UNIVERSITY OF TWENTE.

ORTEC
OPTIMIZE YOUR WORLD

Improving forecast performance of LPG demand

Master Thesis, July 2017

Author:
Suzanne A. Thomasson

Supervisors:
University of Twente:
Dr.ir. L.L.M. van der Wegen
Dr. M.C. van der Heijden

ORTEC:
Dr.ir. G.F. Post
Rogier Emmen

University of Twente
Industrial Engineering and Management - Production and Logistic Management

Preface

In order to complete the master Industrial Engineering and Management, I have been working on my Thesis since February 2017. I had the privilege of doing this project at ORTEC - Optimization Technology. There are a few people that I would like to thank for helping me through this important part of my study.

First of all I would like to thank my supervisors from ORTEC, Rogier Emmen and Gerhard Post. Your positivity and trust in my capabilities encouraged me to go the extra mile and make me enthusiastic for the topic of forecasting. I want to specifically thank Rogier for taking all the time in the world to teach me all there was to know on forecasting at ORTEC, and your 'Code academy' made me realise that programming can be fun after all. Also I would like to thank all colleagues that made my time at ORTEC so much fun that I want to keep working there.

Also, I would like to thank my first supervisor Leo van der Wegen for the great guidance. You always made time to read my thesis and to give critical feedback, which I highly appreciate. Your insights helped me lift my thesis to a higher level. Also, I would like to thank Matthieu van der Heijden who became my second supervisor at one of the last stages of the project. You had to read my thesis on very short notice, but nevertheless, you gave feedback that helped me gain interesting new insights.

Furthermore, I want to thank my parents who always showed unconditional pride and support during the entire course of my study. During the last five years, you were the most important encouragement to pass and make you proud. Also, I want to thank my friends that made my student life unforgettable. Finally, I want to thank my boyfriend for his support, encouragement and care during this sometimes stressful period. Moving to Den Haag together with you made the end of my life as a student a lot better.

Den Haag, July 2017

Suzanne Thomasson

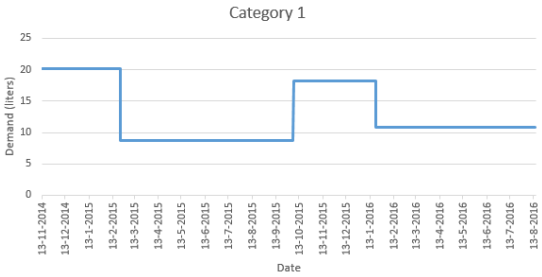
Summary

ORTEC has a customer, Company X, that distributes liquefied petroleum gas (LPG) to clients in the Benelux for which ORTEC should decide *when* to replenish and *how much* to deliver. This is part of the inventory routing product OIR. To be able to do this, ORTEC has a forecasting engine in which generally three methods are used to predict LPG demand: the degree days method (with the yearly script) that is based on the temperature dependency of LPG demand, simple exponential smoothing (SES) with period 1 day, and SES with period 7 days (for datasets that show a within week pattern). The forecast horizon that is required is one week. The time buckets used to predict are one day of length (irrespective of the frequency of observations, weekly or even more infrequent data is disaggregated to daily data). However, they did not know how well this methodology performs and if it is suitable for each client of Company X. We do know that in 38% of the trips, one or more customers did not get their delivery, because the truck was empty before having visited all customers on the planned route. A reason for this could be that the truck driver had to deliver more LPG than planned at the customers earlier on the route due to bad forecasts. Only in 11.5% of the deliveries, the customer received exactly the planned amount of LPG. In order to get insight into this matter, we stated the main research question:

Can, and if so, how can the forecast performance of LPG demand be improved?

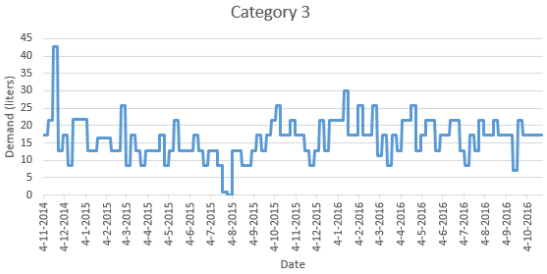
We categorised the clients of Company X into four categories: 'Category 1' for customers with only a few measurements (no telemetry system) and possibly yearly seasonality, 'Category 2' for clients that show a lot of 'negative' usage (the measuring equipment is inaccurate in the sense that it is not able to compensate for volume changes caused by fluctuations in temperature, which leads to the volume being above, and directly after, below a certain threshold resulting in supposedly negative usage), 'Category 3' for clients with weekly data and no seasonality, and 'Category 4' for customers with weekly data and yearly seasonality (sinus shaped). Figure 1 shows what representative datasets of these categories look like and the percentage indicates how many customers fall within each category.

While analysing the data, we found several issues that required solving before we were able to begin forecasting. The most serious data inconsistencies we found are: high, unjust peaks that occur after delivery of LPG, and the fact that all positive usage is unjustly included whereas some should be compensated by negative demand. The first occurs, because of the inconsistent volume measurements the truck driver fills in on a form after delivering LPG and is solved by making those measurements irrelevant and using Cook's distance to remove remaining outliers. The second appears, because the volume of LPG fluctuates with temperature. This causes the input data used to forecast to be 134% of a tank capacity higher than in reality which is solved by sending negative usage to the forecasting engine

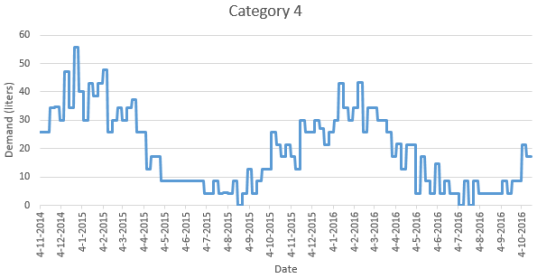


(a) Category 1: few measurements (24%)

(b) Category 2: negative usage (38%)



(c) Category 3: weekly data, no seasonality (3%)



(d) Category 4: weekly data, yearly seasonality (35%)

Figure 1: Categories of datasets

instead of discarding it. After classifying 2284 datasets on category, we found that 38% of them is 'Category 2' which means that the problem is of substantial size. The number of deliveries to 'Category 2' storages can be reduced by 87% when sending negative usage to the forecasting engine.

After solving the data issues, we investigated which forecasting method is most suitable per data category. We found that suitable methods for temperature dependent time series are: Holt-Winters (additive and multiplicative) and linear regression (simple and multiple, using climatological variables as external variables). For the series without seasonality, suitable methods are: simple exponential smoothing (SES) and moving average. For the datasets that show intermittent demand patterns, that result from the inaccuracy of the measuring equipment, appropriate methods are: SES, Croston's method, and the TSB method. Besides, there is proof for the accuracy and robustness of combining forecasts. An important finding is that the performance of the methods should not be expressed in terms of mean average percentage error (MAPE), because it is unreliable for low volume datasets. Instead, the root mean squared error (RMSE) should be used.

The suggested methods are performed on several datasets of the different categories. The best performing methods for 'Category 1' are the methods that exploit the temperature dependency of the LPG demand which improves the current forecast performance by 67%. Interesting is that the current methodology, single exponential smoothing, is one of the worst performing methods. For 'Category 2' datasets, SES turns out to be the best method. 'Category 3' (11.3% improvement of the RMSE) and 'Category 4' are predicted best by the same methods that work for 'Category 1'. The forecast results indicate that simple regression performs better than the degree-days method in most cases (improves the forecast performance by 6.5% for 'Category 4' datasets). Therefore, we recommend to change the degree-days

implementation to simple regression (instead of the current implementation where first the temperature dependency is removed from the series, then the remaining supposedly straight line is predicted with SES, and finally the temperature dependency is added again).

Besides forecasting, we investigated automating model selection. Currently, for each dataset, the user has to choose a forecast script manually. Time could be saved by automating this. We looked at the possibilities of classification. After implementing different methods, we conclude that logistic regression performs best in terms of accuracy, interpretability, and ease of implementation. This method is able to classify the data with an accuracy of 98.4% in WEKA.

Based on these findings, we recommend the following to improve the current forecasting procedure:

- Make the after delivery readings irrelevant in OIR for all storages, except for 'Category 1' datasets
- Forecast 'Category 2' datasets with simple exponential smoothing and the rest with the degree-days method
- Implement Cook's distance before calculating the regression coefficients
- Send the measurements that show negative usage to the forecasting engine instead of discarding them
- Use the RMSE instead of the MAPE as performance indicator
- Implement simple linear regression for the degree-days method
- Compute a tracking signal to monitor whether the forecasting system remains in control using an α of 0.1 and control limits of ± 0.55, but only for 'Category 3' and 'Category 4' datasets
- Use logistic regression as classification method

An important shortcoming of this research is that we know little on the impact of the mentioned problems and improvements. No or little data is available on what happens when the forecasts are inaccurate. When the truck is empty before having visited all customers on the planned route, the last customer(s) have to be visited in another route, and when there is LPG left in the truck at the end of the planned route, another customer should be found to empty the truck at which is both undesirable. In 38% of the routes, one or more customers did not get replenished, because the truck was empty before the end of the planned route. However, we do not know what the implications are in terms of costs and to what extent our recommendations reduce those costs. First of all, this is because not being able to visit a customer caused by the truck being empty before having visited all customers on the planned route, or having LPG left in the truck after finishing the entire route, and a customer running empty before being visited, cannot always be blamed on inaccurate forecasts. Secondly, no data is available on how often all of these situations occur. We do know that for 38% of the storages, the number of deliveries can be reduced by 87% and like mentioned earlier, only in 11.5% of the deliveries, the exact planned amount is delivered.

Contents

Chapter 1

Introduction

The world around us is becoming more and more dynamic. Companies are finding ways to become more efficient and predict the future in order to stay ahead of competition. ORTEC is a leading company in helping companies to achieve this. This report introduces the problem that ORTEC currently has to deal with. Section 1.1 gives some background on the company and introduces the research motive. Section 1.2 gives the goal of the research and states the research questions.

1.1 Background ORTEC and research motive

ORTEC is one of the world's leaders in optimization software and analytic solutions. Their purpose is to optimize the world with their passion for mathematics. Currently, ORTEC has developed a tool that can forecast one time series using at most one external variable, for example sales and the temperature outside. Ice-creams sell better when it is hot outside, than temperature can be used to improve the forecast. However, more and more customers demand forecasting of more variables that influence each other. The topic of research is the situation where several aspects together generate the output to forecast.

Currently, ORTEC has a client, Company X, that distributes LPG to its clients. These clients have one or several LPG bulk tanks (which we call storages from now on). Generally, every one or two weeks, Company X receives inventory levels of the storages belonging to the clients. Since it is inefficient to replenish frequently, ORTEC forecasts the LPG usage in order to predict when to do this, namely, just before the client is out of stock. Also, forecasts are required to determine how much LPG should be delivered to each storage. Company X and its clients have what is called a Vendor Managed Inventory (VMI) system: a supply-chain initiative where the supplier is authorized to manage inventories of agreed-upon stock-keeping units at retail locations (Çetinkaya & Lee, 2000). By this, inventory and transportation decisions are synchronized. This relationship allows Company X to consolidate shipments which means that rides can be combined and less transportation is required. This project is part of the product 'ORTEC Inventory Routing (OIR)', and ORTEC is asked by Company X to determine when to replenish and how much.

An important part of this inventory routing product is forecasting the usage, since that is used to determine the replenishment volume and -moment. In this branch, a bad forecast means that either the truck is empty before having replenished all customers on the route or that the truck still contains LPG at the end of the route which means that another customer must be found to empty the truck at. On a single route, on average seven customers are

visited and the number of customers visited on a route varies from 2 to 20. We do not know how often it occurs that LPG remains after having visited all customers on the route but we do know that **in 38% of the routes, one or more customer could not get its delivery, because the truck was already empty**. Only in 11.5% of the deliveries, the truck driver delivers exactly the planned amount of LPG.

The fact that the inventory level is given every one or two weeks but the outside temperature is given on a daily basis and daily forecasts are needed, makes it challenging to forecast correctly. Currently this forecast is done by simple exponential smoothing and the degree days method. Degree days are a simplified form of historical weather data that are used to model the relationship between energy consumption and outside air temperature. The method uses heating degree days (HDD) which are days that heating was necessary due to the cold weather and cooling degree days (CDD) for when air-conditioning is used when it is hot outside. For example, when the outside air temperature was 3 degrees below the base temperature for 2 days, there would be a total 6 heating degree days. The same holds for CDD, but then the degree days are calculated by taking the number of days and number of degrees that the outside temperature was above that base temperature. Originally, this method is used to determine the weather-normalized energy consumption. Weather-normalization adjusts the energy consumption to factor out the variations in outside air temperature. For example, when a company consumed less energy in one year compared to the year before, weather-normalization can determine whether this was because the winter was a bit warmer or because the company was successful in saving energy. Normalisation is not necessary when forecasting. With the help of historical data on the energy consumption and number of degree days, a regression analysis can be used to determine the expected energy consumption given the number of degree days. The method is explained in more detail later in the report. The advantage of this method is that degree-day data is easy to get hold of and to work with. Besides, it can come in any time scale, so also the one or two weeks that ORTEC has to work with.

Even though this method is easy to work with, ORTEC does not know exactly how good this method performs and if all customers benefit from this method. Also, they want to know whether there are other methods available and if there are other external variables (besides the temperature that is currently used) that could improve the forecast.

1.2 Research goal and research questions

Since ORTEC does not know how well the current forecasting methodology works, and whether other external variables (covariates) could improve the forecast, the main research question is:

Can, and if so, how can the forecast performance of LPG demand be improved?

In order to reach the research objective, several sub questions are answered:

1. What is known in literature on forecasting LPG demand or similar cases? (Chapter 2)
   (a) Which methods are used in literature for forecasting LPG demand or similar cases?
   (b) How can forecast performance be measured?
   (c) How can data automatically be categorized?

Since the current situation requires quite some background information, the first question elaborates on this in Chapter 2. Scientific articles and books on forecasting are used.

2. What is the current situation at ORTEC? (Chapter 3)
    (a) What method is currently used by ORTEC for forecasting LPG demand of the smaller clients of Company X?
    (b) What are the issues of this methodology?
    (c) What are the characteristics of the data?

This second question is answered in Chapter 3. For this question and its sub questions, interviews with the persons currently working on the project have to be performed, which are persons working on the forecast software but also persons that have been working on the business case and have been in contact with Company X. Question 2b is answered by finding out what patterns are present in the data and on which relationship(s) the current methodology is based and investigating how suitable this is. The third sub question is answered by analysing datasets of different customer types.

3. Which methods are eligible for ORTEC? (Chapter 4)
    (a) How should the data be cleaned to be suitable for forecasting?
    (b) How can the current methodology be improved?
    (c) Which method performs best and is most suitable?

To answer these questions, we have to find out which inconsistencies and issues in the data should be corrected. After that, we try to improve the current methodology by making adjustments. Then statistical tools as R, SPSS, and simple visual plots are used in order to find trend and/or seasonality and other data patterns that help determine which other forecasting methods might be suitable for the data. Chapter 4 elaborates on this research question. Which method performs best is selected by using different performance indicators as the MSE, MAPE, and MAD, which are explained later.

4. How can classification methods be used for automatic method selection? (Chapter 5)
    (a) How should the classification methods proposed in literature be used?
    (b) Which classifier performs best?

This last question is answered in Chapter 5 by using the tools WEKA (Waikato Environment for Knowledge Analysis) and R that include a wide range of machine learning techniques and data preprocessing tools.

Chapter 2

Review of existing literature

In order to answer the first research question 'What is known in literature on forecasting LPG demand or similar cases?', this chapter discusses what is available in literature on these topics to get some more insight. The sub questions answered are 'Which methods are used in literature for forecasting LPG demand or similar cases?' and 'How can forecast performance be measured?'.

The first section briefly discusses short-term gas demand and which methods are broadly used in literature to forecast this. There are two common models that are based on projecting forward from past behaviour: moving average forecasting and exponential smoothing. The second section explains moving average and the third explains exponential smoothing and gives several alternative exponential smoothing methods that could be useful for predicting LPG demand. Moving average and exponential smoothing are time series models, which means that the dependent variable is only determined by time and/or previous values of the variable. However, causal models could also be useful for forecasting. In a causal model, external factors (other time series) form the explanatory variables of the dependent variable. For example, in the LPG case, the LPG demand could possibly also be dependent on the outside temperature which is an external factor (also called covariate).
Some of the best-known causal models are regression models, those are discussed in Section 2.5. Currently, ORTEC uses a causal model called the degree-days method. This method is explained in Section 2.6. Besides, there are models that combine time series and causal models. Those are discussed in Section 2.7. Since literature indicates that Artificial Neural Networks (ANNs) could be helpful forecasting LPG demand, Section 2.8 explains this. Section 2.9 explains how combining different forecasting methods could improve forecast performance. In order to determine which of these methods performs best, it is necessary to find out how to measure forecast performance. This is explained in Section 2.10. Section 2.11 elaborates on the sample size required by each method. Section 2.12 addresses automatic model selection by classification. The chapter ends with a conclusion in Section 2.13.

2.1 Short term gas demand

Studies on energy demand have mostly been centred on the electricity sector (Mensah, 2014). The literature that is available on LPG demand is mostly focused on long term prediction instead of modelling short term load like Parikh et al. (2007) and Mensah (2014). Suganthi & Samuel (2012) give an overview. Even this overview, however, focuses on long-term forecasting. Since both LPG demand and electricity demand depend heavily on outside air temperature, the modelling of demand can be done in a similar way. Therefore we elabo-

rate on forecasting electricity demand in this chapter. Literature introduces and tests many electricity demand forecasting methods. This section gives a short introduction on what has been done in literature on this specific topic.

The aim of business forecast is to combine statistical analyses and domain knowledge to develop acceptable forecasts that will ultimately drive downstream planning activities and support decision making (Hoshmand, 2009). In production and inventory control, forecasting is a major determinant of inventory costs, service levels, and many other measures of operational performance (Gardner, 2006). It is used as a tool to make economic and business decisions on tactical, strategic, or operational level. Short-term load (electricity or LPG demand) forecasting is essential in making decisions on all those levels. Many operational decisions are based on load forecasts, under which the decision on when to replenish the LPG storages of several clients which should happen as less often as possible in order to save costs but the client may never run out of stock (Fan & Hyndman, 2010) but also how much to deliver to the customers. In order to do this, with the help of load forecasting, we need to predict when the clients are expected to be out of stock and how much LPG should be delivered.

Various techniques have been developed for electricity demand forecasting. Statistical models as linear regression-, stochastic process- and ARIMA models are widely adopted (Fan & Hyndman, 2010). Recently, machine learning techniques and fuzzy logic approaches have also been used and achieved relatively good performance (Fan & Hyndman, 2010). Exponential smoothing has received more attention since the study of Taylor (2003). Since exponential smoothing is considered an easy in use method that gives relatively accurate results, ORTEC uses this in combination with the degree-days method in order to forecast LPG demand.

Even though mostly electricity, but also natural gas demand, load forecasts are based on outside temperature, there are other exogenous variables that influence demand as working days, weekends, feasts, festivals, cloud cover, and humidity (Kumru & Kumru, 2015). However, the main parameter that heavily influences demand is temperature.

2.2 Moving average method

The moving average approach takes the previous n periods' actual demand figures, calculates the average over these n periods, and uses this average as a forecast for the next period's demand. The data older than the n periods play no part in the next period's forecast and n can be set at any level (Slack et al., 2010). The advantage of this method is that it is very fast and easy to implement and execute. The main assumption in moving average models is that an average of past observations can be used to smooth the fluctuations in the data in the short-run (Hoshmand, 2009). As each observation becomes available, a new mean is computed by leaving out the oldest data point and including the newest observation.

2.2.1 Single moving average

The next equation shows how a moving average is computed:

$$F_t = \frac{Y_{t-1} + Y_{t-2} + ... + Y_{t-n}}{n} \tag{2.1}$$

where
F_t is the forecast value for time t

Y_t is the actual value at time period t
n is the number of terms in the moving average

The choice of n has implications for the forecast. The smaller the number of observations, the forecast is only based on the recent past. The larger the number, the forecast is the average of the recent past and the further past. The first is desirable if the analyst encounters sudden shifts in the level of the series and a large number desirable when there are wide and infrequent fluctuations in the series (Hoshmand, 2009). Moving average is not able to cope with cyclical patterns as seasonality.

2.2.2 Double moving average

The single moving average method as just described is not able to cope with trend, seasonality, or cyclical patterns that could be present in the data. Double moving average is used when the time series data have a linear trend. The first set of moving averages (MA_t) is computed as discussed in Subsection 2.2.1, and the second set is computed as a moving average of the first set (MA'_t).

$$MA_t = F_t = \frac{Y_{t-1} + Y_{t-2} + ... + Y_{t-n}}{n} \tag{2.2}$$

$$MA'_t = \frac{MA_{t-1} + MA_{t-2} + ... + MA_{t-n}}{n} \tag{2.3}$$

The difference between MA_t and MA'_t is computed as follows:

$$a_t = 2MA_t - MA'_t \tag{2.4}$$

Then the slope (trend) is measured by:

$$T_t = \frac{2}{n-1}(MA_t - MA'_t) \tag{2.5}$$

With these, the forecast for x periods into the future can be made by:

$$F_{t+x} = a_t + T_t x \tag{2.6}$$

where
F_{t+x} is the forecast value x periods ahead
n is the number of periods in the moving average
Y_t is the actual value at period t

2.3 Exponential smoothing

In the moving average method, all observations get the same weight. However, exponential smoothing places more emphasis on the most recent observations. This section describes how this method works and describes variants that could be more suitable for the LPG case according to literature. There could be certain patterns in the data that require different methods than simple exponential smoothing. Those are explained in this section.

Exponential smoothing forecasts demand in the next period by taking into account the actual demand in the current period and the forecast that was previously made for the current period, the details are explained later. The method relies on the assumption that the mean is not fixed over all time, but rather changes over time (Hoshmand, 2009). This chapter gives several methods of exponential smoothing. Before explaining the basics of these methods, the different methods are being classified using the method proposed by Pegels (1969) and later extended by Gardner (1985) and again by Taylor (2003). Table 2.1 gives an overview.

In a time series, trend could be present. Trend is defined as 'long-term change in the mean level per unit time'. It can be additive (of constant size from year to year), or multiplicative (proportional to the local mean), or mixed (Chatfield, 2006) (see Figure 2.1). Another aspect that could be present in the time series is seasonality which could also be additive, multiplicative, or mixed. Table 2.1 gives the possible combinations.

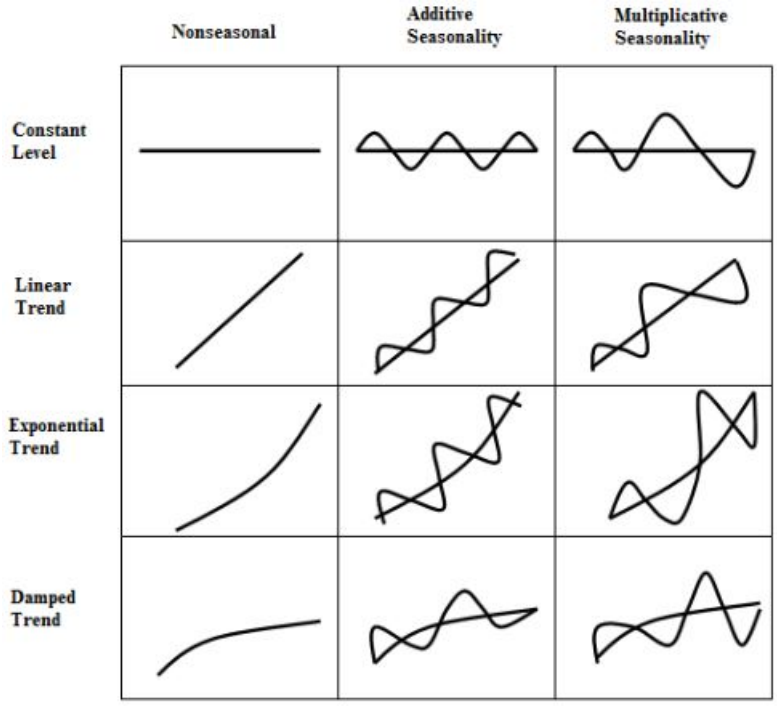


Figure 2.1: Additive and Multiplicative Seasonality (Gardner, 1985)

Trend component	Seasonal component		
	N (None)	A (Additive)	M (Multiplicative)
N (None)	N,N	N,A	N,M
A (Additive)	A,N	A,A	A,M
A_d (Additive damped)	A_d,N	A_d,A	A_d,M
M (Multiplicative)	M,N	M,A	M,M
M_d (Multiplicative damped)	M_d,N	M_d,A	M_d,M

Table 2.1: Classification exponential smoothing (Hyndman, 2008)

For example (M,A) means that the trend component is multiplicative and the seasonal component is additive. OTexts (2017) gives a nice overview, more extensive than the one of

De Gooijer & Hyndman (2006), of the methods described by Hyndman & Athanasopoulos (2014):
(N, N) = Simple exponential smoothing (Subsection 2.3.1)
(A, N) = Holt's linear method (Subsection 2.3.3)
(M, N) = Exponential trend method
(A_d, N) = Additive damped trend method (Subsection 2.3.4)
(M_d, N) = Multiplicative damped trend method
(A, A) = Additive Holt-Winters method (Subsection 2.3.5)
(A, M) = Multiplicative Holt-Winters method Subsection 2.3.5)
(A_d, M) = Holt-Winters damped method (Subsection 2.3.6)

Since ORTEC mentioned that currently simple exponential smoothing is used, that is discussed in the next section. However, in literature, the Holt-Winters method is proposed as well performing exponential smoothing in the specific electricity demand case (Taylor, 2003; Taylor, 2010). Therefore that method is also discussed (Subsection 2.3.5). Since the trend methods assume a constant trend, forecasts using those often tend to over-forecast, especially for long-term forecasts. Therefore, also the damped trend methods are discussed.

One of the biggest advantages of exponential smoothing is the surprising accuracy that can be obtained with minimal effort in model identification (Gardner, 1985). There is substantial evidence that exponential smoothing models are robust, not only to different types of data but to specification error (Gardner, 2006). Many studies have found that exponential smoothing was at least as accurate as Box-Jenkins (ARIMA). However, a disadvantage of exponential smoothing in general is its lack of an objective procedure for model identification (Gardner & McKenzie, 1988). Besides, their usual formulations do not allow for the use of explanatory variables, also called predictors (Bermúdez, 2013).

2.3.1 Simple exponential smoothing

Simple exponential smoothing is a common approach based on projecting forward from past behaviour without taking into account trend and seasonality. It takes into account the actual demand of the current period and the forecast which was previously made for the current period, in order to forecast the value in the next period. It is also possible to forecast more periods into the future (x instead of 1), but this is not desirable since only the presence of a small trend already disrupts the forecast. The most recent observations play a more important role in making a forecast than those observed in the distant past (Hoshmand, 2009). The easiest form of exponential smoothing is:

$$F_t = \alpha Y_{t-1} + (1 - \alpha) F_{t-1} \tag{2.7}$$

where
α is the smoothing constant
Y_{t-1} is the actual value of last period
F_{t-1} is the forecasted value for last period

Parameter α is the weight given to the last piece of information available to the forecaster, and therefore assumed to be most important. This smoothing constant governs the balance between the responsiveness of the forecasts to changes in demand, and the stability of the forecasts (Slack et al., 2010). The method is called 'exponential smoothing', because the weights decrease exponentially as the observations get older which makes observations from

the distant past less important than the more recent ones like mentioned before (Hyndman, Koehler, Ord & Snyder, 2008).

This smoothing constant α must be chosen when using the exponential smoothing method. In this easiest form of exponential smoothing, there is only one smoothing parameter, but later in this report methods containing several smoothing parameters are discussed. This parameters could be chosen based on experience of the forecaster but a more robust method is to estimate them from previous data. A way to do this is by using the sum squared error (SSE). These errors are calculated by:

$$\sum_{t=1}^{n}(Y_t - F_t)^2 \tag{2.8}$$

where Y_t is the actual value and F_t is the forecasted value and n is the number of observations. By minimizing the SSE, the values of the parameter(s) can be estimated (Price & Sharp, 1986). Section 2.10 gives other performance measurements on which the smoothing parameter can be estimated.

2.3.2 Double exponential smoothing

The simple exponential smoothing method is not able to handle trended data. Double exponential smoothing methods on the other hand, are. Let us first discuss Brown's double exponential smoothing, also known as Brown's linear exponential smoothing (LES) that is used to forecast time series containing a linear trend (Hoshmand, 2009). The forecast is done by:

$$F_{t+x} = a_t + xT_t \tag{2.9}$$

where
F_{t+x} is the forecast value x periods into the future
T_t is an adjustment factor similar to a slope in a time series (trend)
x is the number of periods ahead to be forecast

To compute the difference between the simple and the double smoothed values as a measure of trend, we use the following equations:

$$A'_t = \alpha F_t + (1-\alpha)A'_{t-1} \tag{2.10}$$

$$A''_t = \alpha A'_t + (1-\alpha)A''_{t-1} \tag{2.11}$$

where A'_t is the simple smoothed value and A''_t is the double smoothed value. This leads to:

$$a_t = 2A'_t - A''_t \tag{2.12}$$

Besides, the adjustment factor is calculated by:

$$T_t = \frac{\alpha}{(1-\alpha)}(A'_t - A''_t) \tag{2.13}$$

2.3.3 Holt's linear trend method

Brown's method is not the only one that is able to cope with linear trend. Holt's two-parameter method (linear trend method) is too. The difference with Brown's method is that the trend and slope are smoothed by different smoothing constants. This leads to having a

little more flexibility. The shortcoming, however, is that determining the best combination between the two smoothing constants is costly and time consuming (Hoshmand, 2009). The following formula gives the forecast:

$$F_{t+x} = A_t + xT_t \tag{2.14}$$

with

$$A_t = \alpha Y_t + (1-\alpha)(A_{t-1} + T_{t-1}) \tag{2.15}$$

$$T_t = \beta(A_t - A_{t-1}) + (1-\beta)T_{t-1} \tag{2.16}$$

where
A_t is the smoothed value
α is the smoothing constant (between 0 and 1)
β is the smoothing constant for the trend estimate (between 0 and 1)
T_t is the trend estimate
x is the number of periods to be forecast into the future
F_{t+x} is the forecast for x periods into the future

2.3.4 Additive damped trend method

As discussed before, Holt's linear trend model, assumes a constant trend indefinitely into the future. In order to make forecasts more conservative for longer forecast horizons, Gardner & McKenzie (1985) suggest that the trends should be damped (Hyndman, 2014). This model makes the forecast trended on the short run and constant on the long run. The forecasting equation is as follows:

$$F_{t+x} = A_t + (\phi + \phi^2 + ... + \phi^x)T_t \tag{2.17}$$

with

$$A_t = \alpha Y_t + (1-\alpha)(A_{t-1} + \phi T_{t-1}) \tag{2.18}$$

$$T_t = \beta(A_t - A_{t-1}) + (1-\beta)\phi T_{t-1} \tag{2.19}$$

where ϕ is the damping parameter ($0 < \phi < 1$).

2.3.5 Holt-Winters method

The linear trend method can be adjusted when a time series with not only trend but also seasonality must be forecasted. The resulting method is known as the famous Holt-Winters method. The trend formula remains the same, only the formula for A_t (level) and for F_{t+x} change and an equation for seasonality is added (Taylor, 2003).

Additive Holt-Winters method

There are two types of seasonal models: additive (assumes the seasonal effects are of constant size) and multiplicative (assumes the seasonal effects are proportional in size to the local deseasonalised mean). Forecasts can be produced for any number of steps ahead (Chatfield, 1978). The forecast formula for the additive variant is adjusted to the following:

$$F_{t+x} = A_t + xT_t + I_{t-s+x} \tag{2.20}$$

The A_t formula becomes

$$A_t = \alpha(Y_t - I_{t-s}) + (1-\alpha)(A_{t-1} + T_{t-1}) \tag{2.21}$$

and the following formula for seasonality is added

$$I_t = \delta(Y_t - A_{t-1} - T_{t-1}) + (1 - \delta)I_{t-s} \tag{2.22}$$

where
δ is the smoothing constant for seasonality
I_t is the local s-period seasonal index

Multiplicative Holt-Winters method

The forecast equation for the multiplicative variant is as follows:

$$F_{t+x} = (A_t + xT_t)I_{t-s+x} \tag{2.23}$$

and the formula for level A_t is

$$A_t = \alpha\left(\frac{Y_t}{I_{t-s}}\right) + (1 - \alpha)(A_{t-1} + T_{t-1}) \tag{2.24}$$

and the seasonality formula is adjusted to

$$I_t = \delta\left(\frac{Y_t}{A_{t-1} + T_{t-1}}\right) + (1 - \delta)I_{t-s} \tag{2.25}$$

The Holt-Winters method is widely used for short-term electricity demand forecasting because of several advantages. It only requires the quantity-demanded variable, it is relatively simple, and robust (García-Díaz & Trull, 2016). Besides, it has the advantage of being able to adapt to changes in trends and seasonal patterns in usage when they occur. It achieves this by updating its estimates of these patterns as soon as each new observation arrives (Goodwin, 2010).
A disadvantage of the Holt-Winters method (both additive and multiplicative) is that it is not so suitable for long seasonal periods such as 52 for weekly data or 365 for daily data. For weekly data, 52 parameters must be estimated, one for each week, which results in the model having far too many degrees of freedom (Hyndman & Athanasopoulos, 2014). Ord & Fildes (2013) propose a method to make these seasonality estimates more reliable for the multiplicative variant. Instead of calculating the seasonals on individual series level, they calculate the seasonality of an aggregate series. This results in having less randomness in the estimates. For this, series with the same seasonality should be aggregated. For example, in the LPG case, more LPG is used in the winter and less in the summer so we expect that many clients follow the same usage pattern. When similar series are aggregated, individual variation decreases. This does not solve the problem of having to estimate many parameters but makes the estimation slightly more robust.

2.3.6 Holt-Winters damped method

As for Holt's linear model, a damped version exists for the Holt-Winters method. The forecasting equation is:

$$F_{t+x} = A_t + (\phi + \phi^2 + ... + \phi^x)T_t + I_{t-s+x} \tag{2.26}$$

$$A_t = \alpha(Y_t - I_{t-x}) + (1 - \alpha)(A_{t-1} + \phi T_{t-1}) \tag{2.27}$$

with the same equation for trend as the additive damped trend method:

$$T_t = \beta(A_t - A_{t-1}) + (1 - \beta)\phi T_{t-1} \tag{2.28}$$

but adds an equation for seasonality:

$$I_t = \delta(Y_t - A_{t-1} - \phi T_{t-1}) + (1 - \delta)I_{t-s} \tag{2.29}$$

Also here, the damping factor is $0 < \phi < 1$.

2.3.7 Parameter estimation and starting values

In order to implement either of these exponential smoothing methods methods, the user must

- provide starting values for A_t, T_t, and I_t
- provide values for α, β, and δ
- decide whether to normalise the seasonal factors (i.e. sum to zero in the additive case or average to one in the multiplicative case) (Chatfield & Yar, 2010)

The starting- and smoothing values can be estimated in several ways that we describe in Chapter 4.

2.3.8 Conclusion

Concluding, there are many versions of exponential smoothing that are able to cope with either trend, seasonality, or both. The method that is used most in literature for forecasting electricity demand is the Holt-Winters method or a variant of this method. It is used because it is a relatively simple method in terms of model identification that gives surprisingly accurate results. On the other hand, when implemented for series with a long seasonal period (e.g. yearly seasonality with daily or weekly data), many parameters must be estimated which makes the model unstable. Important is, however, to determine which method suits the data best in terms of trend and seasonality. Both patterns could be additive, multiplicative, or neither of those. It is important to try different methods in order to find which is most suitable for the specific dataset.

2.4 Intermittent demand

We show in Section 3.3 that some storages show what is called *intermittent demand*: demand that occurs sporadically, with some time periods showing no demand at all. When demand does occur, the size could be constant or (highly) variable (Teunter, Syntetos, & Babai, 2011). Therefore, variability does not only occur in demand size, but also in the inter-arrival times. Items that show intermittent demand usually are slow movers. When forecasting intermittent demand with traditional forecasting methods as simple exponential smoothing or simple moving average, the fact that intermittent demand patterns are built from two elements: demand size and demand probability (or demand interval) is ignored, which makes those methods unsuitable (Teunter et al., 2011).

2.4.1 Croston's method

A widely used method is Croston's method that differentiate between these two elements by updating demand size (s_t) and interval (i_t) separately after each period with positive demand using exponential smoothing (Teunter et al., 2011). The notation is as follows (Pennings, Van Dalen, & Van der Laan, 2017):

$$\hat{s}_{t+1|t} = \begin{cases} \hat{s}_{t|t-1} & \text{if } s_t = 0 \\ \hat{s}_{t|t-1} + \alpha(s_t - \hat{s}_{t|t-1}) & \text{if } s_t > 0 \end{cases}$$

$$\hat{i}_{t+1|t} = \begin{cases} \hat{i}_{t|t-1} & \text{if } s_t = 0 \\ \hat{i}_{t|t-1} + \beta(i_t - \hat{i}_{t|t-1}) & \text{if } s_t > 0 \end{cases}$$

and demand forecasts follow from the combination of the previous two forecasts:

$$\hat{d}_{t+1|t} = \frac{\hat{s}_{t+1|t}}{\hat{i}_{t+1|t}} \tag{2.30}$$

where
$\hat{d}_{t+1|t}$ is the demand forecast for next period $(t+1)$
$\hat{s}_{t+1|t}$ is the demand size forecast for next period
$\hat{i}_{t+1|t}$ is the interval forecast for next period
α and β are the smoothing constants, $0 \leq \alpha, \beta \leq 1$

2.4.2 SBA method

However, Syntetos & Boylan (2001) pointed out that Croston's method is biased since $E(d_t) = E(s_t/i_t) \neq E(s_t)/E(i_t)$. A well supported adjustment is the SBA method which incorporates the bias approximation to overcome this problem. Equation 2.30 is adjusted to:

$$\hat{d}_{t+1|t} = \left(1 - \frac{\beta}{2}\right) \frac{\hat{s}_{t+1|t}}{\hat{i}_{t+1|t}} \tag{2.31}$$

where β is the smoothing constant used for updating the intervals.

However, as Teunter et al. (2011) point out, some bias remains with this adjustment, indeed there are cases where the SBA method is more biased than the original Croston method. Besides, the factor $(1 - \beta/2)$ makes the method less intuitive which may hinder implementation. Another disadvantage of both the Croston method as SBA is that the forecast is only updated after demand has taken place. When no demand occurs for a very long period of time, the forecast remains the same which might not be realistic (Teunter et al., 2011).

2.4.3 TSB method

Teunter et al. (2011) proposes not to update the inter-arrival time but the probability that demand occurs ($\hat{p}$). Therefore, the equations change to:

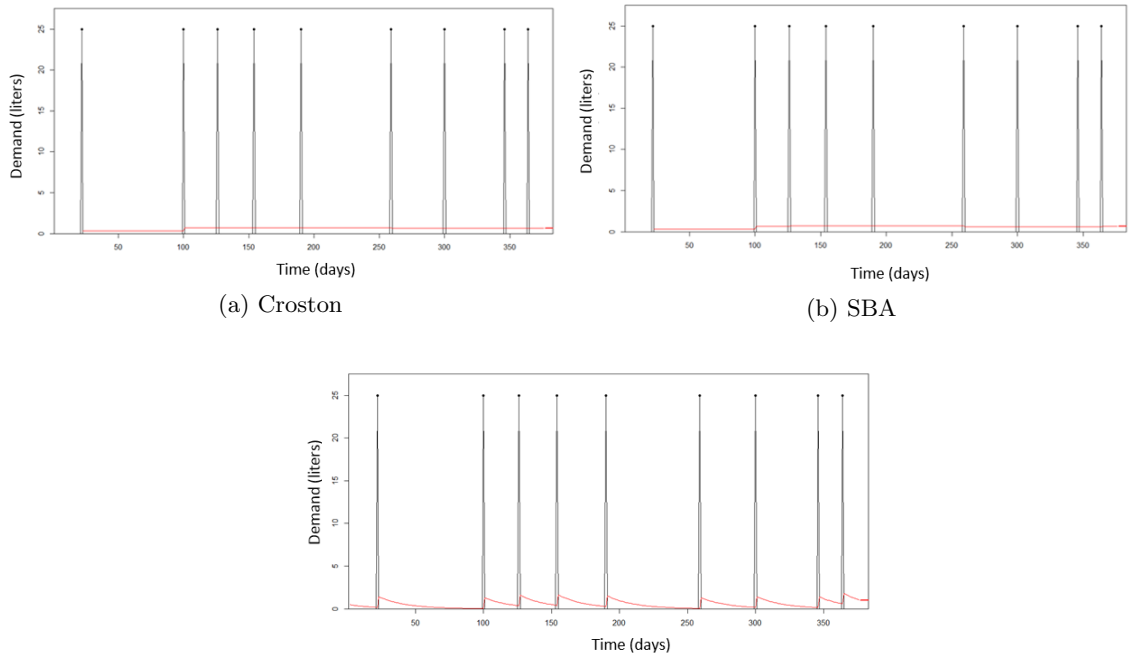
$$\hat{s}_{t+1|t} = \begin{cases} \hat{s}_{t|t-1} & \text{if } s_t = 0 \\ \hat{s}_{t|t-1} + \alpha(s_t - \hat{s}_{t|t-1}) & \text{if } s_t > 0 \end{cases}$$

$$\hat{p}_{t+1|t} = \begin{cases} (1-\beta)\hat{p}_{t|t-1} & \text{if } s_t = 0 \\ (1-\beta)\hat{p}_{t|t-1} + \beta & \text{if } s_t > 0 \end{cases}$$

and the forecast becomes:

$$\hat{d}_{t+1|t} = \hat{p}_{t+1|t}\hat{s}_{t+1|t} \tag{2.32}$$

which is the probability that demand occurs multiplied by the predicted demand size. This method reduces the probability that demand occurs each period with zero demand and this probability increases after non-zero demand occurs. The estimate of the probability of occurrence is updated each period and the estimate of the demand size is updated only at the end of a period with positive demand.



(a) Croston

(b) SBA

(c) TSB

Figure 2.2: Forecasting intermittent demand

Figures 2.2a, 2.2b, and 2.2c show the differences between these methods. The red line represents the forecast. The data used for these figures is from Figure 3.13b but the unjust positive usages are compensated by the negative usages instead of excluding all negative usages. As expected, the forecast provided by SBA (0.533 liter/day) is slightly lower compared to Croston's method (0.676 liter/day). The forecast given by TSB is higher than both (0.972 liter/day) since not so long ago, positive demand occurred.

As the TSB figure shows, the forecast decreases each period since the probability of demand occurrence is updated each period. When no demand occurs for many periods, the probability of demand occurrence becomes zero. Neither of the methods is able to forecast exactly when no demand occurs but gives the forecast of average demand. In the LPG case, in reality this is quite realistic compared to real intermittent demand which occurs in spare parts inventory control of slow moving SKUs (stock keeping units) that really do have sporadic demand. This is because in reality, LPG usage is not zero liters for a couple of

periods and then 25 liters at once but is continuous over time. However, the measurement equipment is not able to measure continuous demand but only measures certain threshold values, for example each percent of the tank. Using these three methods on LPG data must prove which of the three performs best in another situation than the inventory control of spare parts.

2.5 Regression models

In Section 2.2 up and until 2.4, time series methods are discussed. As explained in the introduction, also causal models exist. Causal models assume that the variable to be forecasted (called dependent or response variable) is somehow related to other variables (called predictors or explanatory variables). These relationships take the form of a mathematical model, which can be used to forecast future values of the variable of interest. As mentioned earlier, regression models are one of the best-known causal models (Reid & Sanders, 2005). This section explains two forms: simple linear regression (where one predictor influences the dependent variable) and multiple linear regression (where multiple predictors affect the dependent variable).

2.5.1 Simple linear regression

In the case of simple linear regression, where one predictor or explanatory variable predicts the value of the dependent variable, we are interested in the relationship between these two (X and Y). For example, a shopkeeper might be interested in the effect that the area of the shop (predictor, X) has on sales (dependent, Y) or an employer in the effect that age (predictor, X) has on absenteeism (dependent, Y). This is called a bivariate relationship (Hoshmand, 2009). The simplest model of the relationship between variable X and Y is a straight line, a so called linear relationship. This can both be used to determine if there is a relationship between both variables but also to forecast the value of Y for a given value of X. Such a linear relation can be written as follows:

$$Y = a + bX + \varepsilon \tag{2.33}$$

where
Y is the dependent variable
X is the predictor (independent variable)
a is the regression constant, which is the Y intercept
b is the regression coefficient, in other words the slope of the regression line
ε is the error term (a random variable with mean zero and a standard deviation of σ)

The biggest advantage of this method is its simplicity. However, it is only successful if there is a clear linear relationship between X and Y. An indicator for this linear relationship is the *coefficient of determination* (R^2) which is the fraction of the explained sum of squares of the total sum of squares. This is a statistical measure on how well the regression line approximates the real data.

The simple regression model cannot always be used. The model is based on some assumptions that must be met before being able to use regression:

- Normality
- Linearity

- Homoscedasticity
- Independence of errors

Normality requires the errors to be normally distributed. As discussed before, linearity can be checked by the coefficient of determination R^2 which is the squared correlation coefficient. Homoscedasticity requires the error variance to be constant. This means that when residuals are plot in a scatter plot, no clear pattern should be visible. Independence means that each error ε_t should be independent for each value of X (i.e. the residuals may not have autocorrelation). The Durbin-Watson test checks for this auto-correlation of the residuals.

2.5.2 Multiple linear regression

Many dependent variables do not merely depend on one predictor. In this case, multiple regression can be used for forecasting purposes where one dependent variable is predicted by various explanatory variables. In this way, compared to simple regression, it allows to include more information in the model (Hoshmand, 2009). The regression coefficient is quite similar to that of simple regression:

$$Y = a + b_1X_1 + b_2X_2 + ... + b_nX_n + \varepsilon \tag{2.34}$$

where
Y is the dependent variable
$X_1...X_n$ are the predictors (independent variables)
$a, b_1, ..., b_n$ are the partial regression coefficients
ε is the error term (a random variable with mean zero and a standard deviation of σ)

The regression coefficients a, b_1, b_2,...,b_k must be calculated while minimizing the errors between the observations and predictions. This can be done with what is called the normal equation. Let y be the vector of observations, in the case of Company X this is a vector of actual usage. When m observations are available, y is a m-dimensional vector. X is a matrix that contains the values of the explanatory variables. The first column of this matrix contains only ones and the other n columns contain the values of the covariates where n is the number of predictors. X is a $m \times (n + 1)$ matrix. The vector of regression coefficients (β), containing the constant a and coefficients b_1,...,b_n, is calculated as follows:

$$\beta = (X^TX)^{-1}X^Ty \tag{2.35}$$

This equation is not suitable for n greater than 10,000 since inverting a matrix that large is computationally intensive. Since we do not come close to using this much external variables, this method is suitable for this case.

The R^2 is interpreted similarly as with simple regression but now gives the amount of variation that is explained by several explanatory variables instead of one.

For simple regression, several assumptions were mentioned. Violations of those assumptions may present difficulties when using a regression model for forecasting purposes (Hoshmand, 2009). Since we now have to deal with more than one predictor, an extra assumption is added: no multicollinearity which indicates that the different independent variables are not highly correlated.

- Normality of residuals

- Linear dependency between independent variables and dependent variable
- Homoscedasticity
- Independence of errors (no auto-correlation)
- No multicollinearity

We now explain these assumption by using an example. The data that is used for this is an aggregate series of 21 temperature dependent time series that therefore show yearly seasonality. We choose to do this on an aggregate series in order to reduce variability of individual series.

Normality residuals

First it is important to determine whether the residuals are normally distributed. Figure 2.3 shows the P-P plot of the standardized residuals of the regression. When the expected cumulative probability is equal to the observed cumulative probability, normality can be assumed. This seems to be the case here.
According to the Shapiro-Wilk test, we can assume normality since the test statistic (0.991) is close to 1 which indicates that there is high correlation between the dependent variable demand and ideal normal scores. The test is explained in Appendix B.

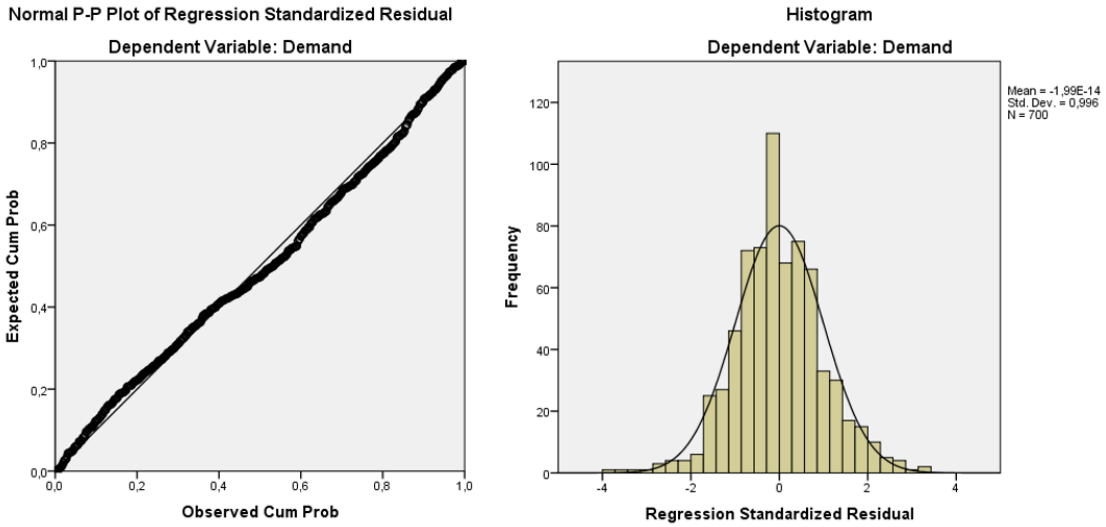


Figure 2.3: Normality of independent variables

Linear dependency between independent variables and dependent variable

Secondly it must be checked whether there is a linear relationship between the independent variables and the dependent variable. This can be checked by making scatter plots. The scatter plots in Figure 2.4 show that the clearest linear relationship is between HDD and demand. Also global radiation shows a linear relation with demand. LPG price, relative humidity, and wind speed show a weak linear relationship with demand.

Homoscedasticity

It is undesirable when the residuals plotted against the dependent variable show a cone-shape. Figure 2.5 shows that there is no cone-shape in the scatter plot of standardised residuals and

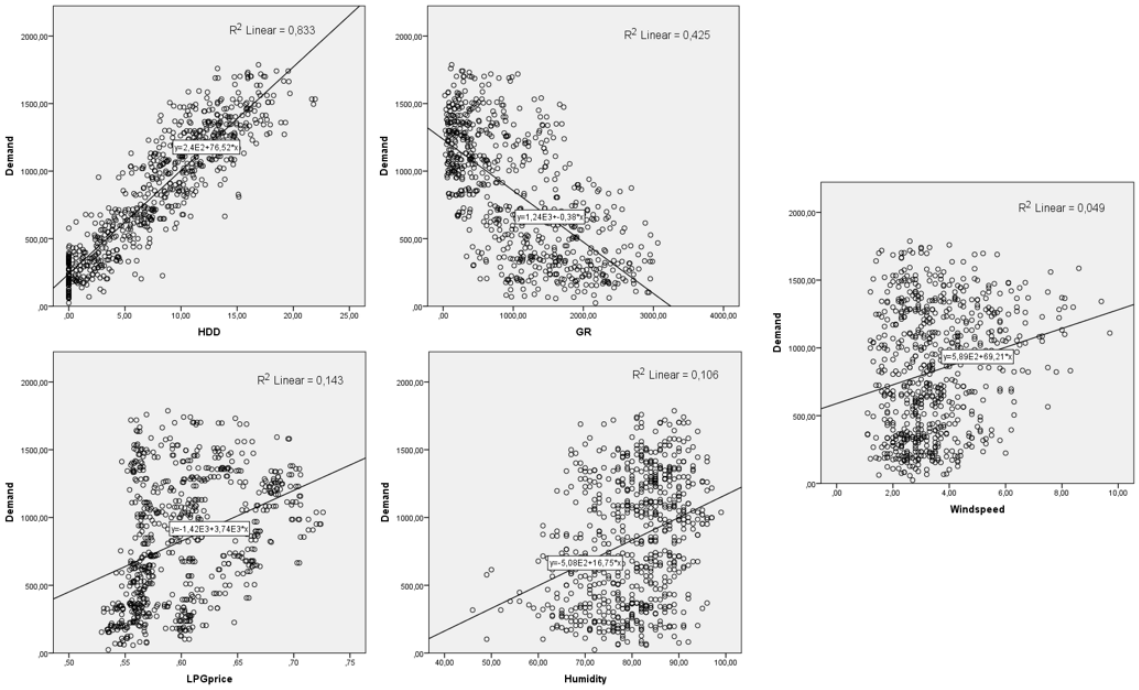


Figure 2.4: Linear relationship between independent variables and dependent variable

LPG demand. When the error terms would have been heteroscedastic, meaning the residuals having different variances, the F-test and other measures that are based on the sum of squares of errors may be invalid (Hoshmand, 2009).

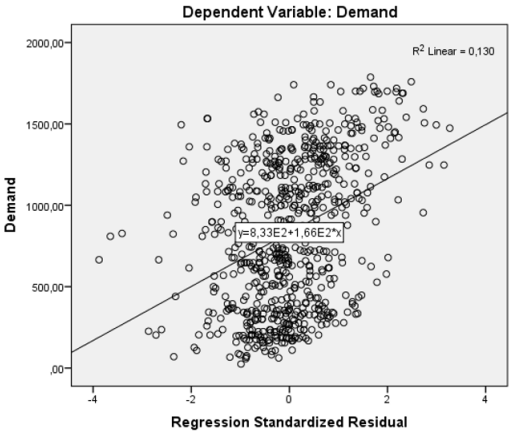


Figure 2.5: Homoscedasticity

Independence of errors

This assumption refers to autocorrelation which means that there is dependence between successive values of the dependent variable Y. This is often present when using time series data since many series move in non-random patterns about the trend. There are two approaches for finding autocorrelation: plot the error terms, and perform the Durbin-Watson test (this test is explained in Appendix B).

Figure 2.6 shows a plot of the residuals. It shows that these are independent. In other words, if the residuals are centred around zero throughout the range of predicted values, they should be unpredictable such that none of the predictive information is in the error. When the latter would have been the case, the chosen predictors are missing some of the predictive information. In our example, $k = 5$ and $n = 700$. The rule of thumb is that the

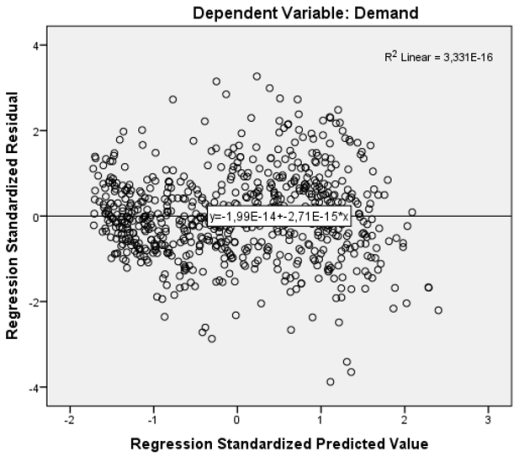


Figure 2.6: Independence of residuals

null hypothesis (the residuals are not autocorrelated) is accepted when $1.5 < d < 2.5$, which is the case here since our d-value is 1.622. However, the critical value table for our k and n values, gives $d_L = 1.864$ and $d_U = 1.887$ which would indicate that our null hypothesis is rejected. It is therefore doubtful whether serious autocorrelation exists in this case.

No multicollinearity

The multicollinearity assumption states that predictor variables may not be highly correlated, in other words, they are independent of each other. It is wise to include the collinearity statistics in SPSS. Severe multicollinearity can result in the coefficient estimates to be very unstable. Multicollinearity implies that the regression model is unable to filter out the effect of each individual explanatory variable on the dependent variable (Hoshmand, 2009). An indicator for this problem is when there is a high R^2 but one or more statistically insignificant estimates of the regression coefficients (a and $b_1, ..., b_k$) are present. This can be solved by simply removing one of the highly correlated variables.
Four criteria must be checked:

- *Bivariate correlations* may not be too high
- *Tolerance* must be smaller than 0.01 ($Tolerance = 1 - R_j^2$ where R_j^2 is the coefficient of determination of predictor j on all the other independent variables)
- *Variance Inflation Factor (VIF)* must be smaller than 10 ($VIF = 1/tolerance$), this would indicate that the variance of a certain estimated coefficient of a predictor is inflated by factor 10 (or higher), because it is highly correlated with at least one of the other predictors in the model
- *Condition indices* must be smaller than 30 (this is calculated by computing the square root of the maximum eigenvalue divided by the minimum eigenvalue which gives an indication of the sensitivity of the computed inverse matrix (that is used in the normal equation, Equation 2.35) to small changes in the original matrix)

When all five independent variables (wind speed, HDD, humidity, global radiation, and LPG price) are forced into the model, the fourth criterion is violated. After removing relative humidity and LPG price from the model, multicollinearity is no problem. Removing those predictors does not jeopardize the R^2 too much, namely from 86.9% to 86.7% which was expected since Hoshmand (2009) mentions that when one or two highly correlated predictors are dropped, the R^2 value will not change much.

A big advantage of regression models is that they can deal with virtually all data patterns (Hoshmand, 2009). Also, it is a relatively easy model when the forecaster wishes to include one or more external variables. The disadvantage is that, as the name indicates, linear regression is only able to cope with linear relationships whereas relationships between variables could be of non-linear nature as well.

2.6 Degree days method

It is broadly agreed upon that the outside air temperature has a large effect on the electricity demand (Kumru & Kumru, 2015; Bermúdez, 2013; García-Díaz & Trull, 2016; Bessec & Fouquau, 2008). Currently, ORTEC uses the degree days method (among others) to include temperature in the LPG forecast. This section explains this method and gives its advantages and shortcomings.

Heating- and Cooling degree days

This method makes a distinction between heating degree days (HDDs) and cooling degree days (CDDs). HDDs come with a base temperature (that should be found by optimising the R^2 when correlating demand with the corresponding HDDs, varying the base temperature) and provide a measure of how many degrees and for how long the outside temperature was below that base temperature (using the average of the minimum- and maximum temperature of a specific day). For example, when the outside air temperature was 3 degrees below the base temperature for 2 days, there would be a total of heating degree days of 6. The advantage of using HDDs over temperature in forecasting is that these HDDs can be aggregated over the time buckets that the user wants to forecast on. CDDs are calculated in a similar fashion, but then the degree days are calculated by taking the number of days and number of degrees that the outside temperature was above that base temperature. This base temperature could be another base temperature than that of HDDs. Moral-Carcedo & Vicéns-Otero (2005) state that it is not trivial whether to use one or two thresholds. Having one threshold indicates that when the threshold temperature is passed, there is a sharp change in behaviour whereas when having two thresholds, it is assumed that in between these two thresholds, there is no appreciable change in demand. In other words, there is a neutral zone for mild temperatures where demand is inelastic to the temperature (Bessec & Fouquau, 2008; Psiloglou, Giannakopoulos, Majithia, & Petrakis, 2009).

In formula form the number of degree days is calculated by:
$HDDs = \sum_{j=1}^{nd} max(0; T^* - t_j)$ and $CDDs = \sum_{j=1}^{nd} max(0; t_j - T^*)$ where nd is the number of days in the period over which the user wants to calculate the number of HDDs, T^* is the threshold temperature of cold or heat, and t_j the observed temperature on day j (Moral-Carcedo & Vicéns-Otero, 2005). With the help of historical data on the energy consumption and number of degree days, a regression analysis can be used to determine the expected energy consumption given the number of degree days.

Base temperature(s)

Several studies indicate that the relationship between demand and temperature is non-linear. This non-linearity refers to the fact that both increases and decreases of temperature, linked to the passing of certain 'threshold' temperatures which we call the base temperature, increase demand. This is caused by the difference between the outdoor- and indoor temperature. When this difference increases, the starting-up of the corresponding heating or cooling equipment immediately raises demand for electricity (Moral-Carcedo & Vicéns-Otero, 2005). The base temperature is the temperature at which electricity demand shows no sensitivity to air temperature (Psiloglou, Giannakopoulos, Majithia, & Petrakis, 2009). The difference between LPG and electricity on this matter is that LPG is primarily used for heating purposes so only one base temperature is required and only HDDs should be considered (Sarak & Satman, 2003). In order to determine this base temperature, the temperature should be plotted against the consumption. This is done in Figure 2.7 for three countries that are categorised as 'warm' (Greece), 'cold' (Sweden), and 'intermediate' (Germany) (Bessec & Fouquau, 2008). The y-axis gives the filtered consumption that isolates the influence of climate on electricity use. We will not go into details because it is of no importance here, the shape of the scatter plot is.

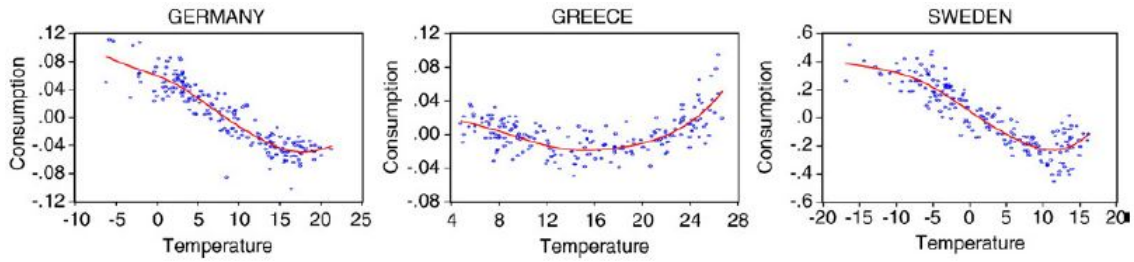


Figure 2.7: Demand versus temperature (Bessec & Fouquau, 2008)

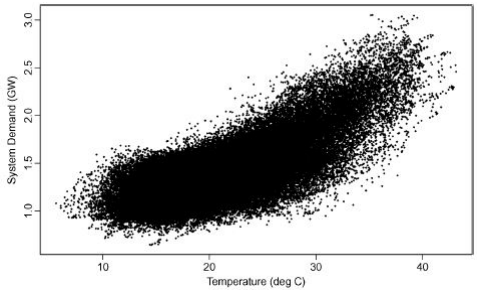


Figure 2.8: Demand versus temperature Australia (Hyndman & Fan, 2010)

A clear U-shape can be seen in the 'warm' country plot which is often seen, also for other warm countries (Moral-Carcedo & Vicéns-Otero, 2005; Pardo, Meneu, & Valor, 2002). However, demand of colder countries is more influenced by the heating effect (Bessec & Fouquau, 2008). Australia, that is even warmer than the countries categorised as 'warm' by Bessec & Fouquau (2008), has a different shape than those in Figure 2.7. Its shape is similar to that of the right part of Greece (see Figure 2.8) which indicates that demand of hot countries is more influenced by the cooling effect.

The zone where demand is inelastic to temperature is around the base temperature. As mentioned before, a decision must be made between one threshold value or two. Having two indicates a temperature interval within demand is unresponsive to temperature variations whereas one indicates a more instant transition between a regime characterised by cold temperatures to a regime corresponding to hot temperatures. Since natural gas (LPG) is used primarily for space heating, using only HDDs is satisfactory which means that only one base temperature is required (Sailor & Muñoz, 1997; Sarak & Satman, 2003).

Shortcomings

A problem of the degree-days method is the determination of an accurate base temperature. In the UK for example, a base temperature of 15.5°C is used since most buildings are heated to 19°C and some heat comes from other sources such as people and equipment in buildings which account for around 3.5°C (Energy Lens, 2016). However, the problem with this is that not all buildings are heated to 19°C, not every building is isolated to the same extent, and average internal heat gain varies from building to building (crowded buildings will have a higher average than a sparsely-filled office with bad isolation and a high ceiling). Energy Lens (2016) states that the base temperature is an important aspect since degree-days-based calculations can be greatly affected by the base temperature used. When the base temperature is chosen wrongly by the forecaster, this can easily lead to misleading results. However, it is difficult to accurately determine whether this base temperature is chosen wrongly since the base temperature can vary over the year depending on the amount of sun, the wind, and patterns of occupancy. Besides, when outside temperature is close to the base temperature, often little or no heating is required. Therefore, degree-days-based calculations are rather inaccurate under such circumstances.

Another important problem is that most buildings are only heated intermittently, for example from 9 to 17 on Monday to Friday for office buildings whereas degree-days cover a continuous time period of 24 hours a day. This means that degree-days often do not give a perfect representation of the outside temperature that is relevant for heating energy consumption. The cold night-time temperatures are fully represented by degree-days whereas they only have a partial effect (when the heating system is off at night), namely on the day-time heating consumption since it takes more energy in the morning to heat the building compared to a less cold night. When the difference between the outside- and inside temperature becomes bigger, as mentioned in Subsection 2.6, the starting-up of the corresponding heating or cooling equipment raises demand for energy (Moral-Carcedo & Vicéns-Otero, 2005). Not only nights are an example but also public holidays and weekends. Moral-Carcedo & Vicéns-Otero (2005) made an adjustment to overcome this problem. They introduce a variable called 'working day effect' which represents the effect of calendar in demand of a particular day as a percentage of electricity demand on a representative day.

There are a couple of suggestions on how to overcome these shortcomings of the degree-days method. The most important one is that an appropriate time scale should be used. In the ORTEC case, the energy consumption is given once every two weeks, degree days should be gained accordingly. For example if only weekly degree days are available, those should be summed in order to make them appropriate. Besides, a good base temperature should be used.

Concluding, the calculations of the degree-days method are rather easy and give fast

results but the combination of the problems stated leads to the overall accuracy of the results being quite low. The results from this method can be used as an approximation of the electricity demand but do not give accurate results.

2.7 Covariates

As described in the introduction, besides time series and causal models, there is also a possibility to combine them. Using covariates results in such a combined model. Taylor (2003; 2010) obtained good results for very short term forecasts in minutes or hours-ahead forecasts (Bermúdez, 2013). However, for short-, medium-, or long-term forecasts, it is important to include covariates in order to cope with for example calendar effects or climatic variables. Besides temperatures, the electricity load is also affected by working days, weekends, feasts, festivals, economic activity, and meteorological variables (Kumru & Kumru, 2015; Bermúdez, 2013; García-Díaz & Trull, 2016). Bermúdez (2013) mentions that unlike in sophisticated methodologies as ARIMA models, in exponential smoothing models the use of covariates is very recent and still infrequent.

According to Bermúdez (2013) there are some authors that use covariates in exponential smoothing models. Wang (2006) proposed to jointly estimate the smoothing parameters and covariate coefficients. The drawback of her method, however, is that she still uses a heuristic procedure to first estimate the initial conditions. Besides, she uses state space models which fall outside the scope of our research. Göb, Lurz, & Pievatolo (2013) refine this model by including multiple seasonalities. Just like Wang's (2006) method, this method requires quite some mathematical skill. Another article that uses covariates in exponential smoothing is that from Hyndman et al. (2008) which introduces covariates into exponential smoothing models when they are expressed as state space models.

A method that is a bit easier, is that from Bermúdez (2013). He adds covariates for the endogenous- and exogenous effects features of which endogenous are seasonal components and exogenous are calendar effects. He does this by adjusting the Holt-Winters model by:

$$A_t = \alpha(Y_t - \omega q_t - I_{t-s}) + (1-\alpha)(A_{t-1} + T_{t-1}) \tag{2.36}$$

$$I_t = \delta(Y_t - \omega q_t - A_{t-1} - T_{t-1}) + (1-\delta)I_{t-s} \tag{2.37}$$

And the forecast now becomes

$$F_{t+x} = \omega q_{t+x} + A_t + xT_t + I_{t-s+x} \tag{2.38}$$

where F_{t+x} is the forecasted value x periods into the future and Y_t is the observed value. ω is the unknown coefficient of the covariate and has to be estimated together with the other unknowns. The equation for trend T_t remains the same. Also a generalization to more than one explanatory variable is also given, ωy_t should be substituted by the linear combination of the number of covariates $\omega' y_t$ where y_t now is a vector with the values of all the covariates at time t and ω the vector of their unknown coefficients. ω cannot be estimated in a similar way as α and δ which values vary between 0 and 1. ω however, can take all kind of values. For this reason, good starting points must be found for these. All the unknowns, so the initial values, the smoothing parameters, and the covariate coefficients, can be jointly estimated by minimising the RMSE in the Excel spreadsheet (Bermúdez, 2013). However, this is a difficult problem to solve since it is a nonlinear optimisation problem. Bermúdez sent me the Excel sheet used for his article and explained that sadly the Excel Solver only looks around the

starting values of the parameters and starting values. Therefore, different starting points should be used. Even then, the method does not give the optimal result that it could have done when estimating the unknowns in a better way. His R-script that is able to determine the optimal parameters and starting values, was not finished.

Bermúdez (2013) states that if too many covariates are introduced into the model, the data could be over-fitted, which produces poor forecasts. Therefore, the first step in the statistical analysis should be the selection of a model which means that a selection of covariates to be used in the analysis should be made.

Concluding, there are some methods that use covariates in exponential smoothing. However, these methods are either relatively difficult because of the fact that the authors use state space modelling which falls outside the scope of our research, or parameter- and initial conditions estimates that are based on heuristics or trial and error.

2.8 Artificial Neural Networks (ANN)

We discussed time series-, causal models, and a combination of the two. Artificial Neural Networks however, can provide both causal- and time series models. ANNs are mathematical tools originally inspired by the way our human brain processes information (Hippert, Pedreira, & Souza, 2001). Just like in a human neuron, the artificial neuron shown in Figure 2.9 receives signals through its dendrites which are the input nodes $x_1, ..., x_4$. Remember

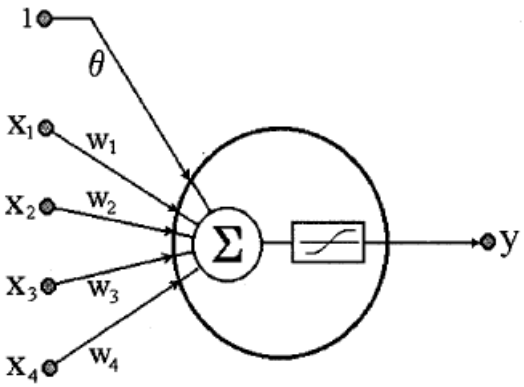


Figure 2.9: Neuron (Hippert, Pedreira, & Souza, 2001)

the multiple regression function (Equation 2.34) and compare this to Figure 2.9. Note that the weights $w_1, ..., w_4$ correspond with the regression coefficients $b_1, ..., b_k$. The information processing of a neuron takes place in two stages: the first is the linear combination of the input values which is in essence the following equation:

$$Y = \theta + w_1 X_1 + w_2 X_2 + ... + w_k X_k \tag{2.39}$$

The second stage is that the result from this linear combination is used as argument of a non-linear activation function. An activation function must be non-decreasing and differentiable, because of which often the identity function ($y = x$) or a sigmoid (logistic) function ($y = 1/(1 + e^{-x})$) is used. When the result of the activation function is above a certain threshold value, it fires. 'Firing' means that the output of this neuron is passed forward to the next layer. The role of an activation function is to make the model nonlinear. One could say that

activation functions answer the following question: 'Some of the input switches are turned on, shall we turn on the output switch?'. In essence, an activation function determines which external variables should be taken into account. Remember that such an activation function is like multiple linear regression, but due to the nonlinear activation function, ANNs are able to model nonlinear relationships as well. Figure 2.10 shows what this logistic activation function looks like. The idea is that such an activation function is binary (0 or 1), it either fires or not.

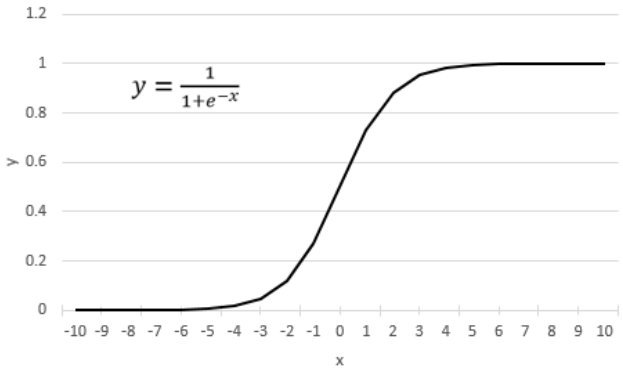


Figure 2.10: Sigmoid (logistic) activation function

The dataset is divided into an estimation, or in this case training set, and a test set. The training set is used for estimating the parameters (weights) and the test set is used for validation (Zhang, Patuwo, & Hu, 1998). This is similar to other forecasting methods and is explained in Section 2.10.

Multilayer Perceptron (MLP)

In load forecasting applications, the multilayer perceptron (MLP) architecture is one of the most popular methods, so this section focuses on that (Hippert, Pedreira, & Souza, 2001). In MLP the neurons are organized in layers. In MLP the first or the lowest layer is the layer where external information is received and the last or the highest layer is the layer where the problem solution is obtained (Zhang, Patuwo, & Hu, 1998). If the architecture is feed-forward, the output of one layer is the input of the next layer. The layers between the input layer and the output layer are called hidden layers. Figure 2.11 shows an example of such a network.

The estimation of the parameters (the weights and the bias θ) is called 'training' of the network and is done by minimizing for example the root mean squared error (RMSE, explained in Section 2.10.

In ANNs, some kind of feedback mechanism is required. Imagine playing basketball. When throwing the ball and seeing that it went too much to the left, next time, you remember the last throw and adjust your movements accordingly in the hope of throwing better this time. That is what an ANN does by the feedback process called backpropagation. This compares the output that a network produces with the output it was supposed to produce and uses the difference between them to modify the weights of the connections between the units in the network. This causes the network to learn, by reducing the difference between the actual and intended outcome.

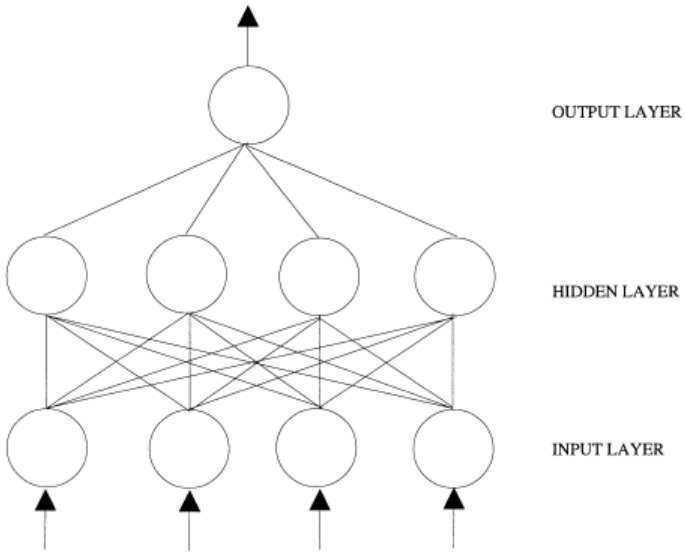


Figure 2.11: Feed-forward neural network (Zhang, Patuwo, & Hu, 1998)

ANNs have several advantages. Firstly, ANNs are non-linear whereas for example regression discussed in Section 2.5 is only able to model linear relationships. Secondly, it has been shown that a network can approximate any continuous function to a desired accuracy (Zhang, Patuwo, & Hu, 1998). This makes ANNs more flexible than the traditional statistical methods described earlier.

Forecasting with ANNs

As stated before, ANNs are able to solve both causal forecasting problems and time series forecasting problems. In the first situation, the vectors of the explanatory variables are the input layer of the ANN. This makes the network quite similar to linear regression except that the network can model non-linear relationships between the explanatory variables and the dependent variable. The relationship estimated by the ANN can be written as $y = f(x_1, x_2, ..., x_n)$ where $x_1, x_2, ..., x_n$ are n explanatory variables and y is the dependent variable. On the other hand, for the time series forecasting problem, instead of having explanatory variables as input, we use past observations. This is written as $y_{t+1} = f(y_t, y_{t-1}, ..., y_{t-p})$. Since the number of input nodes is not restricted, it is also possible to both include past observations and explanatory variables in the model.

Several authors have found satisfactory results when using MLP for predicting LPG usage (Szoplik, 2015; Zhang et al., 1998; Kaytez, Cengiz Taplamacioglu, Cam, & Hardalac, 2015). However, Zhang & Qi (2005) found that neural networks are not able to capture seasonal or trend variation effectively when using ANNs for time series forecasting (not causal). There are some more disadvantages. Firstly, ANNs are black-box methods, there is no explicit form to explain and analyse the relationship between input and output (i.e. what happens inside the black-box). This makes interpreting the results difficult. Secondly, since many parameters must be estimated in ANNs, they tend to overfit. Thirdly, there are no structured methods to identify what network structure can best approximate the function, mapping the inputs and outputs. This is usually done through trial-and-error which can be very time-consuming (Zhang et al., 1998).

Nevertheless, ANNs are appealing because of their ability to model an unspecified non-linear relationship between LPG usage and external variables. This is especially appealing since we found that temperature is an external variable that strongly affects LPG usage. We do not know for sure that this relationship is linear and besides, other covariates could have some influence on LPG usage as well and the relationship (linear or non-linear) is not known yet. Therefore, despite the shortcomings of ANNs, they are definitely worth trying.

There is, however, one big condition for using ANNs: there must be plenty of data. And with plenty of data, think of thousands of observations. ANNs need this to discover patterns correctly without overfitting. A neural network is not a magic black box: it cannot discover a pattern that is not even there. Besides, when the forecaster uses a forecasting method, this choice is based on a certain amount of background knowledge. For example, when choosing Holt-Winters, this choice is based on the idea that the data has some kind of trend and seasonality. When using a neural network, the network has no such background knowledge at all; the network has all the freedom and has no restrictions which makes it extremely difficult to find patterns.

2.9 Combining forecast methods

Often, two or more forecasts are made of the same series in order to decide which one performs best. Bates & Granger (1969) argue that the discarded forecast often contains useful information. Firstly, one forecast is based on variables or information that the other forecast has not considered, and secondly, the forecast is based on different assumptions about the form of the relationship between the variables. There is, however, one condition that both forecast should meet: they should be unbiased. A forecast that consistently overestimates, if combined with an unbiased forecast, leads to biased forecasts: the combined forecast would have errors larger than the unbiased forecast (Bates & Granger, 1969).
The forecasts could be combined by averaging the forecasts but could also be combined using weights:

$$F_t^{combined} = wF_t^{method1} + (1-w)F_t^{method2} \tag{2.40}$$

where w is the weight given to forecasting method 1 and $(1-w)$ the weight given to method 2. Another method of combining forecasts is by using multiple regression with the individual forecast methods as input and the observed demand time series as output. This leads to the weights no longer being constrained to add to one (Deutch, Granger, & Teräsvirta, 1994). The advantages of this method (simple linear combining method) are that it is easy to implement and that it often yields a better forecast than either of the individual methods (Deutch, Granger, & Teräsvirta, 1994).

Hibon & Evgeniou (2005) concluded that the worst performance among individual methods is significantly worse than the worst performance among the combinations. Also they found that the risk (i.e. the difference in post-sample performance between the selected and the best possible) is smaller for choosing a combination instead of an individual method. So, combining forecasts seems to be a good option. However, combining more and more methods seems to worsen the performance (Hibon & Evgeniou, 2005).

2.10 Forecast performance

In previous sections, many statements on which method performs better have been made. But how can different methods be compared on performance and accuracy?

Estimation and validation

The accuracy of a forecasting method is often checked by forecasting for recent periods of which the actual values are known (Hanke & Reitsch, 1998). Data can be held out for estimation validation and for forecasting accuracy. The data that are not held out, are used for parameter estimation (for example the α and β). The model with this parameters is then tested on the data that is withheld for the validation period. When those results are satisfactory, the forecasts for the moments in the future (of which no values are known yet) (Kuchru, 2009). Figure 2.12 visualises this estimation- and validation periods.

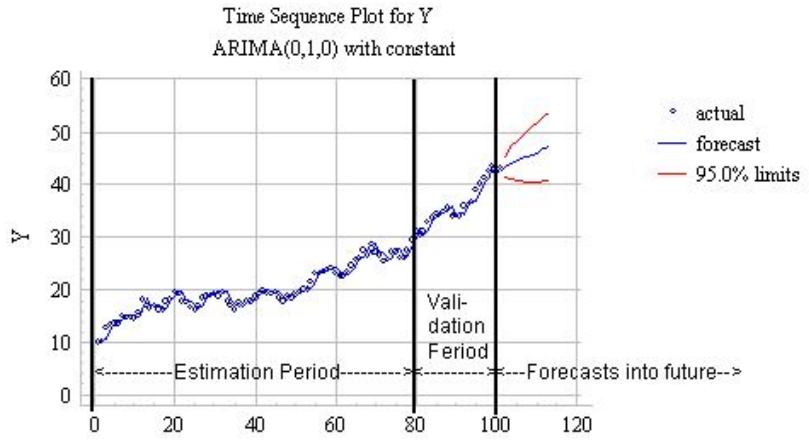


Figure 2.12: Estimation- and Validation period (Nau, 2017)

Withholding data for validation purposes is one of the best indications of the accuracy of the model for forecasting the future. At least 20 percent of the data should be held out for validation purposes (Kuchru, 2009). Normally, a 1-step ahead forecast is computed in the estimation period and an n-step ahead forecast in the validation period. However, in our research, we do this a bit differently. We compute a 1-step ahead forecast in the validation period with what is called a rolling horizon which means that after each 1-step ahead prediction, we do as if the information for that period becomes available (as if the estimation period becomes one step longer and the validation period one step shorter).

Performance indicators

There are several methods that calculate the accuracy of a forecast. Let us define $Y_t - F_t$ as e_t, called the one-step-ahead forecast error. When comparing forecasts on a single series, several common methods could be used; the mean absolute deviation (MAD), the root mean squared error (RMSE), the mean absolute percentage error (MAPE), and the mean error (ME).

- Mean absolute deviation (MAD) $= mean(|e_t|)$
- Root mean squared error (RMSE) $\sqrt{\frac{1}{n}\sum_{t=1}^{n}(e_t)^2}$

- Mean absolute percentage error (MAPE) = $mean(|p_t|)$ where $p_t = 100e_t/Y_t$
- Mean error (bias)

Hyndman et al. (2008), Gardner (1985), Price & Sharp (1986), Taylor (2003) and many others use either the MAPE or MSE or both as accuracy measures. A good model should have small errors in both estimation and validation periods and its statistics in both periods should be similar (Kuchru, 2009). By using the MAPE, the positive and negative errors cancel each other out. The RMSE is more accurate for that reason since it squares the errors and therefore does not let the positive and negative errors cancel each other out. A disadvantage is that the RMSE is scale-dependent (which means that for example an error of one is way worse on an actual observation of two compared to an actual observation of a thousand) so it can only be used to compare forecast performance of different methods on the same time series. The MAPE is similar to the MAD except that it is expressed in percentage terms (Hoshmand, 2009). The advantage of this is that it takes into account the relative size of the error to actual observations. The MAPE also comes with a big disadvantage: it is scale sensitive. Since the actual observation is in the denominator of the equation, the MAPE is not defined when actual usage was zero. Besides, when actual usage is low, the MAPE can take extreme values. Therefore, the MAPE should not be used for low-volume data.

Those error measures are used in three different ways: firstly, for comparison of the accuracy of two different methods. Secondly, to find out whether a method is useful or reliable. Thirdly, it is used to select the optimal technique (Hanke & Reitsch, 1998).

However, errors are not the only aspect to take into account. The choice of model should also be based on the *principle of parsimony* which states that, other things being equal, simple methods are preferable to complex ones (Hoshmand, 2009).

Tracking signal

When a forecast model is chosen, it is important to monitor whether the system remains in control (Trigg, 1964; Gardner, 1983). For example, when SES is chosen, but after a while, a trend appears in the series, the user might want to change the forecasting model or change the value of the parameter(s). In other words, we want to monitor whether biased errors occur. A widely used method for this is to compute a *tracking signal* (Trigg, 1964). The updating equations are as follows:

$$\text{Smoothed error}_t = (1-\alpha)\text{Smoothed error}_{t-1} + \alpha e_t \tag{2.41}$$

$$MAD_t = (1-\alpha)MAD_{t-1} + \alpha|e_t| \tag{2.42}$$

where e_t is the error at period t. The tracking signal is computed as follows:

$$TS_t = \text{Smoothed error}_t / MAD_t \tag{2.43}$$

If the system is so much out of control that all errors have the same sign, this tracking signal will approach plus or minus one (Trigg, 1964). Both Trigg (1964) and Gardner (1983) advice to use $\alpha = 0.1$ since at higher values of α, the performance of the smoothed error signal deteriorates badly. For $\alpha = 0.1$, Trigg (1964) proposes limits of ± 0.55. As long as the tracking signal is between these limits, the system is in control. However, when it is outside these limits, updating is advisable (where updating means either using another model or

updating the parameters).

To give an example, Figure 2.13 shows the forecast made by the degree-days method for a certain time series (Storage 6 in Appendix G) and the corresponding tracking signal. We see that, except for the first couple of observations (the tracking signal needs some warming up, since in the beginning the smoothed error is equal to the MAD), the forecasting system remains in control. What we see is that in the 79^{th} week until the 82^{nd} week, the forecast is structurally below the actual demand which is visible in the tracking signal figure in the sense that the tracking signal value rises in this period and almost hits the control limit.

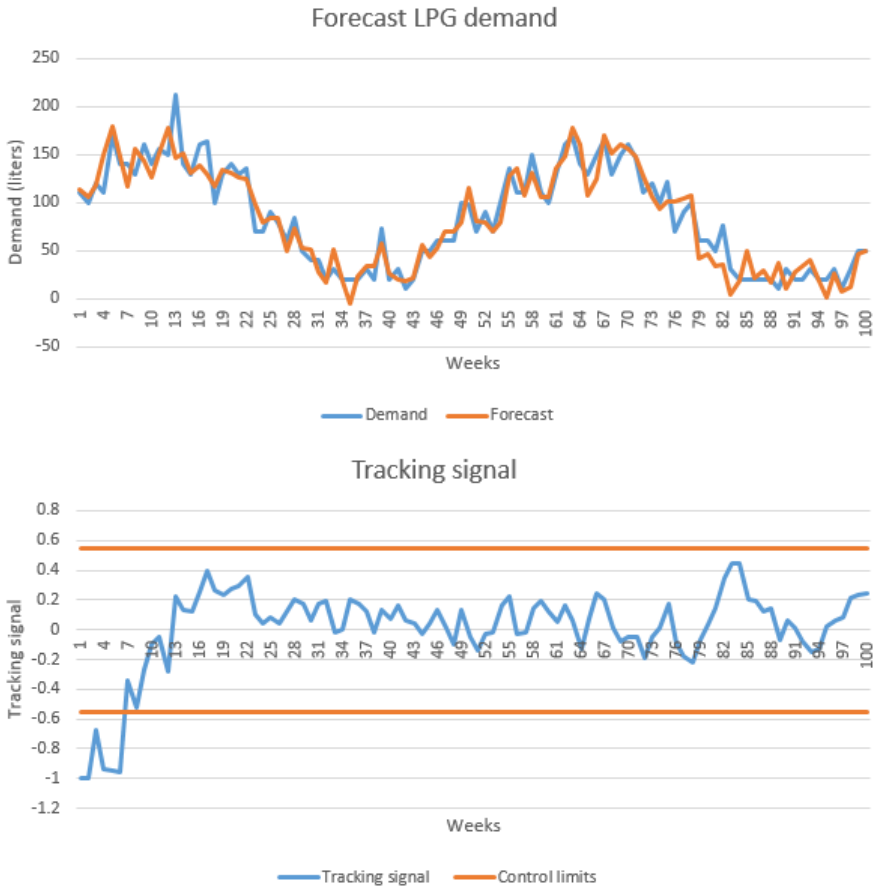


Figure 2.13: Estimation- and Validation period (Nau, 2017)

2.11 Sample size

In Chapter 3, we will discuss that several categories of data occur at customers of Company X, some with many measurements and some with only few. What is the minimum sample size that is required for forecasting models? The required sample size is dependent on at least two things: the number of model coefficients (parameters) to estimate, and the amount of randomness in the data (Hyndman & Kostenko, 2007). For example when considering simple linear regression in Equation 2.33, there are two parameters: a and b. So theoretically, $m+1$ observations, in this case three, are required since it is always necessary to have more observations than parameters (Hyndman & Kostenko, 2007). In the Holt-Winters method for

example, three parameter types for the level, trend, and seasonality are estimated. Besides that, the starting values for level, trend, and seasonality also have to be estimated. The number of starting values for seasonality depends on the seasons per year that are considered. In general, for data with m seasons per year, there are $m + 1$ initial values and three smoothing parameters. Consequently, $m+5$ observations is the theoretical minimum number for estimation (Hyndman & Kostenko, 2007). This is only true however, when little variation is present in the data.

When a lot of variation is present in the data, a lot of data is required in order to accurately estimate a model. When little variation is present, only a few observations are enough. Hanke & Reitsch (1998) give tables with universal minimum data requirements for different forecasting techniques. These are misleading since they ignore the effect variability of the data has on requirements. A way to explain this is with the help of prediction intervals (margins for error) around the point forecasts. The number of observations mentioned earlier are the sample sizes for which the prediction intervals are finite. As sample size increases, the prediction intervals decrease at a rate proportional to the square root of the number of observations. Therefore, by quadrupling the sample size, the prediction intervals are cut in half (Hyndman & Kostenko, 2007).

Concluding, there is no straightforward answer. The only certainty is that it is always necessary to have more observations than parameters. In reality however, a lot of random variation is present in data of practical applications, it is usually necessary to have many more observations than parameters (Hyndman & Kostenko, 2007)

2.12 Classification

Classification of information is an important component of business decision-making tasks (Kiang, 2003). In our case, we would like to classify on forecasting method. A classic task that can be easily formulated as a classification problem is pattern recognition. Different patterns require different forecasting methods. Classification could be useful for the Company X case, since currently a forecasting script must be chosen manually for thousands of clients. Automating this by classification techniques, saves a lot of time. Besides, classification could get insight into the types of data we have to deal with which would be impossible to do manually since there are thousands of datasets.

Classification is quite similar to regression. The biggest difference between regression and classification is that the outcomes are continuous and discrete, respectively (Tan, 2006). Figure 2.14 visualises the process of classification.

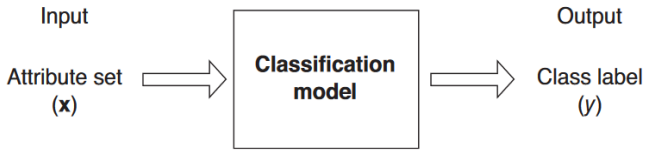


Figure 2.14: Classification maps each attribute set x to one of the predefined categories y (Tan, 2006)

Such a model is useful for generally two purposes: descriptive- and predictive modelling. It

can serve as an explanatory tool to categorise objects of different classes. This could give insight in which features define a time series as being Category 1, Category 2, Category 3, or Category 4. Predictive modelling, however, is more useful in our case. A model can be used to predict the class of unknown records (Tan, 2006). In other words, when new storages or customers come in, the model is able to automatically detect which category, and by that, which forecasting method fits the data.

The general approach for solving classification problems is as follows: the user provides a *training set* that consists of records of which the classes are known (in Table 2.2, the training set consists of the rows that contain a class in the last column). The model that follows from this training set is then applied to the test set. The *test set* consists of records with unknown classes (the rows that have a question mark in the last column in Table 2.2). Table 2.2 gives an example of a number of features that could be used for classification to give an idea of what this looks like.

Table 2.2: Small sample of the Company X dataset with example features

	Features			
Storage	# timed values	zero's ratio	R^2 temperature	Class label
28128	100	0.03	0.818	Degree-days
2373	81	0.78	0.000	SES
22368	101	0.03	0.038	Degree-days
2949	5	0	0.954	Degree-days
295887	3	0	0.981	?
298153	223	0.43	0.189	?

The following subsections give methods that are able to determine the class of the unknown instances given their features. We discuss how they work, and their advantages and disadvantages.

2.12.1 Decision tree methods

Decision trees are powerful tools for classification and prediction that are becoming increasingly popular (Lahiri, 2006; Delen, 2011). Their advantage is that they represent simple rules that humans can understand, unlike methods as neural networks.

Decision tree

A basic decision tree consists of decision nodes, which can be a root node ('Body temperature' in Figure 2.15), or an internal node ('Gives Birth'), specifying some test to be carried out on a single attribute-value, and leaf nodes that represent the possible classes and are at the bottom of the tree (Mammals or Non-mammals). Let us take the flamingo in Table 2.3 as example. Starting at the root node, the question is whether flamingos have a warm or cold body temperature. Since they have a warm body temperature, we arrive at the 'Gives Birth'-node. Since flamingos do not give birth, we arrive at the leaf node that says that flamingos belong to the 'Mammals' class.

Several algorithms exist for decision trees but one of the most popular is C4.5. We decide to elaborate on that one since that algorithm is used in the J48 WEKA implementation. WEKA is an open source data mining tool that makes it easy to implement many data mining and machine learning techniques. C4.5 builds decision trees using the concept of information

Table 2.3: Unlabelled data

Name	**Body temperature**	**Gives birth**	**...**	**Class**
Flamingo	Warm	No	...	?

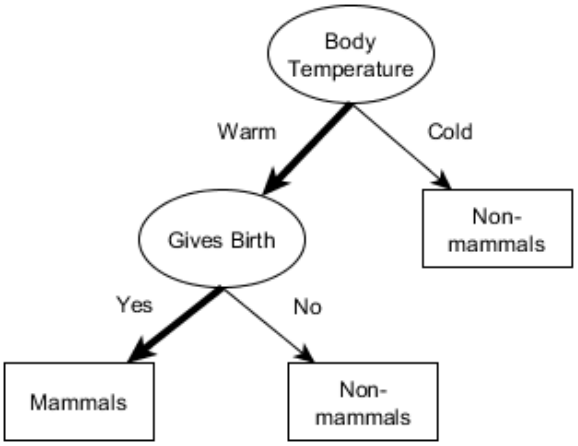


Figure 2.15: Decision tree for the mammal classification problem (Tan, 2006)

gain (difference in entropy).

Let $p(i|t)$ denote the fraction of records belonging to class i at a given node t. Often, we just refer to fraction p_i without referring to a specific node. In a two-class problem, the class distribution at any node can be written as (p_0, p_1) where $p_1 = 1 - p_0$. These distributions can be used to find the best splits, which measures are often based on the degree of impurity of the child nodes. A node with class distribution $(0.5, 0.5)$ is highly impure whereas $(1, 0)$ has zero impurity. The impurity measure that is used in the C4.5 algorithm is entropy which is calculated as follows:

$$Entropy(t) = -\sum_{i=0}^{c-1} p(i|t) \log_2 p(i|t) \tag{2.44}$$

where c is the number of classes. The splitting criterion is the information gain which is the difference in entropy. The characteristic ('Body temperature' or 'Give Birth' in the example from Figure 2.15) with the highest information gain is chosen to make the decision. This is repeated until the entire tree is constructed.

Let us give an example. We have a test set of 14 instances of which 5 are mammal and 9 are not. This gives an entropy of:

$$Entropy(IsMammal) = Entropy(5, 9) = -(0.36 \log_2 0.36) - (0.64 \log_2 0.64) = 0.94 \tag{2.45}$$

To calculate the information gain of an attribute, we subtract the entropy after the split from that before the split. For example, the information gain of the attribute 'Body temperature' is calculated as follows:

$$\begin{aligned} Gain(IsMammal, BodyTemp) &= Entropy(IsMammal) - Entropy(IsMammal, BodyTemp) \\ &= 0.94 - 0.788 = 0.152 \end{aligned} \tag{2.46}$$

For which $Entropy(IsMammal, BodyTemp)$ is computed by:

$$\begin{aligned} Entropy(IsMammal, BodyTemp) &= p(Warm) * Entropy(3,4) + p(Cold) * Entropy(6,1) \\ &= (7/14) * 0.985 + (7/14) * 0.592 = 0.788 \end{aligned} \tag{2.47}$$

Tables 2.4 give the information gain for both the 'Body temperature' and the 'Gives birth' attribute. Here we see that the gain when splitting on 'Body temperature' is a lot higher than when splitting on 'Gives birth'. We choose 'Body temperature' as decision node, since it has the highest information gain. The C4.5 algorithm repeats this process for every branch until at each branch, there is a node with an entropy of 0 (which becomes a leaf node). An entropy of zero indicates that there is no impurity which means that at the leaf node, there are only instances of one class.

Table 2.4: Information gain of the attributes 'Body temperature' and 'Gives birth'

		Mammal	
		Yes	*No*
Body temperature	*Warm*	4	3
	Cold	1	6
Information gain: 0.152			

		Mammal	
		Yes	*No*
Gives birth	*Yes*	3	4
	No	2	5
Information gain: 0.016			

Random forest

A risk of decision tree learning is that too many attributes are included which leads to having the chance that, early in the tree, the tree branches on variables that result in a good information gain, but result in little or no information gain later on. A solution for this problem is feature selection. A widely used technique is random forest. Random forests have been introduced by Breiman (2001). Several authors found that significant improvements in classification have resulted from growing an ensemble of trees and letting them vote for the most popular class. Each tree of the forest is grown as follows:

1. Let N be the number of instances in the *training set*, sample N cases at random (*with* replacement) from the original data
2. If there are M attributes to classify on, a number $m \ll M$ is specified such that at each split, m features are randomly selected out of M and the best split of this subset is used to split the node
3. Each tree is grown to the largest extent possible

In order to classify an unseen instance, it should be ran through each tree in the forest, and when for example 150 out of 200 (the majority) trees classify the instance as being 'Class 1', the instance is classified as being 'Class 1'.

In random forests, the importance of the attributes can be calculated by using the Mean Decreas Accuracy (or permutation importance) (Strobl, Boulesteix, Kneib, Augustin, & Zeileis, 2008). The idea behind this importance measure is that the more the accuracy of the random forest decreases when excluding (or permuting) a single attribute, the more important that variable is regarded.

The advantages of random forests are that they do not overfit due to the Law of Large Numbers and they are quite fast (Breiman, 2001). Besides, generated forests can be saved for future use of other data (Strobl et al., 2008).

2.12.2 *k*-Nearest Neighbour (kNN)

k-Nearest Neighbour is an instance-based learning technique that determines which training instance is closest to an unknown test instance by using a distance function. The class of the nearest training case is predicted for the test case. The distance function is relatively easy to determine when all attributes are numeric (Witten, Frank, Hall, & Pal, 2011).

The distance measure that is most often used is Euclidean (Witten et al., 2011). The distance between two instances with attribute values $a_1^{(1)}, a_2^{(1)}, ..., a_n^{(1)}$ and $a_1^{(2)}, a_2^{(2)}, ..., a_n^{(2)}$ where n is the number of attributes, is defined as:

$$\sqrt{\left(a_1^{(1)} - a_1^{(2)}\right)^2 + \left(a_2^{(1)} - a_2^{(2)}\right)^2 + \cdots + \left(a_n^{(1)} - a_n^{(2)}\right)^2} \tag{2.48}$$

Different attributes are often measured on different scales, so when calculating the distance with the equation above, that uses Euclidean distances, the effect of some attributes might be bigger or smaller than from others. This problem is solved by normalising the attributes values to lie between 0 and 1 by using the following equation:

$$a_i = \frac{v_i - \min\, v_i}{\max\, v_i - \min\, v_i} \tag{2.49}$$

The advantages of instance-based learning are that it is simple and effective. The disadvantage is that it is often slow since for each new instance, it has to calculate the distance with all known instances (Witten et al., 2011).

Let us give an example. We have a set of instances consisting of employees of a certain age, their salary, and the number of vacation days they have. We would like to classify the fifth instance as being satisfied or not, given that the first four instances are already labelled as being either satisfied, or not. We use the normalised attribute values to calculate the distances. The distance between the fifth instance and the first is calculated as follows:

$$\sqrt{\left(0.00 - 0.22\right)^2 + \left(0.00 - 0.22\right)^2 + \left(0.33 - 0.67\right)^2} = 0.454 \tag{2.50}$$

Table 2.5 gives the distances between the instance we want to classify, and the instances with known classes.

Table 2.5: Example k-nearest neighbour (normalised values between brackets)

Instance	**Age**	**Salary**	**Vacation days**	**Satisfied**	**Distance**
1	23 (0.00)	€2,100 (0.00)	25 (0.33)	No	0.454
2	41 (0.56)	€3,500 (0.61)	30 (0.67)	Yes	0.521
3	34 (0.34)	€2,400 (0.13)	20 (0.00)	No	0.684
4	55 (1.00)	€4,400 (1.00)	35 (1.00)	Yes	1.155
5	30 (0.22)	€2,600 (0.22)	30 (0.67)	?	

Now we know the distances, we need to determine the class. When $k = 1$, this means that we

only consider one nearest-neighbour. However, it could also be possible to consider several neighbours and classify the unknown instance as the class that the majority of the considered neighbours have. For example, when $k = 3$, instances 1, 2, and 3 are nearest neighbours of which two are classified as not satisfied and only one as satisfied which leads to instance 5 being classified as being not satisfied. In the training phase, this number of neighbours to consider, can be varied with in order to optimise the accuracy.

2.12.3 Logistic regression

Logistic regression is a generalisation of linear regression (explained in Section 2.5) that is primarily used for predicting binary or multi-class dependent variables (Delen, 2011) when having numeric attributes (Witten, Frank, Hall, & Pal, 2011). The logistic function used to calculate the probability of belonging to class i is (Lahiri, 2006):

$$p_i = \frac{1}{1 + \mathrm{e}^{-y^*}} \tag{2.51}$$

where

$$y^* = \log\frac{p_i}{1 - p_i} = a + b_1 * X_1 + b_2 * X_2 + ... + b_n * X_n \tag{2.52}$$

For classification of instances that belong to more than two classes, multinomial logistic regression is more appropriate (Starkweather & Moske, 2011).

The disadvantage of this method is that the modeller must choose the right independent variables. Besides, this model assumes that the response variable Y is linear in the coefficients of the external variables $X_1, ..., X_n$ (Delen, 2011). The advantage is that the method is interpretable and easy to implement when the regression coefficients are known.

2.12.4 Artificial neural networks

For classification, a multi-layer perceptron (MLP) is often used since it is a strong function approximator for prediction and classification problems (Delen, 2011). Instead of a point forecast, just as with logistic regression, the ANN produces probabilities that a case is assigned to a certain class. The difference with logistic regression is that ANNs do not assume a linear response and it is more of a black box.

2.12.5 Classification performance

The performance of a single classification model is based on the number of test records the model predicts correctly. These counts are summarised in what is called a *confusion matrix*. Table 2.6 shows the general form of a confusion matrix in a 2-class classification problem.

Table 2.6: Confusion matrix of a 2-class problem

		Predicted Class	
		Class = 1	*Class = 0*
Actual Class	*Class = 1*	f_{11}	f_{10}
	Class = 0	f_{01}	f_{00}

f_{ij} denotes the number of instances of class i predicted to be of class j. When $i = j$, the instance is correctly classified and if i is unequal to j, it is not. Therefore, the number of

correctly and incorrectly predicted instances is respectively $f_{11} + f_{00}$ and $f_{10} + f_{01}$. The accuracy is then calculated by the following equation:

$$Accuracy = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} = \frac{f_{11} + f_{00}}{f_{11} + f_{10} + f_{01} + f_{00}} \tag{2.53}$$

And the error rate is $1 - Accuracy$.

Generally, the training set is used to train the model and the test set is used to calculate performance. This means that the records with known classes should be divided into a training- and test set. There are several ways to do this but a widely used method is called *cross-validation*. In this approach, each record is used an equal number of times for training and once for testing. In k-fold cross-validation, the set of records with known classes is divided into k subsets. During each run, one of the subsets is held out as test set and the rest is used for training. This is repeated k times until each subset has been test set once. The total error is calculated by summing op the errors for all k runs. Kohavi (1995) advises to use 10-fold cross-validation.

Adeodato, Vasconcelos, Arnaud, Santos, Cunha, & Monteiro (2004) compared ANNs and logistic regression on a large data set (180,000 examples) and found that MLP performs better than logistic regression. Kiang (2003) tested ANNs, decision tree (using the C4.5 algorithm), logistic regression, and k-nearest neighbour and concludes that it depends on several data characteristics which method performs best, but generally speaking, ANNs and logistic regression are superior to the other methods.

2.13 Conclusion

This chapter tried to answer the question 'What is known in literature on forecasting LPG demand or similar cases?'. Also, we wanted to find out how to measure forecast performance and how classification methods can be used for automatic model selection. Since ORTEC is interested in forecasting using several variables as input, called multivariate forecasting, besides the univariate models, we tried to find methods that are able to incorporate covariates. We conclude the following:

1. Besides the current methodology, the methods that are suitable for temperature dependent data are: Additive- & Multiplicative Holt-Winters, the Holt-Winters damped method, and Simple- & Multiple linear regression
2. For non-seasonal data, simple exponential smoothing and moving average are suitable methods
3. For data that shows an intermittent pattern, suitable methods are: simple exponential smoothing, Croston's method, and the TSB method
4. Combining forecasting methods could improve the accuracy and robustness, but Hibon & Evgeniou (2005) found that combining more and more methods seems to worsen the performance
5. There is a gap in literature on combinations of causal- and time series models: there are some methods that combine these, but are difficult because they use state space modelling which falls outside the scope of this research, or parameter- and initial conditions estimates are based on heuristics or trial and error
6. Causal models that can be used for prediction using external variables are the degree-days method and linear regression
7. It is advisable to compute a tracking signal in order to monitor whether the forecasting system remains in control
8. Suitable classification methods for classifying data patterns are: decision tree, random forest, k-nearest neighbour, ANNs, and logistic regression of which k-nearest neighbour, decision tree, and logistic regression are easiest to interpret whereas the other method are more of a black box

Concluding, there is not always one method that performs best for predicting LPG demand. It is important to look into the data to determine what patterns are present and then decide which methods are suitable for predicting those patterns.

Chapter 3

Current situation

This chapter answers the second research question 'What is the current situation at ORTEC?' by explaining what forecast methodology is currently used for forecasting LPG demand and by finding out whether the assumptions that are made in this methodology are correct. Section 3.1 explains what the data looks like and how it is currently transformed to be suitable for forecasting purposes. Consecutively, the current forecasting procedure is explained in Section 3.2 and the issues that come along are addressed. Section 3.3 describes which data patterns are present in the datasets and how ORTEC currently copes with these. Section 3.4 concludes this chapter and answers the research question.

Let us go back to the LPG case. Company X is a client of ORTEC that is a supplier of liquefied propane gas (LPG) for professional-, agricultural-, and home use. This propane gas is primarily used to heat premises but in this chapter it becomes clear that that is not the only purpose. The clients of Company X generally need replenishment just before running out of gas. In order for Company X to achieve this without having to visit the customers too often, just to be on the safe side, a good forecast is required on when the storages get below their safety stock level. The forecast engine is part of ORTEC Inventory Routing (OIR). Good forecasts are of great importance since the forecast indicates how much LPG should be delivered to each customer. Multiple customers can be visited in a single route. When a route is planned, for each customer a certain order volume is reserved. In reality however, the truck driver fills the tank until it is full. At the end of the planned route, the truck should be exactly emptied out. When the forecasts are inaccurate, two things could happen: the truck is empty before having visited the last customer(s) on its route (which happens in 38% of the trips) or LPG is left after visiting the last customer on the route. When the latter happens, another customer must quickly be found in order to empty the truck anyway. Both should be avoided. Figure 3.1 plots the difference between the planned amount of LPG and the amount that is actually delivered. We see that there is a tendency to deliver more than planned.

3.1 Datasets of storages

Of each storage, a dataset is available. The datasets are of customers from the Netherlands. There are customers in different sectors: camping sites, business to consumer, onion dryers, construction sector, agricultural sector, industrial sector, food and beverage, and other. Figure 3.2 shows a screenshot of some data of a specific customer storage. This contains the readings (stock measurements) and the delivered amounts. The first column gives the date and time of the reading. The second column gives the type of reading: whether the reading

Figure 3.1: Difference between the planned and delivered amount of LPG

Delivery (local)	Status	!	Storage 6764	!	Usage	Usage/hr
6/21/2016 02:00	Reading		2460		9	0.36
6/14/2016 02:00	Reading		2520		4	0.18
6/7/2016 02:00	Reading		2550		0	0
5/31/2016 10:05	Reading after		2550		303	12.62
5/31/2016 10:05	Delivered		882			0
5/31/2016 02:00	Reading		1770		4	0.18
5/24/2016 02:00	Reading		1800		9	0.36
5/17/2016 02:00	Reading		1860		17	0.71
5/10/2016 02:00	Reading		1980		4	0.18
5/3/2016 02:00	Reading		2010		21	0.89
4/26/2016 02:00	Reading		2160		21	0.89
4/19/2016 02:00	Reading		2310		13	0.54
4/12/2016 02:00	Reading		2400		26	1.1
4/6/2016 09:45	Reading after		2550		2	0.06
4/6/2016 09:45	Delivered		1472			0
4/5/2016 02:00	Reading		1080		17	0.71
3/29/2016 02:00	Reading		1200		26	1.07
3/22/2016 01:00	Reading		1380		30	1.25
3/15/2016 01:00	Reading		1590		30	1.25

Figure 3.2: Storage data

was a regular telemetry reading that is done automatically (reading), whether it was a reading after delivery (reading after), or whether it was the amount of LPG that was actually delivered (delivered). The 'usage' column gives the daily usage and the last column gives the usage per hour. What is striking, is that the usage per hour after the delivery on the 31th of May 2016 is way higher than the others (12.62 liters per hour whereas the others are between zero and two liters per hour). After talking with the product managers, it turns out that the truck drivers conduct the reading after and whether the tank is 87% full, 83% or 78%, the truck driver usually fills out 80% full on the form (tanks may not be filled to 100% in order to allow for the volume changes due to changes in temperature). Because of this, the readings after are unreliable.

Besides the table with stock levels, we also have a figure of the tank volume (see Figure 3.3). In this figure, the bottom red line represents the minimum volume and the upper red line the capacity of the tank. A tank may not be totally emptied out. Usually the tank volume may not be lower than about 10% of the tank capacity. The upper pink line represents the maximum to which a tank may be filled, which is usually about 80% and the lower pink line is the safety stock level. The horizontal green line is the average stock. The vertical green line represents today, so the volume after that is the forecast.

Figure 3.4 shows what the series from the data shown in Figure 3.2 looks like when every type of reading is included (green line) and when all measurements that say 'Reading

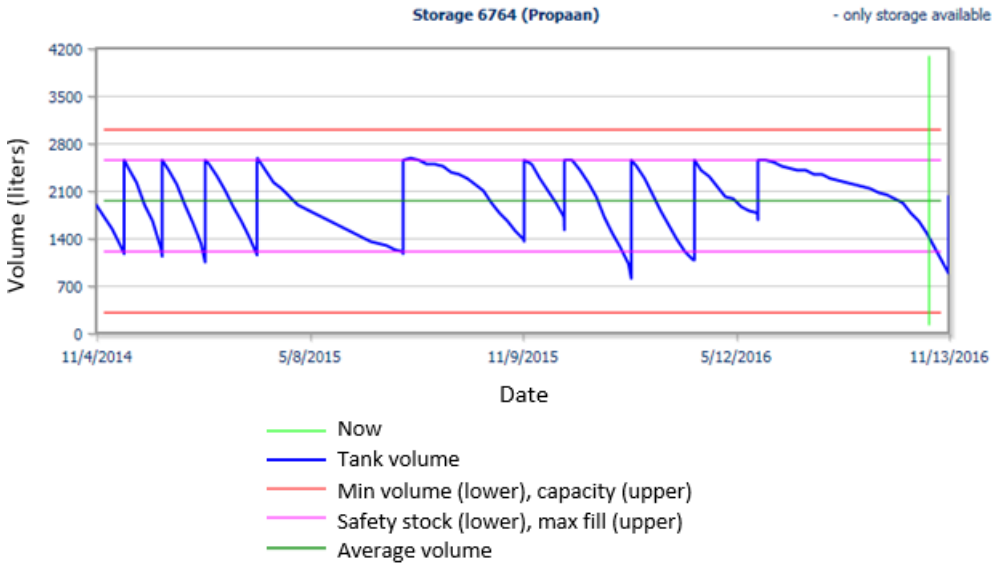


Figure 3.3: Volume tank

after' are excluded (red line). Not only the problem of having peaks is solved by this, also the problem of drops occurring directly after a peak is. Currently, ORTEC forecasts the green line. It would however be more correct to predict the red line. Interesting is that the dependency on temperature doubles when we exclude the readings after (the R^2, which is the coefficient of determination, i.e. the percentage of variation explained by the independent variable(s) between the weekly usage and summed HDDs over weeks changes from 38.3% to 81.8%). Later in this chapter it becomes clear why this is of great importance.

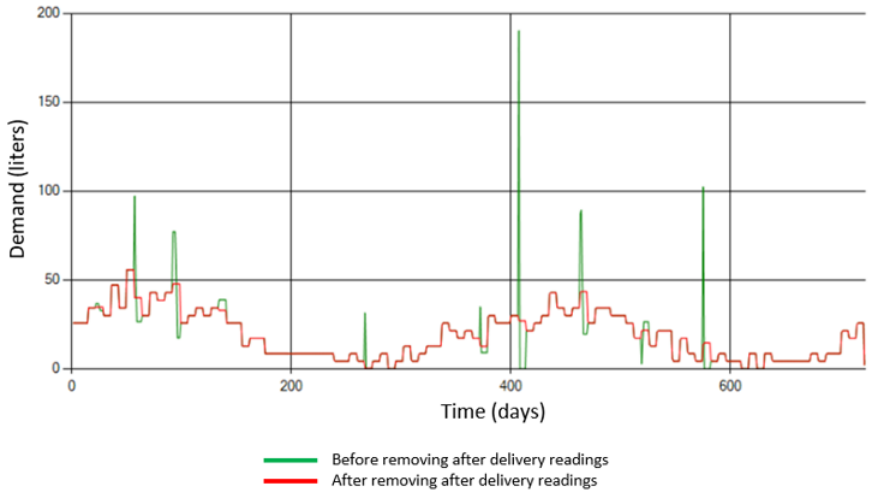


Figure 3.4: Timed value series without 'Reading after'

ORTEC does not forecast storage volume (as shown in Figure 3.3), but they forecast usage (as in Figure 3.4). This usage is not the daily usage in Figure 3.2, but it is calculated differently. In order to do that, the volume measurements are converted to what is called timed values. These have a from- and till date and time and give the usage in between. For example, between the 26^{th} of April and the 3^{rd} of May, the timed value is $2160 - 2010 = 150$. In Figures 3.4 and 3.8 the usage is plotted as if it was constant between every two readings.

These time valued series are input for the forecast engine. When delivery has taken place, the usage is calculated by adding the delivered value to the reading before the delivery and subtracting this by the reading after. The daily usage is then calculated by dividing this by the number of days between the from- and till date and time (the time buckets of the forecast are one day). However, the reading after is unreliable which makes the usage after delivery incorrect. Section 3.2 explains how the time valued series are forecasted.

3.2 Current forecasting procedure

Per dataset, the user is able to choose a forecasting script. There are generally three scripts that can be chosen: simple exponential smoothing, simple exponential smoothing with period 7 days where next Monday is smoothed with last Monday and the Monday before and so on (i.e. this version is able to catch within-week variation, for example when on weekdays the usage is higher than in the weekends), or the degree-days method. SES with period 7 days is computed by making sub series of all Mondays, all Tuesdays, and so on, and performing simple exponential smoothing on those. The idea is that the degree-days methods performs better on temperature dependent datasets. Since that is not entirely sure, the degree-days method has a backup method which is the yearly script. Since the degree-days method is based on the relationship between temperature and usage, we first elaborate on this temperature dependency.

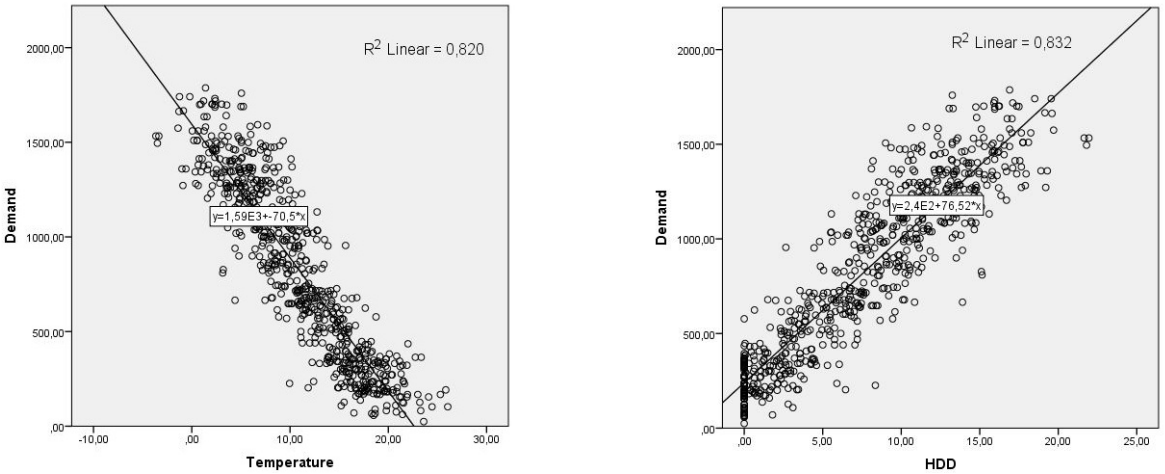
In the exponential smoothing functions that we gave in Chapter 2, no period is present. What is meant by period 7 is that exponential smoothing is performed on all Mondays, Tuesdays, and so on, separately. This is beneficial when a time series shows a within week pattern that other wise cannot be included in SES. However, for the Company X datasets, such a within week pattern never exists or we have not enough data to find that pattern (when we have weekly data).

3.2.1 Dependency on temperature

As is broadly stated in literature, temperature is a major determinant of electricity consumption (Bessec & Fouquau, 2008; Thornton, Hoskins, & Scaife, 2016; Moral-Carcedo & Vicéns-Otero, 2005). Literature on natural gas usage is scarce but there is evidence for the two commodities (electricity and LPG) to have a similar dependence on temperature and on the number of HDDs. To check whether this relationships are also present for the Company X cases, we made scatter plots. For this, we aggregated 21 datasets in order to make the data more reliable and less dependent on accidental variations. Per dataset individually, outliers are removed from demand time series and replaced by their upper- or lower bounds which are calculated by: $mean \pm 2 * STD$ for this scatter plot to be more reliable.

The scatter plot in Figure 3.5a shows a negative correlation between temperature and demand: the higher the temperature, the lower the demand. This is sensible since a higher temperature means that less heating is required. We must note here that LPG in this case is primarily used for heating purposes. Air-conditioning uses electricity and not LPG (Sailor & Muñoz, 1997; Sarak & Satman, 2003). With HDDs we see a positive correlation which is also sensible since when there are more HDDs, the temperature has been below a threshold temperature for more days and/or more degrees so more heating is required.

Figure 3.6 plots the temperature and aggregated demand (the same data are used as



(a) Temperature

(b) Heating degree-days

Figure 3.5: Regression temperature and demand

for the scatter plots in Figures 3.5a and 3.5b). What is interesting is that a peak in air temperature, causes a peak in the opposite direction in demand. Therefore, good weather forecasts are required to forecast LPG demand. For the forecast required to determine how much LPG should be delivered, we can use the realised temperatures which are always reliable. These weather forecasts are imported from the weather institute belonging to the country for which the forecasts are made (KNMI for Dutch weather data, but also customers in for example Australia exist for which other sources are used).

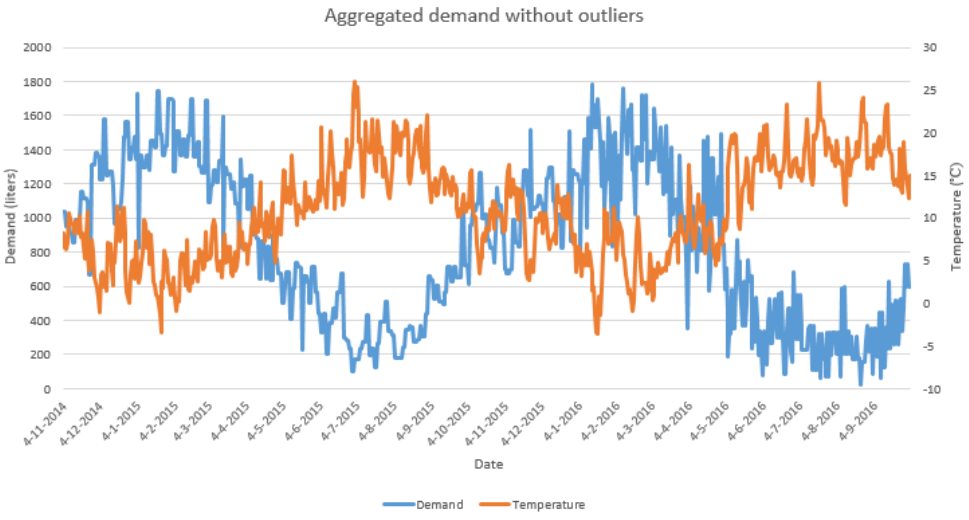


Figure 3.6: Aggregated demand and temperature

What is also interesting, is that this response does not happen directly: the temperature of yesterday influences today's demand. This is found after comparing the R^2 for several shifts of demand in relation to temperature (influence of temperature 2 days ago with regard to demand today, influence of yesterday's temperature to today's demand, and today's temperature influencing today's demand). When using the temperature data and demand data of the same day, the R^2 is 76.4% whereas this is 82.0% when demand is shifted one day

ahead such that today's temperature, influences tomorrow's demand (Figures 3.5a and 3.5b show the correlation of the temperature and HDDs with demand shifted one day ahead). This could be sensible: warmth or cold can remain in a building for quite some time so when it was warm yesterday, some of that heat remains in a building which results in less heating being needed today.

In the forecasting engine, this R^2 when correlating LPG demand with HDDs, is calculated on the level of measurements that are available. When a measurement becomes available every week, the HDDs are aggregated weekly. When this is done for the aggregated weekly demand of fourteen storages, this gives an R^2 of 91.0% which is even higher than the coefficient of determination of daily demand (when daily demand is obtained by assuming constant usage between two measurements). This is sensible since daily demand is nothing more than weekly demand distributed evenly over the days of the weeks which flattens out daily dependency on temperature. Figure 3.7 plots the HDD (right y-axis) and demand (left y-axis). In this figure we can clearly see the dependency of demand on HDDs since the peaks and drops often appear simultaneously. Just as for temperature, the LPG usage is dependent on the HDD of the day before (note that in Figure 3.7, the demand is already shifted one day ahead).

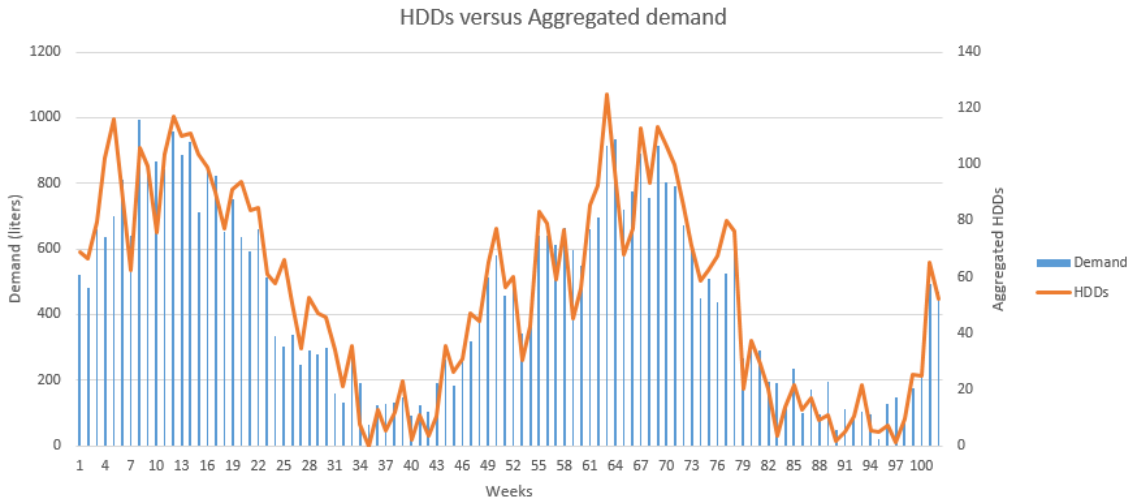


Figure 3.7: Demand versus HDDs

The relationship between LPG usage and HDD or temperature is linear. Besides linear relationships, other relationships could be present as well which might be worthwhile looking at. Also other external variables than temperature based covariates (e.g. humidity, global radiation, and wind) might affect LPG demand in a linear or non-linear way. These should also be looked at.

3.2.2 Degree-days method

We now explain the degree-days method using a representative time series of a storage that shows a nice dependency on temperature and therefore a yearly pattern (this is the same that we used in Figure 3.4 to show the effect of making the after delivery readings irrelevant). The easiest way of incorporating HDDs in the forecast is by using simple linear regression. In a scatter diagram like the one shown in Figure 3.5b, we see what the usage was for each number of HDDs. When you know that next week, 10 HDDs are predicted, about 1000 liters

LPG will be used. However, this was technically difficult to implement in the current framework of ORTEC and, therefore, they decided to do something else: remove the temperature dependency, predict that usage by using simple exponential smoothing, and finally adding the temperature dependency to the series again. Results in Chapter 4 must determine if linear regression gives better results or not.

The degree-days method starts with calculating the correlation between HDDs and LPG usage. When the R^2 when correlating with HDDs is above 40%, the degree-days method and the yearly script are tried. When the R^2 is below 40%, it is assumed that the degree-days method is not an appropriate forecast method. Since the HDDs are based on a certain base temperature, the forecasting engine calculates the coefficient of determination for all base temperatures from 10 to 25°C with increments of one. The base temperature that results in the best R^2 is used.

The method first normalises the time series as if it were always the same temperature (the mean temperature is used for this) by simple linear regression, which theoretically would lead to having a straight line when HDDs would have been the only factor influencing LPG usage. This is because the air temperature shows approximately the same pattern as demand (see Figures 3.6 and 3.7). In other words, it removes the fluctuations caused by temperature. For example, the regression line in Figure 3.5b shows that per HDD, approximately 100 liters extra usage occurs. When the mean number of HDDs is 20, then $20 * 100 = 2000$ liters are subtracted from a week where there were 40 HDDs which gives the usage as if there were 20 HDDs.
As the red line in Figure 3.8 shows, in reality the normalised series is not exactly a straight line (i.e. the temperature does not completely explain the usage). In Figure 3.8, the time is plotted on the x-axis in days, and the y-axis shows demand in liters.

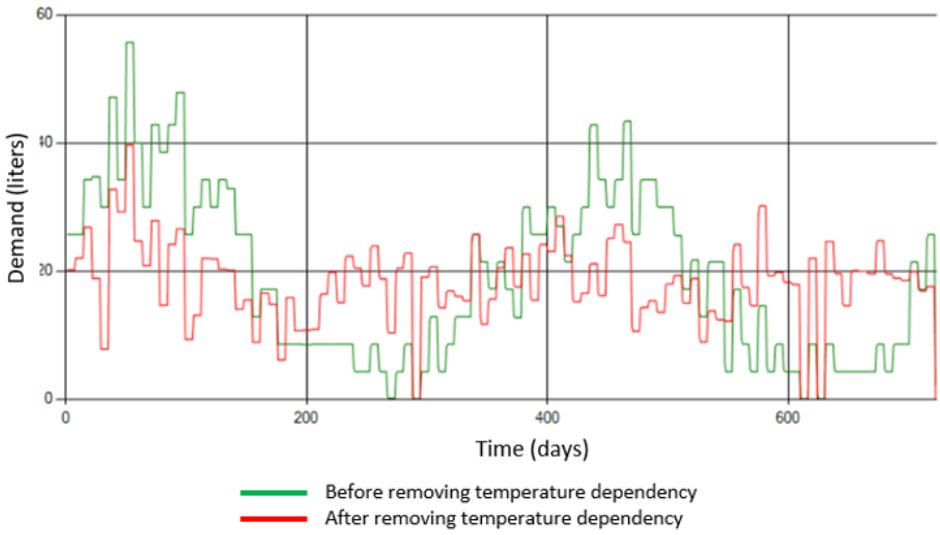


Figure 3.8: Degree-days method

Next, we use exponential smoothing to forecast this 'straight line'. Consecutively, we use the weather forecast (when a forecast of a longer period than a week is required, we use the average of a specific day of previous years) to add the temperature dependency again by simple regression. The orange line in Figure 3.9 shows what this forecast looks like aggregated over weeks (this is done as the realised data is also weekly which makes it easier to compare).

This figure shows the weekly demand distributed evenly over the seven days of the week, that is why every consecutive seven days show equal demand. In reality, the forecast is on daily level, because daily temperature is used. Since weather forecasts do not give reliable information about the far future, forecasts can only be performed one week or at most two weeks ahead, which is also the forecast horizon that we need.

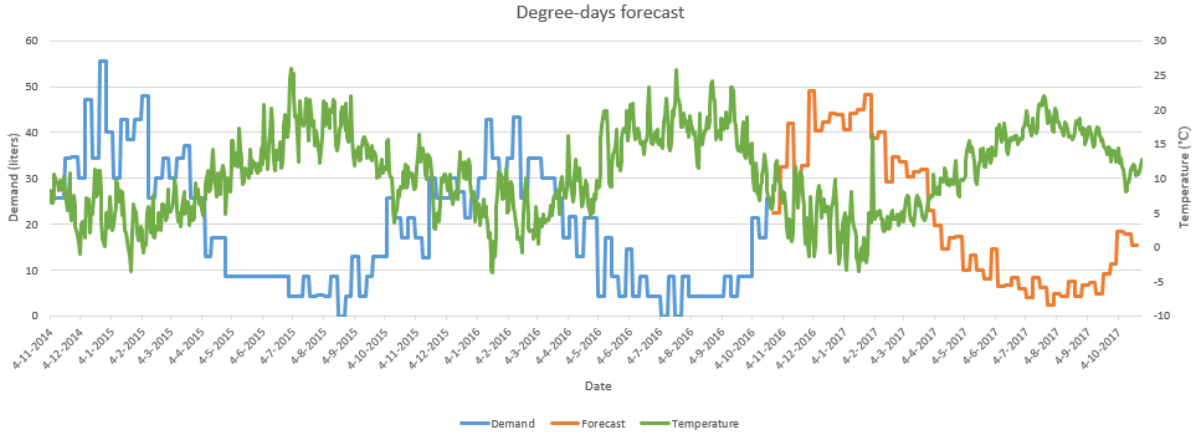


Figure 3.9: Forecast Degree-days method

3.2.3 Yearly script

The method that is used simultaneously with the degree-days forecast, is what is called the yearly script. We use the same time series to illustrate the yearly script as we used for illustrating the degree-days method. The yearly script currently consists of two parts: removing seasonal factors, followed by exponential smoothing. There are four variants: two options in the seasonal factors (weekly and monthly) and two options in exponential smoothing (period 1 day and period 7 days). Let us first explain the two variants of seasonality removal. In the weekly variant, firstly the total usage of a year is calculated and divided by the number of weeks in a year: 52 (to be exact 52.14 weeks, the forecasting engine has a way of coping with this, so we do not miss a part of each year in the forecast). Per week a factor is calculated. For example, when the yearly usage was 1040, the average weekly usage was 20. When in a specific week the usage was 10, the factor assigned to that week would be 0.5. The monthly variant does this for months instead of weeks so the yearly usage is divided by 12 instead of 52.

It turns out that, at least for the storage used as example to illustrate the degree-days method and the yearly script, removing trend before removing seasonality results in a line that is more straight. Figure 3.10 shows the trend in this dataset. The straight line that follows from the trend- and seasonality removal procedure (the red line that remains in Figure 3.11b) is then predicted by exponential smoothing. Then there are two variants of this: with period 1 day and period 7 days. Period 1 means that the direct previous days are smoothed and period 7 means that 7 days ago, 14 days ago and so on are used for exponential smoothing (to smooth all Mondays, Tuesdays, and so on). The latter would be useful when a within-week pattern exists in the data. However, for many storages, no daily data is available which means that this is not an option.

The straight line is then seasonalised again to show the yearly pattern. This is done by taking the average of the factors of the same weeks from previous years (so the factor of week

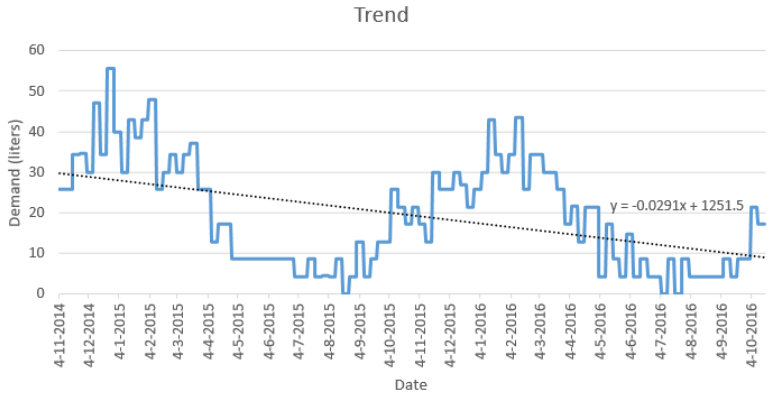
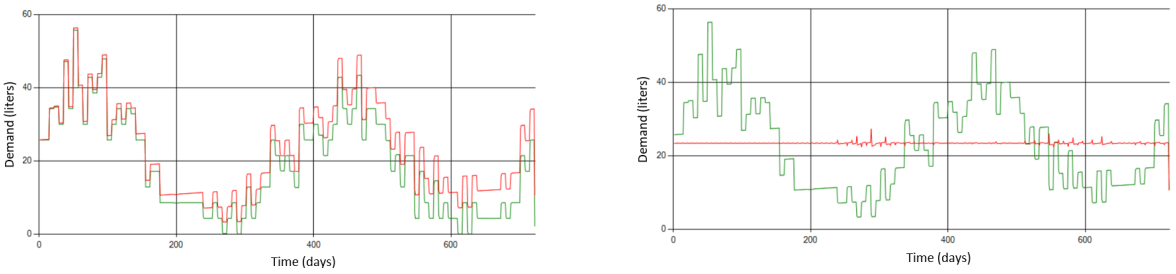


Figure 3.10: Trend



(a) Detrend (b) Deseasonalise

Figure 3.11: Detrending and deseasonalising data in yearly script (Red: before tranformation, Green: after tranformation)

1 from the forecasted year 3 is calculated by averaging the factor of the week 1's of years 1 and 2) which gives 52 seasonal factors. This results in the forecast shown in Figure 3.12 (in this figure, no trend removal transformation has taken place). Or when the user decides to use monthly seasonal factors, this results in 12 factors.

When both methods (degree-days method and yearly script) are executed, the forecast engine chooses the method that results in the lowest mean absolute percentage error (MAPE) in the validation period, which generally is the last 20% of the data. This MAPE is calculated by forecasting weekly. This means that when the validation period is half a year, 26 weekly forecasts are executed. After each weekly forecast, that week is dropped from the validation period and we do as if all information for that week is also available in order to forecast next week. There are however some problems comparing the MAPE of the degree-days method and the yearly script. Firstly, the MAPE of the validation period of the degree-days method is calculated using a different series than is used in the yearly script. Secondly, we use observed temperatures in the validation period of the degree-days method whereas we use predicted temperatures for the actual forecast which results in a slightly optimistic MAPE.

For the storage given as example, the degree-days method turned out to be best (when using the data without readings after delivery and with the MAPE as performance indicator). A reason for this probably is that demand is highly correlated with temperature which is exploited in the degree-days method.

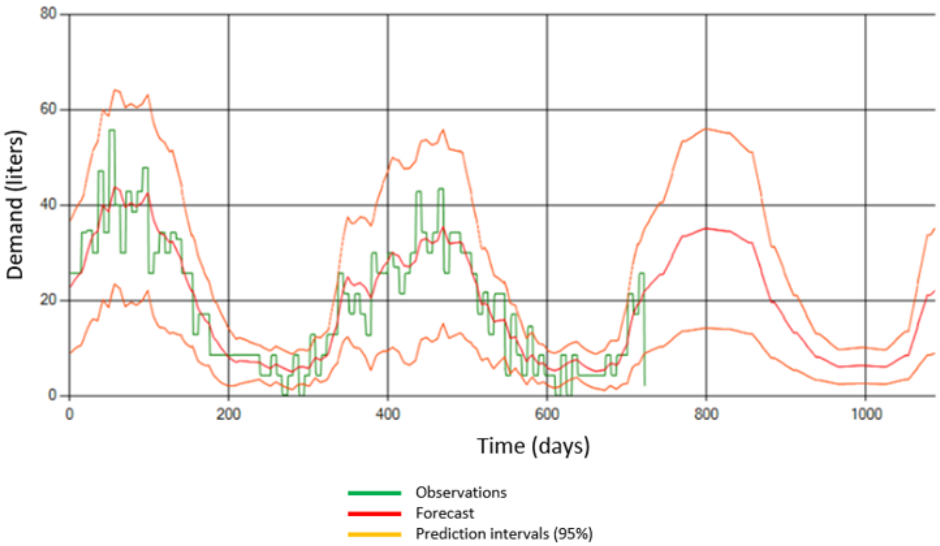


Figure 3.12: Forecast Yearly script

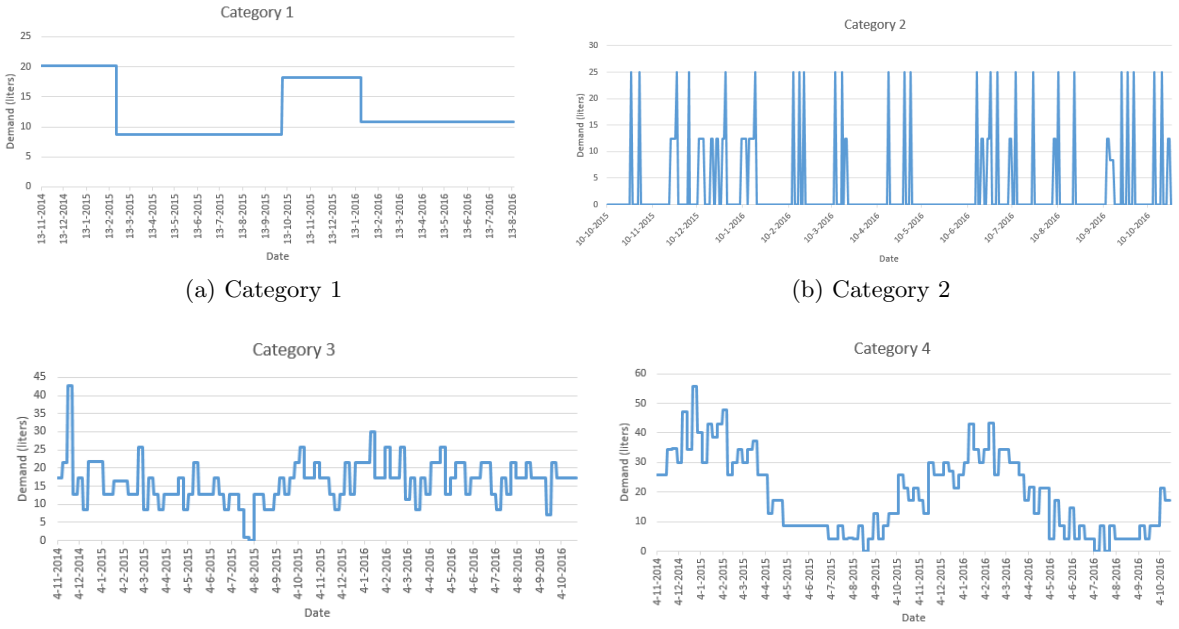
3.2.4 Issues

The problem is that ORTEC currently does not know how well this forecast procedure performs on all different types of datasets and if the decisions that are made are the correct ones.
Firstly, currently the temperature dependent time series are forecasted using the degree-days method or the yearly script. However, we predict all other datasets using exponential smoothing. This might be beneficial for some datasets, but the question is whether exponential smoothing performs well on each of them. Therefore, we need to investigate what patterns are present in the data and which forecasting methods are suitable for each of them.
Secondly, is the MAPE the best performance indicator used for minimisation? The root mean squared error (RMSE) for example penalizes large errors way more than the MAPE does since it squares the errors. An argument for using the RMSE instead of the MAPE in the LPG case is that large errors influence the truck drivers capability of replenishing each customer on the planned route more than a few smaller errors. Moreover, using the MAPE as performance indicator is questionable when assessing low-volume data.
Thirdly, for the dataset used as an example it is clear that the usage is dependent on temperature. The question is however, whether other datasets are as temperature dependent as this one. Besides, there could be other external variables that influence LPG demand. Therefore it should be investigated whether there are, and how they could improve forecast performance, and if those relationships are linear or non-linear (currently only linear relationships are considered). The next section looks at the different patterns that are present in the data.

3.3 Data patterns

As explained before, for some storages, each week on a fixed moment, the usage of that week is obtained. However, this would be the ideal situation. After looking closely at many datasets it turns out that only in a small fraction, such regular measurements are available. And as explained before, even in those datasets, the measurements that occur after delivery are incorrect. Many other clients do not have such an accurate, nicely distributed dataset of

their usage. Let us first define the different categories.



(a) Category 1

(b) Category 2

(c) Category 3

(d) Category 4

Figure 3.13: Types of datasets

There are different types of datasets. Generally they can be separated into four categories. 'Category 1' occurs at customers that do not have a telemetry system, only measurements after delivery exist. Telemetry is an automated communication process that transits measurements and other data to receiving equipment for monitoring. This happens on a regular basis (e.g. weekly or daily). The method where 'reading after' measurements are excluded is not an option for this category, since no measurements remain. Interesting is that due to the small number of measurements, the R^2 when correlating with HDDs is often rather high for this kind of datasets. Therefore, currently, these datasets are predicted using the degree-days method whereas we do not know if this is the right method for this pattern.

'Category 2' is different. Since the volume of LPG is dependent on temperature (as liquids expand when it is warm and shrinks when it is cold), independent of usage, propane volume rises if the temperature rises and falls if it becomes colder, some odd measurements occur in the data. That this is caused by changes in temperature is found after calculating the R^2 of the tank volume belonging to Figure 3.13b and the outside air temperature. The R^2 is rather high: 88.8%. It seems that the level of the tank is sometimes below and directly after that above a certain threshold which leads to having 'negative usage' which is of course impossible. We cannot solve this by correcting that volume for changes in temperature since the measurement equipment is inaccurate in the sense that only steps of usage are observed (for example only steps of 25 liters). What is currently done to make this data usable for forecasting purposes, is that all negative usage measurements are eliminated which leads to showing more usage than there actually is. Figure 3.14 shows that in a year, only about 200 l is used in reality whereas according to the usage given by Figure 3.13b, this would be 1050 l. We must however put this into context: this customer has a relatively low usage compared to other clients that use multiple thousands or even over ten thousand liters per year. Looking

at different datasets showing this pattern should indicate the magnitude of this problem.

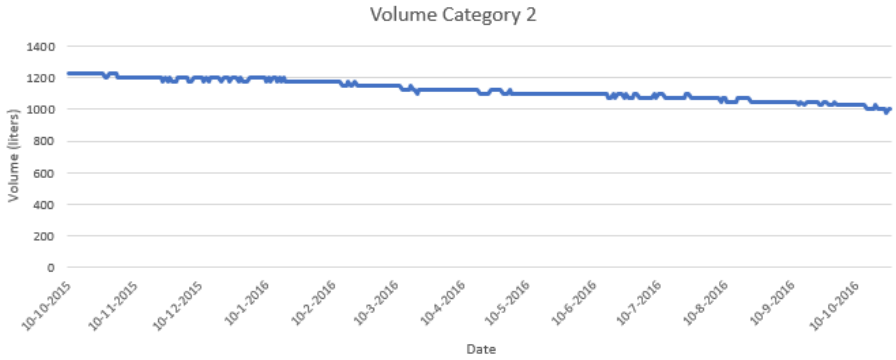


Figure 3.14: Volume storage

'Category 3' does have a telemetry system and demand shows variation, but does not seem to be as clearly temperature dependent as a 'Category 4' dataset. 'Category 4' datasets have regular measurements and show a pattern that looks like a cosine with a peak in the winter and low values in the summer.

In order to find out how much of the clients belong to which category, we classified 2284 datasets with the four categories being the classes. We find that 24% of the datasets belong to 'Category 1', 38% to 'Category 2', only 3% to 'Category 3', and 35% is 'Category 4'.

As mentioned before, LPG volume is dependent on temperature. Volume correction is generally done to a temperature of 15°C. The measuring device in propane delivery trucks includes a volume correction. The temperature is taken into account and the delivery truck makes sure that the delivered amount of propane is the amount that the customer ordered and more importantly: pays for. However, the volumes in the tank are not corrected as noticed in 'Category 2' in the previous paragraph. The ASTM-IP petroleum measurement table shows that at a temperature of 0°C, the volume is already around 4% smaller compared to the net volume at 15°C. Therefore, it could occur that a usage peak is seen on a hot summer day, whereas part of this peak is only there, because the volume of the LPG rose caused by the high temperature. Besides this problem, the telemetry measurements turn out to have a margin of about 5% which also causes unreliability.

Concluding, the data is incorrect in several ways: regular measurements are not temperature corrected, readings after delivery are unreliable and should be excluded, telemetry measurements have a certain degree of unreliability, and some positive usage is unjustly kept whereas some usage should be compensated by negative usage. It becomes clear that before being able to forecast LPG demand, the data must be corrected. Currently, all series are forecasted in a similar way. The distinction that is made is between series that seem to be temperature dependent and the ones that are not. When the R^2 when correlating demand with HDD is above 40%, the degree-days method is tried. However, the other data patterns might not benefit from a method based on the relationship between temperature and demand.

3.4 Conclusion

In this chapter we answered the research question 'What is the current situation at ORTEC?'. We wanted to find out what method(s) are currently used, what the issues are that come up using these, and what characteristics are present in the data. We conclude the following:

1. High, unjust peaks occur when readings after delivery are included
2. All positive usage is unjustly included (some should be compensated by negative demand)
3. Usage measured by the telemetry system is not corrected for volume changes due to temperature
4. The data can be categorised into four categories:
   Category 1: Only a few measurements available, possibly yearly seasonality (24%)
   Category 2: Daily data, many measurements that show negative usage (38%)
   Category 3: Weekly data, no seasonality (3%)
   Category 4: Weekly data, yearly seasonality (35%)
5. Currently, 'Category 4' datasets are predicted using the degree-days method that exploits the temperature dependency of the data, and the other categories are forecasted with simple exponential smoothing
6. It is unclear how the current implementation of the degree-days method performs and whether simple linear regression would yield better results

Chapter 4

Selecting forecasting methods

This chapter answers the third research question: 'Which methods are eligible for ORTEC?'. Since literature gave plenty of suitable methods, we need to find out which method performs best and is most suitable per data category.
First, Section 4.1 describes how the data should be cleaned to be suitable for forecasting since we saw in Chapter 3 that there are some unreliable measurements. Section 4.2 elaborates on parameter estimation. After that, we discuss per data category which methods are suitable according to literature and test those to find out which method(s) perform(s) best. Section 4.3 does this for 'Category 1', Section 4.4 for 'Category 2', Section 4.5 for 'Category 3', and Section 4.6 for 'Category 4'. Section 4.7 concludes per category which method is most promising and whether, and if so, how the current methodology can be improved.

4.1 Data cleaning

In Chapter 3 we saw that two problems in the data should be addressed: the unreliable after delivery readings and negative usage.

After delivery readings

Unreliable after delivery measurements result from the truck driver filling in that he filled the tank to 80% even though this was for example 78% or 83% in reality. There is a possibility in OIR (ORTEC Inventory Routing) to make stock measurements irrelevant which means that the selected measurements are discarded. Since the status of the measurements after delivery is different from the telemetry readings, namely 'Reading after' instead of 'Reading', it is easy to see which readings to make irrelevant. We have tested this for several representative 'Category 4' datasets to see whether the peaks disappear when after delivery readings are made irrelevant. These five datasets are datasets of storages of different customers, that have different usage and therefore different delivery frequencies. Table 4.1 shows the effect of discarding the after delivery measurements. The R^2 is calculated for HDDs (heating degree-days) being the predictor of dependent variable demand. Appendix C visualises these examples and more (the dataset numbers given in Table 4.1 correspond with the numbers given in the top-left corner of the figures in the appendix).

For each dataset, making the unreliable after delivery readings irrelevant improves the temperature dependency. Before cleaning the data, the R^2 of all datasets is close to the threshold of 40% which is used to decide whether to try the degree-days method and the yearly script. After cleaning, the coefficients of determination are far above the threshold.

Table 4.1: Improvement R^2 after data cleaning

	Dataset					
	1	2	3	4	5	6
R^2 Before cleaning	0.233	0.282	0.371	0.080	0.281	0.206
R^2 After cleaning	0.737	0.657	0.808	0.555	0.494	0.496
Improvement	**216%**	**133%**	**118%**	**591%**	**76%**	**141%**
Delivery frequency	High	High	Med	Med	Low	Low

We also see this when plotting the time series before and after cleaning: the strange peaks are gone after cleaning (Figure 3.4 from the previous chapter (Section 3.1) shows an example of such a plot). Therefore, making the 'Reading after' measurements irrelevant, seems to be a good solution.

The big differences in percentage improvement are caused by the number of peaks and the magnitude of the peaks. When more and higher peaks are solved by making the after delivery readings irrelevant, this improves the data more than when there were only a few and/or low peaks to begin with. More peaks are caused by having more deliveries (to give an indication of this, we added the row 'Delivery frequency' to Table 4.1 where low is defined as 0 to 4 deliveries, medium as 5 or 6 deliveries, and high as 7 or more deliveries a year) and the height of the peaks depends on how close the delivery was to a regular (telemetry) reading (when there is only a day in between, the peak is higher compared to when the after delivery reading takes place exactly in between two telemetry readings since then the wrong measurement is smoothed out over more days). For example dataset 4 had one massive peak that caused the R^2 before cleaning to be low.

Not only for 'Category 4' these readings are a problem, but also for 'Category 3' they are since the same peaks occur. For 'Category 1' however, the reading after measurements should not be made irrelevant since clients in this category do not have telemetry which means that reading after measurements are the only measurements their dataset consists of.

For 'Category 3' datasets, improving the data by removing after delivery readings should not be expressed in terms of temperature dependency but we can make it visible. Figure 4.1 shows what two datasets of 'Category 3' look like after making the after delivery readings irrelevant. As for 'Category 4' datasets, most of the strange peaks disappear.

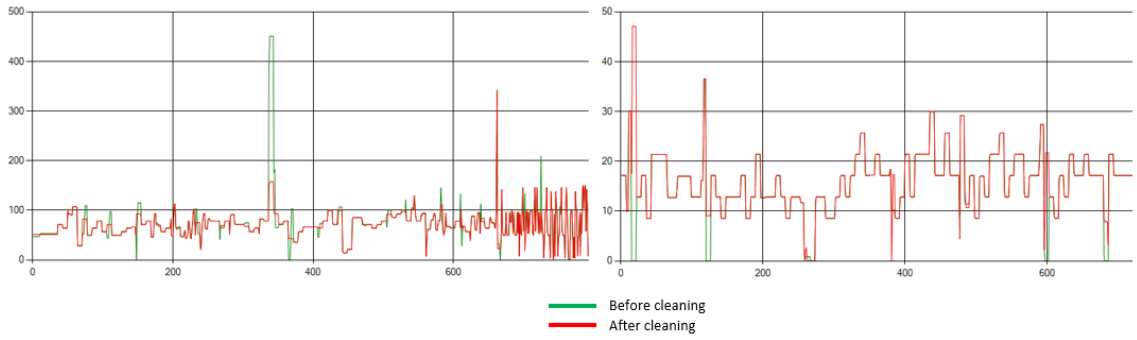
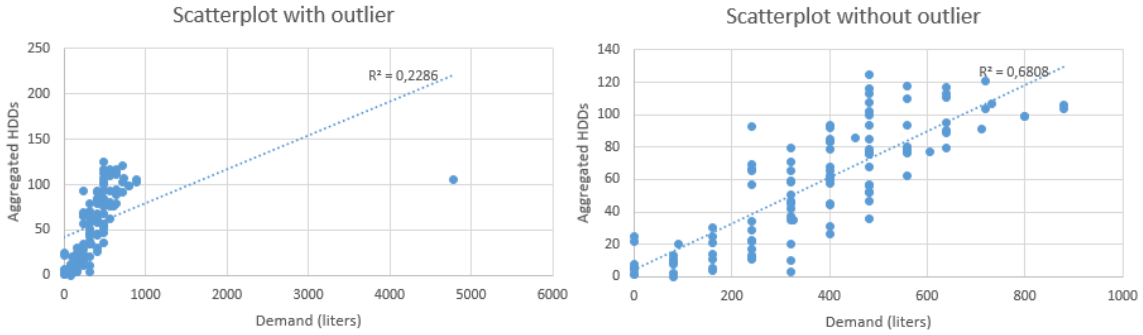


Figure 4.1: Removing after delivery readings for 'Category 3' datasets

However, for both categories, not all peaks disappear (which we see in Appendix C and Figure 4.2). This could indicate that not all incorrect measurements are categorised as 'After

delivery reading' or some peaks have other causes. Since extremely high peaks disrupt the temperature dependency, we need to find a way to exclude these in the regression equation. Figures 4.3a and 4.3b show the correlation of demand with HDDs of one storage that shows an enormous peak (Figure 4.2 shows this storage). We see that when using Cook's distance to remove the remaining peak improves the R^2 when correlating demand with HDDs from 22.9% to 68.1%. In fitting a linear regression model, the regression coefficients can be substantially influenced by one or a few observations (Kim & Storer, 1996).

Figure 4.2: Example of a storage of which not all peaks are removed



(a) With outlier

(b) Without outlier

Figure 4.3: Effect of an outlier on the correlation with temperature

A method to find outliers in a scatter plot is by using Cook's distance. To quantify the effect of individual observations on the fit, Cook proposes influence measures based on deleting observations (Kim & Storer, 1996). Cook's distance summarises how much the values in the regression model (the estimated regression coefficients) change when an observation is left out of the estimation.

When using the normal equation explained in Section 2.5, Equation 2.35, we were able to fit a regression model. The coefficients are calculated by:

$$\beta = (X^T X)^{-1} X^T y \tag{4.1}$$

and the model fit is

$$\hat{y} = Hy \tag{4.2}$$

where

$$H = X(X^T X)^{-1} X^T \tag{4.3}$$

Also, we have the error vector $e = y - \hat{y}$ and the unbiased estimator of σ^2 which is: $s^2 = e^T e/(n-p)$ where n is the number of observations and p is the number of independent variables and the constant (Kim & Storer, 1996). Let $h_i = x_1^T (X^T X)^{-1} x_i$, then Cook's distance is calculated by:

$$D_i = \frac{e_i^2}{s^2 p} \frac{h_i}{(1-h_i)^2} \tag{4.4}$$

Cook & Weisberg (1982) propose the following guideline with regards to outliers: when $D_i > 4/n$, the observation should be excluded when calculating the regression coefficients. The outlier in the example case shown in Figure 4.3a had a Cook's distance value of $D_i = 49.6$ which is far above the threshold value of 0.035.
The great advantage of Cook's distance is that it takes into account the independent variable, temperature in this case, to determine whether an observation is an outlier. For example, there might be a peak in usage which is indicated as outlier when using basic time series outlier detection. However, it might be possible that in the period of the peak, it has been extremely cold outside and when looking at that, the usage perhaps is not that strange.

Negative usage

The second problem we have to address is the one that occurs in 'Category 2' datasets. As we explained in Section 3.3, currently, all negative usages are discarded which leads to showing more usage than there actually is. Instead of discarding all negative usage, we can send all measurements, both positive and negative, to the forecasting engine. We hope that in methods as exponential smoothing, the negative usage will compensate for some of the unjustly measured positive usage. Section 4.4 tests whether this is the case or whether we need to find a method that converts the time series with both positive and negative usage to a time series that only shows the actual positive usage.

4.2 Parameter estimation

In order to implement the several exponential smoothing variants, some parameters must be estimated. The Holt-Winters method (both additive and multiplicative), is the method for which most smoothing parameters must be estimated, three to be precise (alpha, beta, and delta). Many authors elaborate on the magnitude of the smoothing constants. Some suggest values between 0.10 and 0.30, some between 0.05 and 0.50, and some limit α to 0.20 (Ravinder, 2016). Ravinder (2016) himself, however, finds that optimal values of smoothing constants are often outside these ranges. He recommends to ignore the guidelines and vary the smoothing constants between zero and one. Besides, the smoothing parameters are dependent on the time buckets of the forecasts, time buckets of one day will probably lead to lower smoothing constant compared to larger time buckets.

Currently in the forecasting engine of ORTEC the parameters are optimised using grid search. The user can set the minimum and the maximum of the parameters, the step size, and the factor with which the step size is decreased. For example, when the value of smoothing parameter α is varied between 0 and 1 with steps of 0.10, and the best performance indicator results from a value of 0.3, the factor with which the step size is decreased is used to seek around 0.3. Decrease factor 0.10 for example results in testing $\alpha = 0.21, 0.22, ..., 0.39$.

4.3 Category 1

'Category 1' datasets only contain after delivery readings and no telemetry measurement which makes it tricky to calculate the performance of several forecasting methods. The forecasting engine calculates daily predictions. Therefore, we decided to forecast the last reading, represented by the orange line in the example given in Figure 4.4. This figure should be interpreted as follows: the four horizontal lines represent the average daily usage which is calculated by subtracting two subsequent volume measurements and dividing this by the number of days between these measurements (assuming that each day in this period, has had the same usage). Therefore, the length of each horizontal line gives the period between two measurements.
An important remark here is that the validation of the methods for 'Category 1' datasets is extremely unstable since we only predict one measurement. Appendix D shows figures of all the datasets that we forecast. Figure 4.4 serves as a representative example of a 'Category 1' dataset and is equal to Dataset 1 in Table 4.2.

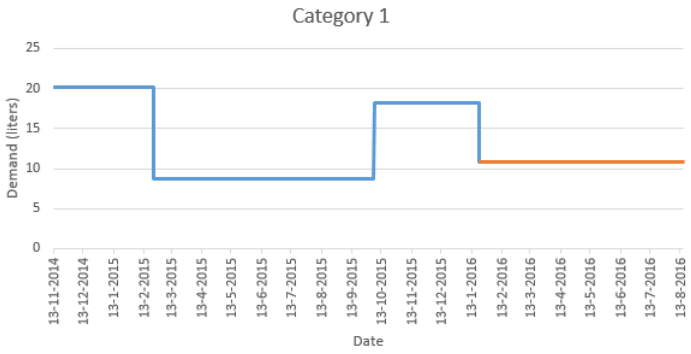


Figure 4.4: Category 1

In order to be able to generalise our findings, we forecast datasets (see Appendix D) that differ in the number of measurements and their dependency on temperature (see Table 4.2). It is striking that the R^2 of the fourth dataset is hundred percent. This is because there are only two observations available to do linear regression on, which always results in a perfect regression line through both data points.

Table 4.2: Characteristics datasets 'Category 1'

	Dataset							
	1	2	3	4	5	6	7	8
R^2 with temperature (%)	92.9	89.6	99.8	100	97.3	88.0	99.1	96.5
Number of measurements	4	5	4	3	8	7	4	5

The methods that we perform on these datasets are single exponential smoothing, the degree-days method, the yearly script, linear regression and moving average. Because these datasets contain only a few measurements, the coefficient of determination when correlating with HDDs for most 'Category 1' datasets is rather high (many have an R^2 above 90%). This is calculated by averaging daily demand per measurement and correlating this with the average daily HDDs per measurement. Because of this high correlation, the forecasting engine often chooses the degree-days method since the R^2 is far above the threshold of 40%.

Simple Exponential Smoothing (SES)

SES is performed on the daily demand since that is how it is currently done in the forecasting engine. Alpha is limited to 0.01 because, when the smoothing value is optimised while minimising the RMSE, alpha is always 1 which replicates the naïve method (demand of today is equal to demand of yesterday) whereas we want the method to smooth demand.

Linear regression

We start by dividing the first readings by the number of days in the period the measurement covered. The HDDs are aggregated on the same periods and also divided by the number of days in the period which represents the average HDDs in these periods. Based on the measurements (without the measurement we ought to predict), the regression coefficients are calculated. The result is a constant and a coefficient which we use to calculate daily demand in the forecasting period. The result is an average daily forecast of the usage which we multiply by the number of days in the period we want to predict. Note here that we use the base temperature to determine HDDs that led to the highest R^2 on the 21 aggregated datasets we discussed in Section 3.2 (which was 18.3°C). This means that for each dataset, the base temperature is the same. However, in the forecasting engine, the base temperature is optimised by grid search so the results we give here, are a bit pessimistic (this is also the case for the other categories where we test linear regression).

Moving Average

Since we work with daily demand, we decided to average all previous demands before the demand that we want to predict. For example for the data in Figure 4.4, we average all the blue demand values. In other words, we average the daily demand multiplied by the number of days in a period of all measurements apart from the one we want to predict, and multiply that average by the number of days in the period of the measurement we predict.

Results Category 1

Since we only predict one measurement, the MAPE gives the percentage the prediction differs from actual demand. Table 4.3 shows the results of the different methods, of which the current method is bold. What is interesting to see is that in all cases, the degree-days method and yearly script perform best or second best. Besides, the current methodology, simple exponential smoothing, belongs to the worst performing methods for all cases. Therefore, we advise to make use of the temperature dependency and use the degree-days method and yearly script since they are always performed together and choose the best performing method. Note here that simple linear regression performs better, but this method is not currently implemented in the framework, because of technical difficulties. The degree-days should be implemented using simple regression instead of the current implementation where firstly the temperature dependency is removed, the 'straight line' is predicted by SES, and finally temperature dependency is added.

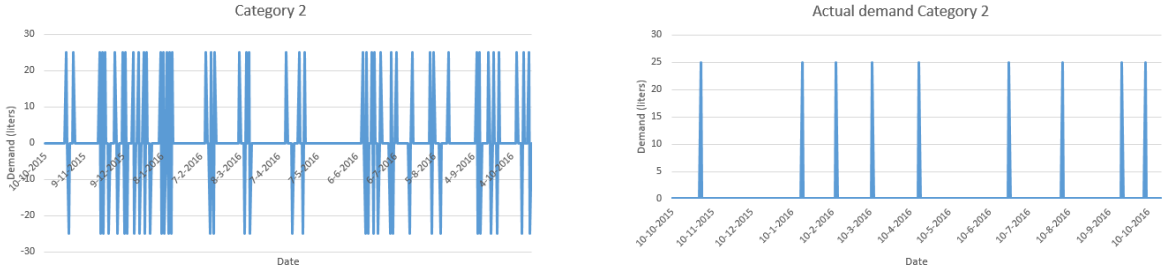
4.4 Category 2

Figure 4.5a shows what the demand looks like of 'Category 2' when negative usage is included instead of discarded. We have daily data of which we hold out 20% for validation purposes. We changed a setting such that now, negative usage is sent to the forecasting engine instead

Table 4.3: Performance Category 1 (the current method is bold)

	Performance (MAPE %)							
Method/Storage	1	2	3	4	5	6	7	8
Degree-days	11.58	3.34	10.46	20.07	26.15	2.96	19.51	32.91
Yearly	7.08	9.62	21.24	21.11	52.56	9.69	26.27	32.24
SES	36.12	23.26	84.32	20.97	80.57	34.55	17.10	57.29
Moving Average	27.05	23.01	75.74	32.83	241.30	47.03	30.23	157.2
Simple regression	8.07	14.30	2.97	30.90	45.11	1.84	4.29	5.56

of being discarded. We hope that the negative usage compensates for some of the positive values in the forecast. For such intermittent demand pattern, Croston's method and TSB are suitable methods according to literature. Single exponential smoothing is currently used so we test that method as well. Since the clients in 'Category 2' are such slow movers, we do not search for a method to convert the usage as shown in Figure 4.5a to the usage shown in Figure 4.5b.



(a) Negative usage included

(b) Actual demand

Figure 4.5: Category 2

Table 4.4 shows the characteristics of the datasets for which we predict the usage (Appendix E gives the plots of both the usage and tank volume). The first row, 'Yearly usage (liters)' gives the actual usage of the first year of data. With actual we mean that the negative usage compensates for the unjust positive usage. 'Discard negative usage' gives the usage of that same year when all negative usage is discarded, as is done currently. What is interesting to see is that the difference between these two is massive. For the storages that we tested, discarding negative usage leads to the usage used as forecast input being 134% of an LPG tank (defined as the maximum volume the tank may be filled with minus the minimum volume) higher than the actual usage. Because of this, the number of deliveries to 'Category 2' storages can be reduced by 87%.
The last row, 'Percentage negative usage', is defined as the number of measurements (timed values) that show negative usage with regards to the total number of measurements. In order to generalise our findings, we forecast datasets with different characteristics.

Croston's method & TSB

This method uses exponential smoothing to predict the demand intervals and the demand sizes separately after each period with demand (both positive and negative). Since Croston's method turned out to be biased, some adjustments are made. The resulting method is called the TSB method. TSB does not update the inter-arrival time, but the probability that

Table 4.4: Characteristics datasets 'Category 2'

	Dataset						
	1	2	3	4	5	6	7
Yearly actual usage (liters)	200	2678	4555	1224	2087	309	740
Discard negative usage (liters)	1050	4967	7450	2384	3162	1409	3120
Percentage 'negative usage'	9.6	16.7	18.1	15.7	12.1	22.0	24.0
Tank capacity	2500	2500	3000	1600	2500	1000	1000
Number of deliveries in 2016	0	2	3	2	1	1	2

demand occurs which is updated each period instead of each period with non-zero demand. The parameters α and β for both methods are calculated using the Excel solver minimising the RMSE in the estimation period.

Simple exponential smoothing

We perform simple exponential smoothing on the usage time series. The alpha's that result from this are rather high: the lowest is 0.31 but the rest are between 0.62 and 1. This is sensible since the usage time series have many values of zero with a sporadic demand peak so more emphasis on the last demand measurement gives the best forecast.

Combining methods

As discussed in Section 2.9, there is proof of the accuracy of combining forecast methods. We combine TSB & SES and Croston's method & SES. The weights are determined using the Excel solver minimising the RMSE in the estimation period. We decided to combine two methods at most since combining more and more methods seems to worsen the performance (Hibon & Evgeniou, 2005).

Results Category 2

Currently, this kind of dataset is predicted using single exponential smoothing. We calculate the MAPE by comparing the forecast made using the time series *with* negative demand (see Figure 4.5a) with the tank volume time series (Figure 3.3 in Section 3.3). We do this by subtracting the predicted demand from the initial volume in order to calculate the predicted volume and compare that with the actual volume. We use the MAPE to base our conclusions on since we compare the volumes and not the usage (which would be very low and make the MAPE unreliable) and the MAPE is more interpretable than the RMSE. Table 4.5 gives the results.

Table 4.5: Performance Category 2 (the current method is bold)

	Performance (MAPE (%))						
Method/Storage	1	2	3	4	5	6	7
Croston	1.69	2.50	12.47	9.67	11.61	23.17	17.81
TSB	0.91	7.61	5.60	12.41	1.32	2.21	3.74
SES	0.74	0.85	3.74	1.24	2.12	2.00	3.05

What is interesting to see is that, almost unanimously, simple exponential smoothing performs best. We therefore advise to remain using this current methodology for 'Category

2' datasets.
Besides the individual forecast methods, we tried improving performance by making combinations. In this case, this did not lead to better results.

4.5 Category 3

Figure 4.6 shows what a typical 'Category 3' dataset looks like: no seasonality and no clear trend. No clear pattern is visible in the time series. We have about two years of weekly data of which 20% is held out for validation purposes. Table 4.6 shows the characteristics of the datasets that we forecast. We only predict three since it turned out that 'Category 3' datasets are rather scarce (only 3% of the datasets are 'Category 3'). Appendix F shows what these datasets look like.

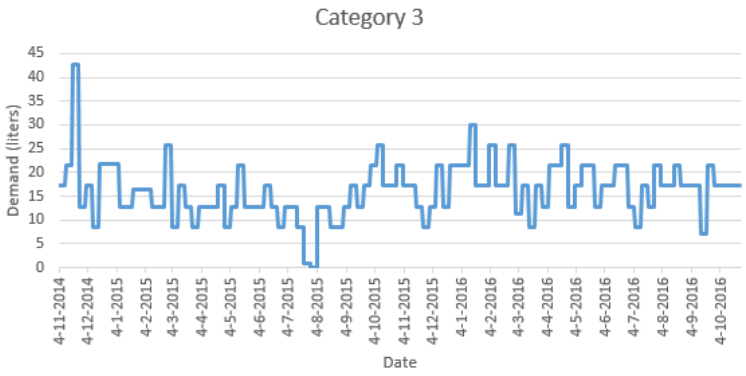


Figure 4.6: Category 3

Table 4.6: Characteristics datasets 'Category 3'

	Dataset		
	1	2	3
R^2 with temperature	1.2%	3.84%	30.3%
Number of deliveries in 2015	3	16	12
Average weekly usage	113	496	955

For this category, we perform moving average (MA) but also double MA, the current methodology (single exponential smoothing), and the degree-days method and yearly script. We also try combining methods.

Single exponential smoothing

Since the current forecasting procedure uses single exponential smoothing (SES) for datasets like 'Category 3', we try this as well in order to compare the different methods. The smoothing parameter α is optimized in the estimation phase while minimising the RMSE. The forecasting engine minimises the MAPE but due to the small volumes that make the MAPE unreliable, we decided to use the RMSE instead. The smoothing parameters for Datasets 1, 2, and 3 are 0.2, 0.1, and 0.3 respectively. These alpha values are sensible since the third dataset has way more peaks and downs compared to Storage 2 that is more stable, so for Dataset 3, more emphasis is on the last observation compared to Dataset 2 that is smoother.

Moving Average

For single MA we average the previous five weeks since this leads to having the best RMSE in the estimation period. For double MA, we average the previous six weeks' MA's for the same reason. Averaging a small number of weeks is desirable when sudden shifts occur in the data which is the case for 'Category 3' (Hoshmand, 2009).

Linear regression

To determine whether this category dataset is surely not dependent on external climatological variables, we try linear regression, both simple and multiple. We use HDDs as external variable for simple regression and multiple regression uses HDDs, global radiation, wind, and relative humidity. Since the latter three are only available as observations of the past and not in the weather forecast, they can only be used te predict for example last weeks' usage. This is useful for the prediction made to determine *how much* LPG the truck driver should deliver at each customer on the planned route since this is based on past observations of the external variables. However, these external variables cannot be used for the forecast required to determine *when* to visit each customer since that forecast is based on future values of the climatological variables.

Combining methods

We try the combination SES with moving average and a combination of multiple regression with SES and multiple regression with MA. Besides giving weights that add up to one to the different methods to be combined, we could also use multiple regression using the two individual methods as input (explanatory variables) and the actual demand as output.

Results Category 3

Table 4.7 gives the results of the individual methods. There is no one clear method that performs best for all 'Category 3' datasets. What we do see is that when looking at the degree-days method and the yearly script, either of the two performs quite well for all datasets. Only for the first dataset, this is not the case but even there, the yearly script does not score that badly. When the degree-days method is chosen, both the degree-days and yearly script are executed and the best of the two is chosen. Therefore, we advise to predict 'Category 3' using the degree-days method. This improves the RMSE on average with 11.3% compared to the current method (SES).

Table 4.7: Performance validation period Category 3 - Individual methods (the current method is bold)

	Performance (RMSE)		
Method	1	2	3
SES	30.17	95.19	211.33
Moving Average	30.85	103.16	208.69
Double Moving Average	36.48	106.70	203.95
Simple regression	34.74	97.99	161.19
Multiple regression	32.87	90.31	173.97
Degree-days	43.01	78.79	176.11
Yearly script	33.36	75.40	177.47

Table 4.8 gives the results of combining forecast methods. The combinations above the double line are made using weights that add up to one and the combinations below the double line are made using multiple regression.

Table 4.8: Performance validation period Category 3 - Combinations

	Performance (RMSE)		
Method	1	2	3
SES & MA	30.15	95.20	208.65
SES & Multiple regression	29.58	89.48	176.13
MA & Multiple regression	29.84	89.26	174.65
Degree-days & Simple regression	39.44	76.41	183.20
Regression: SES & Multiple regression	28.95	88.33	175.92
Regression: MA & Multiple regression	28.95	88.11	174.88

The best scoring combinations score not significantly better than the degree-days and yearly script, only for the first dataset this is the case.

4.6 Category 4

Figure 4.7 shows a representative 'Category 4' dataset. These datasets have regular measurements, and show a nice yearly pattern caused by the dependency on temperature. Appendix G shows the series of the 'Category 4' datasets that we forecast in this section.

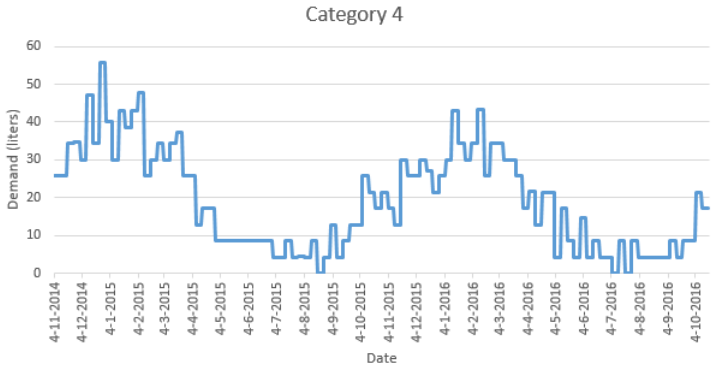


Figure 4.7: Category 4

For testing which methods are most suitable for 'Category 4' datasets, we forecast datasets with different characteristics in order to generalise the results. Table 4.9 summarises these characteristics per dataset. As we see in this table, not all temperature dependency is that high. We can distinguish a 'Category 4' dataset visually by its sinusoid shape, and from the data by its relatively high weekly and yearly autocorrelation.

There are several suitable forecasting methods for seasonal data, which 'Category 4' is. The first are additive- and multiplicative Holt-Winters. Secondly, we test regression models, both simple and multiple. Finally, we combine individual forecasting methods. Also we give the results of the current methodology: the degree-days method and yearly script.

Table 4.9: Characteristics datasets 'Category 4'

	Dataset					
	1	2	3	4	5	6
R^2 with temperature (%)	80.8	49.4	49.6	25.2	73.7	82.1
Average weekly usage (liters)	134	59	91	143	652	85
Number of deliveries in 2015	6	4	3	4	7	11

Additive- and Multiplicative Holt-Winters

For many of the 'Category 4' datasets, we only have weekly data. For weekly data, 52 parameters must be estimated, one for each week, which results in the model having far too many degrees of freedom (Hyndman & Athanasopoulos, 2014). Ord & Fildes (2013) propose a method to make these seasonality estimates more reliable for the multiplicative variant. Instead of calculating the seasonals on individual series level, they calculate the seasonality factors of an aggregate series. This results in having less randomness in the estimates. For now, we use the aggregate series of 21 similar time series that we introduced in Section 3.2.

Table 4.10 gives the resulting parameters (alpha, beta, and delta) for both additive and multiplicative Holt-Winters. Alpha, beta, and delta are the smoothing parameters. When

Table 4.10: Parameters Holt-Winters (left: additive, right: multiplicative)

	Dataset					
	1	2	3	4	5	6
Alpha	0	0	0	0	0	0
Beta	1	1	1	1	1	1
Delta	0.87	0.69	0.75	0.78	0.72	0.87

	Dataset					
	1	2	3	4	5	6
Alpha	0.16	0	0	0.01	0.08	0.09
Beta	0.07	1	1	0.02	0.04	0.11
Delta	0.09	0.10	0.10	0.12	0	0

their value is 1, this means that all emphasis is on the current observation and none on previous ones. Alpha looks at observations, beta is the smoothing constant for the trend, and delta is the smoothing parameter for seasonality. For the additive version we see that alpha is zero in all cases which means that the series is smoothed and little weight is given to the last observation. Beta is one in all cases which means that all emphasis is on the latest trend value. Delta is around 0.75 in all cases so the latest seasonality value has more weight. For all datasets, the parameters are similar. This is not the case for multiplicative Holt-Winters. Dataset 2 and 3 have similar parameters and 1, 4, 5, and 6 have too.

Combining methods

We combine the yearly script & the degree-days method since these two are both executed when the coefficient of determination is above a certain threshold. Secondly, we combine the two best scoring individual methods in terms of RMSE. Some other combinations are tried and Table 4.12 gives the most promising. Just as for 'Category 3', we combine the best performing combination using multiple regression as well.

Results Category 4

Table 4.11 shows the results of the proposed individual methods. What is interesting to see is that the method that was used before the introduction of the degree-days method, simple

exponential smoothing, is one of the worst scoring methods in terms of RMSE. The degree-days method and yearly script both score relatively well. What is striking, is that simple regression performs best in most cases. As explained in Subsection 3.2.2, simple regression is slightly different than the degree-days method in the sense that simple linear regression looks at what is the expected usage given the number of HDDs whereas the current implementation of the degree-days method first removes the seasonality by using simple linear regression, then predicts that 'straight' line by SES, and consecutively adds the temperature dependency again. We also see that additive Holt-Winters is the worst performing method in each case. This was to be expected since we have weekly data and yearly seasonality so many parameters must be estimated. For multiplicative HW this is also the case, but the method that Ord & Fildes (2013) give (calculating the seasonality factors using an aggregated series instead of the single series), seems to robustify the initial values.
What is also interesting is that multiple regression, which we expected to outperform simple regression since more information is included, does not always perform better than simple linear regression. Therefore, there is little or no use in using more than one external variable. The reason for multiple regression performing worse than simple linear regression is overfitting: a higher R^2 does not necessarily result in a better forecast performance.

Table 4.11: Performance Category 4 - Individual methods (the current method is bold)

	Performance validation period (RMSE)					
Method/Dataset	1	2	3	4	5	6
Degree-days	28.93	24.36	38.55	42.46	191.71	16.55
Yearly script	27.80	27.54	43.96	57.55	253.40	11.91
SES	42.26	38.39	61.48	65.69	269.92	17.05
Additive HW	51.17	39.89	71.63	78.65	294.86	19.25
Multiplicative HW	34.10	33.85	68.93	58.38	244.11	16.45
Simple regression (HDDs)	26.76	23.32	34.76	39.17	216.90	12.83
Multiple regression	29.60	24.27	32.41	42.52	210.72	12.93

Table 4.12 shows the results of combining methods. Above the double horizontal line are the combinations made by using weights, and underneath those by using multiple regression. About half of the combinations score better than the best scoring individual method.

Table 4.12: Performance Category 4 - Combinations

	Performance validation period (RMSE)					
Dataset	1	2	3	4	5	6
Yearly & DD	25.92	26.07	39.42	44.74	199.10	11.08
Simple regr. & Yearly	23.32	25.83	37.58	46.41	227.08	10.05
Simple regression & DD	28.76	24.36	35.81	41.45	189.94	15.48
Multiplicative HW & DD	27.03	24.83	38.55	42.07	176.10	12.50
Regression: Yearly & DD	23.46	24.34	36.05	41.80	194.30	10.73
Regression: Simple regr. & Yearly	22.84	23.97	34.78	41.08	202.50	9.73

When looking at the RMSE, the combination that performs best in most cases (simple linear regression & yearly script) performs better than the current methodology (both degree-days and yearly script) in all cases except for dataset 5. The combination simple regression with the yearly script made by using linear regression improves the performance in terms of RMSE 12% on average compared to the degree-days method and 20% compared to the

yearly script. The combination of the degree-days method and the yearly script computed by using regression does not score that differently from the combination of the yearly script with simple regression (also computed by using regression). Since currently the yearly script and degree-days method are always performed together, combining them is not that hard and still leads to a 10% improvement compared to the degree-days method (in terms of RMSE) and an average improvement of 18% compared to the yearly script.

For 'Category 4' datasets we also experimented with artificial neural networks. We implemented both a regular neural network as well as a recurrent neural network with different configurations in terms of hidden layers, learning rate, number of epochs, and different inputs (among which HDDs and the previous observations with different lags). We did not discuss recurrent neural networks in the literature chapter but the idea behind RNNs is that they use sequential information by including feedback loops in the network instead of assuming independency between inputs and outputs. For both ANNs, we see that the model quite easily overfits the data; it finds a perfect way of modelling the training data, but is extremely unstable in the test phase. Sometimes a good forecast occurs but that is more based on coincidence than on a good and stable model. A reason for the instability of ANNs on this data is that we have too few observations. The training data consist of only one and a half year of weekly data, which are 78 observations whereas neural networks usually work with thousands. Moreover, when using a forecasting method, this choice is based on some background knowledge. For example, when choosing Holt-Winters, this implies that there is some kind of seasonality in the data. An ANN however, has no such boundaries which gives it all the freedom to find patterns. Having no restrictions makes it extremely difficult to find the right patterns.

Concluding, for datasets that show a pattern like 'Category 4', the degree-days method and yearly script perform much better than single exponential smoothing. Therefore introducing these methods has improved the forecasts. As discussed earlier, the degree-days method is implemented differently than simple regression would have been, because simple regression would be more difficult technically. We see in the results that simple regression, however, outperforms the degree-days method in almost all cases. Therefore, a trade-off must be made between ease of implementation and performance.

4.6.1 Tracking signal

As described in Chapter 2, it is desirable to incorporate some form of automatic monitoring to ensure that the system remains in control, especially in a routine forecasting system as the one we are dealing with (Trigg, 1964; Gardner, 1983). Since it could occur that the seasonality of the LPG usage of a customer changes, we advice to implement tracking signals in the current framework using the smoothed-error with an α of 0.1, and limits of ± 0.55.

The tracking signal should only be computed for 'Category 3' and 'Category 4' datasets since for 'Category 1' datasets, there are only a few observations and for 'Category 2', between two periods with demand, there is zero demand for which the forecast is a little higher than zero which leads to structurally having negative errors between two periods with demand. When having a series of errors with the same sign (positive or negative), the tracking signal goes out of control quite fast. Besides, since there are generally more peaks in positive direction compared to negative direction, since only some positive usage is compensated by negative usage, the forecast is structurally biased. Especially after a long period with zero demand, the system goes out of control when demand occurs, since then the MAD becomes

relatively small and the smoothed error relatively large. This causes the tracking signal to become either large or small, depending on the direction of the demand (positive or negative). Figure 4.8 gives the forecast in the top figure and the corresponding tracking signal below. Here we see that especially after a relatively long period of zero demand, a period with demand causes the system to immediately be out of control.

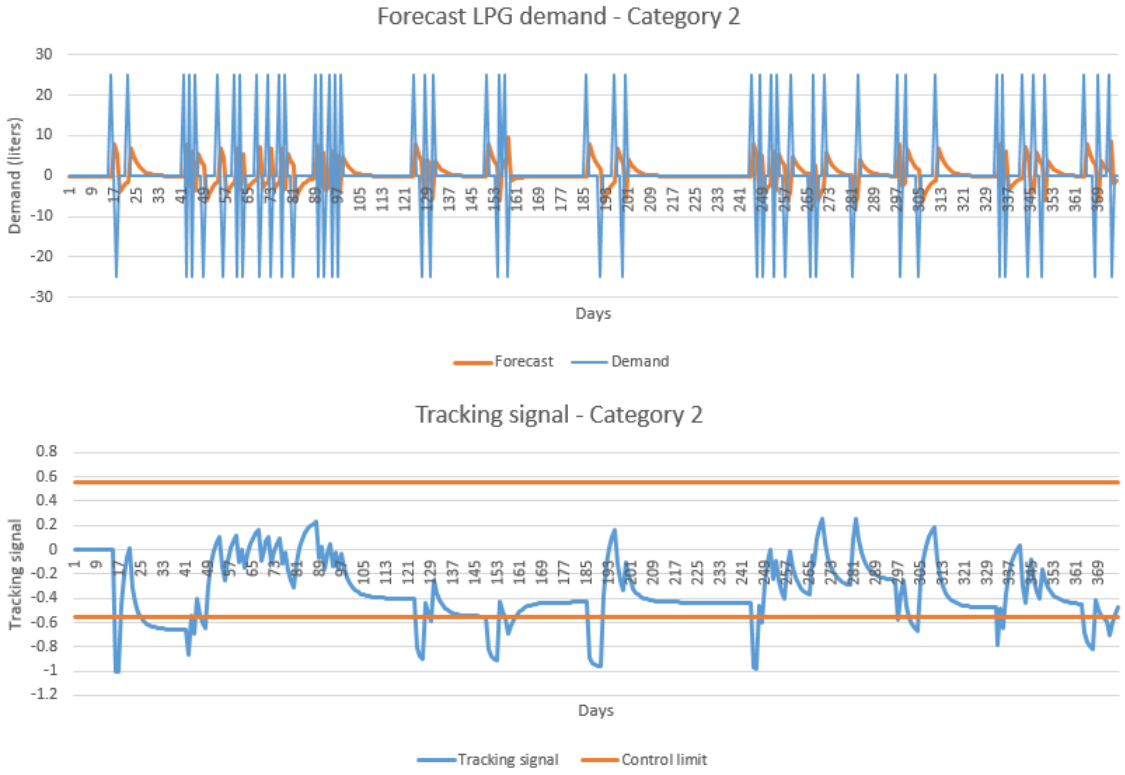


Figure 4.8: Tracking signal 'Category 2'

4.7 Conclusion

In this chapter, we answered the following question: 'Which methods are eligible for ORTEC?'. For this, we wanted to find out how the data should be cleaned to be suitable for prediction, how the current forecasting procedure can be improved (if possible), and which methods perform best. We conclude the following:

1. All after delivery readings should be made irrelevant, because these are unreliable (for all categories except 'Category 1', for which these are the only measurements available)
2. Negative usage should be sent to the forecasting engine instead of being discarded which on average reduces the number of deliveries per year by 87% for 'Category 2' storages
3. For 'Category 2' datasets, SES is the best performing forecasting method, and for the other categories, simple linear regression performs best
4. The MAPE of 'Category 1' datasets on average improves with 67%, and the RMSE of 'Category 3' datasets with 11.3% when changing the forecasting method from simple exponential smoothing to the proposed method (degree-days method using linear regression/yearly script)
5. The RMSE of 'Category 4' datasets on average improves with 6.5% when changing the current implementation of the degree-day method to simple regression
6. The RMSE as performance indicator is more reliable for this data than the currently used MAPE
7. Using more external variables (besides HDDs) does not improve the forecasts, which could be due to overfitting
8. Implement tracking signals in the current framework using the smoothed-error with an α of 0.1, and limits of ± 0.55 to check whether the forecasting system is in control, but only for 'Category 3' and 'Category 4' datasets

Chapter 5

Automatic model selection: Classification

This chapter answers the question 'How can classification methods be used for automatic model selection?'. Until now, the user had to choose a forecasting script manually per dataset. However, since Company X has thousands of clients that need forecasting, it would be more efficient when the forecasting engine automatically detects the most suitable forecasting method by recognizing the patterns in the data. A technique that is able to do this, is called *classification*. Chapter 4 concluded that 'Category 2' datasets should be forecasted with simple exponential smoothing, and the other categories with the degree-days method. Therefore, we have to classify the datasets into two classes (the degree-days method, and SES). There are several reasons for using classification instead of running both forecasting scripts and choosing the one with the best performance indicator. The first is that it simply takes time, especially since there are thousands of storages, running scripts should happen in an efficient way. The second is that we want to find out how much of the cases belong to each of the defined categories. We should make a small remark here, for the latter, we classify on the categories and not on the two forecasting scripts. In order to calculate the percentage of cases per category, we classified 2284 instances on being Category 1, 2, 3, or 4.

Section 5.1 addresses which attributes we choose for classification. Section 5.2 elaborates on how we use the proposed methods on the Company X data, and Section 5.3 yields the results and explains which method is most suitable.

5.1 Attribute choice

The first step in choosing which attributes to use for classification, is to look at the different data categories and the features that characterise them. In essence, we need to separate 'Category 2' from the other categories. The most important characteristics of 'Category 2' datasets are the many timed values of zero, relatively many timed values that show negative usage, and a low R^2 when correlating with HDDs. Based on these features, we choose the following attributes:

- R^2 with HDDs
- Number of timed values
- Autocorrelation of usage (weekly)
- Autocorrelation of usage (yearly)
- Percentage zeros
- Occurrence negative usage (%)

- Occurrence negative usage (number)
- Yearly script quality (MSE)
- Degree-days quality (MSE)
- Regression constant (with HDDs as independent variable)
- Regression slope (with HDDs as independent variable)

Which variables are most important and should be used for classification depends on the method. Therefore, the variable selection procedure is explained for each method separately.

In order to train the model, we manually classified 307 instances of which 186 we labelled as 'Degree-days' and 121 as 'SES'. We did this by first choosing per instance whether it belongs to Category 1, 2, 3, or 4 and then labelling Categories 1, 3 and 4, as 'Degree-days', and Category 2 as 'SES'. We used the characteristics to do this and for doubtful instances, we looked at the plot to determine the category. We use 10-fold cross-validation as validation method and confusion matrices to show the accuracy (we describe both in Section 2.12.5).

5.2 Classification methods

A widely used tool for classification is WEKA (Waikato Environment for Knowledge Analysis). It is designed such that the user can quickly try different methods (the methods that we explained in Section 5.2 and many more) in an easy way. This is a big advantage since the choice of method depends on the actual dataset that is used which makes data mining an experimental science (Witten et al., 2011). Besides the many learning algorithms that WEKA contains, it also includes a wide range of preprocessing tools in which the user for example is able to easily normalise attributes (making the values lie in a range between 0 and 1, as explained in Section 2.12.2). Because of the simplicity and the wide range of possibilities that WEKA offers, we use this tool for our classification problem to determine which classification method is most suitable.

Besides WEKA, R is a widely used data mining tool. R is better in visualising data compared to WEKA and R is often faster compared to WEKA. Also, both implement the algorithms in a slightly different way so it might be interesting to compare the outcomes.

Decision tree methods

This subsection discusses the decision tree method as well as the random forest method.

Decision tree

WEKA and R both use the C4.5 algorithm that uses information gain (difference in entropy) as splitting criterion. A way of preventing overfitting is by *pruning* a tree. Pruning a tree reduces the size of a decision tree by removing sections of the tree that provide little power to classify instances (Drazin & Montag, 2012). Both WEKA and R prune the decision tree by default. In WEKA, we can change more settings compared to how many we can change in R.
Figure 5.1 shows what the resulting decision tree looks like. This figure should be interpreted as follow: in the leaf nodes (node 3, 5, 6, and 7), we can see the accuracy. Node 3 for example, only contains degree-days instances which means that when the percentage negative values is below 0.028 and the percentage zero values is below 0.324, all instances are classified as belonging to the Degree-days class. Node 5 contains mostly SES instances but also some

degree-days which means that it is not entirely pure. Note here that this is the decision tree constructed for the *entire* dataset whereas the accuracy has been calculated on the 10-fold cross-validation for which for each fold, another tree is built. WEKA results in an accuracy of 96.7% and R 95.8%.

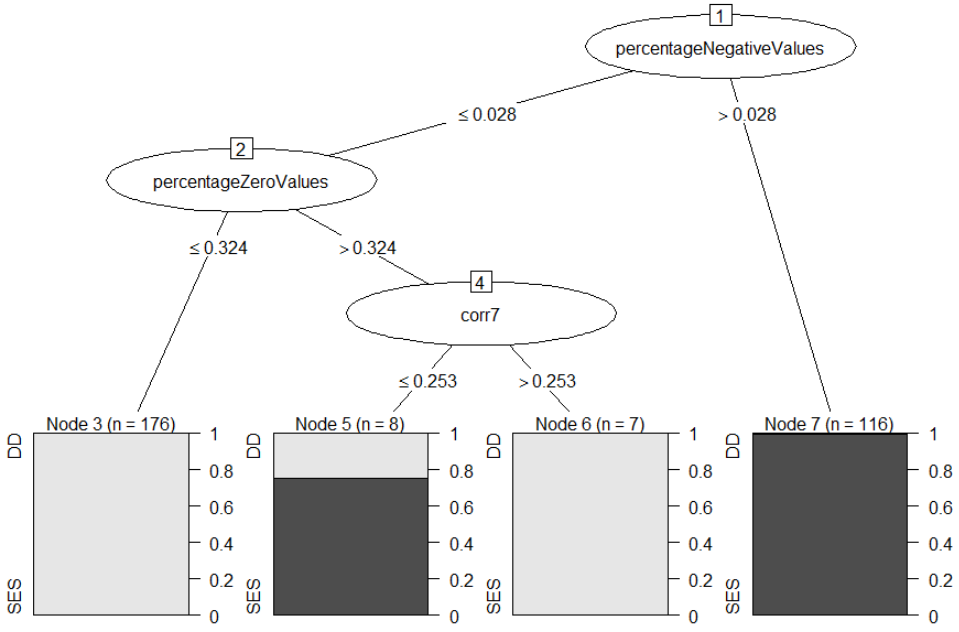


Figure 5.1: Decision tree (Dark grey: SES, light grey; Degree-days)

Random forest

To determine which attributes are most important for the random forest, we use the *mean decrease accuracy* which is based on the fact that the more the accuracy of a random forest decreases when a specific attribute is excluded, the more important the attribute is. The random forest package in R is able to give a visualisation of the mean decrease accuracy. Figure 5.2 shows the ranking of the attributes in terms of importance. We find that the vari-

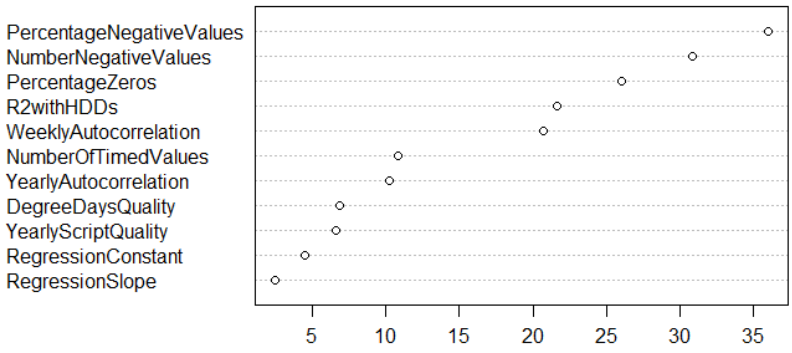


Figure 5.2: Mean decrease accuracy

ables that we described as being important characteristics for 'Category 2' datasets (negative usage, many timed values of zero, and a low R^2), are most important in the random forest.

Just as for the decision tree method, we perform random forest in WEKA as well as in R, because they use slightly different settings. The same differences hold as in decision trees since a random forest is an extension of the decision tree method. Table 5.1 shows the results of both the WEKA and R implementation. Both result in an accuracy of 98.4%.

Table 5.1: Confusion matrix Random Forest

	WEKA	Predicted Class	
		Degree-days	*SES*
Actual Class	*Degree-days*	183	3
	SES	2	119

	R	Predicted Class	
		Degree-days	*SES*
Actual Class	*Degree-days*	183	3
	SES	2	119

k-Nearest Neighbour (kNN)

In WEKA, the kNN method is called IBk (instance-based learning with parameter k). The k parameter determines how many neighbours to consider when classifying a test instance. The outcome is determined by majority. For example when $k = 5$, and 4 of the neighbours are classified as belonging to Class 1 and one of the neighbours is classified as Class 0, then the prediction class is 1. R is a little more advanced in this since the 'knn' package tries different values for k (number of neighbours) and picks the one resulting in the best accuracy. In WEKA this must be done manually.
The number of neighbours that R finds as best is $k = 5$, so we use this parameter as well in WEKA. Table 5.2 shows the results of both the WEKA and the R implementation. WEKA results in an accuracy of 97.7% and R 93.5%.

Table 5.2: Confusion matrix kNN

	WEKA	Predicted Class	
		Degree-days	*SES*
Actual Class	*Degree-days*	181	5
	SES	2	119

	R	Predicted Class	
		Degree-days	*SES*
Actual Class	*Degree-days*	185	1
	SES	19	102

Artificial Neural Networks

In R, we use the 'nnet' function that uses a multilayer perceptron, as explained in Section 2.8. WEKA has a model called 'MultilayerPerceptron'. The advantage of R is that it automatically optimises the number of hidden layers and the decay rate. The latter is used to decrease the learning rate (i.e. how big the steps are you take towards a local minimum) each epoch (one full training cycle on the training set). This is done as follows:

$$\text{LearningRate} = \text{LearningRate} * 1/(1 + \text{decay} * \text{epoch}) \tag{5.1}$$

The advantage of using decay is that each epoch, smaller steps towards the optimum are taken so the chance becomes smaller of stepping over it.
We use the outcome R gives, 5 hidden layers. The decay rate is not a configuration in WEKA, you can only turn it on or off. Using decay improves the results in our problem since without using decay, the accuracy is 96.7% whereas it is 97.7% when we do use decay. WEKA and R both result in an accuracy of 97.7%.

Table 5.3: Confusion matrix ANN

	WEKA and R	Predicted Class	
		Degree-days	*SES*
Actual Class	*Degree-days*	181	5
	SES	2	119

Logistic regression

In WEKA, there are two models that we can use for logistic regression: Logistic, and SimpleLogistic. The most important difference is that SimpleLogistic has a built-in attribute selection, therefore, we use that model. In R, there are two widely used models: 'glm' and 'plr'. 'Glm' is the function used for generalised linear models of which logistic regression is a subtype. 'Plr' is the function for penalised logistic regression in which the estimation of coefficients is optimised by using a certain penalty in order to make sure that fitting the coefficients happens in a stable fashion. Besides, penalised logistic regression also performs attribute selection and the parameters used for penalising and attribute selection are automatically optimised using grid-search. Because of this advantages, we use the 'plr' function for logistic regression in R.

Let us look at the model that WEKA produced in order to understand how to interpret the outcomes. The regression equation that WEKA gives is:

$$\begin{aligned} y^*_{Degree-days} = 0.7407 + (R^2 \times 2.346) + (\text{Weekly autocorrelation} \times 1.083) + \\ (\text{Percentage zeros} \times -2.184) + (\text{Percentage negative usage} \times -26.591) \end{aligned} \tag{5.2}$$

The probability of belonging to this class is calculated as follows:

$$p_{Degree-days} = \frac{1}{1 + \exp^{-y^*}} \tag{5.3}$$

The higher y^*, the higher the probability of belonging to that class. Therefore, the coefficients with which we multiply the attribute values, give an indication on what the effect of a certain attribute is on the probability of belonging to a class. For example, let us look at the first attribute in the regression equation: R^2. We see that the coefficient, β_{R^2}, is a positive number, 2.346, which means that the higher the R^2 when correlating with HDDs, the higher the probability of belonging to class 'Degree-days'. The same holds for the attribute 'Weekly autocorrelation'. However, the other two attributes lower the probability of belonging to this class since they have negative coefficients. So an instance with an R^2 of 73.1%, a weekly autocorrelation of 0.72, zero negative usage measurements, and no zero values has an y^* of 3.215 and therefore, a probability of $\exp(3.215)/(\exp(3.215)+1)$ of belonging to class 'Degree-days', which is 0.961. The probability of belonging to class 'SES' is $1 - p_{Degree-days}$.

The model resulting from the 'plr' function in R includes all 11 attributes. WEKA results in an accuracy of 98.4% and R 97.4%.

5.3 Conclusion

In this chapter, we aim at answering the final research question: 'How can classification methods be used for automatic model selection?'. We wanted to find out how the methods

Table 5.4: Confusion matrix logistic regression

	WEKA	Predicted Class	
		Degree-days	*SES*
Actual Class	*Degree-days*	183	3
	SES	2	119

	R	Predicted Class	
		Degree-days	*SES*
Actual Class	*Degree-days*	184	2
	SES	6	115

should be used and which of them performs best. We implemented the methods in both WEKA and R.
Table 5.5 gives the functions that WEKA and R use for the methods found in literature.

Table 5.5: Performance classification methods

	Method used	
	WEKA	R
Decision tree	J48	J48
Random forest	RandomForest	randomForest
k-Nearest Neighbour	IBk	knn
ANN	MultilayerPerceptron	nnet
Logistic regression	SimpleLogistic	glm

We conclude the following:

1. The advantage of the methods in R is that they automatically optimise the parameters whereas in WEKA, we have to choose them manually
2. In WEKA, generally more configurations can be set which gives the user more freedom, it is easier to adjust the validation method, and there are more data preprocessing possibilities
3. Table 5.6 gives the accuracy of the implemented methods for both WEKA and R For

Table 5.6: Performance classification methods

	Correctly classified (%)	
	WEKA	R
Decision tree	96.7	95.8
Random forest	98.4	98.4
k-Nearest Neighbour	97.7	93.4
ANN	97.7	97.7
Logistic regression	98.4	97.4

   all methods, the methods in WEKA perform best or equal to the methods implemented in R
4. The methods that perform best in WEKA are random forest and logistic regression. In R the best performing method is random forest and secondly ANN. The differences in accuracy are small so we should also look at ease of implementation and understandability. The method that performs best when taking into account those criteria is logistic regression

We therefore recommend to classify the instances with logistic regression using WEKA to determine the regression function since it automatically selects the attributes and R does not.

Chapter 6

Conclusion and recommendations

This final chapter concludes this research and answers the research question. Section 6.1 answers the research question, Section 6.2 proposes several recommendations based on the conclusions, and Section 6.3 gives suggestions for further research.

6.1 Conclusion

The research question we aim to answer in this thesis is:

Can, and if so, how can the forecast performance of LPG demand be improved?

We conclude that the forecast performance can indeed be improved. In general, the solution is threefold: the first focuses on data cleaning, the second on the current forecasting procedure, and the third on automatic model selection by using classification.

The largest improvement is realised by cleaning the data. Especially the solution where we send the negative usage to the forecasting engine instead of discarding it, results in great improvements. For the storages that we tested, discarding negative usage leads to the usage used as forecast input being 134% of an LPG tank (defined as the maximum volume the tank may be filled with minus the minimum volume) higher than the actual usage. Of the 2284 storages that we considered, 38% are 'Category 2' datasets so this data problem is of substantial size and sending negative usage to the forecasting engine reduces the number of deliveries to 'Category 2' storages by 87%.

Another solution we propose, concerns the current forecasting procedure. Currently, for technical reasons, the forecasting method based on heating degree-days is implemented by first removing seasonality, then forecasting the remaining series with simple exponential smoothing, and finally adding the seasonality again. However, from our analysis we conclude that in most cases, simple linear regression with HDDs as predictor, performs better (6.5% improvement of the RMSE of 'Category 4). Besides, currently only 'Category 4' series are predicted using the degree-days method whereas we found that 'Category 1' and 'Category 3' datasets also benefit from this method (on average 67% improvement of the MAPE for 'Category 1', and 11.3% improvement of the RMSE for 'Category 3'). In fact, simple exponential smoothing (the current method), turns out to be one of the worst performing methods for

'Category 1' datasets.

Finally, we addressed classification. Currently, the user has had to pick a suitable forecasting script manually per storage. Since there are tens of thousands of them, this is a cumbersome way of working and automating this process could save time. Performing several forecasting methods and choosing the one that performs best is not an option since it takes too much time. Besides, we needed classification to find out how much storages belonged to which data category. After trying different classification methods, we can conclude that in performance as well as in interpretability and ease of implementation, logistic regression is the best method for this data and classifies the instances with 98.4% accuracy in WEKA.

6.2 Recommendations

Based on these conclusions, we recommend:

- Send the negative usage to the forecasting engine instead of discarding it
- Make the after delivery readings irrelevant, except for 'Category 1' datasets
- Implement Cook's distance before calculating the regression coefficients
- Use the RMSE instead of the MAPE as performance indicator
- Forecast 'Category 2' datasets with simple exponential smoothing and the rest with the degree-days method
- Implement simple linear regression for the degree-days method
- Compute a tracking signal to monitor whether the forecasting system remains in control using an α of 0.1 and control limits of ± 0.55, but only for 'Category 3' and 'Category 4' datasets
- Use logistic regression as classification method

Concerning the order in which we believe these recommendations should be implemented, sending negative usage to the forecasting engine has top priority because this generates the biggest improvement. Also, we would prioritize making the after delivery readings irrelevant. This will lead to the degree-days method being chosen more often since removing the peaks caused by these readings, improves the R^2 when correlating usage with HDDs a lot. After making the input data more reliable, we recommend to forecast each data category with the method we propose.
There are some recommendations that we already implemented in the current framework which are currently under review and when accepted, will be included in the actual product. These are: Cook's distance, and simple linear regression as implementation for the degree-days method.

6.3 Suggestions for further research

An important shortcoming of this thesis is that we did not investigate the actual impact of the problem and our improvements. Naturally, improving forecast performance will improve the inventory routing process, but we do not know how much exactly. We found that bad forecasts lead to not being able to replenish all customers on the planned route (in 38% of the routes) or having to find another customer when LPG remains in the truck after having

visited all customers (we do not know how often this happens). It would be interesting to keep track of this percentage after having implemented the proposed improvements. Moreover, more research is required to determine what the potential cost reduction is of improved forecasts and the other recommendations of this thesis.

Secondly, this research is performed specifically for LPG demand data of Company X. However, ORTEC has more clients similar to Company X within OIR (ORTEC Inventory Routing) for which the findings could also apply. A suggestion for further research would be to see whether these customers show the same data categories and data issues like peaks after delivery and negative usage and therefore could benefit from the findings of this thesis.

Thirdly, while writing the thesis, we also looked at other forecasting cases within ORTEC. Two problems that are essential in automating forecasting processes are trend breaks and holiday effects. After talking to several persons that have been forecasting at ORTEC, these were the two main issues that came up. We found that for a supermarket that wanted ORTEC to predict their bread demand, the effect of certain holidays on the days around that holiday is different depending on the day of the week the holiday falls on. For example, the effect was different when a holiday was on a Tuesday or a Saturday. Besides, there is no solid way to cope with trend breaks. We came across these in the Company X case as well since it could occur that a customer first did not have a telemetry system but got one later which changes the pattern of the data.

Another suggestion is to investigate the sensitivity of the forecasts to changes in temperature. For this research, we used realised temperatures for the degree-days method and linear regression, but in reality, weather forecasts are used. It would be interesting to know to what extent the forecasts are influenced by temperature changes. This is especially interesting when forecasts far into the future are required for which weather predictions are not that accurate.

Bibliography

[1] Adeodato, P. J., Vasconcelos, G. C., Arnaud, A. L., Santos, R. A., Cunha, R. C., & Monteiro, D. S. (2004). Neural Networks vs Logistic Regression: a Comparative Study on a Large Data Set. In *ICPR (3)* (pp. 355-358).

[2] Bates, J. M., & Granger, C. W. (1969). The combination of forecasts. *Or,* 451-468.

[3] Bermúdez, J. D. (2013) Exponential smoothing with covariates applied to electricity demand forecast. *European Journal of Industrial Engineering,* 7(3), 333-349.

[4] Bessec, M., & Fouquau, J. (2008). The non-linear link between electricity consumption and temperature in Europe: a threshold panel approach. *Energy Economics, 30*(5), 2705-2721.

[5] Box, G. E. (1976). Science and statistics. *Journal of the American Statistical Association, 71*(356), 791-799.

[6] Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5-32.

[7] Çetinkaya, S., & Lee, C. Y. (2000). Stock replenishment and shipment scheduling for vendor-managed inventory systems. *Management Science, 46*(2), 217-232.

[8] Chatfield, C. (1978). The holt-winters forecasting procedure. *Applied Statistics*, 264-279.

[9] Chatfield, C., & Yar, M. (1988). Holt-Winters forecasting: some practical issues. *The Statistician,* 129-140.

[10] Chatfield, C. (1998). Durbin-Watson test. *Encyclopedia of Biostatistics.*

[11] Chatfield, C. (2006). What is the 'best' method of forecasting? *Journal of Applied Statistics*, 15(1), 19-38.

[12] Cook, R. D., & Weisberg, S. (1982). *Residuals and influence in regression.* New York: Chapman and Hall.

[13] De Gooijer, J. G., Hyndman, R. J. (2006). 25 years of time series forecasting. *International journal of forecasting*, 22(3), 443-473.

[14] Delen, D. (2011). Predicting student attrition with data mining methods. *Journal of College Student Retention: Research, Theory & Practice, 13*(1), 17-35.

[15] Deutsch, M., Granger, C. W., & Teräsvirta, T. (1994). The combination of forecasts using changing weights. *International Journal of Forecasting, 10*(1), 47-57.

[16] Drazin, S., & Montag, M. (2012). Decision tree analysis using weka. *Machine Learning-Project II, University of Miami,* 1-3.

[17] Energy Lens (2016). Degree Days - Handle with care! Retrieved on January 2nd, 2017, from http://www.energylens.com/articles/degree-days

[18] Fan, S., & Hyndman, R. J. (2010). Forecast short-term electricity demand using semi-parametric additive model. In *Universities Power Engineering Conference (AUPEC), 2010 20th Australasian* (pp. 1-6). IEEE.

[19] Fazeli, R., Ruth, M., & Davidsdottir, B. (2016). Temperature response functions for residential energy demand–A review of models. *Urban Climate, 15*, 45-59.

[20] García-Díaz, J. C., & Trull, Ó. (2016). Competitive Models for the Spanish Short-Term Electricity Demand Forecasting. In *Time Series Analysis and Forecasting* (pp. 217-231). Springer International Publishing.

[21] Gardner, E. S. (1983). Automatic monitoring of forecast errors. *Journal of Forecasting, 2*(1), 1-21.

[22] Gardner, E. S. (1985). Exponential smoothing: The state of the art. *Journal of forecasting,* 4(1), 1-28.

[23] Gardner Jr, E. S., & McKenzie, E. D. (1985). Forecasting trends in time series. *Management Science, 31*(10), 1237-126.

[24] Gardner Jr, E. S., & McKenzie, E. (1988). Model identification in exponential smoothing. *Journal of the Operational Research Society,* 39(9), 863-867.

[25] Gardner, E. S. (2006). Exponential smoothing: The state of the art—Part II. *International journal of forecasting,* 22(4), 637-666.

[26] Gelper, S., Fried, R., & Croux, C. (2010). Robust forecasting with exponential and Holt–Winters smoothing. *Journal of forecasting, 29*(3), 285-300.

[27] Göb, R., Lurz, K., & Pievatolo, A. (2013). Electrical load forecasting by exponential smoothing with covariates. *Applied Stochastic Models in Business and Industry, 29*(6), 629-645.

[28] Goodwin, P. (2010). The holt-winters approach to exponential smoothing: 50 years old and going strong. *Foresight, 19*, 30-33.

[29] Hanke, J. E., Reitsch, A. G. (1998). *Business Forecasting* (sixth edition). Upper Saddle River, New Jersey: Prentice-Hall.

[30] Hibon, M., & Evgeniou, T. (2005). To combine or not to combine: selecting among forecasts and their combinations. *International Journal of Forecasting, 21*(1), 15-24.

[31] Hippert, H. S., Pedreira, C. E., & Souza, R. C. (2001). Neural networks for short-term load forecasting: A review and evaluation. *IEEE Transactions on power systems, 16*(1), 44-55.

[32] Hoshmand, A. R. (2009). *Business Forecasting; A Practical Approach.* Routledge.

[33] Hyndman, R. J., & Athanasopoulos, G. (2014). *Forecasting: principles and practice.* OTexts.

[34] Hyndman, R. J., & Fan, S. (2010). Density forecasting for long-term peak electricity demand. *IEEE Transactions on Power Systems, 25*(2), 1142-1153.

[35] Hyndman, R., Koehler, A. B., Ord, J. K., & Snyder, R. D. (2008). *Forecasting with exponential smoothing: the state space approach.* Springer Science & Business Media.

[36] Hyndman, R. J., & Kostenko, A. V. (2007). Minimum sample size requirements for seasonal forecasting models. *Foresight, 6*(Spring), 12-15.

[37] Kaytez, F., Taplamacioglu, M. C., Cam, E., & Hardalac, F. (2015). Forecasting electricity consumption: A comparison of regression analysis, neural networks and least squares support vector machines. *International Journal of Electrical Power & Energy Systems, 67*, 431-438.

[38] Kiang, M. Y. (2003). A comparative assessment of classification methods. *Decision Support Systems, 35*(4), 441-454.

[39] Kim, C., & Storer, B. E. (1996). Reference values for Cook's distance. Communications in *Statistics-Simulation and Computation, 25*(3), 691-708.

[40] Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai* (Vol. 14, No. 2, pp. 1137-1145).

[41] Kumru, M., & Kumru, P. Y. (2015). Calendar-based short-term forecasting of daily average electricity demand. In *Industrial Engineering and Operations Management (IEOM), 2015 International Conference* on (pp. 1-5). IEEE.

[42] Lahiri, R. (2006). *Comparison of data mining and statistical techniques for classification model* (Doctoral dissertation, Faculty of the Louisiana State University and Agricultural and Mechanical College in partial fulfillment of the requirements for the degree of Master of Science in The Department of Information Systems & Decision Sciences by Rochana Lahiri BE, Jadavpur University, India).

[43] Mensah, J. T. (2014). Modelling demand for liquefied petroleum gas (LPG) in Ghana: current dynamics and forecast. *OPEC Energy Review, 38*(4), 398-423.

[44] Mirasgedis, S., Sarafidis, Y., Georgopoulou, E., Lalas, D. P., Moschovits, M., Karagiannis, F., & Papakonstantinou, D. (2006). Models for mid-term electricity demand forecasting incorporating weather influences. *Energy, 31*(2), 208-227.

[45] Moral-Carcedo, J., & Vicens-Otero, J. (2005). Modelling the non-linear response of Spanish electricity demand to temperature variations. *Energy economics*, 27(3), 477-494.

[46] Nau, R. (2017). Three types of forecasts: estimation, validation, and the future Retrieved on January 26th, 2017, from https://people.duke.edu/∼rnau/three.htm

[47] Ord, K., & Fildes, R. (2013). *Principles of business forecasting.* Cengage Learning.

[48] OTexts (2017). A taxonomy of exponential smoothing methods. Retrieved on January 5th, 2017, from https://www.otexts.org/fpp/7/6

[49] Pardo, A., Meneu, V., & Valor, E. (2002). Temperature and seasonality influences on Spanish electricity load. *Energy Economics, 24*(1), 55-70.

[50] Parikh, J., Purohit, P., & Maitra, P. (2007). Demand projections of petroleum products and natural gas in India. *Energy, 32*(10), 1825-1837.

[51] Pegels, C. C. (1969). Exponential forecasting: some new variations. *Management Science,* 311-315.

[52] Pennings, C. L., van Dalen, J., & van der Laan, E. A. (2017). Exploiting elapsed time for managing intermittent demand for spare parts. *European Journal of Operational Research, 258*(3), 958-969.

[53] Price, D. H. R., & Sharp, J. A. (1986). A comparison of the performance of different univariate forecasting methods in a model of capacity acquisition in UK electricity supply. *International Journal of Forecasting, 2*(3), 333-348.

[54] Psiloglou, B. E., Giannakopoulos, C., Majithia, S., & Petrakis, M. (2009). Factors affecting electricity demand in Athens, Greece and London, UK: A comparative assessment. *Energy, 34*(11), 1855-1863.

[55] Ravinder, H. V. (2016). Determining The Optimal Values Of Exponential Smoothing Constants-Does Solver Really Work?. *American Journal of Business Education (Online), 9*(1), 39.

[56] Reid, R. D., & Sanders, N. R. (2005). *Operations management: an integrated approach.* Hoboken, NJ: John Wiley.

[57] Sailor, D. J., & Muñoz, J. R. (1997). Sensitivity of electricity and natural gas consumption to climate in the USA—methodology and results for eight states. *Energy, 22*(10), 987-998.

[58] Sarak, H., & Satman, A. (2003). The degree-day method to estimate the residential heating natural gas consumption in Turkey: a case study. *Energy, 28*(9), 929-939.

[59] Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika, 52*(3-4), 591-611.

[60] Slack, N., Chambers, S., & Johnston, R. (2010). *Operations Management.* Pearson education.

[61] Soldo, B. (2012). Forecasting natural gas consumption. *Applied Energy, 92,* 26-37.

[62] Starkweather, J., & Moske, A. K. (2011). Multinomial logistic regression. *Consulted page at September 10th: http://www. unt. edu/rss/class/Jon/Benchmarks/MLR_JDS_Aug2011. pdf, 29,* 2825-2830.

[63] Strobl, C., Boulesteix, A. L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC bioinformatics, 9*(1), 307.

[64] Suganthi, L., & Samuel, A. A. (2012). Energy models for demand forecasting-A review. *Renewable and sustainable energy reviews*, 16(2), 1223-1240.

[65] Syntetos, A. A., & Boylan, J. E. (2001). On the bias of intermittent demand estimates. *International journal of production economics, 71*(1), 457-466.

[66] Szoplik, J. (2015). Forecasting of natural gas consumption with artificial neural networks. *Energy, 85*, 208-220.

[67] Tan, P. N. (2006). *Introduction to data mining.* Pearson Education India.

[68] Taylor, J. W. (2003). Short-term electricity demand forecasting using double seasonal exponential smoothing. *Journal of the Operational Research Society,* 54(8), 799-805.

[69] Taylor, J. W. (2010). Triple seasonal methods for short-term electricity demand forecasting. *European Journal of Operational Research,* 204(1), 139-152.

[70] Teunter, R. H., Syntetos, A. A., & Babai, M. Z. (2011). Intermittent demand: Linking forecasting to inventory obsolescence. *European Journal of Operational Research, 214*(3), 606-615.

[71] Thornton, H. E., Hoskins, B. J., & Scaife, A. A. (2016). The role of temperature in the variability and extremes of electricity and gas demand in Great Britain. *Environmental Research Letters, 11*(11), 114015.

[72] Trigg, D. W. (1964). Monitoring a forecasting system. *Journal of the Operational Research Society, 15*(3), 271-274.

[73] Wang, S. (2006). *Exponential smoothing for forecasting and Bayesian validation of computer models* (Doctoral dissertation, Georgia Institute of Technology).

[74] Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2011). *Data Mining: Practical machine learning tools and techniques.* Morgan Kaufmann.

[75] Zhang, G., Patuwo, B. E., & Hu, M. Y. (1998). Forecasting with artificial neural networks: The state of the art. *International journal of forecasting, 14*(1), 35-62.

[76] Zhang, G. P., & Qi, M. (2005). Neural network forecasting for seasonal and trend time series. *European journal of operational research, 160*(2), 501-514.

Appendix A

Correlations external variables

The following correlation diagram (Figure A.1) gives the correlation between the climatological factors and LPG usage (of the 21 aggregated datasets of 'Category 4'). When using 0,700 as criterion, none of the climatological factors are highly correlated. Only sun duration and humidity seem to be a bit correlated. It might be wise to exclude one of these when using regression since otherwise multicollinearity could occur.

Correlations

			Demand	HDD	PrecipitationAmount	PrecipitationDuration	GlobalRadiation	SunDuration	Humidity	Windspeed
Spearman's rho	Demand	Correlation Coefficient	1,000	,914**	,089*	,164**	-,658**	-,383**	,336**	,196**
		Sig. (2-tailed)	.	,000	,018	,000	,000	,000	,000	,000
		N	700	699	700	700	700	700	700	700
	HDD	Correlation Coefficient	,914**	1,000	,026	,118**	-,591**	-,317**	,313**	,090*
		Sig. (2-tailed)	,000	.	,488	,002	,000	,000	,000	,017
		N	699	699	699	699	699	699	699	699
	PrecipitationAmount	Correlation Coefficient	,089*	,026	1,000	,353**	-,144**	-,386**	,272**	,385**
		Sig. (2-tailed)	,018	,488	.	,000	,000	,000	,000	,000
		N	700	699	848	848	848	848	848	848
	PrecipitationDuration	Correlation Coefficient	,164**	,118**	,353**	1,000	-,397**	-,260**	,200**	,200**
		Sig. (2-tailed)	,000	,002	,000	.	,000	,000	,000	,000
		N	700	699	848	849	849	849	849	849
	GlobalRadiation	Correlation Coefficient	-,658**	-,591**	-,144**	-,397**	1,000	,556**	-,552**	-,193**
		Sig. (2-tailed)	,000	,000	,000	,000	.	,000	,000	,000
		N	700	699	848	849	849	849	849	849
	SunDuration	Correlation Coefficient	-,383**	-,317**	-,386**	-,260**	,556**	1,000	-,674**	-,213**
		Sig. (2-tailed)	,000	,000	,000	,000	,000	.	,000	,000
		N	700	699	848	849	849	849	849	849
	Humidity	Correlation Coefficient	,336**	,313**	,272**	,200**	-,552**	-,674**	1,000	-,177**
		Sig. (2-tailed)	,000	,000	,000	,000	,000	,000	.	,000
		N	700	699	848	849	849	849	849	849
	Windspeed	Correlation Coefficient	,196**	,090*	,385**	,200**	-,193**	-,213**	-,177**	1,000
		Sig. (2-tailed)	,000	,017	,000	,000	,000	,000	,000	.
		N	700	699	848	849	849	849	849	849

**. Correlation is significant at the 0.01 level (2-tailed).

Figure A.1: Correlation climatological factors

The correlation diagram in Figure A.2 shows the correlation between LPG demand and LPG price.

The variables that show the highest correlation with LPG demand are: Heating Degree Days, global radiation, LPG price. Precipitation amount and -duration both have a rather low correlation with demand. The correlation in these diagrams is based on the aggregated data of 'Category 4' datasets. We know of these that demand is temperature dependent. It might be possible that even within these datasets there is some difference in dependence on climatological- and economical factors. Firstly, as stated earlier, dependence on climatological factors depends on the ratio of building envelope surface and on the dominance of process

Correlations

			Demand	LPGprice
Spearman's rho	Demand	Correlation Coefficient	1,000	,405**
		Sig. (2-tailed)	.	,000
		N	700	700
	LPGprice	Correlation Coefficient	,405**	1,000
		Sig. (2-tailed)	,000	.
		N	700	700

**. Correlation is significant at the 0.01 level (2-tailed).

Figure A.2: Correlation economical factors

needs of the customers (Fazeli, Ruth, & Davidsdottir, 2016). Secondly, different relationships could be found in the other data categories. Literature indicated that holidays and weekends also could affect LPG demand. This is however hard to test on the Company X datasets since no daily data is available for most of them.

Appendix B

Statistical tests regression models

Durbin-Watson test

When a multiple regression model is fitted to time series data, successive residuals are often found to be dependent and, therefore, autocorrelated (Chatfield, 1998). When autocorrelated errors occur in a model, it might be possible to improve the model to get a better fit and better forecasts. The Durbin-Watson test is a way of checking this.
The hypotheses for the Durbin-Watson test are:

- H_0 = no first order autocorrelation
- H_1 = first order autocorrelation exists

A residual at time t for $t = 1, 2, \ldots, n$, is the difference between observed value Y_t and the predicted value from the regression model and is calculated by:

$$\hat{z} = Y_t - \hat{\beta}_1 x_{1t} - \cdots - \hat{\beta}_k x_{kt} \tag{B.1}$$

where
k is the number of explanatory variables
$\hat{\beta}_1, \ldots, \hat{\beta}_k$ are the fitted coefficients of the regression model
The Durbin-Watson statistic is then given by:

$$d = \frac{\sum_{t=2}^{n} (\hat{z}_t - \hat{z}_{t-1})^2}{\sum_{t=1}^{n} \hat{z}_t^2} \tag{B.2}$$

The null hypothesis of independence depends on the value of k (the number of explanatory variables) and x values, as well as on n (the number of time periods). Therefore, it is not possible to give a single critical value for d, but instead, an upper and lower critical value are given (d_L and d_U) (Chatfield, 1998). In our example, $k = 5$ and $n = 700$. The rule of thumb is that the null hypothesis is accepted when $1.5 < d < 2.5$, which is the case here since our d-value is 1.622. However, the critical value table for our k and n values, gives $d_L = 1.864$ and $d_U = 1.887$ which would indicate that our null hypothesis is rejected. It is therefore doubtful whether serious autocorrelation exists in this case.

Shapiro-Wilk test

This test is a test of normality. The hypotheses of the Shapiro-Wilk test are:

- H_0 = sample comes from a normally distributed population
- H_1 = sample did not come from a normally distributed population

The test statistic is:

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{B.3}$$

where
$x_{(i)}$ is the ith smallest number in the sample
$\bar{x}$ is the sample mean
$(a_1, \ldots, a_n) = \frac{m^T V^{-1}}{(m^T V^{-1} V^{-1} m)^{1/2}}$
where
$m = (m_1, \ldots, m_n)^T$
and $m_1, \ldots, m_n$ are the expected values of standard normal order statistics, and V is the covariance matrix of these (Shapiro & Wilk, 1965).

The value of the W-statistic of our example is 0.991 and the test is significant which means that the null hypothesis is not rejected and we can assume normality.

Appendix C

Data cleaning: reading after

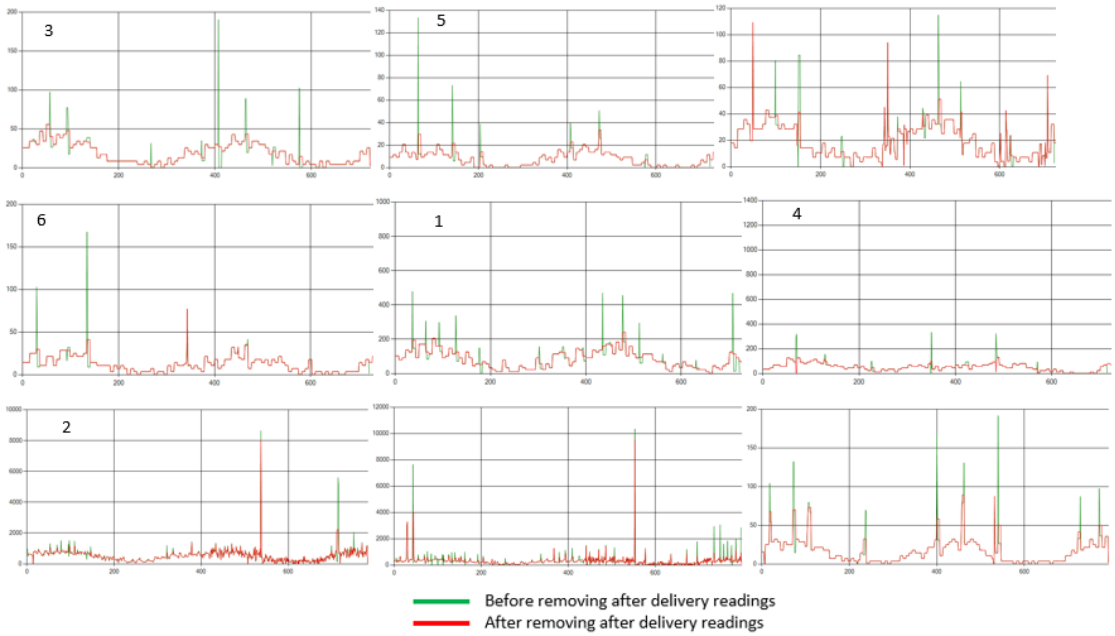


Figure C.1: Reading after made irrelevant

Appendix D

Category 1 forecasting

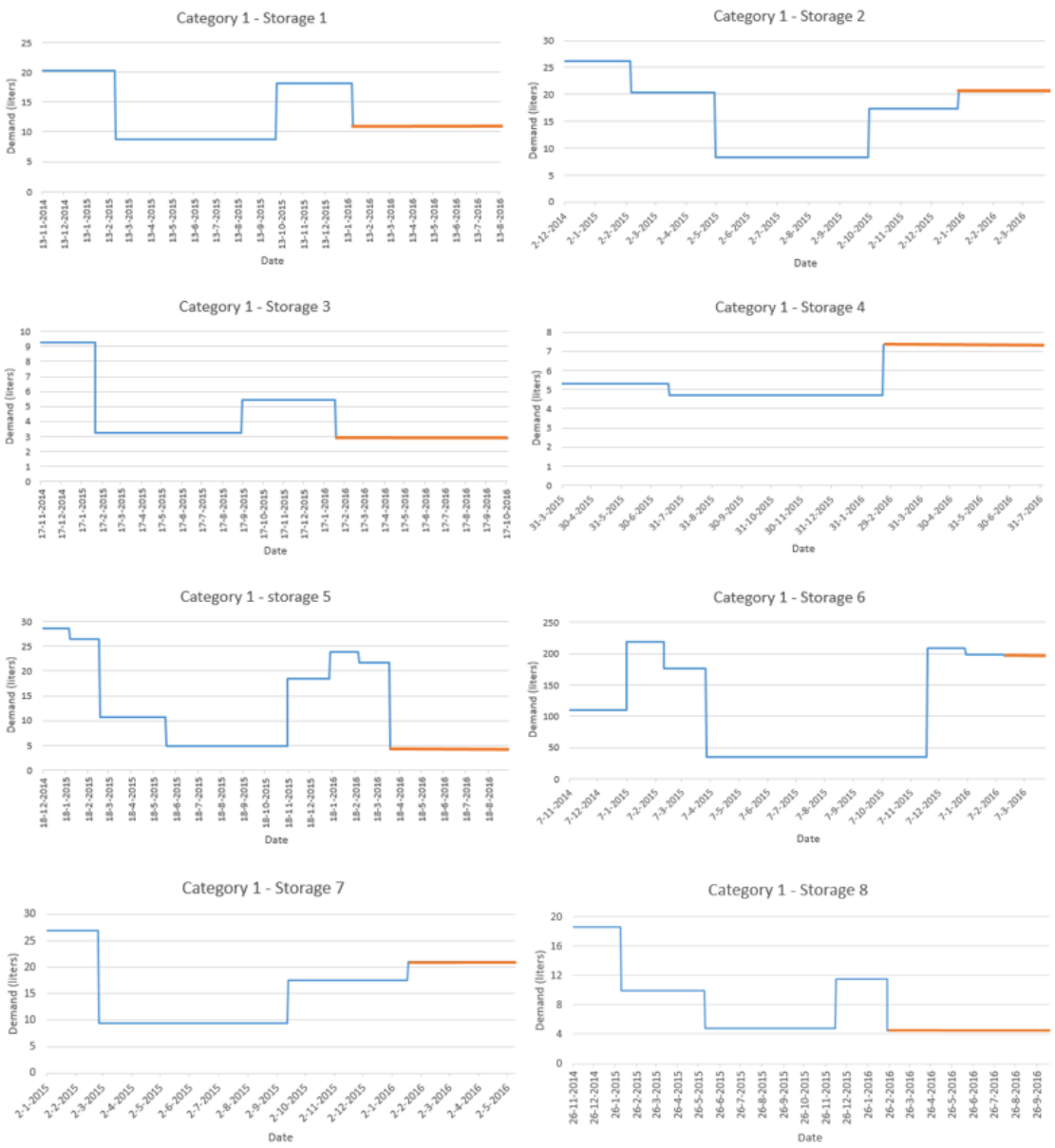


Figure D.1: 'Category 1' datasets

Appendix E

Category 2 forecasting

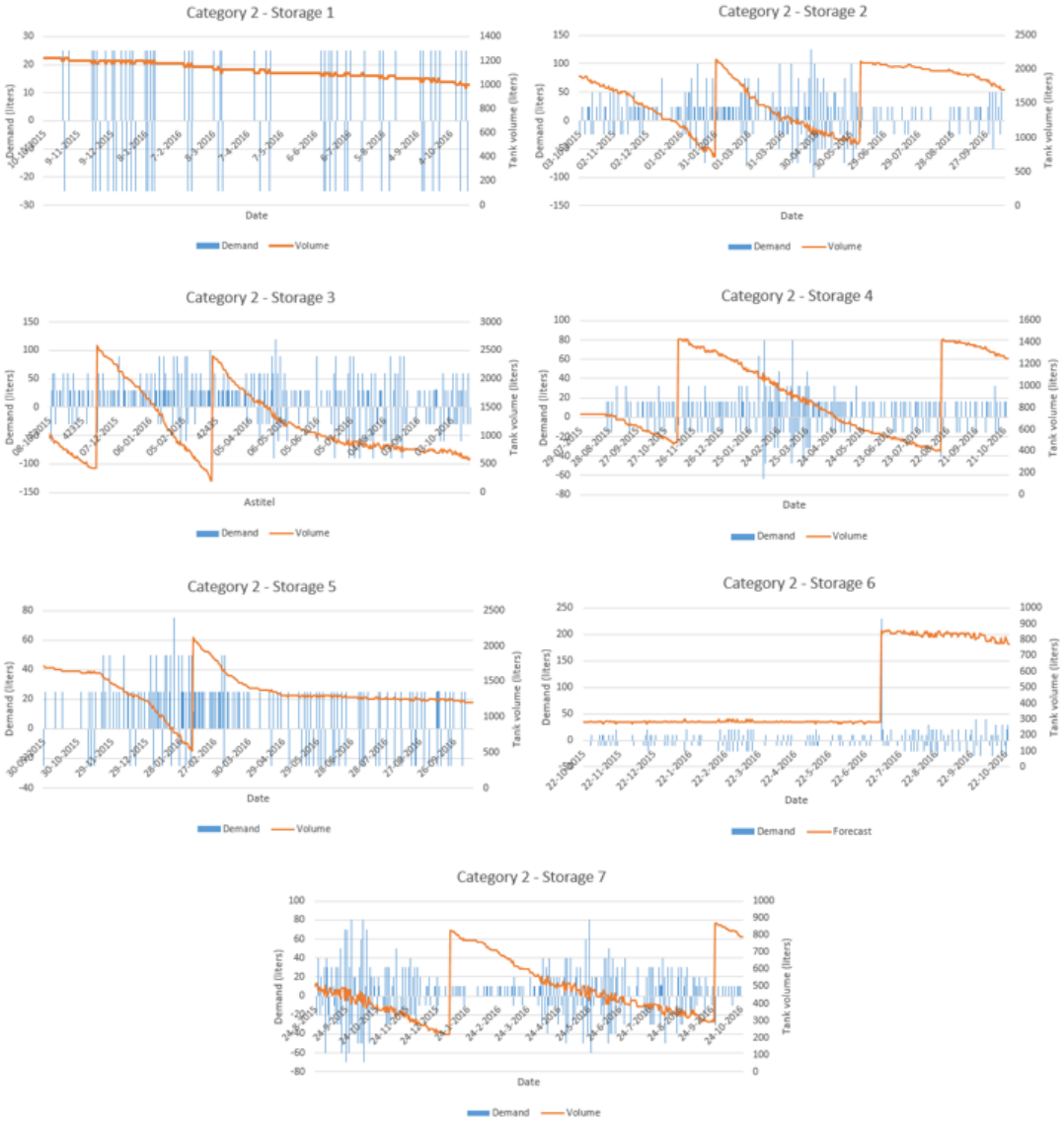


Figure E.1: 'Category 2' datasets

Appendix F

Category 3 forecasting

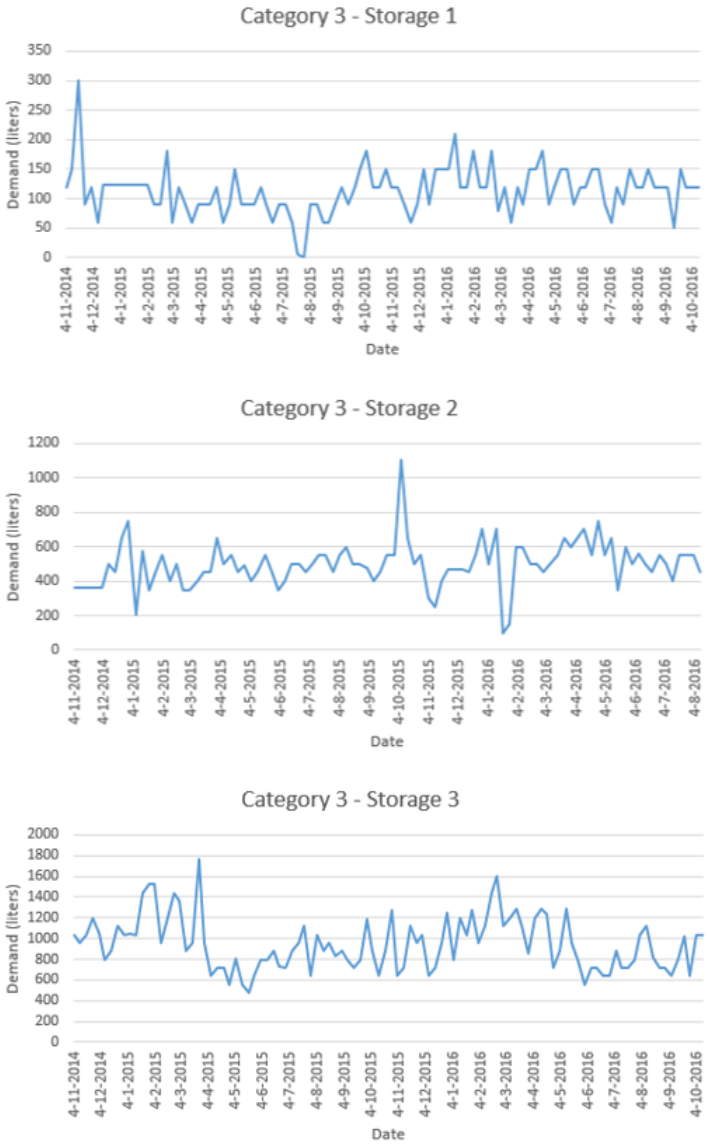


Figure F.1: 'Category 3' datasets

Appendix G

Category 4 forecasting

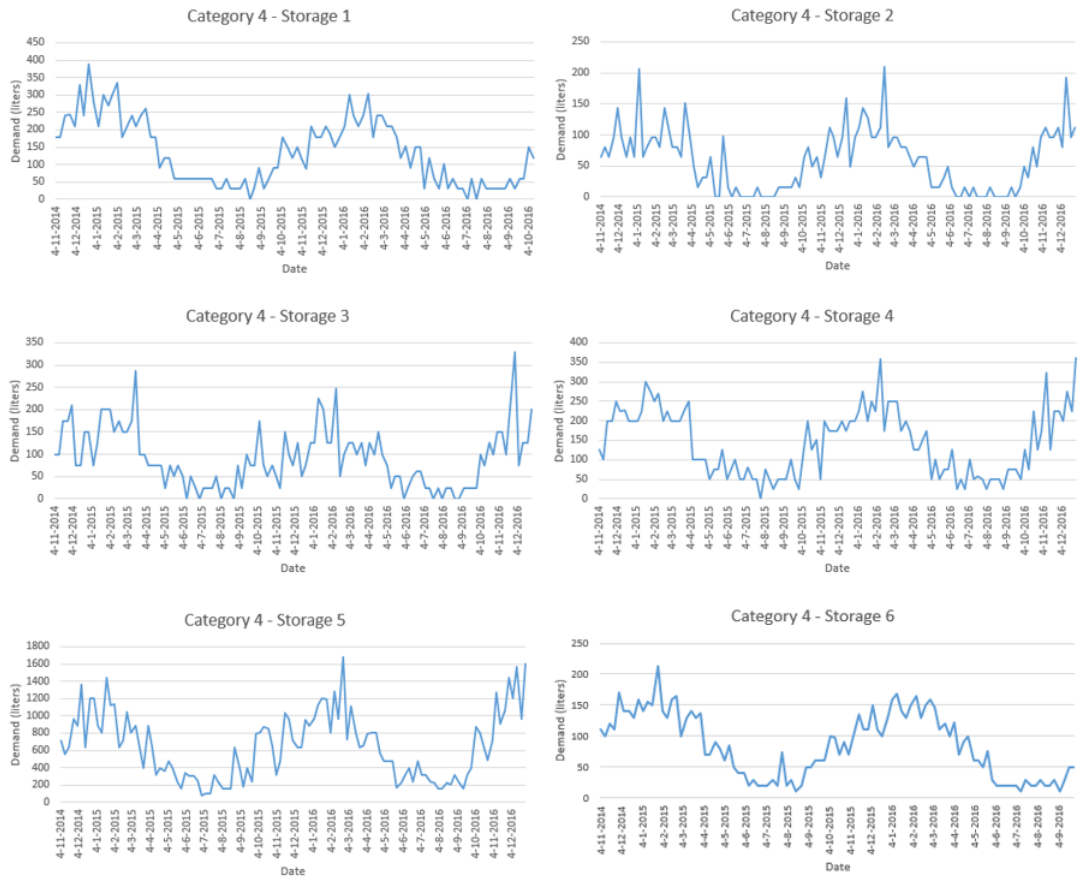


Figure G.1: 'Category 4' datasets