Gaussian process modelling of petrol sales in gas stations in Finland

Aleksandr Makarov

School of Science

Thesis submitted for examination for the degree of Master of Science in Technology.

Espoo 10.10.2020

Thesis supervisor:

Assistant Prof. Pekka Marttinen

Thesis advisors:

Doctoral candidate Onur Poyraz

Ossi Talvitie, Fatman Oy

Aalto University
School of Science

AALTO UNIVERSITY
SCHOOL OF SCIENCE

ABSTRACT OF THE
MASTER'S THESIS

Author: Aleksandr Makarov

Title: Gaussian process modelling of petrol sales in gas stations in Finland

Date: 10.10.2020 Language: English Number of pages: 5+37

Machine learning and data mining

Supervisor: Assistant Prof. Pekka Marttinen

Advisors: Doctoral candidate Onur Poyraz, Ossi Talvitie, Fatman Oy

Petrol delivery is an important challenge for the gas station networks operating hundreds of stations. At the moment, a common approach is to react to the low product level in the tanks using some IoT device and request a delivery based on an alert. The same devices are usually capable of recording the daily sales of petrol. This collected data can be used for analysis and forecasting. The array of daily petrol sales can be analyzed as a time series with modern machine learning methods. The machine learning methods can be used to model sales and accurately predict future demand. With that knowledge, the delivery can be scheduled in advance, minimizing the logistics cost for a compound product delivery problem. Bayesian machine learning, a special class of machine learning algorithms, is also capable of estimating the confidence levels of the prediction, allowing to calculate and account for the risks in case of the series deviation from the predicted values. We propose a Gaussian process model from the Bayesian machine learning methods for forecasting petrol demand for a week forward that can be used to create an optimal delivery schedule to minimize business expenses. Several forms of the model are described, and the model performance is compared to a popular model used in time series analysis SARIMA, which is used as a baseline in this thesis. The models are evaluated on a sample petrol station in Helsinki, Finland.

Keywords: forecasting, Gaussian processes, ARIMA, time series, petrol demand, petrol sales

Contents

1 Abbreviations

AIC	Akaike information criterion
AR	Autoregressive
ARIMA	Autoregressive integrated moving average
ARMA	Autoregressive moving average
BIC	Bayesian information criterion
GDP	Gross domestic product
GP	Gaussian process
GPR	Gaussian process regression
IoT	Internet of things
MA	Moving average
MAE	Mean absolute error
MAP	Maximum a posteriori estimation
MAPE	Mean absolute percentage error
ML	Maximum likelihood
MSE	Mean squared error
RBF	Radial basis function
RMSE	Root mean squared error
RQ	Rational quadratic
SARIMA	Seasonal autoregressive integrated moving average
SMS	Short Message Service
USD	United States Dollar

2 Introduction

2.1 Motivation

2.1.1 Gas station networks

Oil is the largest energy source in the world [3]. Petrol, made of oil, is an essential part of everyday life for the industrial sector and car vehicle owners. According to Statistics Finland more than 5 million vehicles were in traffic use in Finland in 2020. The petrol for those vehicles is most commonly purchased at gas stations owned by one of several gas station networks. Gas station networks are a huge business of the petrol supply chain to the consumers. Nowadays, a number of additional services are provided at numerous gas stations such as shops and canteens. The size of fuel dispenser business gross in 2019 was estimated at USD 2.5 billion and estimated to become USD 3.5 billion by 2027 [2]. There are several highly competitive gas station networks in Finland, with the largest owning hundreds of gas stations all over the country. Each company is interested in keeping its reputation high while reducing supply chain expenses. It is possible to see gas stations owned by different companies next to each other. With the modern market realities, a customer expects constant availability of the essential product at the store. Not having the sufficient supply can be harmful to the company since stops in sales may cost millions a day in direct loss and disappoint the customer, so they switch to a competitor company resulting in more prospect financial loss. Hence delivering petrol to stations in advance to keep the petrol sales process continuous is an important business challenge.

2.1.2 Fuel storage

To store the fuel for sale at the station, a single or multiple fuel tanks are usually deployed underground while local regulations and environmental concerns may require additional restrictions on the petrol storage [1]. Special temperature conditions have to be applied to the fuel storage, making larger fuel storages cost more for additional space and their maintenance. Due to that continuous delivery of the petrol to a moderate storage is desired.

2.1.3 Petrol delivery practice

An essential part of continuous delivery is tracking the level of petrol in the tank at the gas station. With the development of the Internet of Things (IoT), gas stations have been deployed with dynamic monitoring systems of the liquid level. These devices are capable of sending an alert to a system when the product is reaching the minimum safe stock level. A common communication type is SMS. The station manager receiving the alert can schedule a delivery operation to the station with the available resources - a car, driver and petrol. One manager is usually assigned to several stations. With a large number of stations, there are many deliveries scheduled every week, and the logistics costs involving hired drivers and vehicles are significant. However, the delivery routing can be optimized to reduce the logistics cost when the

refilling schedule for the week is known in advance. That is, if some nearby stations need refilling in the same week, one driver can visit them in one route rather than several routes on different days, assuming the vehicle is big enough to carry the petrol for those stations.

2.1.4 Petrol delivery optimization

In the study, we aim to model the petrol demand for stations in Finland based on a sample station. The data was collected by a maintenance resource planning service https://fatman.fi. While forecasting petrol sales has been researched before [33; 5], in this thesis we develop a forecast model to be used for logistics optimization in the petrol supply chain to the gas station. Knowing the petrol demand in advance can help create an optimal delivery schedule utilizing available drivers, cars and the road network. While routing optimization is outside of the scope of this study, it is a hot topic of the research field encompassing a large number of modern industrial challenges [13]. We study a Gaussian process model for petrol sales forecast of a gas station. Gaussian process model was chosen for the study because it was successfully utilized for time series forecast [44; 40; 47] and can estimate the uncertainty of its prediction [11]. We also consider a commonly used forecasting model ARIMA and use it as a baseline for the Gaussian process model.

2.2 Forecasting time series

Time series is a set of observed data points recorded over time [4]. Time series are used to describe how the value of some variable changes over time. Forecasting is the process of predicting the future values of the variable based on the past and present values [4]. In practice, the main aim of forecasting is to be able to predict future values of the time series. This is done via careful study of the past observations in order to create a model which describes the inherent structure of the data. Thus, forecasting can be termed as predicting the future by understanding the past [36]. The time series has a non-deterministic nature, that is, we can not possibly deduce its exact formula as in the real world, we can never predict the future with certainty. One way to model a time series is to treat it as a stochastic process. The process can be assumed to follow some probability distribution, producing the sequence of observations of the series.

ARIMA models Various forecast models can be found in the literature. Autoregressive Integrated Moving Average (ARIMA), proposed by Box et al. [7], is one of the most common stochastic time series models. It is a linear model which assumes the time series to follow a particular known statistic distribution. It has received widespread due to its simplicity and flexibility. However, it is noted that many real-world time series are in fact non-linear [4], and using the ARIMA model is

not always adequate. As a simple yet powerful model, ARIMA is used as a baseline model in this study.

Gaussian processes Gaussian processes utilize a profound Bayesian framework for machine learning problems under which information and uncertainty of the outcomes are well-defined [40]. This property lead to a high spread of the method in the research fields in the last decades. Gaussian processes are used in a multitude of applications, including classification, regression, dimensionality reduction, reinforcement learning, even anomaly detection algorithms [29]. A Gaussian process model is a non-parametric Bayesian model that describes a distribution over functions represented by an infinite number of Gaussians [38]. Mathematically, a Gaussian process model is equivalent to a linear Bayesian model but is as expressive as a neural network and has the benefits of the probabilistic approach [22]. Gaussian processes have been successfully used in forecasting time series, and studies show that Gaussian process models can outperform other models by choosing an appropriate kernel [44].

2.3 Literature overview

There has been a lot of research in petrol sales forecast analysis. Much of it focuses on long term forecasts with monthly, quarterly or yearly values. In those problems, the economic factor can have a meaningful contribution to the model. The purpose of such forecasts is usually for investments, industrial planning or investigating environmental effects.

Azadeh et al. [5] constructs an artificial neural network for the petrol demand regression problem and tests it on several developed and developing countries. Their adaptive intelligent algorithm outperforms the regression model in a long-term monthly gasoline consumption forecast in 5 countries. The economic indicators used in the study are price, GDP (Gross Domestic Production), population, number of vehicles, gasoline demand in the last periods and correlation coefficient between variables. Li et al. [33] compares different models for forecasting automobile petrol demand in Australia. The study showed that a simple quadratic trend model had more predictive accuracy than a sophisticated model such as ARIMA. However, more advanced models such as neural networks or Gaussian processes were not analyzed. Rao and Parikh [37] models petrol demand in India with a linear regression model. Different types of petroleum products are analyzed and which economic factors such as price, Gross domestic product (GDP) and the wholesale price index of petrol have statistical significance in modelling a particular petroleum product's demand. Forouzanfar et al. [21] develops a multi-programming genetic model for forecasting transport energy demand in Iran. With such explanatory variables as GDP, population and number of vehicles, the genetic programming technique shows an improvement of results obtained from neural networks and fuzzy linear regression. The model is trained on data from 1968 to 2002 and tested on 2003-2005. Sun et al. [43] predicts the sales pattern of petrol using a decision tree and k-means for clustering sales patterns in order to reduce the logistics costs. On the first stage, the

sales values during the day for different days are split into several disjoint clusters characterizing different sales profiles. On the second stage, a decision tree is developed which uses exploratory variables such as weather conditions and promotional activity to predict the next day's sales pattern. The model is evaluated on three months sales data from a gas station in Dalian City, China. However, the model performance is not compared to any other generic method.

The logistics optimization for the fuel delivery is also a hot topic in the research field. The problem is known as the petrol station replenishment problem [15]. However, this problem is out of the scope of this thesis.

3 Background

The section describes the methods used for time series analysis and forecast in this study. Time series is defined in subsection 3.1, then ARIMA models and Gaussian Process models are described in subsection 3.2 and subsection 3.3 respectively. The cross-validation method used for evaluation in this study is described in subsection 3.4.

3.1 Time series

3.1.1 Definition of a time series

A time series is a sequential set of data points indexed by time [4]. It is mathematically defined as a set of vectors $y(t), t = 0, 1, 2, \ldots$ where t represents the time elapsed [14]. A time series is called *univariate* if it consists of single observations recorded over equal periods of time. Time series can be discrete and continuous. Continuous-time series are recorded at every instance of time, while discrete time series are measured at discrete points in time. An example of continuous time series is temperature readings. On the other hand, the population of a country or sales of a product can be presented as a discrete-time series. For the purpose of our study, we will be considering observations recorded at equally spaced time intervals of 1 day.

3.1.2 Components of a time series

In forecasting analysis, 4 components are considered which affect the time series: Trend, Seasonal, Cyclical and Irregular [4]. Not all time series contain all of the 4 components, and some forecasting models are aimed at modelling time series with only a subset of these components. Thus, the analysis of the components in given time series is important for an appropriate model choice.

Trend Trend is a general tendency of a time series to increase, decrease or stagnate over a long period of time [4]. Intuitively, a trend can be formulated as a long-term movement in a time series. For example, the term "Global warming" refers to an upward trend in the World temperature, suggesting that the average temperature is increasing every decade. Series relating to population growth also show an upward trend. Long-term stagnation of the time series indicates an absence of a trend.

Seasonal Seasonal variations in a time series are fluctuations within a time period during the season [4]. There are several factors that can cause seasonal variations: climate and weather conditions, traditional habits, customs etc [4]. For example, sales of cold soda drinks increase in summer and restaurants are visited by more customers on weekends.

Cyclical The cyclical variation in a time series describes medium-term changes in the series, which repeat in cycles [4]. A cycle usually extends over a longer period of time. Most of the economic and financial time series show some kind of cyclical

variation [4]. An example of a business cycle was proposed by Lee [32] and consists of 4 phases:

1. *Expansion* increase in production and prices, low-interest rates
2. *Crisis* stock exchanges crash and multiple bankruptcies of firms occur
3. *Recession* drops in prices and in output, high interest-rates
4. *Recovery* stocks recover because of the fall in prices and incomes

Understanding this component is important for modelling financial time series but is out of the scope of this thesis.

Irregular The irregular component of a time series is the residual time series after the trend-cycle, and the seasonal components have been removed [4]. In forecasting these fluctuations are interpreted as unavoidable random errors. In most models, the random errors are assumed to be white noise [14].

3.1.3 Concept of stationarity

Intuitively, stationarity is a property of a time series that its statistical properties such as mean and variance do not change over time [4]. It is an important condition to assess when building a time series forecast model. A simpler model can be designed for a time series with this assumption. In literature, two types of stationary processes are considered.

Strong stationarity A process $\{y(t)\}, t = 0, 1, 2, \ldots$ is *Strongly Stationary* or *Strictly Stationary* if the joint probability distribution function of $\{y_{t-s}, y_{t-s+1}, \ldots, y_t, \ldots, y_{t+s-1}, y_{t+s}\}$ is independent of t for all s. For a strong stationary process, the joint distribution of any possible set of random variables from the process is independent of time [27; 31]. The assumption of strong stationarity is not always needed in practical applications, so a weaker form is considered.

Weak stationarity A stochastic process is said to be *Weakly Stationary* of order k if the statistical moments of the process up to that order depend only on time differences and not upon the time of occurrences of the data being used to estimate the moments [7; 27; 31].

It is intuitive to assume that strong stationarity implies weak stationarity but that is not true. There are conditions under which one stationarity can be infered from another. To be precise, a weakly stationary process following normal distribution is also strongly stationary [14], and any strictly stationary process which has a finite mean and a covariance is also weakly stationary [48]. Dickey and Fuller [16] proposed a statistical test commonly used to detect stationarity in a time series.

3.2 AR and MA-based models

We describe two widely used linear time series models in literature: Autoregressive (AR) and Moving Average (MA) models.

Autoregressive model An AR(p) model can be regarded as a linear regression of the next observation using p past observations and a constant term as exploratory variables [4]. A mathematical expression of the AR(p) model follows [31]:

$$y_t = c + \sum_{i=1}^{p} \phi_i y_{t-i} + \epsilon_t. \tag{1}$$

Here y_t and ϵ_t are respectively the value and random error at the time period t, $\{\phi_i\} i = 1 \dots p$ are the model parameters and c is a constant. The integer constant p is known as the order of the model. A special case of the model with $p = 1$ is a representation of some markov chain models [30]. For estimating parameters of an AR model using the given time series, the Yule-Walker equations [27] can be used.

Yule-Walker equations Consider the general AR(p)

$$y_t = \sum_{i=1}^{p} \phi_i y_{t-i} + \epsilon_t. \tag{2}$$

Multiply both sides of the equation by y_{t-1}

$$y_t y_{t-1} = \sum_{i=1}^{p} \phi_i y_{t-i} y_{t-1} + y_{t-1}\epsilon_t, \tag{3}$$

take expectance

$$\langle y_t y_{t-1} \rangle = \sum_{i=1}^{p} \phi_i \langle y_{t-i} y_{t-1} \rangle + \langle y_{t-1}\epsilon_t \rangle. \tag{4}$$

Here ϕ_i is outside of the expectance operator because its value is determenistic. Also note that $\langle y_{t-1}\epsilon_t \rangle = 0$ because random shock is uncorrelated with the previous values of the process. Divide through by $N - 1$, and use the evenness of the autocovariance, $c_{-l} = c_l$

$$c_2 = \sum_{j=1}^{p} \phi_j c_{j-2} \tag{5}$$

Divide through by c_0

$$r_2 = \sum_{j=1}^{p} \phi_j r_{j-2} \tag{6}$$

Here $c_l = \sum_{i=1}^{N-1} x_i x_{i-l}$ is the autocovariance coefficient and $r_l = \frac{c_l}{c_0}$ is the autocorrelation coefficient. Repeating the process for all lags creates a linear equations system:

$$\underbrace{\begin{pmatrix} r_1 \\ \dots \\ r_p \end{pmatrix}}_{r} = \underbrace{\begin{bmatrix} r_0 & \dots & r_{p-1} \\ \vdots & \ddots & \vdots \\ r_{p-1} & \dots & r_0 \end{bmatrix}}_{R} \underbrace{\begin{pmatrix} \phi_1 \\ \dots \\ \phi_p \end{pmatrix}}_{\Phi} \tag{7}$$

or succinctly

$$R\Phi = r \tag{8}$$

This is a well-posed system with a full-rank and symmetric coefficient matrix R, so the invertability is guaranteed [18]

$$\hat{\Phi} = R^{-1}r \tag{9}$$

More details on the topic can be found in [18].

Moving average model Contrary to an AR(p) model, which regresses against past values of the series, an MA(q) model regresses against past errors. The mathematical expression of the MA(q) model is given by [7; 31; 27]:

$$y_t = \mu + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} + \epsilon_t \tag{10}$$

Here μ is the mean of the series, $\{\theta_j\} j = 1 \dots q$ are the weights of the linear regression, and q is the order of the model. The MA model fitting process is more complicated than for AR model because the random error terms are not initially given. Thus, iterative non-linear fitting procedures need to be used in place of linear least squares [34].

ARMA model Autoregressive (AR) and moving average (MA) models are commonly combined together to form a general class of time series models known as the ARMA models. Mathematically an ARMA(p,q) model is defined as [7; 31; 27]:

$$y_t = c + \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} + \epsilon_t \tag{11}$$

The model orders p, q refer to p autoregressive and q moving average terms respectively.

Usually, ARMA models are manipulated using the lag operator notation [14]. The lag operator, also known as backshift operator, is defined as $Ly_t = y_{t-1}$. Substituting

$$\phi(L) = 1 - \sum_{i=1}^{p} \phi_i L^i \tag{12}$$

$$\theta(L) = 1 + \sum_{i=1}^{q} \theta_j L^j \tag{13}$$

ARMA(p,q) model can be presented as [14]:

$$\phi(L) y_t = \theta(L) \epsilon_t \tag{14}$$

Hipel and McLeod [27] showed that a MA(q) process is stationary for all possible MA parameters [27]. Contrary to that, the autoregressive model is not always stationary as it may contain a unit root [16].

3.2.1 Autoregressive Integrated Moving Average (ARIMA) Models

The ARMA models, described above are applicable to stationary time series data. In practice, a lot of time series show non-stationary behaviour, such as socio-economic processes [27]. Time series with trend and seasonal patterns are non-stationary in nature [20]. Hence ARMA models are not appropriate for a variety of practical non-stationary time series. For this reason an ARIMA model [7] was proposed, which is a generalization of an ARMA model aimed to handle the case of non-stationary time series.

In ARIMA models a non-stationary time series is made stationary by applying finite differencing of the observations. The mathematical expression of the ARIMA(p,d,q) model using lag polynomials is given below [27]:

$$\phi(L)y_t = \theta(L)\epsilon_t \tag{15}$$

$$(1 - \sum_{i=1}^{p} \phi_i L^i)(1 - L)^d y_t = (1 + \sum_{i=1}^{q} \theta_j L^j)\epsilon_t \tag{16}$$

In addition to non-negative integers p and q corresponding to the orders of the autoregressive and moving average parts of the model, the integer d controls the level of differencing. Generally $d = 1$ is enough in most cases [4]. ARIMA(p,0,q) is equivalent to a ARMA(p,q) model, while ARIMA(p,0,0) is the AR(p) model and ARIMA(0,0,q) is the MA(q) model. A special case of ARIMA(0,1,0) $y_t = y_{t-1} + \epsilon_t$ is known as the *Random Walk* [41] model. It is widely used for economic and stock price series [31].

3.2.2 Seasonal Autoregressive Integrated Moving Average (SARIMA) Models

A standard ARIMA model does not model seasonality well [4]. Box et al. [7] have generalized this model to deal with seasonality. Seasonal ARIMA (SARIMA) is an extension of ARIMA, which uses seasonal differencing of appropriate order to remove the seasonal component from the series and make it stationary [7]. For a seasonality period s a first order seasonal difference is calculated as $z_t = y_t - y_{t-s}$. The mathematical formulation of a SARIMA$(p, d, q) \times (P, D, Q)^S$ model in terms of lag polynomials is given by [19]:

$$\Phi_P(L^S)\phi_p(1 - L)^d(1 - L^S)^D y_t = \Theta_Q(L^S)\theta_q(L)\epsilon_t. \tag{17}$$

3.2.3 Box-Jenkins methodology for optimal model selection

Box et al. [7] developed a practical approach to find an optimal fit ARIMA model for a given time series while satisfying the parsimony principle. The principle declares: "the model with the smallest possible number of parameters is to be selected so as to provide an adequate representation of the underlying time series data" [12]. This is a common principle mentioned in various research fields also known as *Occam's razor*. Rasmussen and Ghahramani [39] showed that Occam's razor is embodied in

the application of Bayesian theory.
The Box-Jenkins methodology [7] uses a three-step iterative approach to determine the best model from a general class of ARIMA models.

One of the important aspects of the model selection is determining the optimal model parameters. Widely used measures for model identification are Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) [10]:

$$AIC(p) = n\ln(\sigma_e^2/n) + 2p \tag{18}$$
$$BIC(p) = n\ln(\sigma_e^2/n) + p + p\ln(n) \tag{19}$$

Here n is the number of observations, p is the number of parameters in the model and σ_e^2 is the sum of sample squared residuals. The criteria represent a trade off between the error of the forecast and the flexibility of the model in parameters number. An optimal model is selected via minizing one of the proposed criteria.

3.3 Gaussian processes

In this subsection, we introduce the Gaussian processes. We begin by describing the multivariate Gaussian distribution and its properties. Then we recall the standard linear Bayesian regression setup and introduce functions in replacement of explicit parameter values, moving from a parametric to a non-parametric model.

3.3.1 Gaussian distribution

The Gaussian or normal distribution is one of the most widely used in statistics, proposed by Gauss [23]. It is a type of continuous probability distribution for a real-valued random variable. The general form of its probability density function is given by:

$$f(x) = \frac{1}{v\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-m}{v}\right)^2} \tag{20}$$

The Gaussian distribution is parametrized by mean m and standard deviation v. Gaussian distributions are closed under additions:

$$x_1 \sim \mathcal{N}(m_1, v_1) \tag{21}$$
$$x_2 \sim \mathcal{N}(m_2, v_2) \Rightarrow \tag{22}$$
$$x_1 + x_2 \sim \mathcal{N}(m_1 + m_2, v_1 + v_2) \tag{23}$$

Same is true for any finite number of independent variables:

$$x_i \sim \mathcal{N}(m_i, v_i) \Rightarrow \sum x_i \sim \mathcal{N}(\sum m_i, \sum v_i) \tag{24}$$

In other words any finite sum of Gaussian variables can be formulated as a single Gaussian distributed variable. This property will be useful for the multivariate Gaussian distribution which we describe in next section.

3.3.2 The multivariate Gaussian distribution

Definition A random vector $\mathbf{x} = [x_1, x_2, \ldots, x_D]^T$ is said to have the multivariate Gaussian distribution if all linear combinations of x are Gaussian distributed:

$$y = a^T x = a_1 x_1 + a_2 x_2 + \cdots + a_D x_D \sim \mathcal{N}(m, v) \tag{25}$$

for all $a \in \mathbb{R}^D$, where $a \neq 0$. The multivariate Gaussian density for a variable $x \in \mathbb{R}^D$ is then given by:

$$\mathcal{N}(x|\mu, \Sigma) = (2\pi)^{-\frac{D}{2}} |\Sigma|^{-\frac{1}{2}} \exp -\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu) \tag{26}$$

The density is completely described by its parameters:

1. $\mu \in \mathbb{R}^D$ is the mean vector
2. $\Sigma \in \mathbb{R}^{D \times D}$ is the covariance matrix (positive definite)

Here $(\Sigma)_{ij}$ is the covariance between the i'th and j'th elements x_i and x_j of x. Since the covariance function is commutative, the matrix Σ is symmetric.

We can define a distribution on the function space treating an arbitrary real-values function f narrowed to a set of arguments $(x_1, \ldots, x_n)$ as a random vector $(f(x_1), \ldots, f(x_n))$. A Gaussian distribution on functions would then be a multivariate Gaussian distribution on the function values.

3.3.3 Bayesian linear regression

Let's formulate the general setup for linear regression. We are given a data set: $D = \{x_n, y_n\}_{n=1}^N$. Our goal is to learn some function f such that

$$y_n = f(x_n) + \epsilon_n \tag{27}$$

Here ϵ_n is an error term, since in practice it is not possible to model a random variable with certainty. Assuming f is a linear model:

$$f(x) = w_1 x_1 + w_2 x_2 + \cdots + w_D x_D = \sum_i w_i x_i = \mathbf{w}^T \mathbf{x} \tag{28}$$

Here w are weights. Linear models are linear with respect to parameters, not the data. We can replace x vector components with functions:

$$f(x) = w_1 \phi_1(x_1) + w_2 \phi_2(x_2) + \cdots + w_D \phi_D(x_D) = \sum_i w_i \phi_i(x_i) = \mathbf{w}^T \phi(x) \tag{29}$$

Here ϕ_i can be a non-linear function and is called a *feature* function. We can add an intercept or bias term to the model by assuming $x_0 = 1$:

$$f(x) = w_0 + w_1 x_1 + \cdots + w_D x_D = \mathbf{w}^T \mathbf{x} \tag{30}$$

The Bayesian linear regression model is then given by:

$$y_n = f(\mathbf{x}_n) + \epsilon = \mathbf{w}^T \mathbf{x}_n + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma^2) \tag{31}$$

Here ϵ_i are independent and identically normally distributed random variables with unknown variance σ^2. Likelihood for one data point is given by:

$$p(y_n|\mathbf{x}_n, \mathbf{w}) = \mathcal{N}(y_n|f(\mathbf{x}_n), \sigma^2) = \mathcal{N}(y_n|\mathbf{w}_n^T \mathbf{x}, \sigma^2) \tag{32}$$

and likelihood for all data points in the data set is given by:

$$p(y|\mathbf{X}, \mathbf{w}) = \prod_{n=1}^{N} \mathcal{N}(y_n|f(\mathbf{x}_n), \sigma^2) = \mathcal{N}(y|\mathbf{X}, \sigma^2 I) \tag{33}$$

Since the data is assumed constant, the likelihood is a function of parameters $\mathbf{w}$. The prediction vector for a new input $\mathbf{X}$ is given by $f = \mathbf{X}\mathbf{w}$. Now we introduce a prior distribution $p(\mathbf{w})$ for the weights $\mathbf{w}$. $p(\mathbf{w})$ contains our prior knowledge about $\mathbf{w}$ before we see any data. For example, we might want to restrict $\mathbf{w}$ to only positive values or we suspect that the weight values are unlikely to be too large. With the prior defined, Bayes rule gives us the posterior distribution:

$$\underbrace{p(\mathbf{w}|y)}_{\text{posterior}} = \frac{\overbrace{p(y|\mathbf{w})}^{\text{likelihood}} \times \overbrace{p(\mathbf{w})}^{\text{prior}}}{\underbrace{p(y)}_{\text{marginal likelihood}}} \tag{34}$$

Marginal likelihood $p(y)$ is the joint probability $p(y, \mathbf{w})$ integrated over the weights space:

$$p(y) = \int p(y, \mathbf{w}) \mathbf{dw} = \int p(y|\mathbf{w}) p(\mathbf{w}) \mathbf{dw} \tag{35}$$

The posterior distribution $p(\mathbf{w}|y)$ captures everything we know about $\mathbf{w}$ after seeing the data. The posterior is a distribution over the parameter space. It is a compromise between the prior and the likelihood [46].

Bayesian linear regression is an approach to linear regression widely used in statistics. However, in more complex problems the class of linear functions is too small to adequately model the target function. To increase the class of functions, we are going to replace the parameter vector w with a function ϕ. Instead of learning exact values for every parameter in a function, Gaussian processes employ a Bayesian approach to infer a probability distribution over all possible functions f that fit the data, given a priori [11].

3.3.4 Switching parameters to functions

In the general linear regression setup, our goal is to learn the function f which describes the output y based on the factor variables $\mathbf{x}$.

$$f(x) = \mathbf{w}^T \mathbf{x} \tag{36}$$

So far we have used the distribution on weights $\mathbf{w}$ for inference on y:

$$p(y, \mathbf{w}) = p(y|\mathbf{w})p(\mathbf{w}) \tag{37}$$

Let's now introduce the function to the model in the form of its values at inputs $f = [f(x_1), f(x_2), \ldots, f(x_N)] \in \mathbb{R}^N$. Previously, we described the weights posterior after we have seen the data, and the function probability with certain parameter values $\mathbf{w}$. We will now describe the inference of posterior distribution on the function f based on the values of the function on the inputs. Keeping the model the same, the joint probability with the function values is given by

$$p(y, f, \mathbf{w}) = p(y|f)p(f|\mathbf{w})p(\mathbf{w}) \tag{38}$$

Let's now marginalize the augmented model over the weights:

$$p(y, f) = \int p(y, f, \mathbf{w})\mathbf{dw} = p(y|f) \int p(f|\mathbf{w})p(\mathbf{w})\mathbf{dw} \tag{39}$$

Given

$$p(f) = \int p(f, \mathbf{w})\mathbf{dw} = \int p(f|\mathbf{w})p(\mathbf{w})\mathbf{dw} \tag{40}$$

we can decompose the joint probability as likelihood and prior

$$p(y, f) = p(y|f)p(f) \tag{41}$$

In the inference we represent function f as a random vector consisting of its evaluation on inputs $\mathbf{x}_1, \ldots, \mathbf{x}_N$. The prior distribution on f is hence a N-component multivariate Gaussian distribution:

$$p(f) = \mathcal{N}(f|0, \mathbf{X}\Sigma_p\mathbf{X}^T) \tag{42}$$

After integrating out the weights, the final distribution on a set of outputs $\mathbf{y}$ follows:

$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|0, K) \tag{43}$$

3.3.5 Covariance matrix

Let's have a closer look on the covariance between f_i and f_j:

$$\begin{aligned}
K_{ij} = cov(f_i, f_j) &= cov(f(x_i), f(x_j)) \\
&= cov(\mathbf{w}^T x_i, \mathbf{w}^T x_j) \\
&= E[(\mathbf{w}^T x_i - 0)(\mathbf{w}^T x_j - 0)] \\
&= E[\mathbf{w}^T x_i \mathbf{w}^T x_j] \\
&= E[x_i \mathbf{w}\mathbf{w}^T x_j] \\
&= x_i E[\mathbf{w}\mathbf{w}^T] x_j \\
&= x_i \Sigma_p x_j \\
&\equiv k(x_i, x_j)
\end{aligned}$$

k is called a kernel function. Changing k changes f. The prior is defined by the Gram matrix of kernel function k calculated on all pairs of inputs (x_i, x_j) .
We can now describe how, given a covariance matrix $\mathbf{K}$, Equation 43 specifies a distribution on functions. On a specified set of input points $x = (x_1, \ldots, x_N)$ using $N \times N$ covariance matrix $\mathbf{K}$, we draw a vector $y = y_1 \ldots y_N$ from the Gaussian distribution defined by the Equation 43. This set of points $(x_n, y_n); n = 1 \ldots N$ describes our function evaluated at a finite set of points. Plotting the set of points can give us a general visualization of the result function. In this procedure, we haven't looked at any input data yet. The smoothness properties of the output function will be directly based on the chosen covariance function. Functions sampled from a periodic kernel with some period will have the same period.

3.3.6 Gaussian process

A Gaussian process (GP) is a collection of random variables indexed over space, such that a linear combination of those random variables has a multivariate normal distribution [11]. A Gaussian process can be considered a prior distribution over functions $f : \mathcal{X} \to \mathbb{R}$:

$$f(x) \sim GP(m(x), k(x, x')) \tag{44}$$

A Gaussian process is completely defined by its mean function $m(x)$ and its covariance function $k(x, x')$:

$$E[f(x)] = m(x) \tag{45}$$

$$cov[f(x), f(x')] = k(x, x') \tag{46}$$

The probability of any subset of function values $\{f(x_1), \ldots, f(x_N)\}$ at inputs $\{x_1, \ldots .x_N\}$ is given by:

$$p(f) = \mathcal{N}(f|m, K) \tag{47}$$

where $m = m(x_1) \ldots m(x_N)$ and $[K]_{ij} = k(x_i, x_j)$. As discussed in subsubsection 3.3.5 the model depends only on the covariance matrix K. In a Gaussian process, instead of the weights vector $\mathbf{w}$ we directly specify the joint output covariance K as a function of two inputs [6]. For that purpose we need to define the i, j element of the covariance matrix for any two inputs x_i and x_j. This can achieved using a kernel function $k(x_i, x_j)$

$$[\mathbf{K}]_{i,j} = k(x_i, x_j) \tag{48}$$

Explicitly specifying elements' relations as a matrix into the algorithm instead of using a defined function is known as the kernel trick [38]. It allows us to incorporate a large space of non-linear functions into a linear regression model in order to increase its expressiveness.

3.3.7 Choice and examples of kernel functions

The only requirement for a kernel function is to generate a symmetric positive definite matrix K because the covariance matrix of the Gaussian distribution has to be symmetric positive definite. This must hold for all possible data sets in the

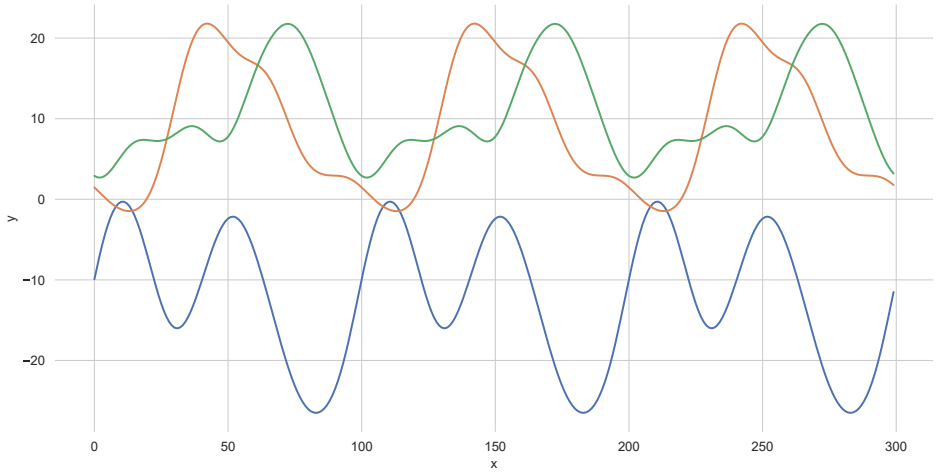

Figure 1: Example of 3 functions sampled from the periodic kernel

input space X. Then the function can be used as a kernel for our model. The kernel incorporates our assumptions about the data into the model so the kernel choice is vital for the model design. This could be the smoothness of a function or its periodicity. The choice of the kernel lets us incorporate domain knowledge into our model. A widely used kernel is Gaussian kernel, however, it is important to understand that any kernel function satisfying the conditions described above will produce a Gaussian process.

We now describe some commonly used kernels and visualize their GP prior samples.

The Exp-Sine-Squared kernel also known as the periodic kernel is given by:

$$k_P(x_i, x_j) = \exp\left(-\frac{2\sin^2(\pi d(x_i, x_j)/p)}{l^2}\right) \tag{49}$$

where $l > 0$ is a length scale parameter and $p > 0$ is periodicity parameter. This kernel has a special meaning in forecasting problems. The kernel allows modelling a seasonal component in the model. Several kernel instances with different periodicity parameters can be used to model multiple seasonalities within the time series. Samples can be seen in Figure 1.

The Squared Exponential kernel widely known as the Radial Basis Function (RBF) kernel or the Gaussian kernel is given by:

$$k_{RBF}(x_i, x_j) = \exp\left(-\frac{(x - x')^2}{2l^2}\right) \tag{50}$$

where $l > 0$ is the length-scale parameter. This kernel has proven to perform well in interpolation tasks and is commonly used in various kernelized learning algorithms. Every function in its prior has infinitely many derivatives. Samples can be seen in Figure 2.

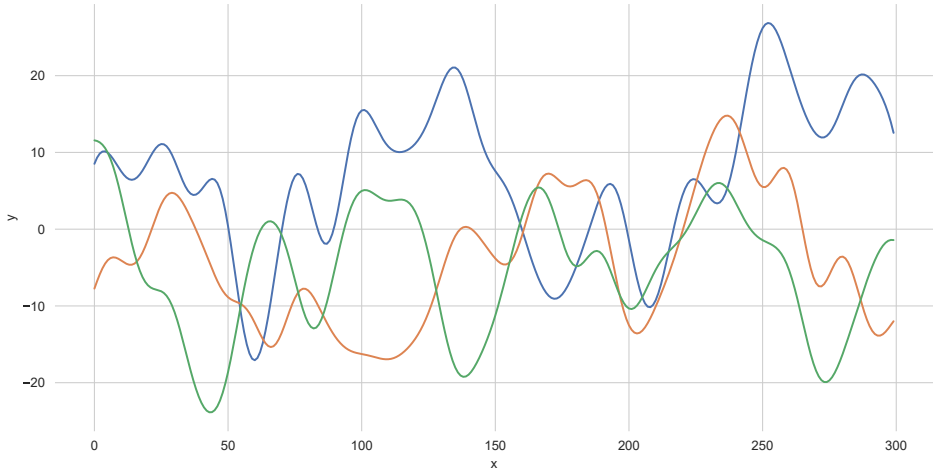

Figure 2: Example of 3 functions sampled from the RBF kernel

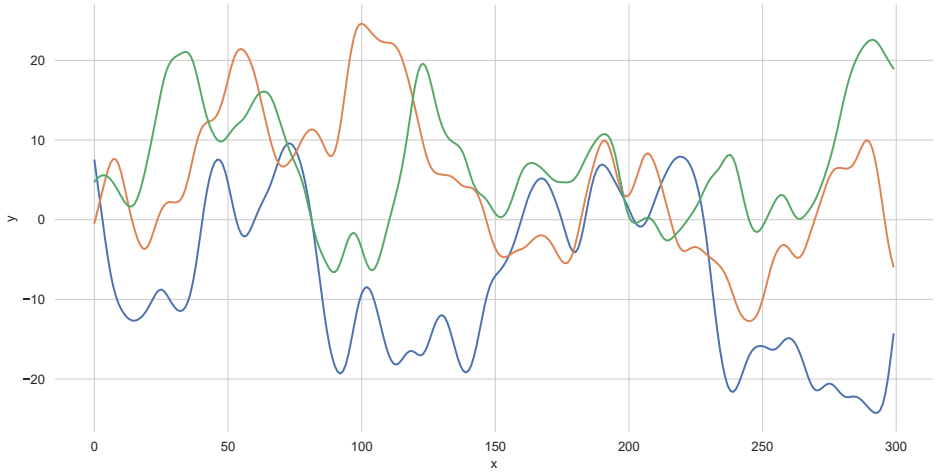

Figure 3: Example of a function sampled from the rational quadratic kernel

The Rational Quadratic kernel is given by:

$$k_{RQ}(x_i, x_j) = \left(1 + \frac{(x - x')^2}{2\alpha l^2}\right)^{-\alpha} \tag{51}$$

where $l > 0$ is a length-scale parameter and α is a scale mixture parameter.

This kernel is equivalent to a scale mixture of RBF kernels with different correlation length distributed according to a Beta distribution [44]. Consequently, GP priors with this kernel will have variable smoothness across different length scales [17]. The parameter α determines the relative weighting of large-scale and small-scale variations [11]. With $\alpha \to \infty$, the RQ becomes identical to the RBF. Samples can be seen in Figure 3.

Kernel properties Since a kernel function is only required to generate a positive semi-definite Gramm matrix, a linear combination of any kernels gives us a kernel:

$$k = k_1 + k_2 \tag{52}$$
$$k = k_1 \times k_2 \tag{53}$$

We can model a quasi-periodic GP by multiplying a periodic kernel by a non-periodic kernel, and we can model multiple seasonal components by adding several periodic kernels.

3.3.8 Hyperparameters

Most common GP kernels rely on one or more hyperparameters. For instance, the periodic kernel contains two hyperparameters: the period and the length scale. These

hyperparameters have to be estimated from the data. Ideally, we would like to put prior distributions on the hyperparameters and compute the posterior:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} \tag{54}$$

but in this case the marginal likelihood is almost always intractable. Instead, this is done via maximum likelihood optimization of the data given the hyperparameters. Since $p(y)$ is constant with respect to θ we can reduce:

$$p(\theta|y) \propto p(y|\theta)p(\theta) \tag{55}$$

The MAP (Maximum a posteriori estimation) estimate is defined as:

$$\theta_{MAP} = \arg\max \ln p(\theta|y) = \arg\max \ln p(y|\theta) + \ln p(\theta) \tag{56}$$

If the prior $p(\theta)$ is uniform meaning we have no prior information on the hyperparameter distribution, the term $\ln p(\theta)$ can be omitten which reduces our estimate to the maximum likelihood:

$$\theta_{MAP} = \arg\max \ln p(y|\theta) + \ln p(\theta) = \arg\max \ln p(y|\theta) = \theta_{ML} \tag{57}$$

3.4 Cross-validation

Cross-validation was originally employed to evaluate the predictive validity of linear regression equations used to forecast a performance criterion from scores on a battery of tests [35]. It was found that the error on the training set was an optimistic estimate of the predictive accuracy of the fit algorithm. Optimizing the accuracy on the validated dataset will lead to choosing a sophisticated model strictly adjusted to solving the given problem on the given dataset. In machine learning, we aim to fit a model that will be able to predict a function accurately on any new entry from the object space. Choosing a complex model in this case violates the parsimony principle and leads to the problem commonly referred to as overfitting [26]. In machine learning it is common that the original data set is split into a train and test datasets. The train set is used to analyze the data and learn the predictor without looking at the test set. Then the predictive validity of the model is evaluated on the test set. It is desirable to have enough data for learning, so the division in percentages is usually 80-20, 60-40, 50-50 etc. Still, a small test set leads to high variance in the evaluation of the predictor performance - that is, the test set error will heavily depend on the actual sample used if it is too small. Attempts to increase the number of observations available for training purposes resulted in a rotational estimation method [35]. In this method, the dataset is split into many batches, and each batch is used as a test set through a repeated cycle. In the simplest form, in a dataset with N samples, there can be N rounds which $N-1$ samples in the train set and 1 in the test set. An average of the N rounds is then used. This method is called "leave-one-out". The method can be extended to leaving more than one element out as shown by Geisser [24].

Definition Cross-validation is any of various similar model validation techniques for assessing how the results of a statistical analysis will generalize to an independent data set. [45; 8; 42]

Linear regression Let's consider a simple case of linear regression with output values $y_1, \ldots, y_n$ and an p-dimensional vector covariates $x_1, \ldots, x_n$. Using least squares to fit a function in the form of a hyperplane $y = a + \beta^T x$ to the data $(x_i, y_i) 1 \leq i \leq n$ yields a model with mean squared error (MSE) of the form

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - a - \boldsymbol{\beta}^T \mathbf{x}_i)^2 \tag{58}$$

Under certain conditions it can be shown that the expected value of the MSE for the training set is $\frac{n-p-1}{n+p+1} < 1$ times the expected value of the MSE for the validation set [45]. Thus the MSE on the training set gives us an optimistic biased estimate of the error on a potential test set. The biased estimate is called the in-sample estimate of the fit and the cross-validation estimate is called an out-of-sample estimate [45].

The formula gives us a direct way to check whether the model was overfit during learning process. If the MSE in the validation set greatly exceeds the anticipated values of the out-of-sample estimate, then the model fit should be reconsidered. A common technique for preventing overfitting in linear regression models is regularization. A regularization term in the error function penalizes the model from being too complex [9]. The regularized model then is a compromise with some coefficients between the likelihood (the data) and the regularization term (the generalization). The cross-validation evaluation can help choose the optimal weight for the regularization term.

Unfortunately, in most regression models, it is not feasible to compute the expected out-of-sample fit. However, cross-validation can still be used as a general way to estimate the performance of a model on potential data.

Date	Opening Volume	Metered Sales	Deliveries	Observed error
2013-01-01	14840	1929	0	15
2013-01-02	12926	2610	0	-4
2013-01-03	10312	2618	16593	50
2013-01-04	24337	2526	0	-13
2013-01-05	21798	2106	0	-10

Table 1: Dataset sample. Date is the date of the observation. Opening volume describes the product level in the tank at the time of observation. Metered sales describes sales in litres since the previous observation. Deliveries describe product deliveries since the previous observation. Observed error describes the estimated error of the sales observation.

4 Experiments

4.1 Data description

The data used for this work comes from a station in Helsinki with installed devices with fuel monitoring systems. The system collects sales of petrol in the tank each night. It also tracks the refilling of the tank and the current product level. At around 2 a.m. a daily report is generated and saved to the database. In the current business process, the device detects when the tank product level is lower than a specified threshold and sends an alert usually communicated through SMS to the station manager.

4.1.1 The columns explanation

An example part of the dataset can be seen at Table 1. There is a timestamp, daily sales of the product in litres, opening volume of the tank, and how much it was filled that day. In the table, delivery was done on 3rd Jan as indicated by the number of litres refilled to the tank in the Deliveries column. The columns definition is described as follows:

1. *Date* Date of the observation. Observations are taken a bit after midnight
2. *Opening Volume* Product level in litres at the time of observation
3. *Metered Sales* Recorded product sales in litres from the previous observation
4. *Deliveries* Recorded product in litres deliveries from the previous observation
5. *Observed error* The difference between directly measured sales and consumed petrol calculated by subtracting opening volumes. The value represents measurement error and external interference.

Metered Sales column is the target variable used for modelling in this study. Rows with high observed error were treated as outliers, but otherwise, the column was not

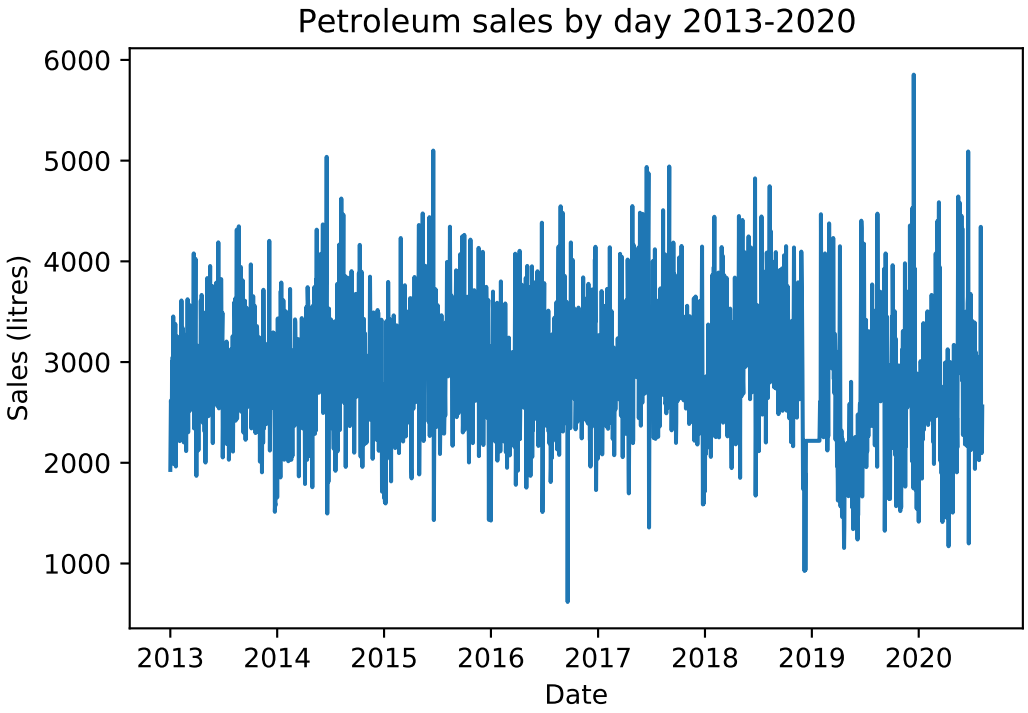

Figure 4: Petrol sales in a sample Helsinki station

used in the data processing.

A full graph of the sales collected for 7 years from the station in Helsinki is demonstrated in Figure 4

The yearly plot in Figure 5 shows that the time series has no trend since the mean does not change between different years.

4.1.2 Outliers and data preprocessing

The data was collected from an old distributed measurement system that inherently has noise due to hardware, software failures and other factors. Some days are missing data, and some days show values that are outside of accepted bounds. Point outliers in this study are considered to be values outside of $(0.15, 0.95)$ quantiles. These days and missing days values were replaced with an average of adjacent days.

Another factor is the first COVID19 lockdown period which was enforced in Finland from 16th March to 10th July. Its effect on the sales can be seen in Figure 6. This time period had to be excluded from the dataset because it differs from the normal sales pattern.

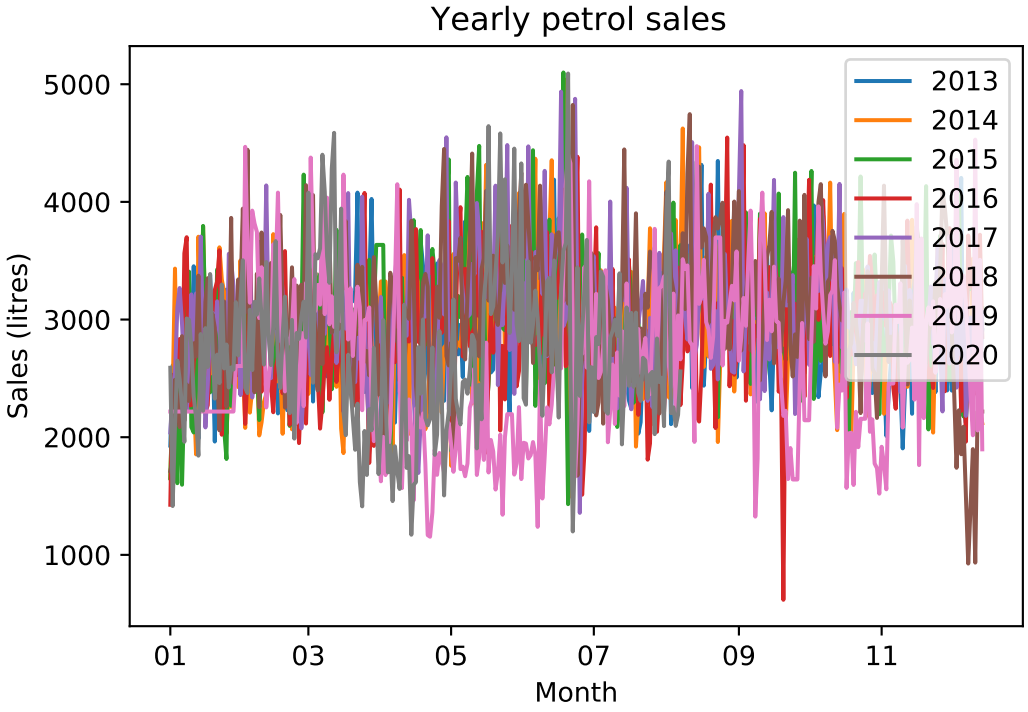

Figure 5: Petrol sales in a sample Helsinki station year comparison

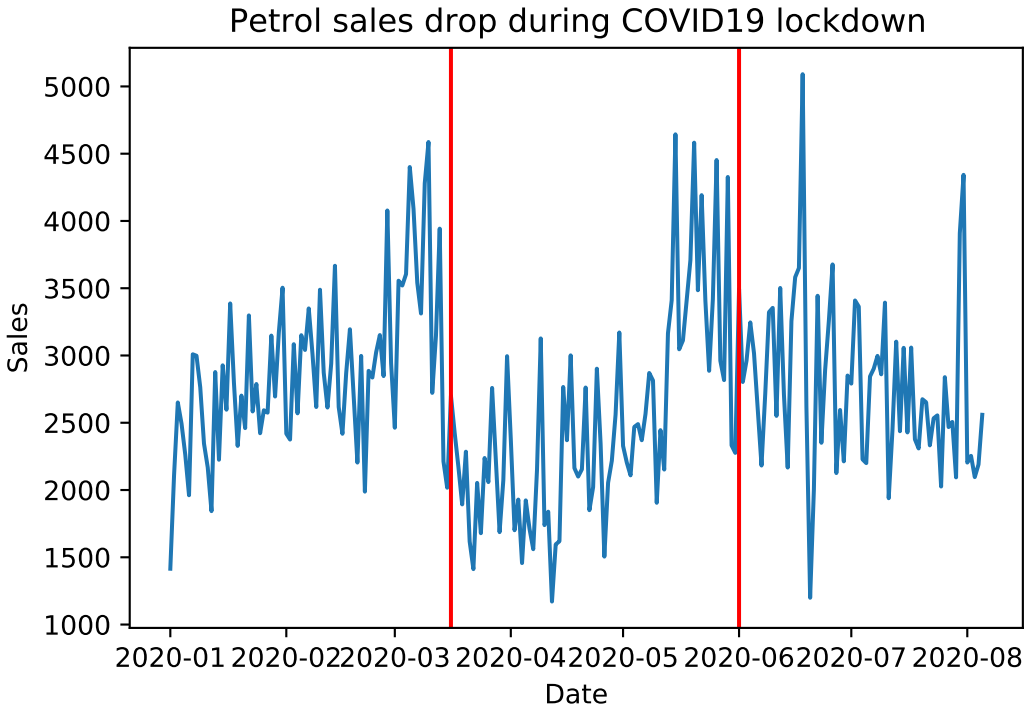

Figure 6: Petrol sales during first COVID19 lockdown. Red lines draw the lockdown period.

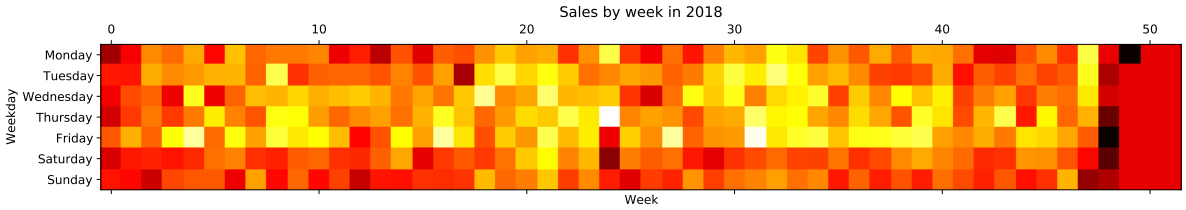

Figure 7: Heatmap of weekly sales in 2018. Red tone means smaller sales

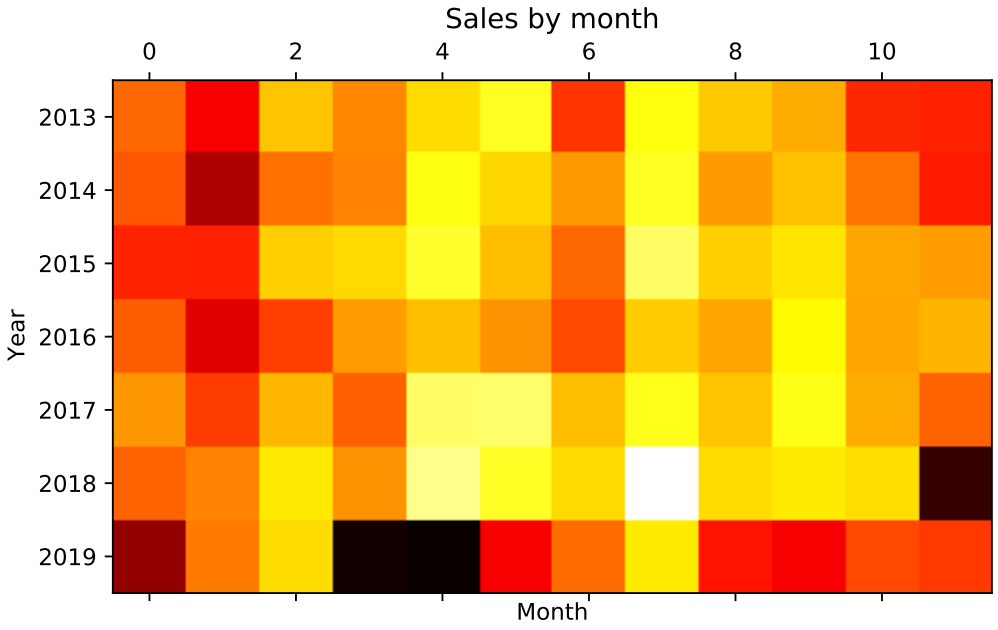

Figure 8: Heatmap of monthly sales. Red means smaller sales

4.1.3 Seasonality

From the domain knowledge, there is an assumption that petrol sales carry a seasonal pattern as is common in mass-consumer goods sales. Usually, it would be weekly, monthly, or yearly. Most customers' gas petrol refilling patterns are likely to be weekday-based, while summer is a more common season for rides which should result in increased sales. Recognizing the seasonality pattern in the time series is essential in specifying a good forecast model.

By placing yearly sales on top of each other, we can see a general sales pattern throughout the year. For example, by the end of each year, sales decrease, as explained in the first paragraph.

The heatmap in Figure 7 demonstrates weekly sales. Each column represents a week, and each row represents one of 7 days of the week, lighter colour is a larger sales number. A contrasting line between Friday and Saturday is clearly visible, with weekends having significantly fewer sales than weekdays. This brings us to the conclusion that the same weekdays have similar sales suggesting a weekly seasonality

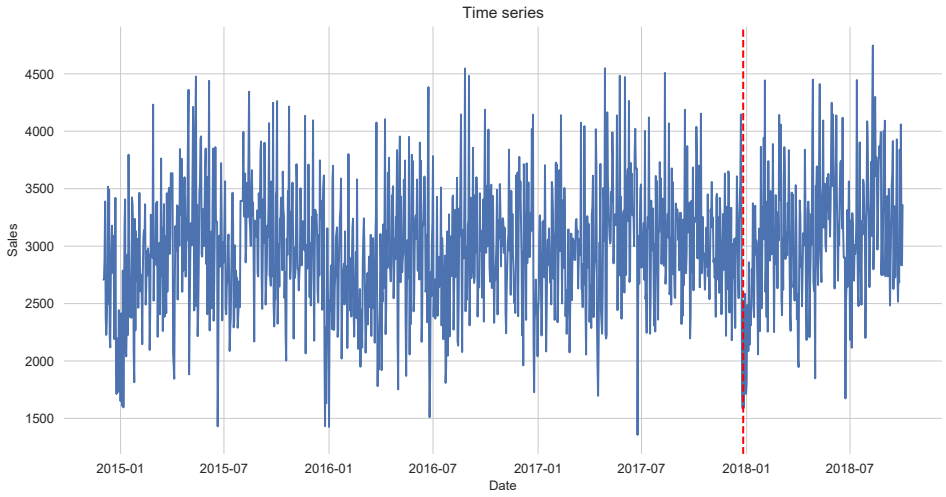

Figure 9: Train-test split of the dataset

pattern. It can be noted that the weeks at the beginning of the year and at the end of the year have fewer sales than in the middle, which suggests a yearly pattern. It is also visible on the years' comparison graph in Figure 5 and the heatmap for months in Figure 8.

4.1.4 Training and test sets

Evaluation of the models included in the study was made using a dataset covering 3 years, from December 12, 2014 to October 1st, 2018, containing daily sales data of a single station in Helsinki. The training-test ratio was selected as 80:20. The visualization of the series is given in Figure 9

4.2 Evaluation

The models in this study are compared using several evaluation metrics, which are described in this section. Here $y_{\text{test}} \in \mathbb{R}^N$ is the test data and $y_{\text{forecast}} \in \mathbb{R}^N$ is the forecast result given by the model.

Mean absolute error (MAE) is given by:

$$\text{MAE} = \frac{\sum_{i=1}^{N} |y_{\text{test}}(i) - y_{\text{forecast}}(i)|}{N} \tag{59}$$

MAE is a common evaluation metric used in time series analysis [28]. It is a scale-dependent accuracy measure and has the same scale as the data. We can use a scale-independent accuracy measure by diving the error by the test value. That accuracy measure is known as mean absolute percentage error (MAPE) given by:

$$\text{MAPE} = \frac{\sum_{i=1}^{N} \frac{|y_{\text{test}}(i) - y_{\text{forecast}}(i)|}{y_{\text{test}}(i)}}{N} * 100\% \tag{60}$$

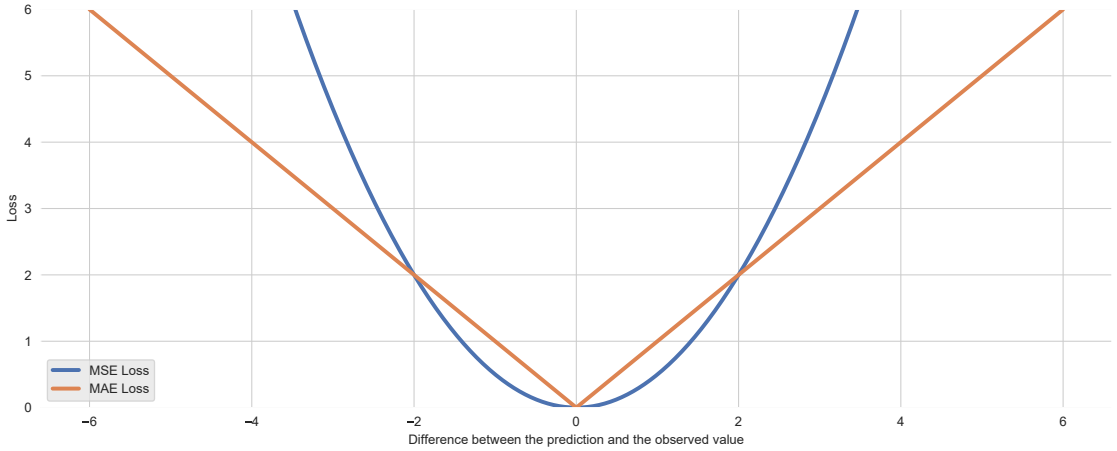

Figure 10: Comparison of MSE and MAE losses

An advantage of MAPE is intuitiveness for a broad audience such as business. That is 15% MAPE would mean that on average, the model is off the true value by 15%. Another important accuracy measure in machine learning is the root mean squared error (RMSE) given by:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(y_{\text{test}}(i) - y_{\text{forecast}}(i))^2}{N}} \tag{61}$$

MSE is the most accurate measure for machine learning problems where the target variable has a Gaussian noise. Historically, the RMSE has been popular largely because of its theoretical relevance in statistical modelling. However, it is more sensitive to point outliers than MAE, which has led some authors to recommend against their use in forecast accuracy evaluation [25]. The comparison of the error functions is visualized in Figure 10.

While the model was designed to predict daily sales, the business case is mainly interested in the next week's forecast to schedule the tank refill. Thus it makes sense to group days into weeks and calculate the error on full week sales. These measures are included as Week MAE and Week MAPE.

4.3 SARIMA

For the SARIMA model the search of optimal parameters was done using grid search. The enumerated parameters set is described below:

1. $p = 0 \dots 6$
2. $d = 0 \dots 2$
3. $q = 0 \dots 2$

p	d	q	P	D	Q	AIC	MAE	MAPE	week MAE	week MAPE
3	1	1	1	1	1	10232.57	483.25	14.70%	1792.63	8.08%
0	1	1	1	1	1	10252.23	457.50	14.80%	1803.27	8.11%
2	1	1	1	1	1	10244.47	463.30	14.71%	1811.70	8.12%
3	1	0	3	0	0	10289.62	525.60	14.19%	1812.54	8.13%
3	1	1	2	1	0	10270.77	471.93	14.36%	1841.77	8.18%
1	1	1	2	1	0	10298.83	471.68	14.34%	1842.45	8.18%
2	1	1	2	1	0	10284.13	471.78	14.36%	1843.19	8.19%
3	1	0	0	0	1	10663.78	528.19	15.33%	1859.50	8.24%
2	1	1	3	1	0	10154.91	468.57	14.35%	1849.14	8.26%
3	1	0	2	0	0	10414.15	528.65	16.07%	1892.26	8.27%
1	1	1	0	1	1	10244.50	459.06	14.02%	1859.87	8.31%
1	1	1	3	1	0	10170.79	469.32	14.40%	1860.80	8.33%
2	1	1	0	1	1	10246.49	459.33	14.04%	1863.47	8.34%
3	1	0	0	0	0	10822.41	530.71	16.12%	1907.85	8.34%
3	1	1	3	1	0	10141.73	469.93	14.44%	1962.42	8.34%
3	1	0	1	0	0	10577.72	528.80	16.52%	1965.02	8.35%
3	1	0	2	0	1	10373.13	459.95	14.14%	1961.42	8.37%

Table 2: ARIMA results with different parameters

4. $P = 0 \dots 2$
5. $D = 0 \dots 2$
6. $Q = 0 \dots 2$

From the initial data analysis the most significant periods in the time series were found to be weekly and yearly. However, a seasonality period of 365 for SAMIRA is too large for the complexity of the learning to be able to search through a reasonable parameters set. So the seasonality period was set to 7. Then all the possible parameters combinations from the parameters set were evaluated. Due to a large number of the combinations, only the best results are shown at Table 2

During learning ARIMA model calculates the Akaike information criterion (AIC) based on the fit. However, we use the same evaluation metric as with GPR models to be able to compare the ARIMA with other models. An example of the fit is shown in Figure 11

4.4 Gaussian Process Regression

4.4.1 Setting the hyperparameters

For using GPR model in our sales prediction we need to specify the kernel we are going to use. The kernel families presented in subsubsection 3.3.7 and their sum and products are considered. The hyper parameters for the kernels were estimated from the data via maximizing the log marginal likelihood. $\mathcal{L}(\theta)$ is non-convex so the iterative optimization may not converge to the global maximum. A common

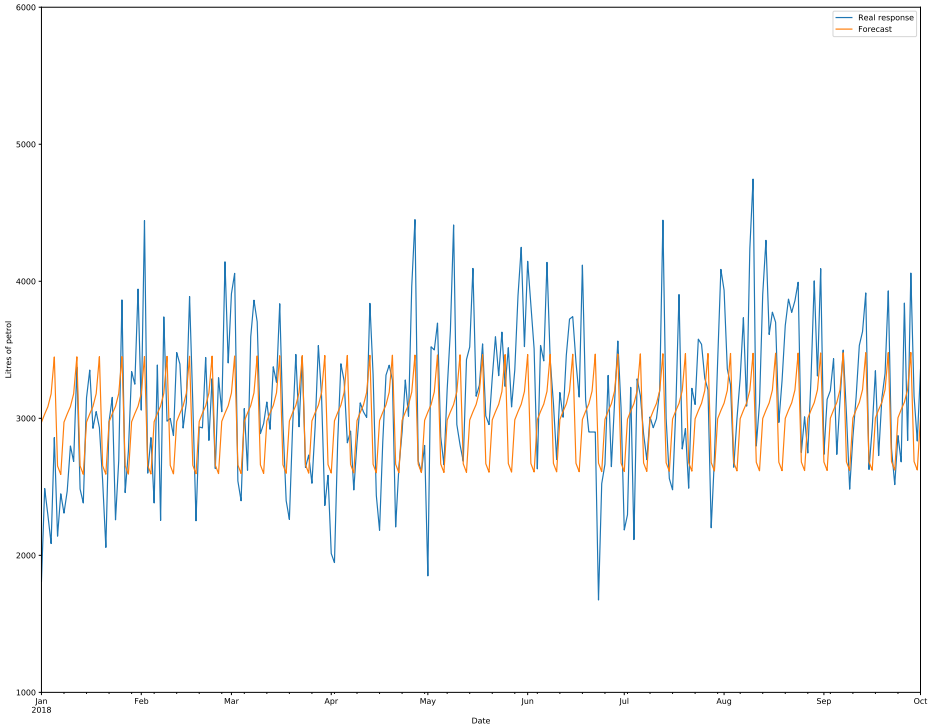

Figure 11: SARIMA fit

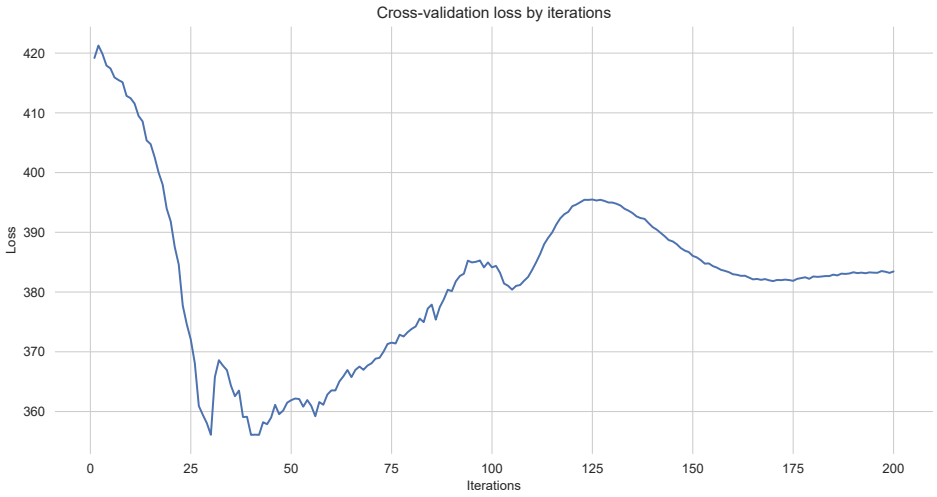

Figure 12: Cross-validation error by the number of iterations

way to approach this challenge is to restart the optimization process from several different starting points randomly. For periodic kernels, the values are fixed and set to assumed seasonal patterns of sales according to the investigation on the data - weekly and yearly.

4.4.2 Number of training iterations

Implementation of the experiments in the study uses a python library GPyTorch. It is a high-performance implementation of the Gaussian process model that can also utilize GPU. Experiments were done on an Nvidia 2080 SUPER GPU and AMD Ryzen 3900X CPU.
The training process is done using gradient descent towards reducing the training loss. Usually, the first iterations reduce loss by a lot then converge to some value. This training process can be stopped when a certain tolerance on further loss reductions is reached, or a satisfying number of iterations is completed. However, the training loss minimization for an excessive number of iterations can cause overfitting, in particular, worse performance on the test dataset. To estimate the optimal number of iterations, a cross-validation learning procedure was applied. The testing period of a year was divided into weeks, and each week forecasting performance was evaluated by training on all the historical data until that week. Evaluation result was recorded after each performed training iteration. The evaluation metric value for a certain number of iteration is the average of performances on the weeks. The results are shown in Figure 12. Experiments with different kernels showed that with the given data, it is enough to only do up to 200 iterations and take the best result.

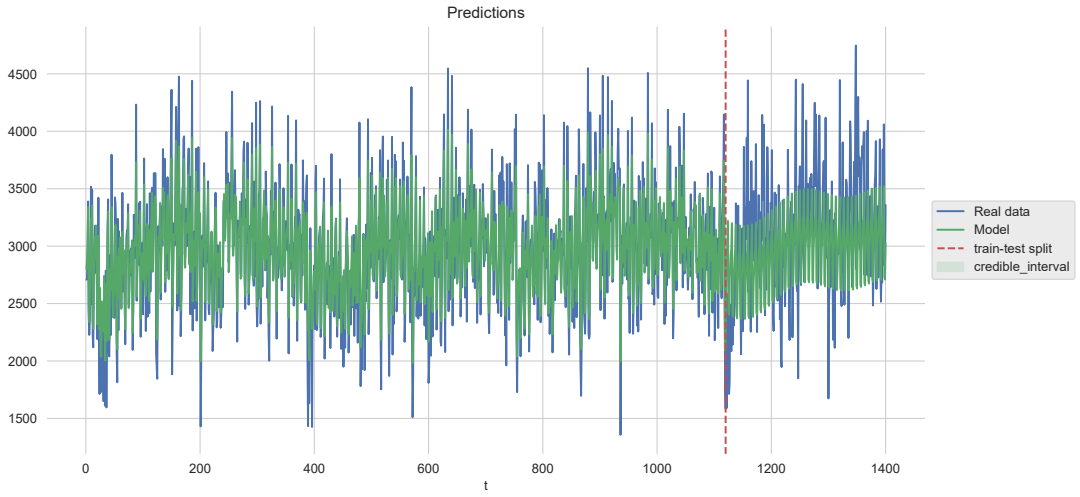

Figure 13: $k_{weekly} + k_{yearly} + k_{rbf}$ fit. t represents days

Kernel configuration	MAE	MAPE	RMSE	Week MAE	Week MAPE
$k_{\text{weekly}} + k_{\text{yearly}}$	356.1	11.43%	437.86	1500.22	6.76%
k_{weekly}	405.48	13.35%	485.67	1909.02	8.89%
$k_{\text{weekly}} \times k_{\text{RQ}}$	398	12.88%	476.63	1957.61	9.05%
$k_{\text{yearly}} \times k_{\text{weekly}}$	381.41	12.51%	461.1	1722.41	8%
$k_{\text{weekly}} \times k_{\text{RQ}} + k_{\text{yearly}} \times k_{RQ}$	397.84	12.86%	475.13	1950.87	8.99%

Table 3: GPR results with different kernels

4.4.3 Forecasting

Table 3 shows the results of the GPR model for several selected kernel combinations and accuracy measures. Several kernels and their combinations were attempted. Firstly, we used each of the kernels individually: RBF, RQ, weekly periodic and yearly periodic kernels. In order to model both seasonalities the weekly and yearly periodic kernels were summed to create a weekly+yearly kernel whose fit is shown in Figure 15. More complex kernel combinations were attempted, including a quasi-periodic kernel weekly multiplied by RQ which performed well in [44]. RBF is generally good for interpolating smooth functions however it did not do much for our seasonal time series forecasting problem as seen in Figure 14. We attempted to add a smooth component to weekly+yearly kernel with RBF but it negatively affected the performance of the model. The fit can be seen in Figure 13.

The best performing kernel, according to the experiments, is a sum of weekly and yearly kernels. An example fit of the kernel can be seen in Figure 15. Some basic kernels were tested but had significantly worse performance and were not included in the results.

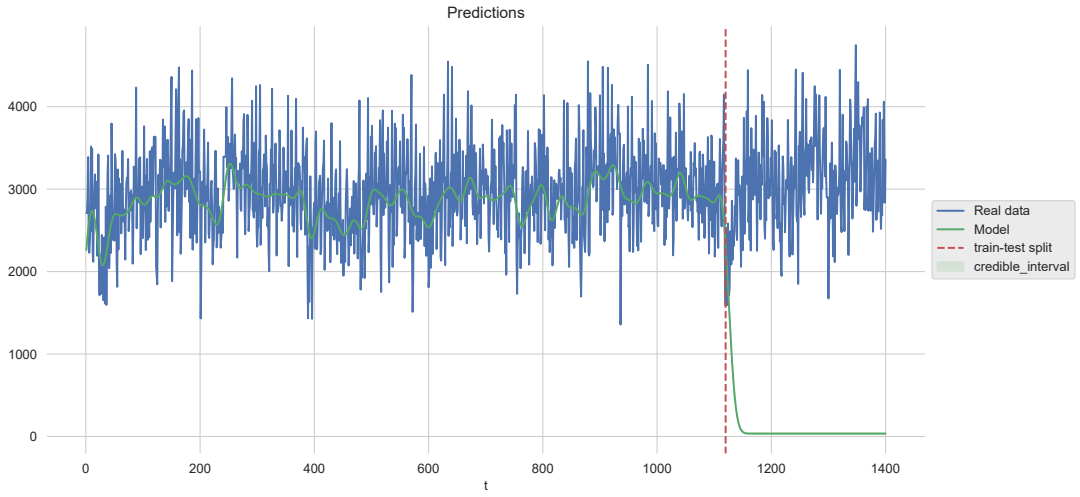

Figure 14: k_{rbf} fit. t represents days

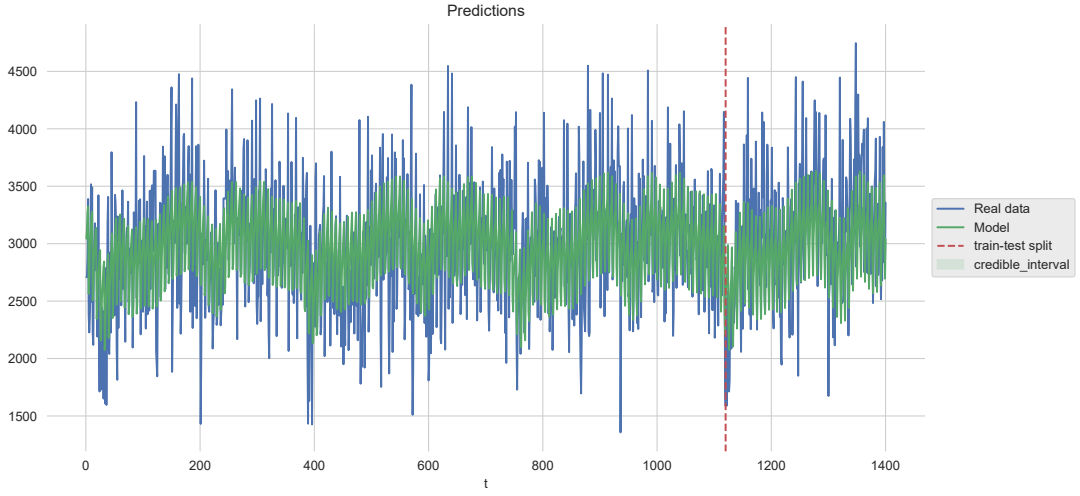

Figure 15: $k_{weekly} + k_{yearly}$ fit. t represents days

Model	MAE	MAPE	Week MAE	Week MAPE
GPR($k_{\text{weekly}} + k_{\text{yearly}}$)	356.1	11.43%	1500.22	6.76%
GPR($k_{\text{yearly}} \times k_{\text{weekly}}$)	461.41	12.51%	1722.41	8%
SARIMA(3,1,1,1,1,1,7)	483.25	14.70%	1792.63	8.08%

Table 4: GPR and SARIMA comparison

4.5 GPR and SARIMA Comparison

Table 4 compares the performance of GPR and SARIMA models on the dataset. The best performing models were selected for comparison. In the case of GPR both sum and product of periodic kernels outperformed the SARIMA model, which attempts to model only the weekly period. As the figures demonstrate, SARIMA uses a static weekly pattern of sales to explain the graph while GPR naturally captures the sales pattern which forms throughout the year. It is also worth noting that not only GPR performs better on a long-term prediction of 3 months, but also in auto-regressive prediction for the next week, which was used in the cross-validation process, which created the metrics for the table. This is most likely due to the weekly pattern still being dependent on the time of the year as discussed in subsubsection 4.1.3, so GPR was able to use that information to provide better results. However, to be fair on the short-term weekly prediction SARIMA result is not much worse than GPR with the difference of only 1.32% in week MAPE.

5 Conclusions and discussion

The thesis develops a forecast model for petrol sales in gas station networks. Different kernels were compared in the GPR model, and the sum of the weekly and yearly periodic kernel was found out to perform the best. Both assumed seasonalities made an impact on the forecasting model since the sum of the kernels outperforms both simple kernels. Unlike in the study by Tolba et al. [44], more complex quasi-periodic kernels created by multiplying kernels did not perform better than a simple sum of periodic kernels. This can be related to 2 simultaneous seasonalities in the series. It is possible that there is a different way to incorporate 2 periodic kernels to achieve higher forecasting accuracy. The performance of the best GPR model is an improvement on the attempted SARIMA models caused by its ability to model both seasonalities in a simple formulation. This case highlights the flexibility advantage of the GP model. It is likely that in cases when more seasonalities or other domain knowledge can be expressed in kernels, the gap between the model performances would be larger.

It is worth mentioning that performance on the individual day sales forecasting (MAE) is not always monotonous with the performance on the week's sales forecasting (Week MAE). This is clearly demonstrated at Table 2. Consequently, a different model would perform optimally for a particular task - predicting next day's sales or predicting cumulative week's sales. It is also shown that the week error is always smaller since day errors can partially compensate each other.

We believe that a forecasting model with less than 7% relative error can be useful for the formulated problem of estimating the next week's petrol sales and preparing an optimal route plan for product deliveries to the stations in the network. Moreover, the confidence interval lets us explain to the business how reliable the forecast is and implement the necessary risk assessment procedures to the logistic processes.

The model can further be improved by incorporating marketing factors that can affect the sales in short term such as price changes and discounts. The model we specified only uses the values of the time series into account and relies on the weekly and yearly seasonalities which makes it simple enough to apply to nearly any petrol station. An overview of the model performance on a huge mass of stations can be carried out in the future.

References

[1] Natural gas transmission and storage facilities: National emission standards for hazardous air pollutants. https://www.epa.gov/stationary-sources-air-pollution/natural-gas-transmission-and-storage-facilities-national-emission. Accessed: 2021-04-15.

[2] Fuel dispenser market size, share & covid-19 impact analysis, by fuel type (petrol/gasoline, diesel, cng, and others), by dispenser system (submersible system and suction system), by flow meter (mechanical and electronic), and regional forecast, 2020-2027. https://www.fortunebusinessinsights.com/industry-reports/fuel-dispensers-market-100431. Accessed: 2021-04-15.

[3] Iea (2020), key world energy statistics 2020, iea, paris. https://www.iea.org/reports/key-world-energy-statistics-2020. Accessed: 2021-04-15.

[4] Ratnadip Adhikari and Ramesh K Agrawal. An introductory study on time series modeling and forecasting. *arXiv preprint arXiv:1302.6613*, 2013.

[5] A. Azadeh, R. Arab, and S. Behfard. An adaptive intelligent algorithm for forecasting long term gasoline demand estimation. *Expert Systems with Applications*, 37(12):7427–7437, 2010. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2010.03.009. URL https://www.sciencedirect.com/science/article/pii/S0957417410001909.

[6] David Barber. *Bayesian reasoning and machine learning*. Cambridge University Press, 2012.

[7] George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

[8] Michael W Browne. Cross-validation methods. *Journal of Mathematical Psychology*, 44(1):108–132, 2000. ISSN 0022-2496. doi: https://doi.org/10.1006/jmps.1999.1279. URL https://www.sciencedirect.com/science/article/pii/S0022249699912798.

[9] Peter Bühlmann and Sara Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

[10] Kenneth P Burnham and David R Anderson. A practical information-theoretic approach. *Model selection and multimodel inference*, 2, 2002.

[11] C.E.Rasmussen and C.K.I.Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

[12] Chris Chatfield. Model uncertainty and forecast accuracy. *Journal of Forecasting*, 15(7):495–508, 1996.

[13] Lu Chen, Yuyi Chen, and André Langevin. An inverse optimization approach for a capacitated vehicle routing problem. *European Journal of Operational Research*, 2021. ISSN 0377-2217. doi: https://doi.org/10.1016/j.ejor.2021.03.031. URL https://www.sciencedirect.com/science/article/pii/S0377221721002575.

[14] John H Cochrane. Time series for macroeconomics and finance. *Manuscript, University of Chicago*, pages 1–136, 2005.

[15] Fabien Cornillier, Gilbert Laporte, Fayez F Boctor, and Jacques Renaud. The petrol station replenishment problem with time windows. *Computers & Operations Research*, 36(3):919–935, 2009.

[16] David A Dickey and Wayne A Fuller. Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American statistical association*, 74(366a):427–431, 1979.

[17] David Duvenaud. The kernel cookbook: Advice on covariance functions. *URL https://www. cs. toronto. edu/˜ duvenaud/cookbook*, 2014.

[18] Gidon Eshel. The yule walker equations for the ar coefficients. *Internet resource*, 2:68–73, 2003.

[19] Tingting Fang and Risto Lahdelma. Evaluation of a multiple linear regression model and sarima model in forecasting heat demand for district heating system. *Applied Energy*, 179:544–552, 2016. ISSN 0306-2619. doi: https://doi.org/10.1016/j.apenergy.2016.06.133. URL https://www.sciencedirect.com/science/article/pii/S0306261916309217.

[20] Julian Faraway and Chris Chatfield. Time series forecasting with neural networks: a comparative study using the air line data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 47(2):231–250, 1998.

[21] Mehdi Forouzanfar, A Doustmohammadi, Samira Hasanzadeh, et al. Transport energy demand forecast using multi-level genetic programming. *Applied Energy*, 91(1):496–503, 2012.

[22] C. Fyfe, Tzai-Der Wang, and Shang Chuang. Comparing gaussian processes and artificial neural networks for forecasting. 01 2006. doi: 10.2991/jcis.2006.7.

[23] Carl Friedrich Gauss. *Theoria motus corporum coelestium in sectionibus conicis solem ambientium*, volume 7. FA Perthes, 1877.

[24] Seymour Geisser. The predictive sample reuse method with applications. *Journal of the American statistical Association*, 70(350):320–328, 1975.

[25] Paul Goodwin, J.Keith Ord, Lars-Erik Öller, Janet A Sniezek, and Mike Leonard. Principles of forecasting: A handbook for researchers and practitioners: J. scott armstrong (ed.), (2001), boston: Kluwer academic publishers, 849 pages.

hardback: Isbn: 0-7923-7930-6; 190, 133, 210.00, paperback: Isbn: 07923-7401-0; 95; 66.50, 105. *International Journal of Forecasting*, 18(3):468–478, 2002. ISSN 0169-2070. doi: https://doi.org/10.1016/S0169-2070(02)00034-1. URL https://www.sciencedirect.com/science/article/pii/S0169207002000341.

[26] Douglas M Hawkins. The problem of overfitting. *Journal of chemical information and computer sciences*, 44(1):1–12, 2004.

[27] Keith W Hipel and A Ian McLeod. *Time series modelling of water resources and environmental systems*. Elsevier, 1994.

[28] Rob J. Hyndman and Anne B. Koehler. Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4):679–688, 2006. ISSN 0169-2070. doi: https://doi.org/10.1016/j.ijforecast.2006.03.001. URL https://www.sciencedirect.com/science/article/pii/S0169207006000239.

[29] K. Kowalska and L. Peel. Maritime anomaly detection using gaussian process active learning. In *2012 15th International Conference on Information Fusion*, pages 1164–1171, 2012.

[30] CD Lai. First order autoregressive markov processes. *Stochastic processes and their applications*, 7(1):65–72, 1978.

[31] Junsoo Lee. Univariate time series modeling and forecasting (box-jenkins method). *Econ 413, lecture 4*.

[32] Maurice Wentworth Lee. *Economic fluctuations: an analysis of business cycles and other economic fluctuations*. RD Irwin, 1955.

[33] Zheng Li, John M. Rose, and David A. Hensher. Forecasting automobile petrol demand in australia: An evaluation of empirical models. *Transportation Research Part A: Policy and Practice*, 44(1):16–38, 2010. ISSN 0965-8564. doi: https://doi.org/10.1016/j.tra.2009.09.003. URL https://www.sciencedirect.com/science/article/pii/S0965856409000998.

[34] Luca Marinai. Gas-path diagnostics and prognostics for aero-engines using fuzzy logic and time series analysis. 2004.

[35] Charles I Mosier. I. problems and designs of cross-validation 1. *Educational and Psychological Measurement*, 11(1):5–11, 1951.

[36] Thanapant Raicharoen, Chidchanok Lursinsap, and Paron Sanguanbhokai. Application of critical support vector machine to time series prediction. In *Proceedings of the 2003 International Symposium on Circuits and Systems, 2003. ISCAS'03.*, volume 5, pages V–V. IEEE, 2003.

[37] Raghavendra D Rao and Jyoti K Parikh. Forecast and analysis of demand for petroleum products in india. *Energy policy*, 24(6):583–592, 1996.

[38] Carl Edward Rasmussen. Gaussian processes in machine learning. In *Summer school on machine learning*, pages 63–71. Springer, 2003.

[39] Carl Edward Rasmussen and Zoubin Ghahramani. Occam's razor. *Advances in neural information processing systems*, pages 294–300, 2001.

[40] S. Roberts, M. Osborne, M. Ebden, S. Reece, N. Gibson, and S. Aigrain. Gaussian processes for time-series modelling. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1984):20110550, 2013. doi: 10.1098/rsta.2011.0550. URL https://royalsocietypublishing.org/doi/abs/10.1098/rsta.2011.0550.

[41] Frank Spitzer. *Principles of random walk*, volume 34. Springer Science & Business Media, 2013.

[42] M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):111–133, 1974. doi: https://doi.org/10.1111/j.2517-6161.1974.tb00994.x. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1974.tb00994.x.

[43] Lijun Sun, Xiuwu Xing, Yaxian Zhou, and Xiangpei Hu. Demand forecasting for petrol products in gas stations using clustering and decision tree. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, 22(3):387–393, 2018.

[44] Hanany Tolba, Nouha Dkhili, Julien Nou, Julien Eynard, Stéphane Thil, and Stéphane Grieu. Ghi forecasting using gaussian process regression: kernel study. *IFAC-PapersOnLine*, 52(4):455–460, 2019. ISSN 2405-8963. doi: https://doi.org/10.1016/j.ifacol.2019.08.252. URL https://www.sciencedirect.com/science/article/pii/S2405896319305890. IFAC Workshop on Control of Smart Grid and Renewable Energy Systems CSGRES 2019.

[45] Lorenzo Trippa, Levi Waldron, Curtis Huttenhower, Giovanni Parmigiani, et al. Bayesian nonparametric cross-study validation of prediction methods. *Annals of Applied Statistics*, 9(1):402–428, 2015.

[46] Mike West and Jeff Harrison. *Bayesian forecasting and dynamic models*. Springer Science & Business Media, 2006.

[47] W. Yan, H. Qiu, and Y. Xue. Gaussian process for long-term time-series forecasting. In *2009 International Joint Conference on Neural Networks*, pages 3420–3427, 2009. doi: 10.1109/IJCNN.2009.5178729.

[48] Roy D Yates and David J Goodman. Probability and stochastic processes. *John Willey & Sons*, 1999.