National College of Ireland

Understanding the Impact of COVID-19 on Electrical Demand

MSc Research Project
Data Analytics

Stephane Nichanian
Student ID: 18202632

School of Computing
National College of Ireland

Supervisor: Dr. Rashmi Gupta

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Stephane Nichanian
Student ID:	18202632
Programme:	Data Analytics
Year:	2020
Module:	MSc Research Project
Supervisor:	Dr. Rashmi Gupta
Submission Due Date:	21/09/2020
Project Title:	Understanding the Impact of COVID-19 on Electrical Demand
Word Count:	9010
Page Count:	22

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	27th September 2020

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Understanding the Impact of COVID-19 on Electrical Demand

Stephane Nichanian
18202632

Abstract

The COVID-19 pandemic has fundamentally changed our society's behaviour. Along with these changes, electrical consumption has also been impacted in ways never expected before. This research has highlighted the main changes in electrical consumption statistics during lockdown: the general decrease of up to 20% electrical demand, the assimilation of weekend and weekdays and the shift of daily activities towards later hours of the day. From a technical standpoint this study has explored two powerful regression techniques, Generalized Additive Models and ARIMA polynomial regression. The choice of regression techniques is justified by the control it gives over predictor variables and the possibility for manual tuning of model parameters. An important choice of forecasting methodology was to create a separate model for each hour of the day, thus creating accurate representation of the complex intra-daily seasonality. Additionally, the introduction of lagged demand variables was instrumental in reducing autocorrelation as well as increasing model accuracy. The study has found GAM models to perform slightly better on short-term forecast (24 days ahead) with a total MAPE of 2.24 over the 5 dates considered against 2.92 for ARIMA regression. A mid-term forecast was also implemented for a period of 3 months where the GAM model significantly outperformed the ARIMA regression with an MAPE of 2.31 against 4.08 for ARIMA regression. We do however note, significantly longer processing times for GAM models due to complex smoothing functions. We also note that there is no substantial degradation of prediction accuracy from a one day ahead forecast to a 3-month forecast, which validates the usage of GAM models for the lockdown period simulation.

1 Introduction

In December 2019, the very first case of Coronavirus Disease 2019 (COVID-19), a highly infectious disease that affect lungs and airways, was contracted in the Chinese region of Wuhan. The first patient consequently spread the disease at an exponential rate which eventually reached a worldwide scale and caused a global pandemic. Each country has adopted laws and restrictions to address this disease and limit its outbreak thus protecting those who are most at risk such as the elderly and people with poor health. In multiple countries strict measures have been adopted that put the entire population in a lockdown state. In such instances all restaurants, bars, schools and shops are closed, and people are only allowed to exit their homes for essential shopping. These measures although strict and conservative are necessary to address the serious consequences of COVID-19.
One notable consequence of these measures, is the change in energy consumption from a residential, commercial and industrial standpoint. Whilst people are asked to stay at home more, we can expect the residential electrical consumption to increase. However, the commercial and industrial premises where people would usually work are now empty, therefore

we can expect the commercial electrical load to decrease. This paper aims to understand how energy figures have been impacted by the Covid-19 pandemic on short and mid-term time-frames.
By using the insights derived in this analysis, key players of the energy industry can adapt and prepare solutions for similar instance in the future. For example, electricity production companies can use this information to understand when peak and low electricity demand happens (in relation to phased lockdown states) for demand side management (Ayan & Turkay 2018). Similarly, grid level electricity companies can make better decisions on transmission and distribution planning (Akbari & Moghaddam 2020). Finally, this study can be used by energy traders, that hugely benefit in understand the peak demand of electricity at various times and geographical locations for energy purchasing and selling (Li et al. 2019).
According to a review paper on load forecasting by Mustapha et al. (2015) and as illustrated in Figure 1 , the prediction range can be classified into 4 categories. Very Short-Term load forecasting (VSTLF) predicts values from seconds to an hour and is used for trading and short term operations management, Short-term load forecasting (STLF) usually predicts half-hourly values from one hour to two weeks and is useful for trading, operations management and planning. Medium-term load forecasting (MTLF) usually predicts half-hourly or daily values between two weeks to three years, whereas Long-term load forecasting (LTLF) is for time periods longer and is used for long term financial and operations planning. It should be noted that each of these categories benefit from different prediction methods and vary in complexity. The objective of the prediction can also vary, in STLF typically the prediction would be the half-hourly electrical consumption. In MTLF the prediction could be hourly values or daily averages, whereas in LTLF the prediction target is weekly or monthly peak demand. The following diagram (figure 1) shows a summary of the time frames in electrical forecasting and how they are used in the industry as previously explained.

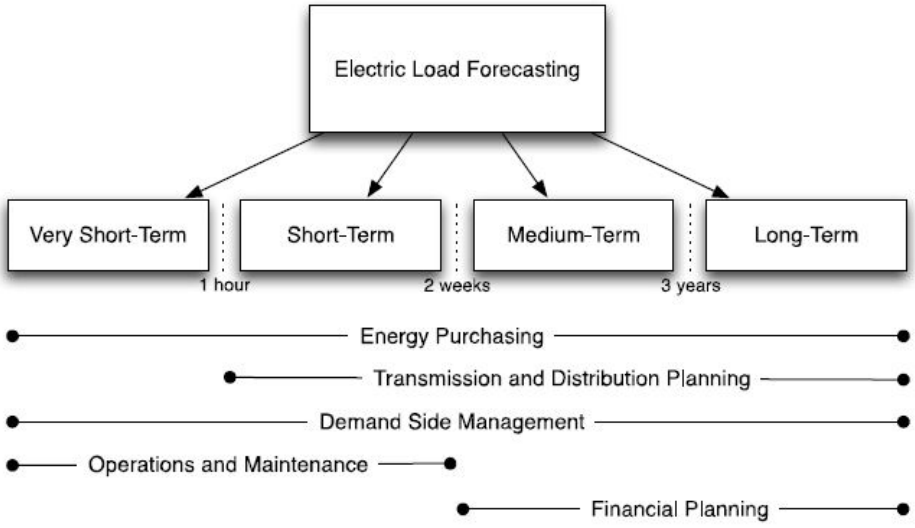

Figure 1: Electrical forecasting timeframe and industry application

The data used for this project is for the french region of Ile-De-France which contains the city of Paris. In this region, strict lockdown rules were instated from March 17th 2020. Therefore, the electrical forecast target for this project is from March 17th to June 29th 2020 (date at which the research is conducted). By forecasting the electrical consumption during this period, we can compare the current pandemic statistics to the simulated environment without the pandemic.
The details of the electrical forecasting are included in the methodology chapter and the predictive model uses weather, calendar and lagged variables information from 2016 to present.

The data mining task associated with electrical load forecasting is based on work done by Mohamed et al. (2006) which provides the data collection, pre-processing, prediction and evaluation techniques in electrical forecasting. The measure of accuracy for the model is Mean Absolute Percentage Error (MAPE), Percentage Error (PE) along with graphs for visualisation.
The contribution of this paper in academic literature comes from the knowledge derived from the application of electrical forecasting in times of catastrophic events such as pandemics. Due to the very recent nature of the COVID-19 pandemic there is a strong gap of academic literature in this field and potential for scientific discovery which in turn provides a framework for future cases of pandemic or catastrophic events.
This paper also aims to improve existing methods of electrical forecasting by providing a novel implementation of an existing predictive analysis technique.

1.1 Research question

This research aims to answer two questions:
Q1. What are the most adapted techniques in regression for short-term and mid-term electrical forecasting?
Q2. What is the impact of the COVID-19 pandemic on electrical consumption?

The remainder of this study will be organised as follows: Related Work, Research methodology, Implementation, Evaluation, Conclusion and Discussion.

2 Related Work

Forecasting in general is a popular technique to predict the expected values of a variable given their past historical data. This method can be applied to a wide variety of fields such as stock market prediction, supply chain management, sales forecasting and weather forecasting (Sharma et al. 2017).
In this paper, the proposed objective is to forecast the electrical consumption in a specific geographical area. This is a popular study that has been implemented multiple times using a great variety of techniques that we will analyse in this chapter. The existing work ranges from regression techniques to time series models and neural networks.
The objective of this paper is a 3 month load forecasting of hourly electrical consumption values, therefore it falls under the medium-term load forecasting category. However, this paper is different than other existing studies due to the fact that we are forecasting hypothetical past historical values (to compare to COVID-19 lockdown situation). Therefore, a number of actual exogenous factors such as weather or economic variables are already known and can be integrated in this study. Therefore, although the time-frame of the analysis is medium-term, the resources available for the forecast are those typically used in short-term forecasting. Therefore, we will review both short-term and medium-term forecasting literature to select an appropriate method that can be used in both cases.
A study done by Kuster et al. (2017), which is a review of existing work for electrical consumption forecasting, has shown that a majority of papers use regression models (multiple regression or multivariate regression) making up for 41% of all papers considered. Artificial neural networks were used in 38.5% of the papers and time series analysis present in 30.8% (some papers combine the methods). In the following subsections we shall therefore consider all three of these techniques and select the most appropriate for this study.
An important concept to keep in mind while comparing existing paper's results is that all these studies use the same performance measurement metrics such as Mean Absolute Percentage Error (MAPE) and Root Mean Square Error (RMSE) for different time-frames or objectives. Indeed, one study might measure the monthly peak demand whereas the other

study might measure the half hourly consumption. Therefore the same MAPE metric can measure two different electrical demand targets, but cannot be used to compare to each other.
The literature review is separated in Regression techniques followed by Time-Series techniques and Artificial Intelligence techniques for short-term and medium-term electrical forecasting.

2.1 Regression techniques

Regression techniques are a common method of electrical demand prediction and are useful to represent the relationship between demand and external factors. They are also very useful in understanding which factors contribute the most in electrical consumption and allow consumers to make better driven decisions. The predictor's choice depends on the target of the prediction (single building, district level or grid level) and geographical area (cold weather or tropical with rainy seasons). The most prominent method of forecasting in existing literature is multiple linear regression (MLR) with integration of non-linear behaviour. It was noted in an electrical forecasting review paper by Kuster et al. (2017) that regression techniques offer very strong performances given that the exogenous conditions are known, however the difficulty relies in identifying the correct model that best fits the conditions of the study.

According to Mustapha et al. (2015) the predictors in regression techniques for electrical load forecasting can be separated in three categories: time related (hour of the day, day of the week, season of the year), environmental (weather related) and socio-economic (income, GDP, holidays).
The following paragraph focuses on the time related factors in electrical forecasting. Souzanchi-K et al. (2010) have used classical Multiple Linear Regression with the main prediction parameters being days of the week. As demonstrated in this paper, the load on weekdays is different than the load on weekends and therefore a separation between weekdays and weekends is essential. It has also been shown by Souzanchi-K et al. (2010) that the effect of weekend days is so important, that it also affects the surrounding days of Monday and Friday, which usually have electrical load profiles different than Tuesday, Wednesday and Thursday. Therefore, for our analysis we shall consider a predictor that specifies the weekday as a categorical variable. An interesting extension of this paper would have been to consider separate models for weekday and non week-days and verify if the final results is improved.

Using the same time related approach by Amral et al. (2007) the time of the day component was separated in 3 time-frames, from 1 to 6h, from 7 to 17h and from 18 to 24h. This separation allows to accurately reflect the load fluctuation during a typical day with the night time load being the lowest and evening load being the highest. This approach proved to be efficient and will be used in our paper. However, the country considered is India and could typically have different load profiles than France. Therefore, further analysis is necessary to decide the correct hourly time-zone separation. It is interesting to note that due to the extreme variations in meteorological conditions due to the rainy season in India, the authors have decided to create two separate models for each season. We will study this approach in more detail in the environmental chapter. The Mean Average Percentage Error (MAPE) achieved in this paper is 3.52% for dry season and 4.34% for rainy season.

The next aspect to consider in time related approach is the yearly seasonal variation of electricity consumption. As can be expected, the electrical consumption in winter would be different than the one in summer and yearly electrical load curves show strong seasonality patterns. Lin et al. (2018) have implemented a Least Squares Support Vector Regression (LS - SVR) model to forecast the daily maximum electrical demand using accumulated

weather effect. To achieve this task, multiple models were created for each season of the year. Although this study predicts daily maximums (and not hourly values), it shows the importance of seasonality within the year and how they can affect the prediction. Further study is required to determine whether one hot encoding the seasons is sufficient of separate models have to be built for better accuracy.

The last time related component in regression electrical forecasting that must be considered is the trend factor. The trend component is specifically important for mid-term to long-term forecasts at will have a significant impact on the demand evolution. This was highlighted by Imtiaz et al. (2006) where multiple regression analysis based on least squares method was used to forecast electrical demand over a 10 year period. The trend component in this study is introduced by encoding a time variable that increases by 1 for each observation. In this paper the predictor variables were population, electricity consumption per capita, number of electricity consumers, peak electrical demand and GDP. The study could be further expanded by including environmental factors. Using this study, we will implemented a time variable that will reflect the trend of the electricity consumption data.

After reviewing the time related analysis using regression techniques, we can now study the electrical load forecasting literature which use environmental factors as main predictors for the regression. A study by Ching-Lai Hor et al. (2005) focuses on identifying the significant weather variable in electrical consumption using multiple linear regression. It was established in this paper that the contributing weather variables are temperature (by far the most significant) followed by rainfall, humidity, windspeed and sunshine. This paper also notes that the relationship between temperature and electricity consumption is non-linear. The final MAPE with the weather variables is 2%. Again, it should be noted that one prediction is made for each month, therefore having a total of 36 predictions (over 3 years) it is easier to achieve a low MAPE.

The same non-linearity issue between temperature and electricity was dealt differently by Hong et al. (2010) where a quadratic term in the polynomial equation was introduced to represent the temperature. The regression equation therefore becomes a polynomial equation of degree 2 to model the temperature non-linearity. Hong et al. (2010) also introduced interaction effects in the MLR equation which allows to achieve significant improvement in the final MAPE figure. Along the weather effects, the authors have also included time index and socio-economic. The final MAPE with interaction and quadratic polynomial equation is 4.6% MAPE.

There are two remaining popular approaches to deal with the temperature and electricity non-linearity used by Abu-Shikhah et al. (2011) and Feng & Wang (2019). The former (Abu-Shikhah et al. 2011) proposes to apply exponential smoothing to the temperature variable to transform it's form to a linear component. The latter (Feng & Wang 2019) proposes to create two separate models that each represent a linear part of the temperature-electricity curve. The cut-off point is at 25 degrees Celsius where the non-linear behavior is apparent. Abu-Shikhah et al. (2011) uses an exponential regression technique to compare linear, polynomial and exponential regression to forecast mid-term hourly and weekly load on Jordanian electrical data. The authors conclude that for hourly load prediction the exponential regression technique under-performs and suggest to use the polynomial regression method.

This chapter will focus on the Socio-economic factors in regression based techniques for electrical forecasting. Supapo et al. (2017) have focused on mid to long-term power forecasting using MLR considering Socio-Economic factor. For this study, the authors have considered the historical electricity consumption data, the consumer base and growth statistics, GDP and the commercial and industrial development plans for the concerned area of Palawan. The overall MAPE over the 5 years of prediction is 2.26%. However it should be noted that the study makes one prediction for a full year. Although this differs from out papers objective, it shows which predictors are to be considered the socioeconomic part of

our study, namely population and GDP.

A last approach that should be studied was implemented by Bruhns et al. (2005) and uses advanced regression techniques namely Generalised Additive Models (GAM) to create two separate models for the final prediction. One model uses the non-weather related predictors and the other model uses the weather related predictors only. The non-weather predictors are time related predictors such as trend, seasonal factors and calendar factors (holidays). The weather model includes all standard weather predictor and deals with the temperature non-linearity by exponential smoothing followed by double averaging with "apparent temperature". The distinction between both models is justified by the fact that weather related predictors have a very different behaviour than other predictors and therefore including them both in the same model could lead to errors in prediction. This paper is an internal study for the EDF utility company and provides high level technical implementation built on years of experience and experimental forecasting development. Although some of the proposed solution seem technically complex and not viable for a Master's thesis, the two model approach will be considered to verify the potential improvements on MAPE. The final MAPE of this study is 1.88%, figure which we will use for comparison to our model. Furthermore, an hourly MAPE is also shared which we will also use to compare to our model.

After reviewing the main regression techniques existing in literature, the popular predictors and models, we can now focus on the time series analysis and decide whether it is appropriate for our study.

2.2 Time Series Analysis

Time series analysis created by Box and Jenkins in 1970 is one of the oldest method of electrical forecasting. They are separated in two main classes, uni-variate and multi-variate models. Uni-variate analysis is appropriate for forecasts up to 6 hours ahead whereas multi-variate models can be used in any timeframe.

Due to the inherent non-linear nature of electrical load (with multiple seasonality) the time series models also need to be non-linear models (ARMA and ARIMA). ARIMA is better suited for data with complex non-linear behaviour and therefore are more appropriate for electrical load forecasting (Maniatis 2017).

Angelaccio (2019) implemented an ARIMA model on Italian cities energy consumption. The authors created a total of 6 models with varying (p,d,q) parameters ranging from (0,0,0) to (2,2,2). The training set is from 2016 to 2018 and the data set is further separated into 5 datasets following a rule of similar months. For each model, the optimal (p,d,q) parameter is selected. The months of January, February and March are in the same model because the electrical consumption data is similar during those months. The final MAPE is 8.5% across all models. This high figure is justified by the fact that the electricity consumption is done over multiple counties all over Italy that inherently display different electrical consumption behaviour. This paper proposes a Seasonal ARIMA time series analysis that offers good results for a uni-variate analysis over multiple cities. The natural extension of this work, that would lower the final MAPE figure would be to integrate exogenous variables to implement a multi-variate time series analysis called Seasonal AutoRegressive Integrated Moving Average eXogenous (SARIMAX).

Hutama et al. (2018) chose the SARIMAX method to forecast the power consumption of the Bali island. The objective is a medium-term forecasting of daily average loads for the 2017 year. The SARIMAX parameters are (1,0,1) and the exogenous factors coded in the equation are: day of the week and holiday dates. The MAPE for average daily load

was 2.68% which is a good value for this type of analysis. Again it should be noted that average daily values are easier to predicted than hourly values implemented in our paper. The limitations of paper lie in the restricted number of exogenous predictors chosen. However an interesting aspect of this paper, is that the authors also implemented an Artificial Neural Network with the same data and have found the SARIMAX model to perform better.

An inherent problem with time series analysis in electrical forecasting is that with increasing complex seasonal components (daily, weekly, monthly...) the time series equation quickly increase in complexity as well and can lead to long formulas difficult to understand. This problem is not present in regression where the relationship between predictors and dependant variable is clearly defined and quantifiable.

2.3 Artificial Intelligence Techniques

Neural Networks are a popular method of electrical forecasting and have proven to achieve good results in prediction accuracy of Short-Term timeframes (Sharma et al. 2017), although they can highly vary in complexity, computational time and pre-processing time. According to Sharma et al. (2017) the advantage of using Neural Networks for electrical forecasting lie in the fact that there are a wide variety of neural networks that can be adapted to each specific study and are able to capture complex variations of data. Additionally, Support Vector Machines, neuro-fuzzy systems and Genetic Algorithms are also popular techniques considered Artificial Intelligence techniques.

A basic Multi Layer Perceptron (MLP) Neural Network has been used by Yi et al. (2019) with the addition of extra hidden layers to form a Deep Belief Network (DBN). The DBN is combined with a Nonlinear Auto Regressive (NAR) dynamic neural network. The NAR is used to forecast the time series component of the data whereas the DBN is used to forecast and reduce the residual errors of the first model. This method not only incorporate seasonal time components but also weather factors including, temperature, illuminance and humidity. Although the load forecasting focuses on PC output generation, it shows that by combining two neural networks of different structure the accuracy of the model can be improved whilst considering weather information. This paper also puts in evidence the self-learning capability of neural networks and how they can be used to automatically detect seasonality in electrical load.

A comparative study of multiple Neural Networks done by Tao et al. (2019) aims to predict the electrical load on the European Intelligent Technology Network (EUNITE). The dataset is made of half-hourly electricity data, daily average temperatures, holiday dates and working days. The training set is from January 1997 to November 1998 and the testing set is over the month of December. The targeted prediction is daily average values and therefore easier to predict than hourly values. The authors have considered four different Neural Networks, namely Back Propagation (BP), Radial Basis Function (RBF), Elman Network (EN), and Long-Short Term Memory (LSTM). The LSTM model was found to be the most accurate with a MAPE of 3.46% followed by RBF with 3.69% whilst Elmann and BP both had MAPEs higher than 5%. The limitations of this study rely in the fact that the predictions are only made on average daily values. The extension of this work would be to predict hourly values over a month period. As suggested in the paper, this could be done by building 24 Neural Networks that would each predict one hour of the day.

In conclusion, Artificial Intelligence techniques are well adapted for electrical load forecasting as they can incorporate exogenous factors as well as capture complex variations of data. Neural Networks are considered to achieve to lowest MAPEs in STLF. As shown by Kuster et al. (2017) Artificial Intelligence techniques in existing literature mostly focus

on STLF with very few implementations on MTLF which are mainly done by regression techniques.

Regression techniques are the preferred forecasting method for our papers because unlike black-box systems (Neural Networks and SVM) it allows us to control the predictor variables and understand their relationship with electrical demand. MLR will also allow us to specifically analyze the causes of electrical consumption variation in relation to COVID-19 and other exogenous factors such as temperature, rainfall, holidays... The final choice of the model that achieves the lowest MAPE will be explained in the methodology section and various models will be considered with their outputs. The literature review summary table can be found in Appendix C.

3 Methodology

The research done in this study follows the Knowledge Discovery Database (KDD) methodology. This Chapter aims to explain the various tasks and processes followed within the KDD methodology for the implementation of electrical forecasting.

3.1 Original dataset

The electrical consumption dataset is obtained from the french website of the company RTE, https://www.rte-france.com/.The data obtained from the RTE website, contains the quarter hourly electrical demand readings for the region of Ile-De-France from the 1st of January 2016 to the 29th June 2020. The RTE website shares their data in separate excel files for each year. Therefore, the initial data pre-processing task was to convert those files to separate csv files and merge them by row in R studio.
The next task is to transform the data into hourly values (as quarter hourly readings offer a level of detail that is not necessary for the study). This is done by simply omitting the observations that are not full hourly observations. The final task for data cleaning of the original dataset is to detect and replace missing values. There were a total of 12 missing values. We will use the Moving Average method, that uses the before and after observations and averages them (with a window size of 2). This is an acceptable solution as we expect the electricity consumption to evolve linearly between two times of a day.
Figure 2 shows the electrical consumption graph from 2016 to 2020 obtained from the original dataset.

We can observe the general spikes and lags in the data which represent the summer and winter months, where electricity is mainly affected by temperature. We also note a downward trend in the last two years during the summer months. This is again due to temperatures being lower than previous years and therefore reducing electrical consumption.

3.2 Explanatory variables

The explanatory variables are the following: Weather variables, time variables and lagged variables. The weather variables are obtained from an API called Dark Sky API. The information sent by the API contains a total of 20 weather variables which will be reduced by feature selection. It was found that 3 entire days were missing from the API weather data (March 15th 2016, November 30th 2017 and August 17th 2019). Those missing values are dealt with the last observation carried forward (LOCF) technique.

The time related variables must accurately represent any behaviour of electrical demand that has a seasonal variation. For example, we know that electrical demand on weekends is different than electrical demand on weekdays. Therefore, we will create a variable that

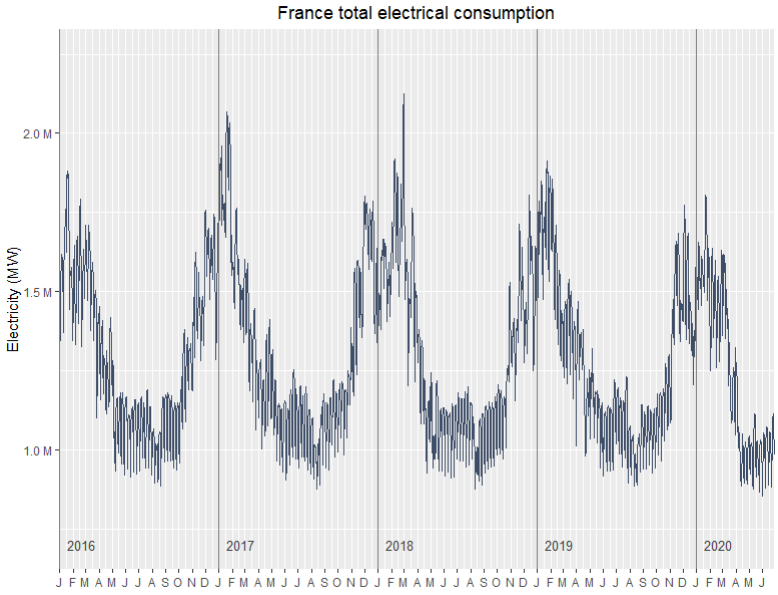

Figure 2: France total electrical consumption

specifies the day of the week. The month of the year is also used as a predictor variable. Similarly, a holiday or bank holiday dates have a strong impact on the electricity consumption, therefore we import a csv file, which contains all the holiday dates (coded as 0 and 1) for the Ile-De-France region from 2016 to 2020. The trend component, which represents the long term evolution of electrical demand is represented by a variable that increases gradually over time with 1 incremental for each hour. It should also be noted, that it is not necessary to add a variable that represent the hour of the day (as this would normally contain high seasonality) because a separate model for each hour of the day is created.

The relationship between temperature and electricity is not only non-linear but also affected by inertia. As explained by Mustapha et al. (2015), due to the fact that heat takes a while to dissipate, a lag is created between temperature and electricity demand. We therefore introduce lagged temperature variables, that link the current electrical consumption to past temperature values. The lagged temperatures variables include the following: maximum temperature in the last 24 hours, minimum temperature in the last 24 hours, last temperature observed 24 hours ago, last temperature observed 48 hours ago, mean temperature over the last 7 days and the temperature differential from hour to hour. The minimum and maximum temperature represents the lagged behaviour and shows the extent to which heating or cooling had to be used to deal with the temperature extremas. The last temperature observed 24 and 48 hours ago is necessary since we create one model for each hour. The mean temperature over the last 7 days gives a measure of the total thermal inertia accumulated over the course of the past week. Finally, the temperature differential describes the changes in temperature which is associated with people heating or cooling their houses.

Time series data is known to create auto-correlation in the forecast residuals. One way to deal with this is to introduce lagged demand variables, which will be used as a baseline for the following observation forecast and therefore minimize correlated behaviour of errors. Using lagged demands also increases the accuracy of the prediction by allowing the model to base it's current prediction on previously known values (important for time series analysis). The lagged demand variables include: maximum demand in last 24 hours, minimum demand in last 24 hours, last demand observed 24 hours ago, last demand observed 48 hours ago

and mean demand in last 7 days.

3.2.1 Dynamic forecast update

An important task for this project is the dynamic update of lagged demand variables at each step of the prediction. This task requires the lagged demand variables to be updated, each time a prediction is done by the model. By doing so, the lagged demands calculations are made on the forecasted values and not the actual observed values. This is particularly important when the range of the forecast is long (3 months in our paper) and after a few predictions, the lagged demand observations are entirely based on forecasted values. If this step was not implemented, the model would base it predictions on lagged demands of actual values, which would not normally be available. This type of forecasting is called simulated forecast where future predictions are based on past predictions and building up as a step by step forecast. In R, this is done by calling a dynamic updating function, each time a prediction is made by the model. This function updates the lagged demand values in the table (according the the newly made prediction) and returns the updated table for the next prediction.

Both predictive models used in this project do not require the predictor variables to be normally distributed or evenly distributed in the case of multi-level categorical variables. No feature scaling tasks (such as normalization) are required as both predictive methods use statistical models and are mathematically robust. Finally, The factor variables (categorical) can be used by the GAM model without transformation, however the ARIMA regression requires categorical variables to be one hot encoded into 0 and 1 format.

3.3 Predictive models

As justified in the related work section 2, this study focuses on regression techniques in electrical forecasting. We have therefore applied two regression models; General Additive Model (GAM) and ARIMA regression which will be explained in this chapter.

3.3.1 General Additive Model (GAM)

Generalised Additive Model (GAM) (Hastie & Tibshirani 1986) are the semi-parametric extensions of Generalised Linear Models (GLM). GAMs have the advantage over linear models of allowing complex non-linear relationship between the response and explanatory variables (as we have established the non-linear relationship of electricity and temperature). As observed in the report, many of the interactions between dependant and independent variables happen to be non-linear (such as temperature and demand). Therefore, the final regression model, must incorporate non-linear components. These variables are included by using smoothing functions (see Chapter 4.1.1). However, some of the other independent variables do have linear relationship with electrical demand and can be represented as a simple identity function. Therefore, the final model for the prediction will include a mix of identity functions and smooth functions which composes the general additive model (GAM) used for the prediction. To ensure that the model does not overfit the data a Generalized Cross Validation process as been used to create the GAM model (see Chapter 4.1.3).

The general equation for the GAM model can be written as:

$$g[E(Y)] = \beta_0 + f_1(x_1) + f_2(x_2) + ... + f_n(x_n) \tag{1}$$

where:

E is the family of distribution of the response variable, Y is the response variable, in this case electrical demand, x_1 to x_n are the independent variables, the $g()$ is the link function (such as a log transformation), the f_1 to f_n functions are the independent variables transformation function and can be non-parametric, semi-parametric or parametric.

This regression is composed of 24 GAM models with different predictor coefficients and gam equation, for each hour of the day.

3.3.2 Regression with ARIMA errors

Classical regression methods, when applied to time series analysis are known to violate assumption of independent and identically distributed errors. Although measures in the previous GAM model have been taken to minimize this violation of regression assumptions, due to the high collinearity of observations (one per hour for our dataset) we cannot ensure complete lack of autocorrelation in errors. Therefore, a second model is considered which specifically addresses this issue by applying an Auto Regressive Integrated Moving Average (ARIMA) model on the regression errors. Therefore, the classical regression equation given by:

$$y_t = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \eta_t \tag{2}$$

where:
β_0 is the intercept, $\beta_1 ... \beta_n$ are the regression coefficients, x_1 to x_n are the independent variables, η_t is the error term with auto-correlation,

In this case the error term η_t is allowed to be auto-correlated. The ARIMA model then is applied on the error term η_t and in the instance of a ARIMA(1,1,1) is given by:

$$(1 - \phi_1 B)(1 - B)\eta_t = (1 + \theta_1 B)\epsilon_t \tag{3}$$

where:
B is the backshift operator, ϕ_1 is the auto-regressive (AR) coefficient, θ_1 is the Moving Average (MA) coefficient, η_t is the regression error term with auto-correlation, ϵ_t is the ARIMA error term with no auto-correlation.
We can therefore see how by applying a ARIMA model on regression residuals, we eliminate any autocorrelation in the model. Therefore the electrical forecast regression can be considered to be the Best Linear Unbiased Estimator (BLUE). It should also be noted that each hour of the day will have a separate ARIMA regression, and the parameters (p,d,q) of the ARIMA process will be different for each hour.

3.4 Testing and evaluation methodology

3.4.1 Training and testing sets

In order to evaluate both models fairly, five different testing sets of one day each evenly spread through the year 2019 have been considered. Each models will make predictions for all five dates and the results will be compared to understand which model performs better throughout these different times of the year. This is particularly important as various times of the year have specific characteristics such as cold or hot weather, falls in holiday dates or may be a weekend day or not.
Additionally, the final objective of our study is electrical forecasting over a 3 month period from March to June 2020. Therefore, an additional testing set of 3 complete months from March to June 2019 shall also be considered to understand how well the model predicts over this longer period of time.
The training set ranges from the first observation (January 1st 2016 at midnight) until the last observation before the testing set.

3.4.2 Evaluation methodology

In order to compare nested GAM models, the Akaike information criterion (AIC) statistic is used to measure the quality of the model and the adjusted R squared statistic is used for goodness of fit. To ensure no overfitting on the training data, a Generalized Cross Validation method is used whilst building each GAM model and the corresponding GCV score is reported. The lowest GCV model is selected which ensures that the smoothing functions do not overfit the training data (see Chapter 4.1.3)
The Durbin Watson statistic is used to test for auto-correlation within each model. Finally, each prediction on the testing set is measured through Mean Absolute Percentage Error (MAPE). This is the most common metric in electrical forecasting and can be used to easily compare with other relevant papers. However, it should be noted that the MAPE stastics fails to capture in which direction the error is made (underestimate or overestimate electrical consumption),to counteract this issue, we will also visualize results on graphs which show the direction of the error (over or under the actual curves). Furthermore, when comparing different papers, the MAPE statistic can be misleading as a percentage of a small value would be penalized more than a percentage of a higher value.

3.4.3 Methodology summary

The diagram in figure 3 provides a high level summary of the project methodology described in the methodology chapter 3.

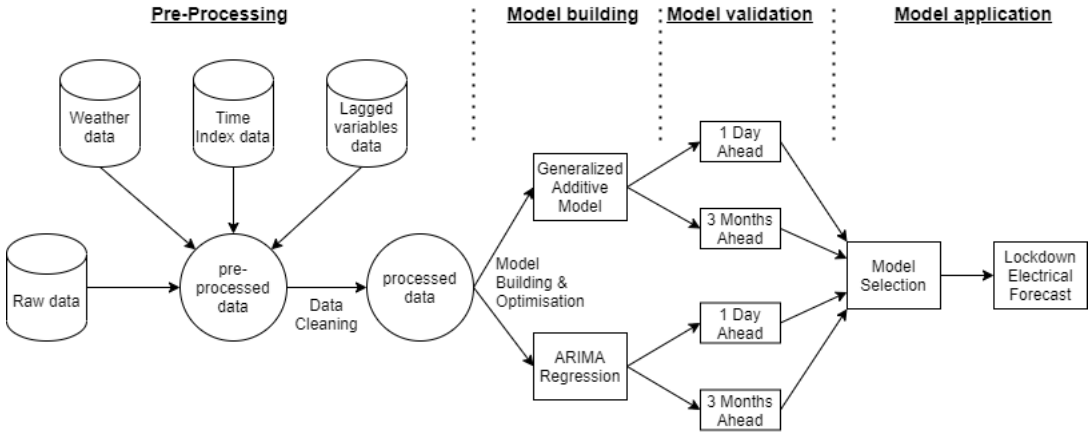

Figure 3: Process diagram of project methodology

4 Implementation

This Chapter describes how the predictive models were built, the essential checks on model assumptions and feature selection process for both models.

4.1 General Additive Model (GAM) implementation

As explained in Chapter 3.3.1, the GAM is composed of non-linear smoothing functions, linear identity functions and categorical variables. We will therefore first study how the smoothing functions were chosen (Chapter 4.1.1) and optimized, followed by feature selection (Chapter 4.1.2) as the model has more than 35 variables and finally we will verify the GAM assumptions (Chapter 4.1.3) in regards to our data.

4.1.1 Smoothing Functions

Non-linear predictors are applied to smoothing functions and are then included in the GAM. Smooth splines, propose an automated smoothing application where the user only needs to input a parameter λ for penalty of the least squares regression residuals. The λ parameter, determines the wiggliness of the interpolation regression curve and therefore is also referred as the smoothing parameter. There are a number of other smoothing functions compatible with GAMs, such as thin plate regression splines, P-splines, Markov Random Fields which provide additional methods of non-linear data approximation. Each have been tested using a Generalized Cross Validation process and the Smooth Spline technique has achieved the best results. The following figure shows the smooth spline function interpolation of the temperature data points with 5% confidence intervals.

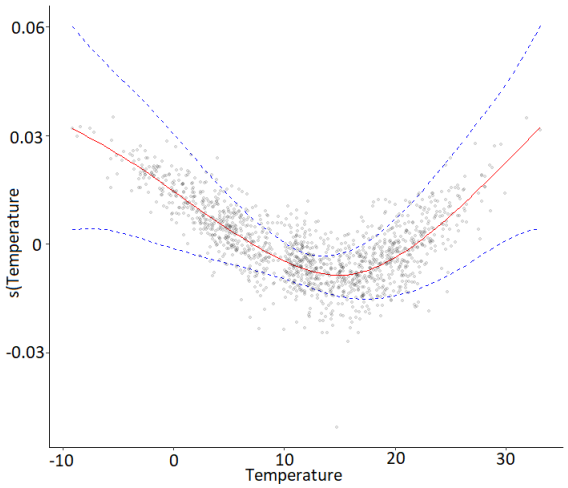

Figure 4: Smooth Spline of Temperature variable

4.1.2 Feature selection

The GAM function used in this project from the mgvc package doesn't offer a front, back or stepwise feature selection function as is common with normal linear models. Therefore, to find the optimal subset of predictors required for the analysis we refer to work done by Marra & Wood (2011). In their work, they suggest that by introducing an additional penalty parameter, the wiggliness of the variables which have no effect on the model are shrunk to 0 and are effectively left out of the model.
Additionally, a manual feature selection process was implemented, to understand the importance of the variables considered. Table 1 was obtained.

	Full Model	No Lag Temps	No Lag Demands	No Interaction Terms
R square	0.993	0.991	0.984	0.991
AIC	-9148	-8938	-8014	-8953
DW	1.7	1.57	0.87	1.65

Table 1: Manual feature selection statistics

We read from the table that the model without lagged temperature variables has a 0.2% lower R square score compared to the full model. The interaction terms also contribute similarly with an increase of 0.2% R square. The lagged demands have a very strong impact on the model, where the model with no lagged demand has a 0.9% lower R square and a significantly lower Durbin-Watson statistic. Therefore we conclude that lagged demands are essential to lower auto-correlation and for model accuracy.

4.1.3 GAM assumptions

As seen in Chapter 3.3.1 the GAM function requires the selection of a distribution and link function of the response variable. In order to approximate the response variable "electricity" to a normal distribution, we first perform a log base 10 transformation and then check the distribution of the new variable. We have obtained the plots shown in figure 5 in regards to the distribution of log(Electricity).

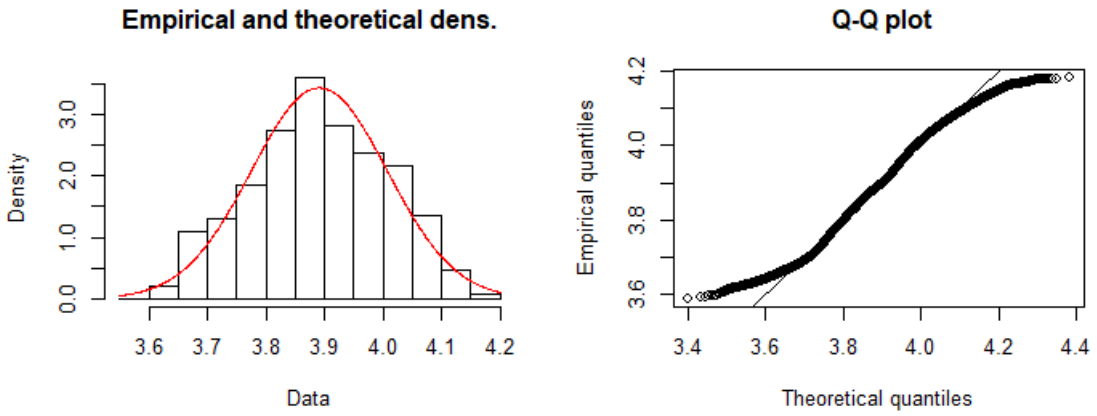

Figure 5: Distribution of response variable

As we can observe from figure 5, the log transformation of the response variable is approximately normally distributed. This justifies the choice of the Gaussian distribution and log transformation for the link function in the GAM formula 1.
The assumptions of a GAM model follow the regular assumptions about residuals for any regression models. These assumptions are particularly sensitive in a time series problem such as the one intended in this study. In order to ensure the results of the model are correct we must first ensure the following assumptions about the residuals of the model: Independence, constant variance and normality. These assumptions can be verified with the residuals plots for the model created at 22:00h shown in figure 6.

As the qqplot and histogram of the residuals show, they are normally distributed. The residual vs linear prediction graph shows a scatterplot of values which is centered around the zero line with randomly distributed residuals which follow no particular pattern. Figure 6 leads us to believe that the GAM assumptions are met. The Durbin-Watson test is used to check for autocorrelation of residuals. The Durbin-Watson test statistic for the GAM model at 22:00h is 1.66 (the DW statistic of all 24 models are reported in the evaluation Chapter 5). A result showing no auto-correlation at all should be around 2, therefore, this model is showing signs of auto-correlation of residuals. However, due to the highly correlated nature of the data (1 pbservation per hour) a DW statistic of 1.66 is an acceptable figure. We also check for influential points by computing Cook's distance and verifying the distance is always under 1.

4.1.4 Interaction terms

As shown by Hong et al. (2010) the inclusion of interaction effects in regression models for electrical forecasting can have a drastic improvement on the prediction accuracy. In the context of our study, we observe that each month of the year has a different temperature to load profile and the direction of this relationship changes with the month of the year. Therefore, it is necessary to include the interaction term of month with temperature. There is furthermore interaction between temperature and hour of the day, however, this is already captured by having one model for each hour.

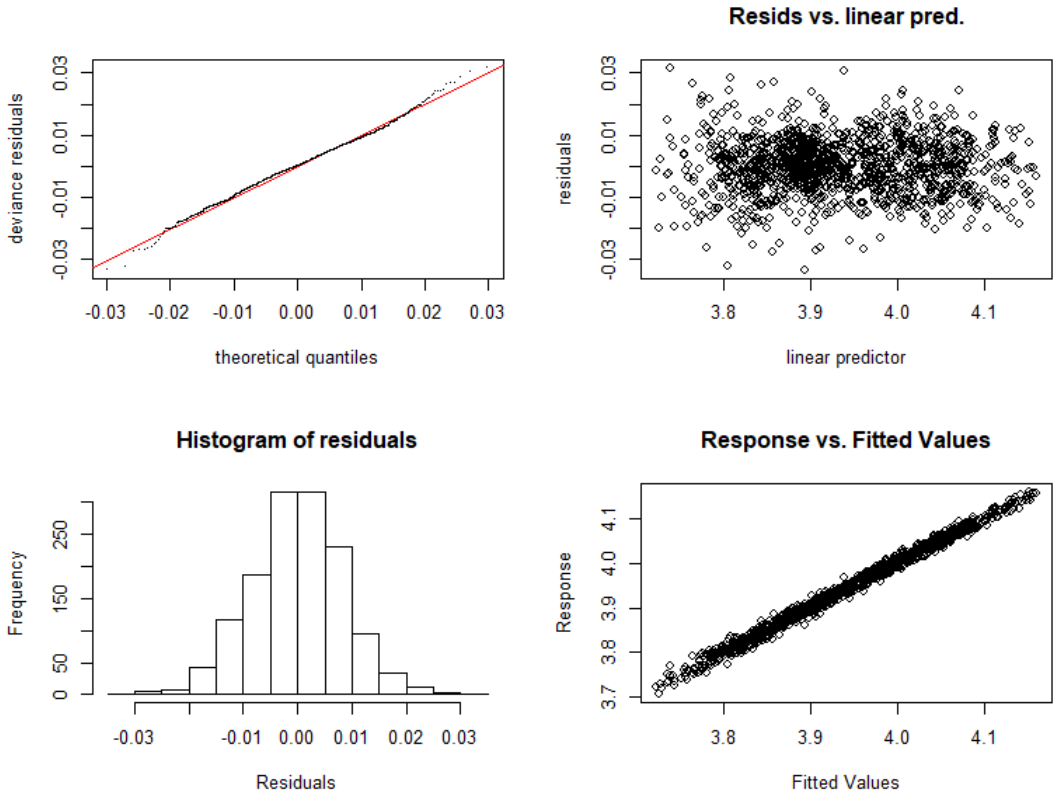

Figure 6: Residual plots of the GAM model

4.2 Arima regression implementation

In order to address auto-correlation issues observed in the GAM model, we propose to consider a second predictive technique ARIMA regression. By applying ARIMA method on the residuals of a standard regression we intend to reduce the autocorrelation and approximate the residuals to white noise. We fit a separate ARIMA regression model for each of the 24 hours in the day. By using this method a specific (p,d,q) parameter are fitted for all 24 models.

4.2.1 Arima regression predictors

The variables used for the ARIMA regression are the same as the GAM model and the dynamic forecast update method is also applied. We have introduced a quadratic term for every predictor variable that possess a non-linear behavior. Therefore, the regression technique is a polynomial regression.

4.2.2 Arima regression residuals

Figure 7 shows the acf plot from the polynomial regression model on the left and the acf plot of the ARIMA model applied on the residuals of the polynomial regression on the right. This acf plot was obtained from the 22:00 hour model with the corresponding ARIMA order of (1,0,1).

The ARIMA model on the residuals has clearly reduced the auto-regressive spikes. Finally, a Durbin-Watson (DW) test was made on both model and showed an improvement from 0.4 for the polynomial regression to 2 for the ARIMA regression. The DW statistics

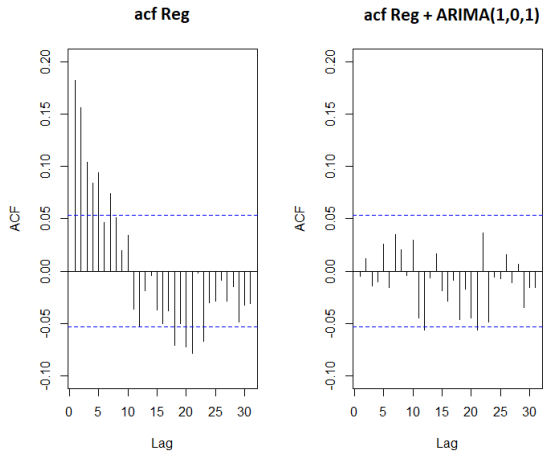

Figure 7: acf plots of regression vs ARIMA regression models for 22h

for each hour model are shown in the evaluation section 5.

5 Evaluation

5.1 Day Ahead Forecast / Case Study 1

Both models as detailed in the implementation Section 4 were applied to five dates evenly spread throughout the year of 2019: January 1st, March 15th, May 30th, July 15th and October 1st. For conciseness, only the table for October 1st is included with MAPE score and Durbin Watson statistic (Appendix A), the remaining forecasts are summed up in table 2. An internal study for the EDF utility company done by Rob J Hyndman and Shu Fan Bruhns et al. (2005) are used as the state of the art reference level for MAPE values. This paper is used because the authors forecast one day ahead values of electrical consumption with GAM models, which is exactly in line with this paper's objective.
We can observe in Appendix A that GAM and ARIMA regression models perform well and achieve a final MAPE of 1.08% and 1.37% respectively. This is a good result which is more accurate than the reference paper. The performance of the GAM model in our paper over the reference paper can be explained by multiple factors: we have proposed to introduce interaction terms in the GAM equation as well as smoothing spline instead of cubic splines as done in the reference paper. The trend variable is also modelled as a non-linear variable which was not the case in the reference paper.
The Durbin Watson statistics table in Appendix A shows that the ARIMA model has values much closer to 2 than the GAM model. Whilst the ARIMA regression has taken care of auto-correlation issues, these results have shown that this does not necessarily translate into more accurate predictions. We can therefore conclude that the auto-correlation of error assumption is somewhat flexible and Durbin-Watson statistics of 1.5 (out of 2) can still translate into accurate predictions.
Table 2 reports the results obtained from both models at each 5 dates.

Overall, the GAM model performs slightly better than the ARIMA model over the 5 dates. For the date of July 15th the ARIMA model has an MAPE of 5.23 whilst the GAM model has an MAPE of 1.62. This is due to the fact that GAM model's smoothing functions are better adapted at representing high temperature's complex non-linearity than the simple polynomial regression used in the ARIMA model. We also note that the predictions made for the date of January 1st 2019 have high MAPE values for both models. This is due to the fact that during this day the electrical consumption is highly perturbed due to New Year

	GAM MAPE	ARIMA MAPE
Jan 1st	5.69	5.14
Mar 15th	1.20	1.43
May 30th	1.64	1.45
July 15th	1.62	5.23
Oct 1st	1.07	1.37
Total	**2.24**	**2.92**

Table 2: GAM and ARIMA regression results

celebration.
The GAM model in this paper outperforms the reference paper in 4 out of the 5 dates selected (with the exception of January 1st).

5.2 Three months forecast / Case Study 2

The models as described and created in the implementation Chapter 4 were applied to a 3 month period ranging from March 17th 2019 to June 29th 2019. These dates are selected because they represent the same dates as lockdown period, one year before.
Appendix B shows the lineplot of GAM and ARIMA MAPEs for this period. We observe that the GAM model MAPE is constantly lower than the ARIMA regression MAPE curve. Moreover, the ARIMA regression makes more extreme errors with a maximum MAPE of 11% whilst the GAM model has a maximum MAPE of 6%. A very encouraging observation, is that both models do not display an increase of MAPE over time, which shows the robustness of both predictive models in short term forecasting and medium term forecasting as well as a correct choice of methodology for the forecast.
Table 3 shows the overall MAPE of both models for the 3 month period and model building times.

March 17th - June 29th 2019		
	GAM model	ARIMA model
MAPE	2.31	4.08
Time	2:05:00"	1:13:00"

Table 3: MAPE and time statistics for 3 months forecast of GAM and ARIMA models

The MAPE of GAM is significantly lower than the ARIMA model which means that GAM model is better suited for medium term forecasting. The processing time to build the ARIMA model is 1 hour and 13 minutes while the processing time to build the GAM model is 2 hours and 5 minutes. The gam() function from mgvc package in R is more computationally demanding than auto.arima() methods as each variable is fitted with a complex smooth spline function which takes longer to compute.
In accordance with the results obtained from and table 3, we will use the GAM model for the 3 months forecast of COVID-19 lockdown.
Figure 8 shows the three month forecasted values against the actual values of electricity for the GAM model.

Generally, we observe the predictions to be closely following the actual values. The lows between each period are the weekends where electrical consumption is at a lower value. It is also interesting to note how the electricity drastically decreases from march to may. This is due to the change of temperature going from the cold months to the hotter months where less electricity is used. Finally we observe an irregular pattern in actual values in early May where electrical consumption doesn't resemble any other weeks. This is due to the fact that

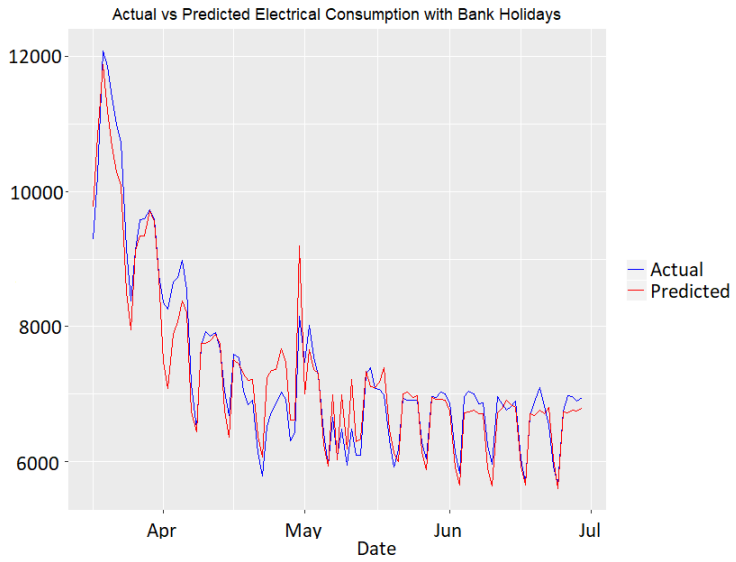

Figure 8: GAM model actual vs forecast 3 months forecast

May 8th and May 10th are both bank holidays in France which has a strong impact on electrical consumption. We note that the GAM model does somewhat capture these bank holidays by predicting lower values for May 8th and 10th (represented by the up and down spikes during those dates).
Considering the MAPE achieved during the 3 month period of 2% along with the satisfying actual vs predicted graph, we can conclude that our model is well adapted for the pandemic lockdown electricity prediction and can be applied in the last case study.

5.3 Lockdown simulation / Case Study 3

The GAM model has been applied from March 17th 2020 to June 29th 2020. The forecast values are plotted against the actual values and we report the graph in figure 9.

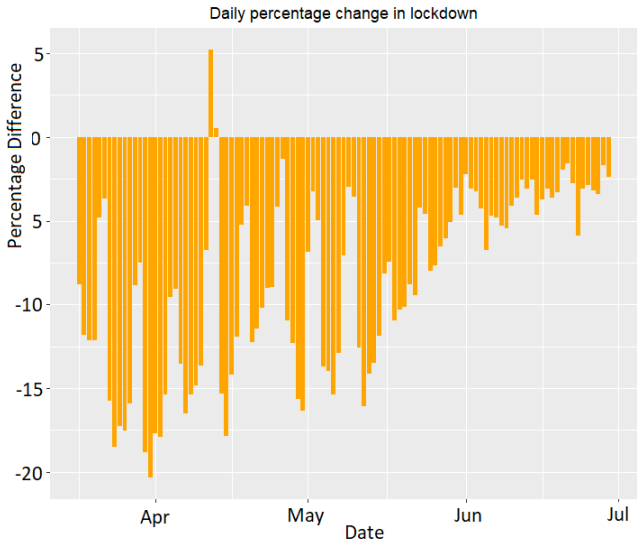

Figure 9: Barchart of lockdown vs simulated no lockdown percentage difference

The first observation is that the difference between lockdown and no lockdown statistics is almost exclusively negative, this means that the power consumption during lockdown has

decreased. We also note the percentage difference was at a maximum (20% difference) at about half a month after the lockdown start date and gradually returns to normal values in the next two months. Weekend days display lower percentage differences (marked by recurrent two day dents in the barplot) due to the fact that in normal conditions the electrical consumption is naturally lower (therefore the difference is also lower). Finally, on April 12 and 13 the power demand is actually higher than under normal conditions. These dates correspond to Easter Sunday dates, where during normal conditions the power would be significantly lower due to people staying at home.
Figure 10 shows the daily electrical consumption of two days in April against its forecasted normal condition.

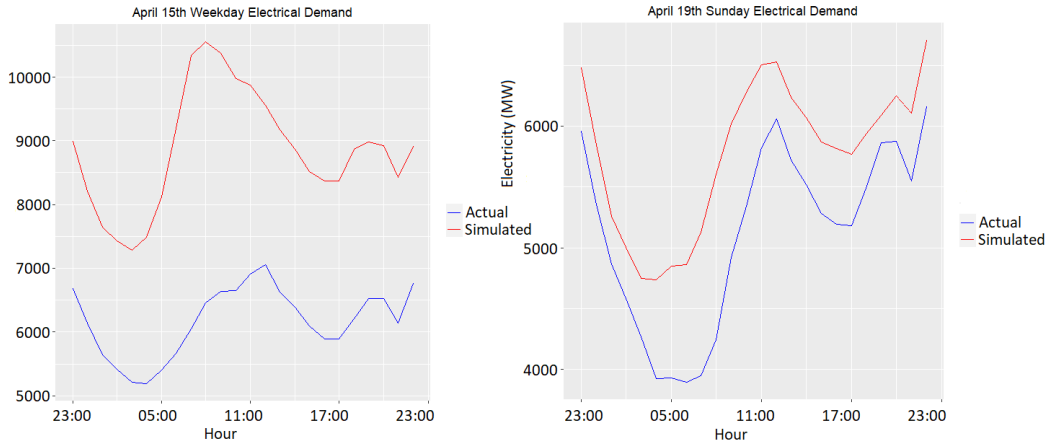

Figure 10: Daily Electrical demand of weekday and weekend day in April

Figure 10 shows that weekdays are marked with a strong decrease of electrical demand whilst weekend days are somewhat similar.
We also note that, during weekdays the actual curve (with lockdown) seems to be generally shifted towards the right (later hours of the day). In the morning there is a 1 hour shift which means people usually wake up later and start their daily activities 1 hour later. The normal daily maximum is usually around 9pm whereas under lockdown conditions it happens around 12pm, which again signifies a delay in standard daily activities. The weekend graph however does not show any shifting which means the weekend activities have not changed in lockdown phase.
Finally, an interesting observation is that the maximum electrical power demand during weekdays (7000MW) and the maximum electrical demand during weekend days (6000MW) during lockdown are very close. This is a significant change than normal conditions where weekday power is usually much higher than weekend power. This is due to the fact that during lockdown the population is asked to stay at home, therefore the power consumption during weekend and weekdays is similar.

6 Conclusion and Future Work

The first part of the conclusion aims to answer the following research question: Q1. What are the most adapted techniques in regression for short-term and mid-term electrical forecasting?
The study done in this paper has shown that Generalized Additive Models (GAM) and ARIMA regression models can both be used to accurately forecast electrical consumption over short-term time periods whilst GAM models are better suited for medium-term forecasts. In one day ahead forecasts, GAM models achieved a 2.24% MAPE slightly outperforming

the ARIMA regression of 2.92%. However, it was shown for both models that calendar dates with extreme changes in electricity demand were not accurately forecasted on days such as New Year Day. The short-term analysis also provided interesting information on auto-correlation and showed good results for regression models with moderately correlated errors.
Temperature is widely considered to be the most important predictor in electrical consumption. However, it has been shown that the addition of lagged temperature variables models thermal inertia and increases the prediction accuracy. This analysis also highlighted the importance of lagged electrical demand predictors in reducing auto-correlation and increasing final accuracy. Interaction terms were considered as well as non-linear trend model to accurately represent real-world situation. One notable limitation stems from the absence of economical and industrial predictors which can impact the prediction over longer periods of prediction.
The mid-term analysis (over 3 months) highlighted the weakness of GAM models, in terms of computational burden with longer model training times compared to ARIMA regression. However, on the 3-months prediction the GAM model significantly outperformed the ARIMA regression model with respective MAPEs of 2.31% and 4.08%. These results have shown that GAM models are adapted for both short-term and mid-term forecasting.

The second part of the conclusion aims to answer the following research question: Q2. What is the impact of the COVID-19 pandemic on electrical consumption?
After studying graphs 9 and 10 we conclude that the power consumption during lockdown phase is significantly lower during weekdays. This difference is less noticeable during weekend days although still present. There is a general shift of human activities towards later hours of the day specially in morning and mid-day activities during week days. We also note that the power consumption is gradually returning to normal values during the last two months of lockdown.
Finally, we note that during lockdown the power consumption statistic between weekend and week days are very similar.
One limitation comes from the fact that the original dataset combines all power usage types, therefore is it sometimes difficult to interpret exactly the reasons of the changes in electricity consumption.

6.1 Future Work

For future work, this study could be segregated in power usage types, such as residential, commercial or industrial. Each of these industries have been impacted differently by the lockdown and could benefit from further analysis.
We also note higher MAPE values for specials days, in future work this could be remedied by adding variables to represent special days or build separate models for said days. The weather variables could be improved by importing weather information for multiple locations across the Ile-De-France region and combining them by their population weighted statistic for a more accurate representation of their effect on electrical consumption. Additionally, the incorporation of economic variables is an important factor that should be considered for future studies.
Finally, an extension of Generalized Additive Models for time series data with auto-correlation, is Generalized Additive Mixed Models (GAMM) and allows to model the non-random behaviour of the errors through smooth functions control. This method could be considered for future studies, although it is known to be computationally heavy.

Acknowledgements

I would like to thank my supervisor Dr. Rashmi Gupta who's technical and moral support has been instrumental in the completion of this project.

References

Abu-Shikhah, N., Elkarmi, F. & Aloquili, O. (2011), 'Medium-term electric load forecasting using multivariable linear and non-linear regression', *Smart Grid and Renewable Energy* **2**.

Akbari, T. & Moghaddam, S. Z. (2020), 'Coordinated scheme for expansion planning of distribution networks: a bilevel game approach', *IET Generation, Transmission Distribution* **14**(14), 2839–2846.

Amral, N., Ozveren, C. S. & King, D. (2007), Short term load forecasting using multiple linear regression, *in* '2007 42nd International Universities Power Engineering Conference', pp. 1192–1198.

Angelaccio, M. (2019), Forecasting public electricity consumption with arima model: A case study from italian municipalities energy data, *in* '2019 International Symposium on Advanced Electrical and Communication Technologies (ISAECT)', pp. 1–3.

Ayan, O. & Turkay, B. (2018), Domestic electrical load management in smart grids and classification of residential loads, *in* '2018 5th International Conference on Electrical and Electronic Engineering (ICEEE)', pp. 279–283.

Bruhns, A., Deurveilher, G. & Roy, J.-S. (2005), 'A non linear regression model for mid-term load forecasting and improvements in seasonality', *15th Power Systems Computation Conference, PSCC 2005* .

Ching-Lai Hor, Watson, S. J. & Majithia, S. (2005), 'Analyzing the impact of weather variables on monthly electricity demand', *IEEE Transactions on Power Systems* **20**(4), 2078–2085.

Feng, Y. & Wang, Q. (2019), A new calendar effect and weather conditions based day-ahead load forecasting model, *in* '2019 IEEE Power Energy Society General Meeting (PESGM)', pp. 1–5.

Hastie, T. & Tibshirani, R. (1986), 'Generalized additive models', *Statistical Science* **1**(3), 297–310.
URL: *http://www.jstor.org/stable/2245459*

Hong, T., Gui, M., Baran, M. E. & Willis, H. L. (2010), Modeling and forecasting hourly electric load by multiple linear regression with interactions, *in* 'IEEE PES General Meeting', pp. 1–8.

Hutama, A. H., Akbar, S. & Catur Candra, M. Z. (2018), Medium term power load forecasting for java and bali power system using artificial neural network and sarimax, *in* '2018 5th International Conference on Data and Software Engineering (ICoDSE)', pp. 1–6.

Imtiaz, A. K., Mariun, N. B., Amran, M. M. R., Saleem, M., Wahab, N. I. A. & Mohibullah (2006), Evaluation and forecasting of long term electricity consumption demand for malaysia by statistical analysis, *in* '2006 IEEE International Power and Energy Conference', pp. 257–261.

Kuster, C., Rezgui, Y. & Mourshed, M. (2017), 'Electrical load forecasting models: A critical systematic review', *Sustainable Cities and Society* **35**, 257 – 270.

Li, F., Luo, Z., Zhang, M., Liao, J., Yang, H. & Wang, Y. (2019), Power purchasing decision of grid company considering power deviation under renewable energy quota, *in* '2019 IEEE 3rd Conference on Energy Internet and Energy System Integration (EI2)', pp. 688–692.

Lin, Q., Wang, Q., Zhang, G., Shi, Y., Liu, H. & Deng, L. (2018), Maximum daily load forecasting based on support vector regression considering accumulated temperature effect, *in* '2018 Chinese Control And Decision Conference (CCDC)', pp. 5199–5203.

Maniatis, P. (2017), 'A taxonomy of electricity demand forecasting techniques and a selection strategy', *International Journal of Management Excellence* **8**, 881.

Marra, G. & Wood, S. N. (2011), 'Practical variable selection for generalized additive models', *Computational Statistics and Data Analysis* **55**(7), 2372 – 2387.
URL: *http://www.sciencedirect.com/science/article/pii/S0167947311000491*

Mohamed, H. K., El-debeiky, S. M., Mahmoud, H. M. & El Destawy, K. M. (2006), Data mining for electrical load forecasting in egyptian electrical network, *in* '2006 International Conference on Computer Engineering and Systems', pp. 460–465.

Mustapha, M., Mustafa, M. W., Khalid, S. N., Abubakar, I. & Shareef, H. (2015), Classification of electricity load forecasting based on the factors influencing the load consumption and methods used: An-overview, *in* '2015 IEEE Conference on Energy Conversion (CENCON)', pp. 442–447.

Sharma, A., Bhuriya, D. & Singh, U. (2017), Survey of stock market prediction using machine learning approach, *in* '2017 International conference of Electronics, Communication and Aerospace Technology (ICECA)', Vol. 2, pp. 506–509.

Souzanchi-K, Z., Fanaee-T, H., Yaghoubi, M. & Akbarzadeh-T, M. (2010), A multi adaptive neuro fuzzy inference system for short term load forecasting by using previous day features, *in* '2010 International Conference on Electronics and Information Engineering', Vol. 2, pp. V2–54–V2–57.

Supapo, K. R. M., Santiago, R. V. M. & Pacis, M. C. (2017), Electric load demand forecasting for aborlan-narra-quezon distribution grid in palawan using multiple linear regression, *in* '2017IEEE 9th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management (HNICEM)', pp. 1–6.

Tao, Y., Zhao, F., Yuan, H., Lai, C. S., Xu, Z., Ng, W., Li, R., Li, X. & Lai, L. L. (2019), Revisit neural network based load forecasting, *in* '2019 20th International Conference on Intelligent System Application to Power Systems (ISAP)', pp. 1–5.

Yi, P., Jianyong, Z., Yun, Y., Rui, Z., Cheng, Z. & Tian, S. (2019), An electricity load forecasting approach combining dbn-based deep neural network and nar model for the integrated energy systems, *in* '2019 IEEE International Conference on Big Data and Smart Computing (BigComp)', pp. 1–4.

Appendix A MAPE Tables

1st October 2019			
Hour	GAM MAPE	ARIMA MAPE	Ref MAPE
0:00	1.28	3.07	1.84
1:00	1.23	1.31	2.10
2:00	1.38	2.88	1.72
3:00	0.53	1.55	1.62
4:00	0.37	0.95	1.42
5:00	1.26	1.13	1.56
6:00	3.06	0.10	2.21
7:00	2.20	2.49	2.18
8:00	0.06	1.82	2.01
9:00	0.42	0.39	1.81
10:00	0.57	0.83	1.88
11:00	0.71	0.46	1.89
12:00	0.57	0.75	2.00
13:00	0.16	1.07	1.91
14:00	0.36	0.14	2.02
15:00	0.32	1.04	1.99
16:00	0.19	0.60	2.23
17:00	1.08	0.64	2.14
18:00	2.16	3.28	1.83
19:00	1.85	0.49	1.50
20:00	2.35	1.71	1.58
21:00	0.06	1.63	1.50
22:00	2.59	3.43	1.68
23:00	1.11	1.01	1.62
Total	**1.08**	**1.37**	**1.88**

1st October 2019		
Hour	GAM DW	ARIMA DW
0:00	1.58	1.98
1:00	1.66	1.96
2:00	1.58	1.99
3:00	1.62	2
4:00	1.64	2
5:00	1.59	2.03
6:00	1.57	1.99
7:00	1.66	2
8:00	1.77	1.68
9:00	1.73	1.72
10:00	1.75	1.97
11:00	1.82	2
12:00	1.84	2.01
13:00	1.84	2
14:00	1.85	2
15:00	1.85	2.02
16:00	1.90	2.01
17:00	1.87	2
18:00	1.88	2
19:00	1.67	2
20:00	1.63	2.01
21:00	1.59	2
22:00	1.67	1.99
23:00	1.72	2
Total	**1.72**	**1.97**

Table 4: MAPE and DW statistics for October 1st, GAM, ARIMA and reference models

Appendix B MAPE Lineplot

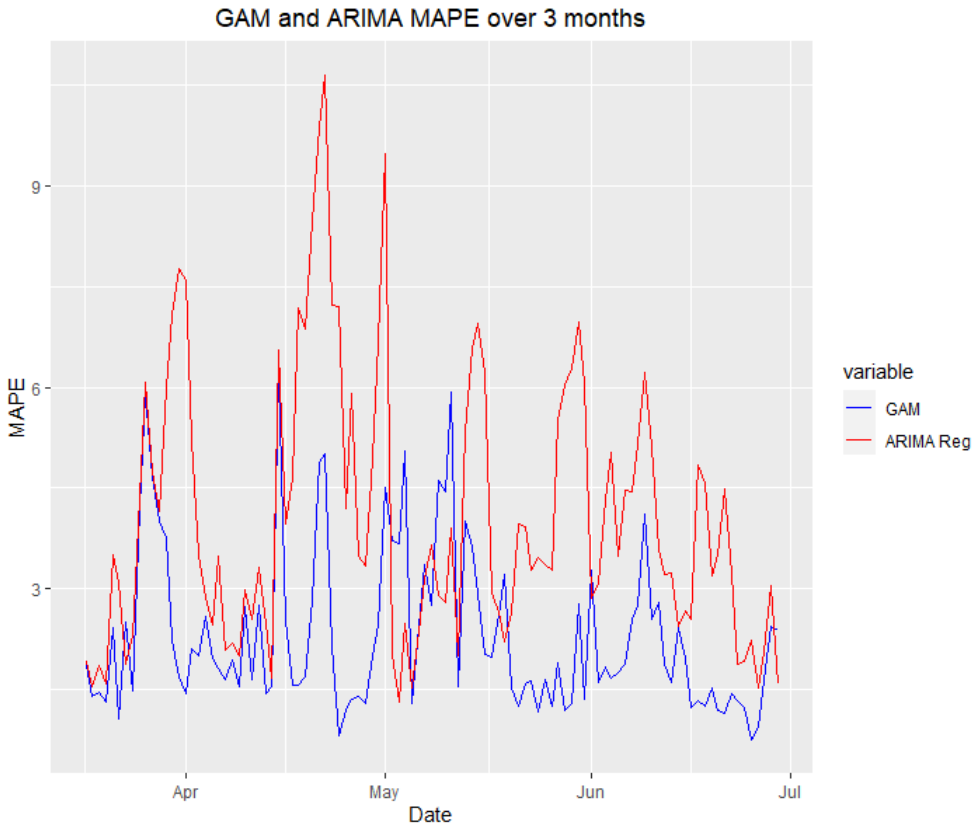

Figure 11: GAM and ARIMA models MAPE values

Appendix C Literature Review Summary Table

Authors	Paper name and year	Objective	Method	Findings & Results	Limitations
T. Hong, M. Gui, M. E. Baran, and H. L. Willis	"Modeling and forecasting hourly electric load by multiple linear regression with interactions" (2010)	Hourly electrical consumption for a 1 year period	Multiple linear regression with polynomial terms using weather (temperature and humidity), time components (weekends vs weekdays) and socio-economic factors. The most significant interaction terms between variables have also been added to the MLR.	The temperature - electricity non-linearity is dealt by introducing a quadratic term (temperature square) to capture the temperature curve. Workdays and weekend days are also considered. The final MAPE is 4.6%. This value is used as a baseline for our study (same timeframe and objective)	Localized to one transformer (very specific and small area), no holiday factors
A. Bruhns, G. Deurveilher, and J.-S. Roy	"A non linear regression model for mid-term load forecasting and improvements in seasonality" (2005)	Half-hourly electrical demand, seven days time-period in the Australian National Electricity Market	Generalized Additive Model (GAM)	The temperature non-linearity is dealt with cubic regression splines. The introduction of lagged temperature and lagged demands highly reduces the MAPE and solves autocorrelation issues. Finally, for optimal accuracy, each hour of the day can be used as a separate model. Final MAPE not shared.	No workday or holiday variable
A. H. Hutama, S. Akbar, and M. Z. Catur Candra	"Medium term power load fore-casting for java and bali power system using articial neural network and sarimax" (2018)	Medium-term forecasting of daily average loads for the 2017 year	Seasonal ARIMA with exogenous factors (SARIMAX), parameters (1,0,1) using day of the week and holiday dates	The addition of exogenous factors in ARIMA method significantly reduce the MAPE to 2.68% and capture important forecasting information.	Did not consider temperature factor, vital for hourly predictions
Y. Tao, F. Zhao, H. Yuan, C. S. Lai, Z. Xu, W. Ng, R. Li, X. Li, and L. L. Lai	"Revisit neural network based load forecasting" (2019)	Daily average values of electrical demand over a 1 month period.	Comparative study of multiple neural networks: Back Propagation (BP), Radial Basis Function (RBF), Elman Network (EN), and Long-Short Term Memory (LSTM). days.	LSTM and RBF methods have the highest accuracy with 3.5% MAPE and Elmann and BP are lower with MAPE > 5%.	Extend study to hourly predictions as neural networks can predict accurately