3	Animal boarding	24	Coworking place	45	Post office
4	Animal shelter	25	Crematorium	46	Prison
5	Arts centre	26	Dentist	47	Pub
6	ATM	27	Doctors	48	Recycling
7	Bank	28	Drinking water	49	Restaurant
8	Bar	29	Embassy	50	School
9	BBQ	30	Fast food	51	Shower
10	Bench	31	Fountain	52	Social centre
11	Bicycle Parking	32	Fuel	53	Social facilities
12	Bicycle Rental	33	Gambling	54	Swingerclub
13	Biergarten	34	Grave yard	55	Taxi
14	Bus station	35	Hospital	56	Theatre
15	Cafe	36	Kindergarden	57	Toilets
16	Car Rental	37	Library	58	Townhall
17	Car Sharing	38	Marketplace	59	University
18	Car wash	39	Nightclub	60	Vending machine
19	Casino	40	Pharmacy	61	Veterinary
20	Cinema	41	Place of worship		
21	Clinic	42	Planetarium		

Table A.1: `chrg_spatial`'s features names

Bibliography

[1] Natascia Andrenacci, Roberto Ragona, and Gaetano Valenti.
A demand-side approach to the optimal deployment of electric vehicle charging stations in metropolitan areas.
Applied Energy, 182:39–46, 2016.

[2] Komisja Europejska.
Memo-clean power for transport–frequently asked questions.
`http://europa.eu/rapid/press-release_MEMO-13-24_en.htm`, 2013.
Accessed: 2018-07-17.

[3] Netherlands Enterprise Agency.
Electric transport in the Netherlands - highlights 2016.
Technical report, Netherlands Enterprise Agency, April 2017.

[4] Ilyès Miri, Abbas Fotouhi, and Nathan Ewin.
Electric vehicle energy consumption modelling and estimation,Äîa case study.
International Journal of Energy Research, 45(1):501–520, 2021.

[5] Andrew Higgins, Phillip Paevere, John Gardner, and George Quezada.
Combining choice modelling and multi-criteria analysis for technology diffusion: An application to the uptake of electric vehicles.
Technological Forecasting and Social Change, 79(8):1399–1412, 2012.

[6] JR Helmus, JC Spoelstra, N Refa, M Lees, and R van den Hoed.
Assessment of public charging infrastructure push and pull rollout strategies: The case of the netherlands.
Energy Policy, 121:35–47, 2018.

[7] Xiaomin Xi, Ramteen Sioshansi, and Vincenzo Marano.
Simulation–optimization model for location of a public electric vehicle charging infrastructure.
Transportation Research Part D: Transport and Environment, 22:60–69, 2013.

[8] Richard Curtin, Yevgeny Shrago, and Jamie Mikkelsen.

Plug-in hybrid electric vehicles.
Working paper. University of Michigan, 2009.

[9] AP Robinson, PT Blythe, MC Bell, Y Hübner, and GA Hill.
Analysis of electric vehicle driver recharging demand profiles and subsequent impacts on the carbon content of electric vehicle trips.
Energy Policy, 61:337–348, 2013.

[10] Nasrin Sadeghianpourhamami, Nazir Refa, Matthias Strobbe, and Chris Develder.
Quantitive analysis of electric vehicle flexibility: A data-driven approach.
International Journal of Electrical Power & Energy Systems, 95:451–462, 2018.

[11] Rina Foygel Barber and Emmanuel J. Candès.
Controlling the false discovery rate via knockoffs.
The Annals of Statistics, 43(5):2055–2085, October 2015.

[12] Erotokritos Xydas, Charalampos Marmaras, Liana M Cipcigan, Nick Jenkins, Steve Carroll, and Myles Barker.
A data-driven approach for characterising the charging demand of electric vehicles: A uk case study.
Applied energy, 162:763–771, 2016.

[13] Yong Bing Khoo, Chi-Hsiang Wang, Phillip Paevere, and Andrew Higgins.
Statistical modeling of electric vehicle electricity consumption in the victorian ev trial, australia.
Transportation Research Part D: Transport and Environment, 32:263–277, 2014.

[14] Nederland Elektrisch.
`https://nederlandelektrisch.nl/home`.
Accessed: 2018-06-14.

[15] ElaadNL.
`https://www.elaad.nl/`.
Accessed: 2018-06-14.

[16] Robert Tibshirani.
Regression shrinkage and selection via the LASSO.
Journal of the Royal Statistical Society. Series B (Methodological), 58, 1996.

[17] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani.
An introduction to statistical learning: with applications in R.
Number 103 in Springer texts in statistics. Springer, New York, 2013.
OCLC: ocn828488009.

[18] Jerome Friedman, Trevor Hastie, and Robert Tibshirani.
The elements of statistical learning, volume 1.
Springer series in statistics New York, 2001.

[19] Bruce Ratner.
Variable selection methods in regression: Ignorable problem, outing notable solution.
Journal of Targeting, Measurement and Analysis for Marketing, 18(1):65–75, March 2010.

[20] Matteo Sesia, Chiara Sabatti, and Emmanuel Candès.
Gene hunting with knockoffs for hidden markov models.
arXiv preprint arXiv:1706.04677, 2017.

[21] Yasunori Ogura, Denise K Bonen, Naohiro Inohara, Dan L Nicolae, Felicia F Chen, Richard Ramos, Heidi Britton, Thomas Moran, Reda Karaliuskas, Richard H Duerr, et al.
A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease.
Nature, 411(6837):603, 2001.

[22] Juha Reunanen.
Overfitting in making comparisons between variable selection methods.
Journal of Machine Learning Research, (3):1371–1382, 2003.

[23] Patrenahalli M. Narendra and Keinosuke Fukunaga.
A branch and bound algorithm for feature subset selection.
IEEE Transactions on computers, (9):917–922, 1977.

[24] George M Furnival and Robert W Wilson.
Regressions by leaps and bounds.
Technometrics, 16(4):499–511, 1974.

[25] Bradley Efron and Trevor Hastie.
Computer Age Statistical Inference: Algorithms, Evidence and Data Science.
Cambridge University Press, 1 edition, July 2016.

[26] Carl M. O'Brien.
Statistical learning with sparsity: The LASSO and generalizations.
International Statistical Review, 84(1):156–157, April 2016.

[27] Arthur E Hoerl and Robert W Kennard.
Ridge regression: Biased estimation for nonorthogonal problems.
Technometrics, 12(1):55–67, 1970.

[28] Stephen Boyd and Lieven Vandenberghe.
Convex optimization.
Cambridge university press, 2004.

[29] Hui Zou and Trevor Hastie.
Regularization and variable selection via the elastic net.
Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(2):301–320, 2005.

[30] Sture Holm.
A simple sequentially rejective multiple test procedure.
Scandinavian journal of statistics, pages 65–70, 1979.

[31] Yosef Hochberg.
A sharper bonferroni procedure for multiple tests of significance.
Biometrika, 75(4):800–802, 1988.

[32] Yoav Benjamini and Yosef Hochberg.
Controlling the false discovery rate: a practical and powerful approach to multiple testing.
Journal of the royal statistical society. Series B (Methodological), pages 289–300, 1995.

[33] John D. Storey.
The positive false discovery rate: a bayesian interpretation and the q-value.
The Annals of Statistics, 31(6):2013–2035, December 2003.

[34] Koen J.F. Verhoeven, Katy L. Simonsen, and Lauren M. McIntyre.
Implementing false discovery rate control: increasing your power.
Oikos, 108(3):643–647, March 2005.

[35] Weijie Su, Małgorzata Bogdan, Emmanuel Candes, et al.
False discoveries occur early on the LASSO path.
The Annals of Statistics, 45(5):2133–2150, 2017.

[36] Asaf Weinstein, Rina Barber, and Emmanuel Candès.
A power and prediction analysis for knockoffs with LASSO statistics.
arXiv:1712.06465 [stat], December 2017.
arXiv: 1712.06465.

[37] Ran Dai and Rina Foygel-Barber.
The knockoff filter for fdr control in group-sparse and multitask regression.
arXiv preprint arXiv:1602.03589, 2016.

[38] Ming Yuan and Yi Lin.
Model selection and estimation in regression with grouped variables.
Journal of the Royal Statistical Society: Series B (Statistical Methodology), 68(1):49–67, 2006.

[39] Rina Foygel-Barber and Emmanuel Candès.
A knockoff filter for high dimensional selective inference.
arXiv preprint arXiv:1602.03574, page 32, 2018.

[40] Eugene Katsevich and Chiara Sabatti.
Multilayer knockoff filter: Controlled variable selection at multiple resolutions.
arXiv:1706.09375 [stat], June 2017.
arXiv: 1706.09375.

[41] Emmanuel Candès, Yingying Fan, Lucas Janson, and Jinchi Lv.
Panning for gold: model-x knockoffs for high dimensional controlled variable selection.
Journal of the Royal Statistical Society: Series B (Statistical Methodology), 80(3):551–577, June 2018.

[42] R Core Team.
R: A Language and Environment for Statistical Computing.
R Foundation for Statistical Computing, Vienna, Austria, 2017.

[43] RStudio Team.
RStudio: Integrated Development Environment for R.
RStudio, Inc., Boston, MA, 2015.

[44] Google maps.
`https://www.google.com/maps/@51.4317935,5.4284751,3a,60y,47.55h,87.76t/data=!3m6!1e1!3m4!1sxv8V3kbRSUZzbPTuvg8jkg!2e0!7i13312!8i6656`.
Accessed: 2018-07-07.

[45] Central Bureau of Statistics.
`https://www.cbs.nl/`.
Accessed: 2018-07-09.

[46] Map of 100 meters by 100 meters with statistics (2016).
`https://bit.ly/2zSWZuJ`.
Translated to English from Dutch. Accessed: 2018-07-09.

[47] QGIS Development Team.
QGIS Geographic Information System.
Open Source Geospatial Foundation, 2009.

[48] Tableau Software, Seattle, USA.
Tableau Desktop, 2018.

[49] Hadley Wickham, Romain François, Lionel Henry, and Kirill Müller.

dplyr: A Grammar of Data Manipulation, 2018.
R package version 0.7.6.

[50] Jerome Friedman, Trevor Hastie, and Robert Tibshirani.
Regularization paths for generalized linear models via coordinate descent.
Journal of Statistical Software, 33(1):1–22, 2010.

[51] Evan Patterson and Matteo Sesia.
knockoff: The Knockoff Filter for Controlled Variable Selection, 2017.
R package version 0.3.0.

[52] Hinrich Helms, Martin Pehnt, U Lambrecht, and A Liebich.
Electric vehicle and plug-in hybrid energy efficiency and life cycle emissions.
In *18th International Symposium Transport and Air Pollution*, volume 5, pages 113–124. Citeseer, 2010.