Stockholms universitet

Using Time Series Analysis to Forecast Daily Municipal Water Demand

Rickard Strandberg

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2015:24
Matematisk statistik
September 2015

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Mathematical Statistics
Stockholm University
Bachelor Thesis **2015:24**
http://www.math.su.se

Using Time Series Analysis to Forecast Daily Municipal Water Demand

Rickard Strandberg*

September 2015

Abstract

In this report, the daily water demand of the Swedish municipality of Vetlanda is analyzed, for the purpose of forecasting. By regressing on calendar effects found in the data, and assuming a seasonal ARIMA structure for the errors, a suitable model is selected through a combination of the Ljung-Box test, the Akaike Information Criterion, and backtesting procedures. The resulting ARIMA$(1,0,5)(1,1,2)_7$ error model with t(5.4)-distributed innovations is then estimated, and a 90 day forecast is provided.

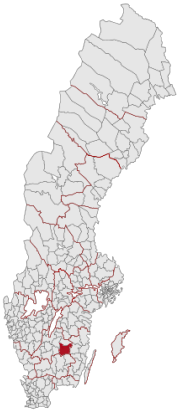
*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: RickardJS@gmail.com. Supervisor: Joanna Tyrcha, Mathias Lindholm, Karl Rökaeus.

Acknowledgments

I would like to thank my project supervisors Joanna Tyrcha, Mathias Lindholm, and Karl Rökaeus for their advice and help on this report; as well as Daniel Mattsson from VETAB for providing me with the data.

1 Introduction

The importance of water in our lives dates back thousands of years—to the dawn of irrigation and sanitation—and has become an integral part of society, influencing everything from public health, to energy, manufacturing, and recreation. Thus, providing access to clean water, and understanding the demand for it, is of great interest. This report aims to take the daily water demand of one particular municipality, and attempt to predict the future demand, with the help of time series modeling.



The purpose of this report is to forecast the water demand of Vetlanda, a 1 500 m^2 municipality in Sweden having almost 27 000 inhabitants (according to SCB, 2014). Being able to predict the demand is useful for the water providers, since it can help them plan future operations, such as the upgrade of facilities, or the changing of water prices. It also helps predict the impact of, say, a new factory opening, or a new residential area being built. Finally, it can also help detect anomalies, such as water leaks and faulty equipment.

Several different methods for analyzing and forecasting water demand have been suggested, including principal component analysis [9] [10], transfer functions [13], artificial neural networks [3], and univariate time series modeling [6]. A variant of the last of these will be the focus of this report.

Previous work in this field have primarily focused on places with warmer climates, i.e. Australia [9] [18], Spain [6], South Korea [10], and the southern U.S. [13]. In such areas rainfall was often considered to have a significant effect on the water demand—and [13] even included temperature—which meant that in the warmer and dryer summer months, these areas saw an increase in demand. However, since the opposite seemed true for Vetlanda, these variables were not considered. Instead, the various weekends and holidays seemed to reduce the demand significantly, and a regression was made using these as dummy variables—with an ARIMA model fitted for the regression errors.

2 Theory

2.1 Time Series

A *time series* is a sequence of observations (of some quantity) taken over an extended period of time. Usually these observations are made at a set time interval—hourly, daily, monthly, etc. Such examples include the hourly number of cars arriving at an intersection; the daily closing price of a stock; and the monthly revenue of a store.

This report will focus on a time series that displays a serial correlation, i.e. that past observations are correlated with the present one. This is a common property of many time series, and thus literature on the subject is abundant (see [5] and [16] for example).

The following concepts and tests are all to serve us in working with this kind of correlation.

2.2 Stationary and weakly stationary time series

A time series $\{Y_t\}$ is said to be a *stationary* process if any sequence $(Y_{t_1}, Y_{t_2}, ..., Y_{t_n})$ has the same joint distribution as $(Y_{t_1+l}, Y_{t_2+l}, ..., Y_{t_n+l})$ for any integer l. In other words, we have the same distribution no matter the starting point of our observations—it is time invariant. This is a strong condition (see [15]), so a less restrictive condition is often used (section 1.2.1 of [5]). A time series is *weakly stationary* if it has a constant mean and the covariances between observations is time invariant.

2.3 Backshift Operator

We can establish an alternative notation for our time series by defining an operator B on a time series $\{Y_t\}$, with the property

$$BY_t = Y_{t-1}.$$

That is, an operator which changes an observation to the one from the previous time step. This B is called the *backshift operator*. You can go back further by using powers of B, since in general

$$B^k Y_t = Y_{t-k}$$

for any integer k. This will provide a useful notation for our time series models, as we shall soon see. For more on the properties and implications of the backshift operator, see [14].

2.4 ARMA and ARIMA

The *AutoRegressive Moving Average* model [5][16] is a time series model of the form

$$Y_t - \sum_{i=1}^{p} \phi_i Y_{t-i} = \varepsilon_t - \sum_{i=1}^{q} \psi_i \varepsilon_{t-i}, \tag{1}$$

where $\{Y_t\}$ is the observation series; $\{\varepsilon_t\}$ is the series of innovations; $p, q \geq 0$ and $\phi_1, \phi_2, ..., \phi_p, \psi_1, ..., \psi_q$ are parameters. This is called the ARMA(p,q) model. The *autoregressive* part of the name refers to the left-hand side, where we regress on past observations; and the *moving average* part refers to the right-hand side, where the past (random) innovations are allowed to influence the expectation of the current observation (they randomly "move the average").

We can rewrite this with the help of the backshift operator by writing

$$(1 - \sum_{k=1}^{p} \phi_k B^k) Y_t = (1 - \sum_{k=1}^{q} \psi_k B^k) \varepsilon_t,$$

where B is the backshift operator; $p, q \geq 0$, and $\phi_1, ..., \phi_p, \psi_1, ..., \psi_q$ are the same as in (1).

A way to generalize this model a bit is to consider the ARIMA(p,d,q) model, where the 'I' stands for *integrated*. In backshift notation it looks like this:

$$(1 - \sum_{k=1}^{p} \phi_k B^k)(1 - B)^d Y_t = (1 - \sum_{k=1}^{q} \psi_k B^k)\varepsilon_t.$$

What is new is the factor $(1 - B)$ on the left-hand side, and what it does is transform the data by letting $Z_t = Y_t - Y_{t-1}$, generating a new time series $\{Z_t\}$. This is referred to as taking the first difference of the series, and the d means that this is done d times. The model is therefore called 'integrated' because the model is still written in terms of $\{Y_t\}$, with the differencing being a part of the model itself.

Let us simplify our notation, by introducing the following naming conventions (slightly modified from [5]):

$$\begin{aligned}
\nabla :=& (1 - L), && \text{the (first) difference operator,} \\
a_p(B) :=& (1 - \sum_{k=1}^{p} \phi_k B^k), && \text{the AR(p) backshift polynomial,} \\
m_q(B) :=& (1 - \sum_{k=1}^{q} \psi_k B^k), && \text{the MA(q) backshift polynomial.}
\end{aligned}$$

With these, we can write our ARIMA(p,d,q) model as

$$a_p(B)\nabla^d Y_t = m_q(B)\varepsilon_t.$$

This will be convenient, and almost necessary, as we expand our model in the following section.

2.4.1 Seasonal ARIMA

In many time series, we encounter seasonal patterns. For daily data there could be a weekly pattern (all Mondays are similar, etc.), or for monthly data a yearly pattern could exist (more sales in a store each December, for instance). For such a time series we might consider the following model:

$$(1 - \sum_{k=1}^{P} \Phi_k (B^s)^k)(1 - B^s)^D Y_t = (1 - \sum_{k=1}^{Q} \Psi_k (B^s)^k) e_t.$$

What we have here is, in some sense, an ARIMA model, but where all lags are multiples of some positive integer s called the *seasonality*. For weekly patterns in daily data, $s = 7$; for yearly in monthly data, $s = 12$; etc. We can name the parts of this model similarly to the regular ARIMA, by letting

$$\begin{aligned}
\nabla_s :=& (1 - B^s) \\
A_P(B^s) :=& (1 - \sum_{k=1}^{P} \Phi_k (B^s)^k) \\
M_Q(B^s) :=& (1 - \sum_{k=1}^{Q} \Psi_k (B^s)^k).
\end{aligned}$$

Once again, we can get a similar, compact notation in

$$A_P(B^s)\nabla_s^D Y_t = M_Q(B^s)e_t \tag{2}$$

Now, there is no reason to expect that the series $\{e_t\}$ is uncorrelated, so a natural step is to continue by fitting a regular ARIMA model to the innovations, i.e.

$$a_p(B)\nabla^d e_t = m_q(B)\varepsilon_t, \tag{3}$$

where $\{\varepsilon_t\}$ is (hopefully) an iid series. Substituting for e_t in (2) and (3), we can write the so called *Seasonal ARIMA model* as

$$a_p(B)A_P(B^s)\nabla^d\nabla_s^D Y_t = m_q(B)M_Q(B^s)\varepsilon_t,$$

and refer to it as the ARIMA(p,d,q)(P,D,Q)$_s$ model, which contains all necessary information—namely the degrees of all polynomials, both degrees of differencing, and the seasonality s.

2.5 The Autocorrelation Function

If we apply the usual definition of correlation to a time series $\{Y_t\}$, we can define the *autocorrelation* between observations at times t and $t-k$ as

$$\rho_k := \rho_{Y_t,Y_{t-k}} = \frac{Cov(Y_t, Y_{t-k})}{\sqrt{Var(Y_t)Var(Y_{t-k})}}.$$

If we can assume that $\{Y_t\}$ is weakly stationary, the autocorrelation becomes

$$\rho_k = \frac{Cov(Y_t, Y_{t-k})}{Var(Y_t)},$$

for *any* t, making it exclusively a function of k. We denote this the *lag-k autocorrelation*.

Given an observed time series $\{y_t\}_{t=1}^T$, we estimate this as one does correlations in general, through

$$\hat{\rho}_k := \frac{\sum_{t=k+1}^T (y_t - \bar{y})(y_{t-k} - \bar{y})}{\sum_{t=1}^T (y_t - \bar{y})^2},$$

where $\bar{y} = \frac{1}{T}\sum_{t=1}^T y_t$. This is called the *autocorrelation function* (ACF), and is crucial in time series analysis.

Under the null hypothesis that the autocorrelation for a given lag is zero, this estimate is asymptotically normal, with mean zero, and variance 1/T (see [5], section 2.1.6). We can therefore create an approximate rejection region for this hypothesis as $|\hat{\rho}_k| > 1.96/\sqrt{(T)}$, and use this to quickly determine which autocorrelations are significantly different from zero.

2.6 The Partial Autocorrelation Function

In addition to the ACF, we can design another nifty tool for specifying the orders of ARIMA processes. We will call it the *partial autocorrelation function* (PACF), and there are several ways of defining it. A more technical definition

is given in section 3.2.5 of [5], but we will choose another, courtesy of 2.4.2 in [16].

Consider the following AR models (in linear notation):

$$\begin{aligned}
Y_t =& \phi_{1,1} Y_{t-1} + \varepsilon_t \\
Y_t =& \phi_{2,1} Y_{t-1} + \phi_{2,2} Y_{t-2} + \varepsilon_t \\
Y_t =& \phi_{3,1} Y_{t-1} + \phi_{3,2} Y_{t-2} + \phi_{3,3} Y_{t-3} + \varepsilon_t \\
Y_t =& \phi_{4,1} Y_{t-1} + \phi_{4,2} Y_{t-2} + \phi_{4,3} Y_{t-3} + \phi_{4,4} Y_{t-4} + \varepsilon_t \\
& ...
\end{aligned}$$

These parameters can be estimated using the least squares method, for these are just multiple linear regression models. We define the estimates of $\phi_{k,k}$ ($k = 1, 2, 3, ...$) as the lag-k PACF of our Y_t. This can be interpreted as the effect Y_{t-k} would have, if added to an AR(k-1) model.

We would then expect that a true AR(p) process would show large PACF for lags up to, and including, lag-p; but PACF close to 0 beyond p. This gives us a way to determine the appropriate order of AR model to use: We simply look for the lag after which the PACF "cuts off", and use that as our p (see section 6.2 of [5]).

2.7 Ljung-Box Test

We would like to have some kind of test to determine whether our ARIMA model provides an adequate fit. The Ljung-Box test was first proposed in [12], and consists of using the statistic

$$Q(m) = n(n+2) \sum_{k=1}^{m} \frac{\hat{r}_k^2}{n-k},$$

where r_k is the lag-k autocorrelation of the residual series. It tests the hypothesis $H_0 : r_1, r_2, ..., r_m = 0$, against $H_1 : r_k \neq 0$, for some $k = 1, 2, ..., m$. This statistic is then approximately $\chi^2(m)$ distributed for large n [2]. In other words, we can test if there is any autocorrelation unaccounted for in our model, provided we choose a large enough m. For our purposes, $m = 20$ will suffice.

2.8 AIC

Whenever we try to fit a model to data, we are only making guesses. We can qualify these guesses by studying the data, and using various tests and criteria, but we can seldom hope to find the true structure of our data.

So, what are we to do when our tests and criteria can produce multiple qualified guesses? It is proposed in [1] to use the following statistic for directly comparing statistical models:

$$AIC := 2k - 2\text{LogL},$$

where k is the number of model parameters that need estimation, and LogL is the maximum value of the log-likelihood function [11] (which is at the parameter estimates by definition).

This is called the *Akaike Information Criterion*, and asserts that the adequate model with the smallest AIC is preferable.

3 Data

The data set under consideration consists of the daily water demand in the municipality of Vetlanda, Sweden. More specifically it is the amount of water measured (in m^3), that leaves the central water distribution center.

The supplied data can be seen in Figure 1. Upon inspection, we can see a clear weekly seasonal pattern. It would appear that the demand is lower by several hundred m^3 during Saturdays and Sundays. Furthermore, we see each year in July, a similar reduction in demand that persists for several weeks. There is also a brief drop the last week of every year.

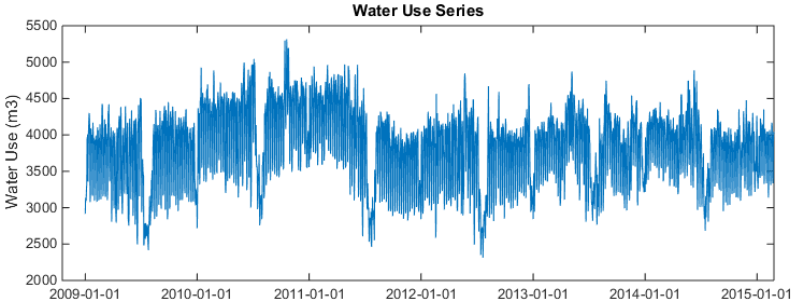


Figure 1: Time series plot of daily water demand for Vetlanda.

Naturally, this is no coincidence, and we can quite easily make the connection between these patterns and industry. Indeed, it is common for factories to close down (or at least reduce capacity) during weekends and holidays, in addition to four weeks each July (sometimes referred to as the Industrial Vacation in Sweden).

So, what about other days where water demand is low? Referencing [17], which lists and describes all national holidays, we can in fact see similar reduction in demand for those dates. An important thing to note, however is that many of these are moving holidays. While Christmas Eve is always on December 24, Easter always starts on a Friday (and not always in April!), so some holidays are not necessarily 365 days apart. So, while [6] had grounds to assume a yearly seasonal effect, we could in our case make a different one; namely that of a *calendar effect*.

This motivates us to introduce two dummy variables: One x_{1i} for weekends and holidays, as described in [17]; and one x_{2i} for the industrial vacation each July. The reason we separate the two can be seen in Figure 1. During the industrial vacations, water demand seems to drop more than during a normal weekend. We could reason that this is due to the would-be factory workers going on vacation outside the municipality during this time. We can also see that a smaller weekend effect persists, so having both together is reasonable.

Lastly, it should be noted that when a holiday occurs on a Tuesday or a Thursday, it is customary to extend the weekend to include that Monday or Friday, respectively. This is reflected in the data also. For a full list of which dates were considered weekends or holidays, see the Appendix.

4 Analysis

Before we begin the process of fitting a suitable model for forecasting, we must first sketch out the procedure. Following the general one detailed in [5] (specifically chapter 9), we recognize the following three steps to specify our model:

1. Remove calendar effects by fitting a regression model to our data.
2. Remove the seasonal effect from the residuals by fitting an appropriate ARIMA(0,0,0)(P,D,Q)$_s$ model.
3. Consolidate the seasonal part of this model, with an appropriate ARIMA(p,d,q) model to remove the remaining autocorrelation.

Once we have accomplished this, we can estimate the model using the exact maximum likelihood method described in [5], and forecast future water demand.

4.1 Calendar Effects

Following the discussion of section 3, regarding holidays and vacations, we would like to propose the following regression model to account for these calendar effects:

Let $X_1 = (x_{11}, x_{12}, ..., x_{1T})'$ and $X_2 = (x_{21}, x_{22}, ..., x_{2T})'$ be vectors of zeros and ones, with ones indicating holidays in X_1 and industrial vacation in X_2; and let $\{y_t\}$ be the water demand series. Our proposed regression model is then

$$Y_t = \alpha + \beta_1 x_{1t} + \beta_2 x_{2t} + \beta_3 x_{1t} x_{2t} + e_t,$$

where $\alpha, \beta_1, \beta_2, \beta_3$ are unknown parameters, and $\{e_t\}$ is some autocorrelated time series. We have included the product of the dummy variables due to the smaller weekend effects during industrial vacations.

To proceed, we need to study $\{e_t\}$. But how, when we all we can do is work with the residuals

$$\hat{e}_t = y_t - \hat{\alpha} - \hat{\beta}_1 x_{1t} - \hat{\beta}_2 x_{2t} - \hat{\beta}_3 x_{1t} x_{2t}$$

of some estimate (i.e. least squares) of our regression model? Fortunately, according to [8], these residuals have autocovariances which are asymptotically equivalent to those of e_t, so in terms of determining the time series structure of $\{e_t\}$, we have access to all our usual tools, such as the ACF, PACF, and Ljung-Box test.

What we now need is the ordinary least squares estimates for our parameters, which result in the model

$$y_t = 4163 - 834 x_{1t} - 1012 x_{2t} + 580 x_{1t} x_{2t} + \hat{e}_t,$$

where we have assumed for the moment that the $\hat{e}_t$ are iid normal and uncorrelated.

We can quickly note that the estimates are just about what we expected from studying Figure 1. On regular business days we get about 4000m^3, which goes down by some 800m^3 during weekends. Once the industrial vacation starts it drops by $1000m^3$ during week days, and another $800 - 600 = 200m^3$ those weekends.

Let us now return to the residuals. In Figure 2, we have plotted first the fitted values and the actual data; and second the residuals $\hat{e}_t$. For ease of inspection, only the year 2014 is shown. We can see that the fitted model follows the data quite well, making appropriate jumps when expected.

Looking at the residual plot in Figure 2, we see significant autocorrelation, including some semblance of weekly seasonality. This is unfortunate, in the sense that we would have been significantly closer to our solution; but very much expected, looking at the data and the nature of our problem.

Figure 2: Fitted values and residuals of our basic regression model.

In Figure 3 below, we have a normal QQ plot of the residuals. In it, we can actually see that the residuals follow the normal distribution quite well. This is a useful property, since it gives us confidence in various asymptotic results, such as the efficiency of the least-squares estimators, or the normality of the autocorrelations (see [2]). These properties are necessary for the various tests we will employ later, and we can now reasonably assume that our sample size is sufficient for these to hold.

4.2 Seasonality

Equipped with the residual series $\{e_t\}$, we now begin the process of finding an appropriate ARIMA model for our regression errors. We start by plotting the residuals. In Figure 4 we have plotted the residuals, as well their ACF and PACF. We can see strong autocorrelations in all three plots, but we pay special attention to the spikes that occur at lag multiples of 7 in both the ACF and PACF. This seems to suggest weekly seasonality, and we address it immediately.

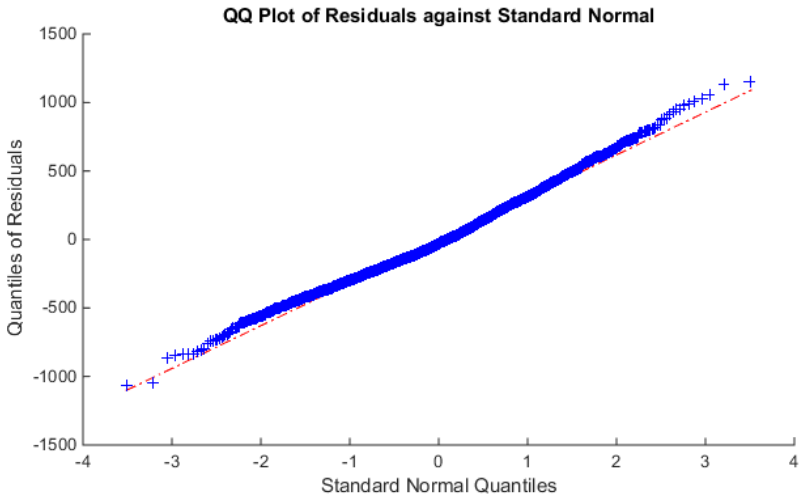


Figure 3: Quantile-Quantile plot for the residual series. Adherence to the straight line implies normal distribution.

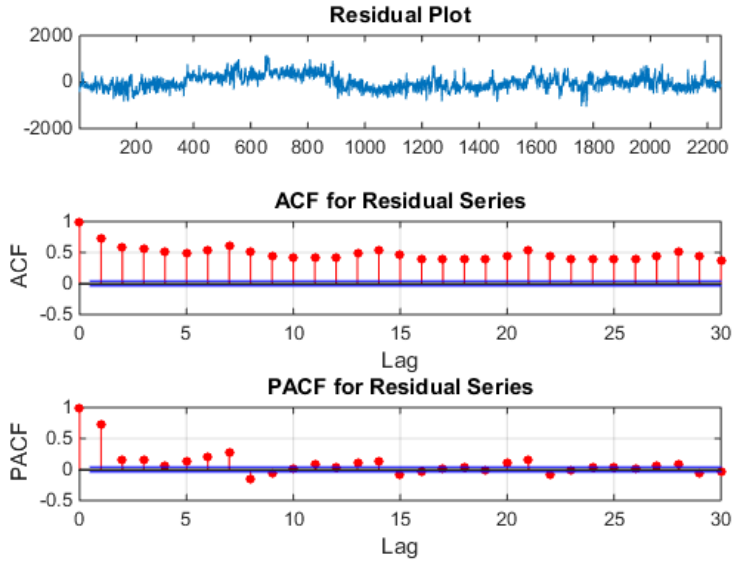


Figure 4: Diagnostics plots for the residual series. These show significant autocorrelation, particularly lag 7 seasonal.

Taking the seventh difference of our residual series, we get the rather trivial ARIMA(0,0,0)(0,1,0)$_7$ model

$$\nabla_7 e_t = \varepsilon_t,$$

where $\{a_t\}$ supposedly is iid. We can then find out if there is any autocorrelation left, by creating similar plots as in Figure 4 for the new ARIMA(0,0,0)

$(0,1,0)_7$ residuals (which, in reality, is just the differenced series).

This results in Figure 5, which shows a vast improvement. The new residuals now seem centered around 0, and the ACF tends to 0 relatively quickly. But we can still see significant spikes in the PACF at the same lag multiples of 7, and even the ACF has a hefty spike at lag 7. This all suggest that some seasonality may still persist, and so we look to include some seasonal AR and MA terms in our model.

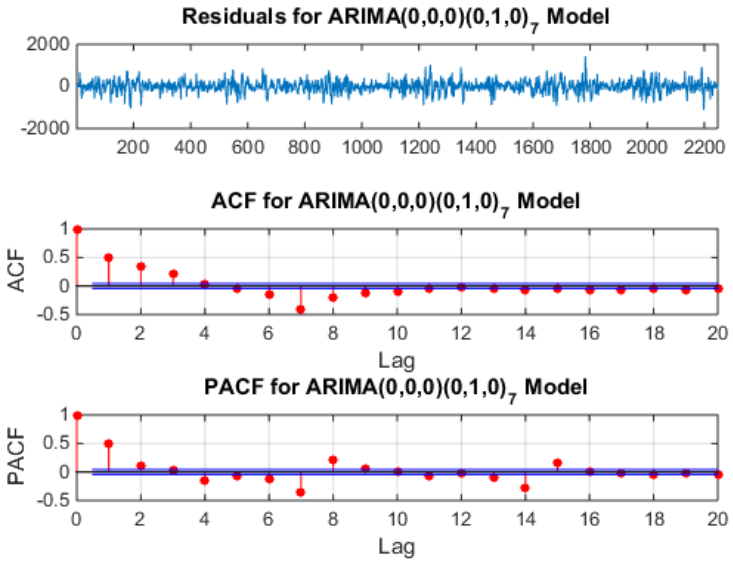


Figure 5: Residual, ACF, and PACF for ARIMA$(0,0,0)(0,1,0)_7$.

After trying some low order ARIMA$(0,0,0)$(P,1,Q)$_7$ models, we find that (P,Q)=(0,1),(0,2),(1,1),(1,2) handle the remaining seasonality quite well. Because the four different variants turn out to be graphically indistinguishable, we only show the plots for ARIMA$(0,0,0)(1,1,2)_7$ in Figure 6, but for the sake of completeness, the rest can be found in the Appendix. In the figure we can see that all seasonal effects have disappeared from the ACF, and while lags 7 and 14 still stand out slightly in the PACF, increasing P and Q does not seem to help us. So we will be content with these lower order models as we proceed to the next step. Hopefully, once we are done with that, these lags will also be close to 0.

While we would like to proceed with only one of the seasonal parts above, we do not have to. As we will determine and estimate full models using a computer algorithm in the proceeding section, we can simply repeat it for all four variants (at the expense of computational time only).

4.3 Completing the Model

From Figure 6 we still see significant autocorrelation. So our next step will be to specify the nonseasonal part of our ARIMA model, and our approach will

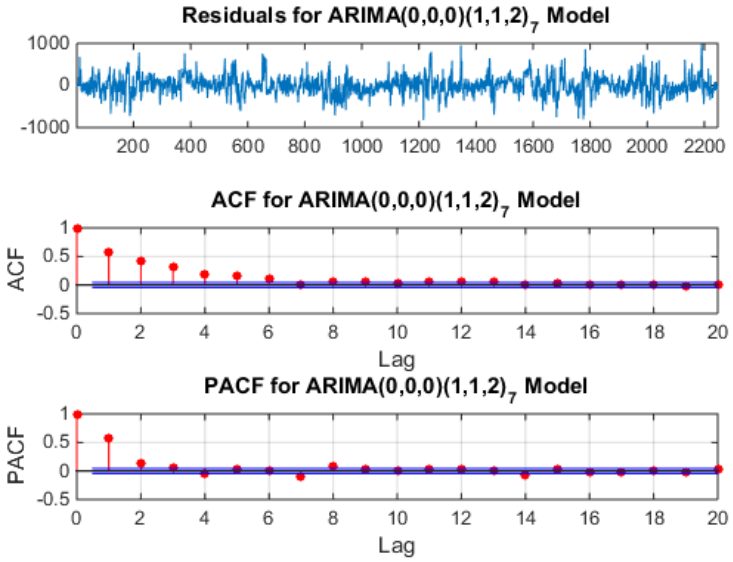


Figure 6: Residual, ACF, and PACF for ARIMA$(0,0,0)(1,1,2)_7$.

be one of brute force. If we suppose that the best ARIMA models will contain one of the four seasonal parts developed in the previous section, and that the non-seasonal part will not involve lags greater than nine, then we will have a total of $4 \cdot 10 \cdot 10 = 400$ possible candidates. That is, we look at the models

$$y_t = \alpha + \beta X_t + e_t, \quad a_p(B)A_P(B^7)\nabla_7 e_t = m_q(B)M_Q(B^7)\varepsilon_t. \tag{4}$$

where $p, q \in \{0, 1, ..., 9\}, P \in \{0, 1\}$, and $Q \in \{1, 2\}$. We can see from Figure 6 that no non-seasonal differencing is necessary (zero-mean residuals, and the ACF tends to zero), and we pick nine as the maximum lag because it was the maximum lag for the resulted model in [6].

In order to narrow down our search, we will employ two criteria for choosing what models we want to proceed with. They are:

1. The residuals of a good model should pass a Ljung-Box test for lag 20. That is, $Q(20) = 0$ for any such model.

2. The best of these should be among the five models with the lowest AIC.

While the particular boundaries are arbitrary, they are arguably reasonable. The downside to this is that we still need to estimate all 400 models in order to use these criteria. However, with the aid of a computer this becomes a mere matter of time, rather than effort. We can even go back to considering the time series $\{y_t\}$ itself, and estimate the entire regression model—now with ARIMA errors—without additional work on our end.

There is one significant problem with this type of estimation, however. It turns out that—because we have seasonally integrated ARIMA models—the regression constant is unidentifiable, and thus cannot be estimated. To understand why, see Appendix. Our solution to this predicament will be to fixate

the constant to be that of the least squares estimate in Section 4.1. It is, after all, the mean of the values for which the other regressors are zero, which is a reasonable choice at least. We make no claims that this is the best estimate of this constant, but it allows us to proceed.

In order to build the maximum-likelihood functions used for estimating these models, we need a distribution for the innovations ε_t. The easiest would be to assume a normal distribution, but we need to verify this.

Using the ACF and PACF of the $\{\hat{e}_t\}$ series acquired in Section 4.1, we can find that an ARIMA$(1,0,4)(1,1,2)_7$ model—under the assumption of normally distributed innovations—passes our Ljung-Box test. Looking at Figure 7 however, where we have made a QQ plot and a histogram for this model's residuals, we can see that—while they do not follow a normal distribution—they do make out a bell shape, with zero mean. It would then seem like we are better off assuming a $t(\nu)$ distribution, for some parameter ν we can estimate along with the rest of our models.

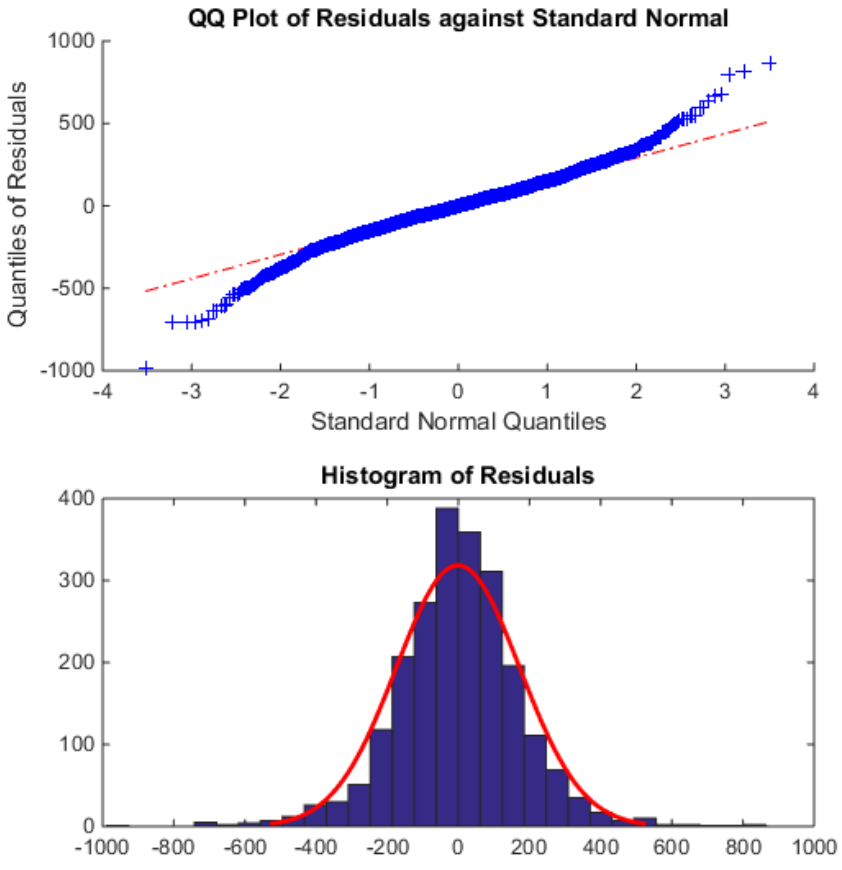


Figure 7: QQ plot and histogram for the ARIMA$(1,0,4)(1,1,2)_7$ model residuals.

ARIMA Model	AIC
$(6,0,5)(0,1,1)_7$	29443
$(1,0,5)(1,1,2)_7$	29456
$(5,0,1)(0,1,2)_7$	29456
$(1,0,4)(0,1,2)_7$	29456
$(5,0,1)(1,1,2)_7$	29457
$(2,0,4)(1,1,2)_7$	29457
$(1,0,5)(0,1,2)_7$	29457
$(1,0,6)(1,1,2)_7$	29457
$(6,0,1)(1,1,2)_7$	29457
$(6,0,1)(0,1,2)_7$	29458
$(7,0,1)(1,1,2)_7$	29458
$(2,0,5)(1,1,2)_7$	29458
$(1,0,4)(1,1,1)_7$	29459
$(7,0,1)(0,1,2)_7$	29459
$(1,0,7)(1,1,2)_7$	29459
...	...

Table 1: The top of the list of ARIMA models for which $Q(20) = 0$, sorted by lowest AIC.

We are now ready to perform maximum-likelihood estimation for our 400 potential models, and employ the criteria we just established (For an explanation of how this is done, see Appendix). This results in 132 models passing the Ljung-Box test, with the ones with lowest AIC presented in Table 1. We can see that the ARIMA(6,0,5)$(0,1,1)_7$ model has significantly lower AIC than the rest, but instead of rejecting all other models outright, we will stay true to our plan, and proceed with the top five.

4.4 Selection Through Backtesting

Since we are interested in fitting a model for the purpose of forecasting future water demand, a reasonable final selection would be based on the competing models' ability to forecast. We will thus end our analysis with a form of *backtesting*. For us, this means we will look at how well the models would have performed if we had implemented them a year ago, and let them forecast up until now. It would then stand to reason, that the model that performed best then, would have the best conditions to perform well going forward.

We design the following metric for measuring forecasting performance:

$$[\text{Forecast Performance}] := \frac{\sum_{i=1}^{n}(y_{t+i} - \hat{y}_{t+i})^2}{n}, \tag{5}$$

where $\hat{y}_{t+i}$ is the i:th forecasted day from the starting time t, and n is the total number of days forecasted in our backtest. This is a metric similar to the *mean squared forecast error* (MSFE) [16], and a smaller value means better performance.

What the best performing model is might depend on the type of forecasting we use. One model might be better for short-term prediction, while another would be better for longer periods. Since we do not know how the contents of

this report would be used, we can without much additional effort study several different forecasting periods. We therefore select the four most reasonable, where forecasting is done every 1, 7, 30, or 90 day(s). For the periods larger than 1 we make sure to only forecast each date *once*.

The backtesting procedure will thus be as follows, where we have a given model and forecast period k:

1. Go back one year (365 days) from the most recent data point (which in this case is 2014-02-15).
2. Take this point and the five years (1825 days) leading up to it as our data sample.
3. Use this sample to estimate the parameters of the model in question.
4. Predict the value of the next k time steps recursively by plugging in the sample data and its inferred innovations.
5. If our total data allows us to forecast another period, take the day of the k:th forecast and go back to step 2. Otherwise, continue to step 6.
6. Take the total of n forecasted values and their respective observed data points, and calculate the forecast performance according to (5).

Performing this procedure for our five models and four forecasting periods yields Table 2.

ARIMA Model	**Forecast Period**			
	1 Day	**7 Days**	**30 Days**	**90 Days**
$(6,0,5)(0,1,1)_7$	33710	39603	50644	<u>54180</u>
$(1,0,5)(1,1,2)_7$	<u>32135</u>	38611	<u>43961</u>	54234
$(5,0,1)(1,1,2)_7$	32411	38726	45174	55833
$(5,0,1)(0,1,2)_7$	32262	<u>38041</u>	45025	54385
$(1,0,4)(0,1,2)_7$	32409	39158	44547	54868

Table 2: The forecasting performance of the different models, for several different forecast periods. The lowest are underlined.

In it we can see that indeed, different models performed better at different forecast periods. We first note that the ARIMA(6,0,5)$(0,1,1)_7$ model—while having the lowest AIC of all—performed far worse than the rest in all but the 90 day forecasts, in which it did only slightly better than the ARIMA(1,0,5)$(1,1,2)_7$. In fact, the ARIMA(1,0,5)$(1,1,2)_7$ model had arguably the best performance overall, being at least second best for all four periods. While we could choose a different model depending on which forecast period we want to use, we will be making the conclusion that the ARIMA(1,0,5)$(1,1,2)_7$ model is universally preferred, and will be the sole model presented in the results.

5 Results

After considering the various calendar effects, and seasonal structure of the data—as well as backtesting over multiple forecast periods—we arrive at the

conclusion that the regression model with ARIMA(1,0,5)(1,1,2)$_7$ errors is the best among the ones considered, for the purpose of forecasting future water demand in this particular municipality. Using the entire data set to estimate the parameters, we get

$$\begin{aligned} y_t &= 4163 - 661x_{1t} - 649x_{2t} + 505x_{1t}x_{2t} + e_t, \quad \text{and} \\ a_1(B)A_1(B^7)\nabla_7 e_t &= m_5(B)M_2(B^7)\varepsilon_t, \end{aligned}$$

where x_{1t} and x_{2t} are indicator variables, indicating if day t is part of a weekend/holiday or the industrial vacation respectively; and

$$\begin{aligned} a_1(B) &= 1 - 0.942B, \\ A_1(B^7) &= 1 + 0.201B^7, \\ m_5(B) &= 1 - 0.375B - 0.111B^2 - 0.054B^3 - 0.088B^4 + 0.024B^5, \\ M_2(B^7) &= 1 - 0.672B^7 - 0.242B^{14}, \\ \varepsilon_t &= 139.79z_t \end{aligned}$$

where B is the backshift operator, and z_t is estimated to be $t(5.4)$-distributed.

Using this estimated model, we can provide a 90 day forecast, which is perhaps a more useful result than these particular parameter estimates (which should be re-estimated with every use, anyway). This forecast can be found listed in the Appendix, but is plotted here in Figure 8.

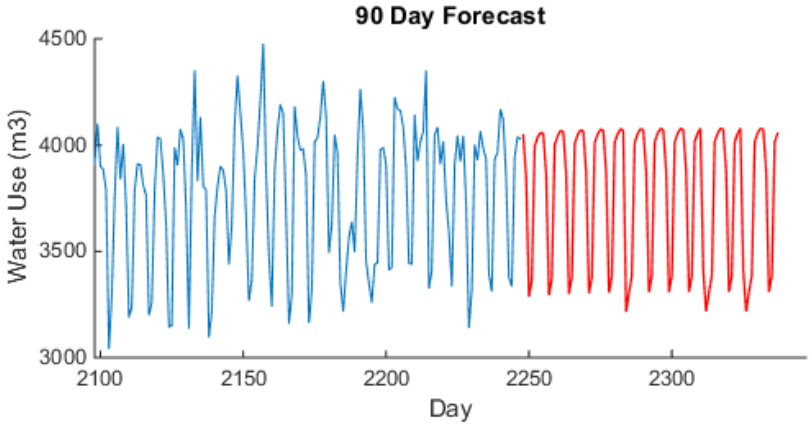


Figure 8: A 90 day forecast using a regression model with ARIMA(1,0,5)(1,1,2)$_7$ errors.

6 Discussion

While we arrived at a reasonably well performing model, it comes with its share of caveats. We made a deliberate choice of regressors, which is certainly open to criticism. We chose a rather simple regression model, focusing on the impact industry had on the water demand, but there is no reason to suspect that this

is the only factor of note. Even so, the way we *did* account for industry is very binary and simple. We could for instance have looked into the fact that some factories shut down early on Fridays (something which is actually reflected in the data, with Fridays having a lower mean than the total).

There is also the matter of our estimate of the regression constant, and the fact that we could not estimate it as a part of our likelihood optimization. But since the least squares estimate we used instead represents the mean of the regular non-holiday, non-vacation days; letting this be the mean of our series, and letting the indicators shift the water demand down from this mean, seems quite reasonable. So, by fixing this as the constant, we can be confident that we are not committing any grave errors.

As far as further analysis is concerned, we can see from Figure 9 that, while no significant autocorrelations persist, the residual variance seems far from constant. In particular, the larger residuals seem clustered, which would suggest that some sort of heteroscedastic model for the variance is in order—such as the ARCH model of [7], or the GARCH model introduced in [4]. While outside of the scope of this report, such analysis would open up other powerful tools for forecasting, such as reliable confidence intervals and simulation possibilities.

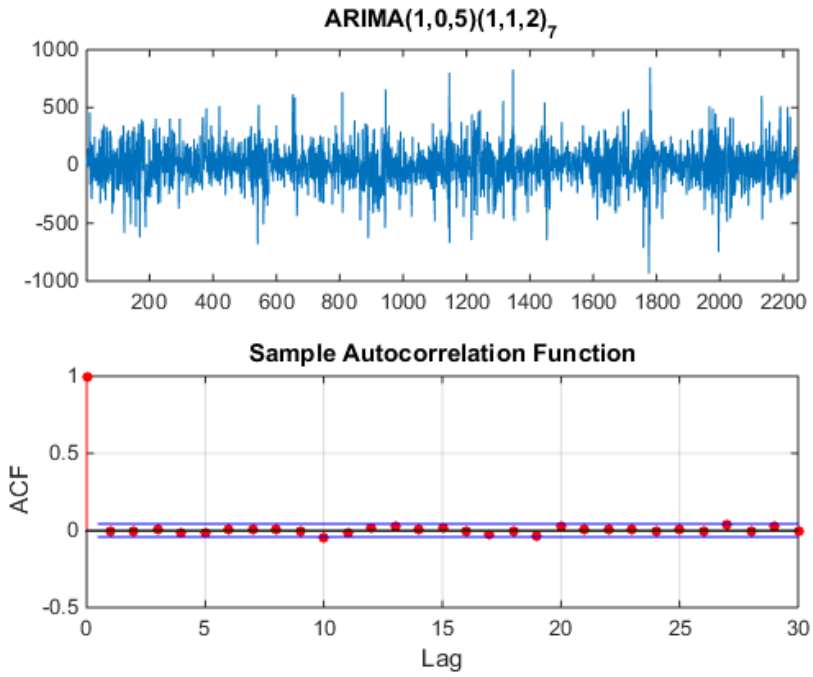


Figure 9: Diagnostic plots for our regression model with ARIMA$(1,0,5)(1,1,2)_7$ errors.

7 Appendix

7.1 Identifiability Concerns

Suppose we have a regression model with ARIMA errors, i.e.

$$Y_t = \alpha + \beta X_t + e_t, \quad \text{where} \quad D(B)e_t = E(B)\varepsilon_t,$$

for some backshift polynomials D and E specified by the model. The likelihood function is based on the distribution of ε_t, and solving the above for ε_t gives us

$$\begin{aligned}\varepsilon_t =& E^{-1}(B)D(B)(y_t - \alpha - \beta X_t) \\ =& E^{-1}(B)D(B)(y_t - \beta X_t) - E^{-1}(B)D(B)\alpha,\end{aligned}$$

where $E^{-1}(B)$ is an inverse filter of $E(B)$ used by software to estimate ARIMA models. Now, should our model have any sort of integration, we can write $D(B) = g(B)(1-B)^k$, for some other backshift polynomial g, and some integer $k > 0$. We further note—since α is constant—that $B\alpha = \alpha$. Putting all this together, we get

$$\begin{aligned}\varepsilon_t =& E^{-1}(B)D(B)(y_t - \beta X_t) - E^{-1}(B)g(B)(1-B)^k\alpha \\ =& E^{-1}(B)D(B)(y_t - \beta X_t) - E^{-1}(B)g(B)(1-B)^{k-1}(\alpha - B\alpha) \\ =& E^{-1}(B)D(B)(y_t - \beta X_t).\end{aligned}$$

So the likelihood function will not depend on α, and it is thus unidentifiable.

7.2 Maximum Likelihood Estimation

In this section we will use the ARMA(1,1) model as an example, to illustrate how to use maximum likelihood to estimate parameters. The general case of ARIMA(p,d,q) is much the same, but considerably more convoluted.

The difficulty of estimating an ARIMA model's parameters lies in the fact that the model is a function of both past values *and* past innovations—the latter not being observable. Also, if we consider the first observation y_0; in the ARMA(1,1) model this will depend on the observation y_{-1} and innovation ε_{-1} from the previous time step—which we do not have access to! The solution is to let these remain unknown, and treat them as parameters to be estimated. This is called the *unconditional* or *exact* likelihood method [16]. (Note, however that the estimates themselves are not interesting, they are only going to be a part of the likelihood optimization.)

We can by solving recursively for the innovations. We write the ARMA(1,1) model in linear notation:

$$y_t - \phi y_{t-1} = \varepsilon_t - \psi\varepsilon_{t-1},$$

and solve for ε_t:

$$\varepsilon_t = y_t - \phi y_{t-1} + \psi\varepsilon_{t-1}.$$

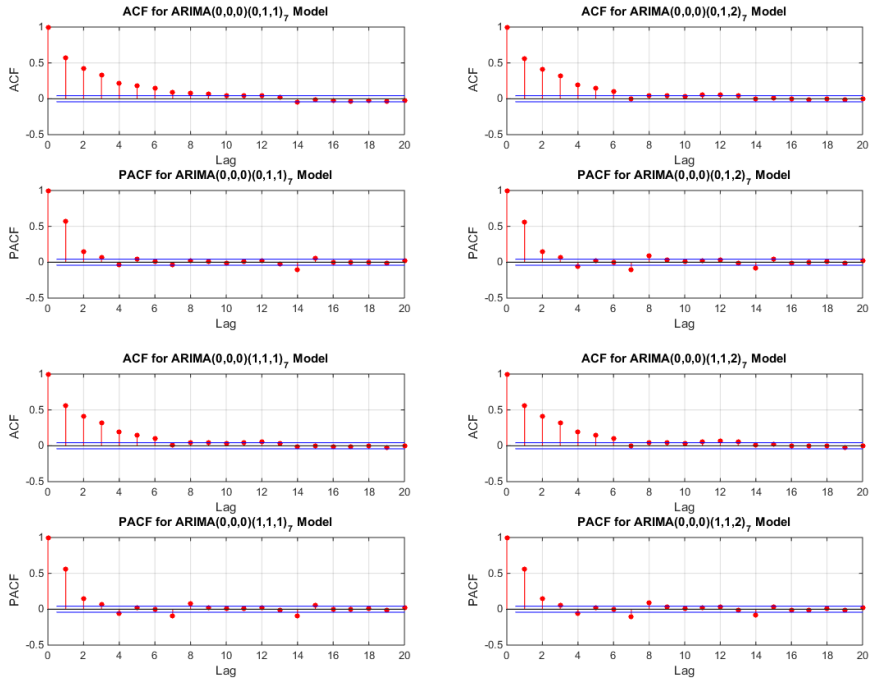
Doing this recursively, starting with ε_0, we get

$$\begin{aligned}
\varepsilon_0 &= y_0 - \phi y_{-1} + \psi\varepsilon_{-1} \\
\varepsilon_1 &= y_1 - \phi y_0 + \psi\varepsilon_0 = y_1 + (\psi - \phi)y_0 - \psi\phi y_{-1} + \psi^2\varepsilon_{-1} \\
\varepsilon_2 &= y_2 - \phi y_1 + \psi\varepsilon_1 = y_2 + (\psi - \phi)y_1 + \psi(\psi - \phi)y_0 - \psi^2\phi y_{-1} + \psi^3\varepsilon_{-1} \\
&\vdots \\
\varepsilon_t &= y_t + (\psi - \phi)\sum_{i=1}^{t} \psi^{i-1} y_{t-i} - \psi^t \phi y_{-1} + \psi^{t+1}\varepsilon_{-1}
\end{aligned}$$

which gives us each innovation as a function of the observations $\{y_t\}$, the regular parameters ϕ and ψ, and the additional parameters y_{-1} and ε_{-1}. We can then build the log-likelihood function by assuming some distribution for these innovations, usually the normal- or t- distribution.

7.3 ACF and PACF of Seasonality Models

Below are the ACF and PACF plots for the seasonality models under consideration in section 4.2.



7.4 Holidays and Vacations

Directly below are the dates—other than the Saturdays and Sundays—that were considered holidays for the purpose of our regression. Further down are the dates for the industrial vacations.

2009-01-01	2009-12-31	2011-01-01	2012-01-06	2013-03-29	2014-04-21
2009-01-02	2010-01-01	2011-01-06	2012-04-06	2013-04-01	2014-05-01
2009-01-05	2010-01-06	2011-04-22	2012-04-09	2013-05-01	2014-05-29
2009-01-06	2010-04-02	2011-04-25	2012-05-01	2013-05-09	2014-05-30
2009-04-10	2010-04-05	2011-05-01	2012-05-17	2013-05-10	2014-06-06
2009-04-13	2010-05-01	2011-06-02	2012-05-18	2013-06-06	2014-06-20
2009-05-01	2010-05-13	2011-06-03	2012-06-06	2013-06-21	2014-12-24
2009-05-21	2010-05-14	2011-06-06	2012-06-22	2013-12-24	2014-12-25
2009-05-22	2010-06-06	2011-06-24	2012-12-24	2013-12-25	2014-12-26
2009-06-06	2010-06-25	2011-12-24	2012-12-25	2013-12-26	2014-12-27
2009-06-19	2010-12-24	2011-12-25	2012-12-26	2013-12-27	2014-12-28
2009-12-24	2010-12-25	2011-12-26	2012-12-27	2013-12-28	2014-12-29
2009-12-25	2010-12-26	2011-12-27	2012-12-28	2013-12-29	2014-12-30
2009-12-26	2010-12-27	2011-12-28	2012-12-29	2013-12-30	2014-12-31
2009-12-27	2010-12-28	2011-12-29	2012-12-30	2013-12-31	2015-01-01
2009-12-28	2010-12-29	2011-12-30	2012-12-31	2014-01-01	2015-01-02
2009-12-29	2010-12-30	2011-12-31	2013-01-01	2014-01-06	2015-01-05
2009-12-30	2010-12-31	2012-01-01	2013-01-06	2014-04-18	2015-01-06

Table 3: Holiday dates.

Start Dates	End Dates
2009-07-11	2009-08-09
2010-07-10	2010-08-08
2011-07-09	2011-08-07
2012-07-07	2012-08-05
2013-07-06	2013-08-04
2014-07-05	2014-08-03

Table 4: Start and end dates for the industrial vacations.

7.5 90 Day Forecast

Date	m3	Date	m3	Date	m3
2015-02-26	4046	2015-03-28	3305	2015-04-27	4015
2015-02-27	3852	2015-03-29	3381	2015-04-28	4054
2015-02-28	3287	2015-03-30	4012	2015-04-29	4078
2015-03-01	3364	2015-03-31	4051	2015-04-30	3411
2015-03-02	3994	2015-04-01	4075	2015-05-01	3217
2015-03-03	4037	2015-04-02	4070	2015-05-02	3309
2015-03-04	4058	2015-04-03	3215	2015-05-03	3384
2015-03-05	4055	2015-04-04	3306	2015-05-04	4015
2015-03-06	3863	2015-04-05	3382	2015-05-05	4054
2015-03-07	3293	2015-04-06	4013	2015-05-06	4078
2015-03-08	3371	2015-04-07	4052	2015-05-07	4073
2015-03-09	4002	2015-04-08	4076	2015-05-08	3878
2015-03-10	4042	2015-04-09	4071	2015-05-09	3309
2015-03-11	4067	2015-04-10	3877	2015-05-10	3384
2015-03-12	4062	2015-04-11	3307	2015-05-11	4015
2015-03-13	3868	2015-04-12	3383	2015-05-12	4054
2015-03-14	3299	2015-04-13	4014	2015-05-13	4078
2015-03-15	3376	2015-04-14	4053	2015-05-14	3412
2015-03-16	4007	2015-04-15	4077	2015-05-15	3217
2015-03-17	4046	2015-04-16	4072	2015-05-16	3309
2015-03-18	4071	2015-04-17	3877	2015-05-17	3384
2015-03-19	4066	2015-04-18	3308	2015-05-18	4015
2015-03-20	3872	2015-04-19	3384	2015-05-19	4054
2015-03-21	3303	2015-04-20	4014	2015-05-20	4078
2015-03-22	3379	2015-04-21	4053	2015-05-21	4073
2015-03-23	4010	2015-04-22	4077	2015-05-22	3878
2015-03-24	4049	2015-04-23	4072	2015-05-23	3309
2015-03-25	4073	2015-04-24	3878	2015-05-24	3385
2015-03-26	4068	2015-04-25	3308	2015-05-25	4015
2015-03-27	3874	2015-04-26	3384	2015-05-26	4054

Table 5: A 90 day forecast using a regression model with ARIMA(6,0,5)$(0,1,1)_7$ errors.

References

[1] Akaike H., *A New Look at the Statistical Model Identification.*
IEEE Transactions on Automatic Control (1974), AC-19, 6, p. 716-723.

[2] Anderson T.W., Walker A.M., *On the Asymptotic Distribution of the Autocorrelations of a Sample From a Linear Stochastic Process.*
Annals of Mathematical Statistics (1964), 35, 3, p. 1296-1303.

[3] Behboudian S., Tabesh M., Falahnezhad M., Ghavanini F.A., *A long-term prediction of domestic water demand using preprocessing in artificial neural network.*
Journal of Water Supply: Research and Technology (2014), 63, 1, p. 31-42.

[4] Bollerslev T., *Generalized Autoregressive Conditional Heteroskedasticity.*
Journal of Econometrics (1986), 31, p. 307-327.

[5] Box G.E.P., Jenkins G.M., Reinsel G.C., *Time Series Analysis: Forecasting and Control* 4th ed.
John Wiley & Sons, Inc. (2008).

[6] Caiado J., *Performance of combined double seasonal univariate time series models for forecasting water consumption.*
Munich Personal RePEc Archive (2009) `http://mpra.ub.uni-muenchen.de/15242/`

[7] Engle R.F., *Autoregressive conditional heteroskedasticity with estimates of the variance of U.K. inflation.*
Econometrica (1982), 50, p. 987-1008.

[8] Fuller W.A., *Introduction to Statistical Time Series.* 2nd ed. Section 9.3.1.
John Wiley & Sons, Inc. (1996).

[9] Haque M.M., Rahmana A., Hagarea D., Kibria G., *Principal Component Regression Analysis in Water Demand Forecasting: An Application to the Blue Mountains, NSW, Australia.*
ResearchGate (2008) `https://www.researchgate.net/publication/268209977_Principal_Component_Regression_Analysis_in_Water_Demand_Forecasting_An_Application_to_the_Blue_Mountains_NSW_Australia`

[10] Koo J.Y., Yu M.J., Kim S.G., Shim M.H., Koizumi A., *Estimating regional water demand in Seoul, South Korea, using principal component and cluster analysis.*
IWA Publishing (2005), Water Supply vol. 5, p. 1-7.

[11] Liero H., Zwanzig S., *Introduction to the Theory of Statistical Inference.* Section 3.1.
CRC Press (2012).

[12] Ljung G.M., Box G.E.P., *On a measure of lack of fit in time series models.*
Biometrika (1978), 65, 2, p. 297-303

[13] Maidment D.R., Miaou S.P., *Daily Water Use in Nine Cities.*
Water Resources Research (1986), 22, 6, p. 845-851.

[14] Nau R., *The mathematical structure of ARIMA models* (2014).
`http://people.duke.edu/~rnau/Mathematical_structure_of_ARIMA_models--Robert_Nau.pdf`.

[15] Ross S.M., *Introduction to Probability Models* 10th ed. Section 10.7.
Academic Press (2010).

[16] Tsay R.S., *Analysis of Financial Time Series* 3rd ed.
John Wiley & Sons, Inc. (2010).

[17] Various authors, *Swedish Code of Statutes* (Svensk författningssamling).
SFS 1989:253.

[18] Zhou S.L.,McMahon T.A., Walton A., Lewis J., *Forecasting daily urban water demand: a case study of Melbourne.*
Journal of Hydrology (2000), 236, p. 153–164.