Cleaning and Forecasting Population Data
## Contents

1 Introduction .................................................. 1

2 Data ............................................................ 2
   2.1 Descriptives .............................................. 2
   2.2 Monthly means and reporting effects ...................... 5
   2.3 Christmas effect ......................................... 6
   2.4 Weekends and other holidays ............................. 14
   2.5 Results .................................................. 16

3 Theory .......................................................... 22
   3.1 Naive Methods ............................................ 22
   3.2 (S)ARIMA-Models ......................................... 22
   3.3 Applying (S)ARIMA-Models .............................. 25

4 Forecasts ....................................................... 27
   4.1 Births ........................................................ 27
   4.2 Deaths ...................................................... 29
   4.3 Immigrants ............................................... 31
   4.4 Departures ................................................ 33
   4.5 Main residences growth rate ............................. 34

5 Summary ......................................................... 36

References ....................................................... 37

6 Appendix ......................................................... 38
   6.1 Birth plots ................................................. 38
   6.2 Death plots ............................................... 40
   6.3 Immigrant plots .......................................... 42
   6.4 Migrates plots ............................................ 44
   6.5 Main residence growth rate plots ......................... 46
   6.6 Cleaned data .............................................. 48
1 Introduction

It is obvious that population data is an important source for policy makers. For example, the Zensus 2011 had a huge impact on the inter-state fiscal adjustment which depends on the true population of a state. But also on a smaller scale information on the population is a very important source to determine for example the tax amount, planning the local public transport or - which is obviously an imminent problem in Munich - the need for flats or the rent development in a city.

To analyze or even forecast such data shows sometimes quite interesting patterns which are clearly not part of the data generating process: Reporting effects. These reporting effects will be a focus of this work. The second focus is on forecasting the data. Here we deal with monthly data on the number of births, deaths, influxes, departures and the main residence population in Munich from 1987 until 2013. At first, we describe the data we got from the Statistical Office in Munich in section 2. In this section we also model the reporting effects resulting from the Christmas holidays, weekends or other holidays at the end of the month. In section 3 we want to use a little more sophisticated modeling method and try to forecast the series without cleaning the data of the reporting effects. Section 4 concludes.
2 Data

2.1 Descriptives

The dataset of the Statistical Office of Munich incorporates time series of the births, deaths, influxes and departures of Munich lasting from June 1987 until December 2012. The data was compiled on a monthly basis. The births and deaths are related to the inhabitants of Munich solely. The births exhibit an upward trend which seems to increase over time. The outlier of December 2012 at the end of the time series is probably caused by a reporting effect which we will discuss later.

Figure 1: Series of births, deaths, influxes and departures, primary data.

As seen in figure 1, the deaths decline over time. There is a temporary increase of
the deaths between May and September 2006 possibly resulting from the introduction of a secondary residence tax.

The influxes exhibit an interesting pattern: after reaching a peak in the early 1990s they remain relatively constant. Since 2005 the influxes exhibit an upward trend. Increased immigration from East Germany and refugees of the Yugoslav War caused the peak of influxes in the early 1990s. The cyclical pattern also originates from the fact that students move to Munich in September or October every year for their studies.

With the exception of the late 1980s, the departures seem to have a constant expected value over time. However, there are several outliers and an abnormally low number of departures at the end of the 1980s. The high number of departures in March 2009, with more than 21,000 departures, is the result of the introduction of a tax identification number. The outlier of July 1990 is possibly caused by an adjustment prior to the Bundestag election of 1990.

In addition, the dataset includes monthly time series of the population with main residence, with secondary residence and the total population of Munich lasting from August 1999 to December 2012. The total population is the sum of the population with main residence and the population with secondary residence.

The population of main residence in figure 2 exhibits an upward trend over the time. There is a temporary drop in March 2009 caused by an adjustment caused by the introduction of the tax identification number.

![Figure 2: Series of main and secondary residence population.](image)

The time series of the population with a secondary residence is mainly characterized by the drop between June and July of 2006. This drop was due to the introduction of
the tax on secondary residences in June 2006.

The total population of Munich in figure 3 is the sum of the population of main residence and secondary residence. As in the case of the population of main residence, the drop in March 2009 is caused by an adjustment resulting from the introduction of tax identification number. In general, the monthly change of the total population is not exactly equal to the balance of the births, deaths, influxes and departures.

![Graph showing population changes over years](image)

Figure 3: Series of total residence population.

Using this dataset we explore the relationship between the deviation of the values of the months of January and December of the time series of births, deaths, influxes and departures and the Christmas holidays. Reporting effects resulting from the Christmas holidays shift the date of data entries from December to January. We show that the strength of this Christmas effect depends on how the 24th of December falls each year and adjust for this reporting effect with regards to the weekday of the 24th of December. The adjustment of the Christmas effect results in lower monthly averages for January and higher ones for December because we correct the artificial shift of data entries from December to January. Furthermore, we find other significant report effects resulting from weekends and other holidays like Whit-Monday and adjust for them.

Overall, we show that the adjustment of these reporting effects reduced the amount of outliers and the variance of each time series. Moreover, the adjustment improves the (P)ACFs of the detrended time series of the births and deaths by a remarkable
reduction of the existing lag 12. The ACFs of the adjusted series exhibit a pattern of
damped sinusoids.

This section is structured as follows: Subsection 2 presents the monthly averages for
the time series of the births, deaths, influxes and departures and explains the existing
reporting effects which shift data entries for late December to January. Subsection 3
analyses the Christmas effect and explains the adjustment of the time series for this
effect. In subsection 4 we study the reporting effects of other holidays and weekends
and adjust for these. Finally, subsection 5 presents the adjusted time series and the
(P)ACFs of the adjusted and detrended time series of births and deaths.

### 2.2 Monthly means and reporting effects

Table 1 presents the monthly averages of births, deaths, influxes and departures of the
corresponding time series:

<table>
<thead>
<tr>
<th></th>
<th>births</th>
<th>deaths</th>
<th>influxes</th>
<th>departures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan</td>
<td>1121</td>
<td>1185</td>
<td>7994</td>
<td>7192</td>
</tr>
<tr>
<td>Feb</td>
<td>992</td>
<td>1029</td>
<td>6811</td>
<td>6415</td>
</tr>
<tr>
<td>Mar</td>
<td>1085</td>
<td>1055</td>
<td>7759</td>
<td>7543</td>
</tr>
<tr>
<td>Apr</td>
<td>993</td>
<td>966</td>
<td>7463</td>
<td>6255</td>
</tr>
<tr>
<td>May</td>
<td>1042</td>
<td>965</td>
<td>7143</td>
<td>5702</td>
</tr>
<tr>
<td>Jun</td>
<td>1066</td>
<td>919</td>
<td>7563</td>
<td>6612</td>
</tr>
<tr>
<td>Jul</td>
<td>1154</td>
<td>970</td>
<td>8831</td>
<td>8378</td>
</tr>
<tr>
<td>Aug</td>
<td>1123</td>
<td>943</td>
<td>8640</td>
<td>8302</td>
</tr>
<tr>
<td>Sep</td>
<td>1068</td>
<td>907</td>
<td>10029</td>
<td>8146</td>
</tr>
<tr>
<td>Oct</td>
<td>1143</td>
<td>937</td>
<td>10265</td>
<td>7305</td>
</tr>
<tr>
<td>Nov</td>
<td>1062</td>
<td>1017</td>
<td>8111</td>
<td>7049</td>
</tr>
<tr>
<td>Dec</td>
<td>917</td>
<td>852</td>
<td>6566</td>
<td>6593</td>
</tr>
</tbody>
</table>

Table 1: Monthly averages.
At first, we notice that the average values for December are lower than the values for January in births and deaths as well as in influxes and departures. In particular there is a remarkable difference in the line chart in figure 4 of the births and deaths. The January values of the births are approximately more than 22.2 per cent higher and the ones of the deaths are about 39.0 per cent higher than the corresponding monthly average of December.

Thereeto, the Statistical Office of Munich acknowledged that the time series of births, deaths, influxes and departures may contain reporting effects as a result of the Christmas holidays.

### 2.3 Christmas effect

In this section we analyze the relationship between the strength of this reporting effect and the annual Christmas holidays. If the employees are not at the office during late December and the work remains undone because of the Christmas holidays, they enter up the births, deaths, influxes, departures of late December at the beginning of January. Thus, this reporting effect artificially increases the quantities in January and decreases the ones in December. We aim to adjust the time series for this Christmas effect.

At first, we compute the ratios of January and the previous December as a proxy-variable for the strength of the reporting for the period 1988 to 2011. The nominator
corresponds to January, the denominator to December. The year is related to January. We exclude the ratio of December 2012 and January 2013 because the time series only last until December 2012.

Furthermore, we create the variable weekday which measures the weekday of the 24th of December for each year. Thus, it can attain values between 1 and 7. For instance, 1 corresponds to Monday and 7 to Sunday. The Panels A-D of figure 5 present the scatter charts of the Christmas effect with the variable weekday on the x-coordinate and ratios for January and December on the y-coordinate.

![Figure 5: Scatter plots of the Christmas effect.](image)

The scatter charts exhibit an interesting pattern: if the 24th falls on a Monday and
weekday equals 1, the ratios as a proxy-variable of the Christmas effect are unusual high, indicating a strong Christmas effect. If the 24th falls on a Friday, the ratios are remarkably lower. In Germany the 25th and the 26th of December are public holidays. If the 24th is on a Monday, then the public holidays fall on the workdays resulting in a higher ratio (and a stronger Christmas effect). If the 24th falls on a Friday, the public holidays fall on the weekend and the Christmas effect is less pronounced as expected. If the 24th is on a Sunday, the Christmas effect is higher than if it were on a Saturday because the 25th and 26th fall both on workdays now. The scatter plots indicate that the employees take their holidays also depending on how Christmas falls each year.

Formally, we estimate the Christmas effect for each time series by the following dummy variable regression:

\[
\text{ratio}_i = \beta_0 + \beta_1 d_{1i} + \beta_2 d_{2i} + \beta_3 d_{3i} + \beta_4 d_{4i} + \beta_5 d_{5i} + \beta_6 d_{6i} + u_i
\]

The dependent variable \( \text{ratio}_i \) denotes the ratios of the January values and December values for each year \( i \). The independent variables are six dummy variables for the days Monday to Saturday. We only use 6 variables to avoid perfect multicollinearity. Thus, the intercept measures the strength of the Christmas effect when the 24th falls on a Sunday. The following table 2 shows the adjusted R-squared of the regression for each time series:

<table>
<thead>
<tr>
<th>series</th>
<th>adj. R-squared</th>
</tr>
</thead>
<tbody>
<tr>
<td>births</td>
<td>0.828</td>
</tr>
<tr>
<td>deaths</td>
<td>0.639</td>
</tr>
<tr>
<td>influxes</td>
<td>0.763</td>
</tr>
<tr>
<td>departures</td>
<td>0.438</td>
</tr>
</tbody>
</table>

Table 2: Adjusted \( R^2 \) for each time series.

Given that we only use the weekday of the 24th for the dummy variables, the adjusted R-squared is relatively high, particularly the one of births and influxes. We only use this regression to adjust for the Christmas effect.

We adjust the time series for the Christmas effect by subtracting a certain amount from January which we add to December for each year because the Christmas effect artificially increases the quantities in January and decreases the ones in December. Our starting point for the adjustment is \( \text{ratio}_i \) which describes the ratio \( \frac{\text{January}_i}{\text{December}_{i-1}} \) for each
year \( i \). We can write for a particular \( \text{ratio}_i \) \( \frac{1.58}{1} \) instead of 1.58. Thus, the nominator corresponds to January, the denominator which is normalized to 1 to December. We assume that the monthly quantities of December and January are equal without the Christmas effect. Thus, the target value of the adjusted \( \text{ratio}_i \) is 1. We adjust the ratios by the following formula:

\[
\text{adjustedratio}_i = \frac{\text{ratio}_i - \text{estimated}_i + 1}{1}
\]

The variable \( \text{estimated}_i \) is the estimated ratio of the dummy variable regression. Therefore, \( \text{estimated}_i \) can attain 7 different values in dependency of the weekday of the 24th of the year. We subtract \( \text{estimated}_i \) from \( \text{ratio}_i \) which results in a value next to zero. Thus, we add 1 in order to reach the desired \( \text{adjustedratio}_i \) with a value of circa 1. Then we adjust the values of January and the corresponding December with the following formula:

\[
\text{adjustedJanuary}_i = \frac{\text{sum}_i}{\text{ratio}_i - \text{estimated}_i + 2} \cdot (\text{ratio}_i - \text{estimated}_i + 1)
\]
\[
\text{adjustedDecember}_i = \frac{\text{sum}_i}{\text{ratio}_i - \text{estimated}_i + 2} \cdot 1
\]

\( \text{sum}_i \) is the sum of the absolute values of January and the December for each year \( i \). \( \text{ratio}_i - \text{estimated}_i + 2 \) is the sum of nominator and denominator of \( \text{adjustedratio}_i \). For the adjustment of January we multiply this fraction with the nominator of \( \text{adjustedratio}_i \), for the adjustment of December we multiply the fraction with the denominator of \( \text{adjustedratio}_i \). By this method the sum of the quantities of January and December for each year stays unchanged as required. In other words, we subtract the exact amount from January which we add to December.

Our adjustment corrects these reporting effects in dependency of the weekday of the 24th of December for each year. The Panels A and B of figure 6 present the monthly averages after the adjustment for the Christmas effect.
As expected, we can see a significant decline in the monthly averages of January and an increase in the monthly averages of December resulting from the adjustment. The adjusted monthly averages of January and December are approximately equal because our assumption of the adjustment is that the monthly averages of January and December are equal without the Christmas effect. The Panels A and B of Figure 7-10 incorporate a before and after comparison of the adjustment of the Christmas effect in the form of scatter charts and in absolute terms. The y-coordinate measures the difference between January and the previous December values for each year, while the x-coordinate describes the weekday of the 24th of December. Without the Christmas effect the differences are as high as in Panel B of figure 7-12.
Figure 7: Figure Before and after comparison births.

Figure 8: Before and after comparison deaths.
Figure 9: Before and after comparison influxes.

Figure 10: Before and after comparison departures.

After the adjustment, the differences now seem to be independent of the weekday and of a similar quantity over the weekdays.
Meanwhile the data of the births, deaths, influxes and departures are now available online as of January 2013 on the website of the Statistical Office. The 24th of December 2012 is a Monday which should result in a very strong Christmas effect. Figure 11 and 12 presents the time series of births and deaths from August 2012 to March 2013.

Figure 11: Births including as of 2013. Taken from http://www.mstatistik-muenchen.de/datamon/datamon.jsp?thema=C07

Figure 12: Deaths including as of 2013. Taken from http://www.mstatistik-muenchen.de/datamon/datamon.jsp?thema=C07

Table 3 presents a comparison of the new ratio (January 2013 / December 2012) and the estimated ratio when the 24th December is a Monday. The new ratio of the
births is higher than our estimation of the Christmas effect, while the new ratios of the
deaths, influxes and departures are close to the estimated ratios.

As seen in figure 11 and 12, there is a remarkable difference between the quantities
of December 2012 and January 2013 with the January value being considerably higher.
A visitor of this website might be surprised at that. Yet, this makes sense within the
context of the strength of the Christmas effect depending on the weekday of the 24th
of December.

<table>
<thead>
<tr>
<th>series</th>
<th>January (2013)/December (2013) estimation (weekday = 1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>births</td>
<td>2.09 1.55</td>
</tr>
<tr>
<td>deaths</td>
<td>1.83 1.86</td>
</tr>
<tr>
<td>influxes</td>
<td>1.53 1.49</td>
</tr>
<tr>
<td>departures</td>
<td>1.45 1.55</td>
</tr>
</tbody>
</table>

Table 3: Ratio comparison

2.4 Weekends and other holidays

In this section we look for a reporting effect resulting from weekends and other holidays.
We aim to adjust for these effects. Thereto, we assume that these need to fall exactly
at the end of a month to cause a reporting effect.

The value of a month can be influenced by reporting effects in two ways: non-
working days at the end of the month and non-working days at the end of its previous
month. Thereto, we create the variable noworkdays for each month by subtracting the
amount of non-working days at the end of the previous month from the amount of
the non-working days at the end of that month. Non-working days include Saturdays
and Sundays and Bavarian public holidays like Easter Monday, Good Friday or Whit-
Monday. For this computation we set the non-working days at the end of December at
zero because we already adjusted the Christmas effect in subsection 2.3.

The variable noworkdays can attain values between -4 and 4. For instance, if it
is 0, then there are either no non-working days at the end of this month and of the
previous one or there is an equal amount of non-working days at the end of this month
and the previous one. It can reach the value 4 for March if the 31th of this March is
Easter Monday because of the preceding weekend and Good Friday while there are no
non-working days at the end of the previous February. We estimate the reporting effect
of a non-working day for each time series by the following regression:

\[ y_i = \beta_0 + \beta_1 \text{noworkdays}_i + \beta_2 \text{duration}_i + u_i \]
The dependent variable $y_i$ represents one of our time series which were adjusted for the Christmas effect. $i$ corresponds to the number of the observation. The parameter of interest is $\beta_1$, the coefficient of the independent variable $\text{noworkdays}$. The independent variable $\text{duration}$ controls for the month length. $u_i$ represents the error term. Again, we only do this regression to estimate coefficients for adjustments.

Before doing the regression we delete the first and last values of all the time series (June 1987 and December 2012) because they are outliers. The entries of June 1987 are the first values of the time series. We suppose that December 2012 is an outlier due to the Christmas effect. Moreover, we exclude the months between June 1987 and March 1989 of the time series of departures due to the abnormally low departures for that time.

The following table 4 presents the estimated coefficients which describe the estimated effect of a non-working day on the monthly value:

<table>
<thead>
<tr>
<th>series</th>
<th>estimated coefficient</th>
<th>average amount per day</th>
</tr>
</thead>
<tbody>
<tr>
<td>births</td>
<td>-21.13</td>
<td>35.5</td>
</tr>
<tr>
<td>deaths</td>
<td>-21.33</td>
<td>32.66</td>
</tr>
<tr>
<td>influxes</td>
<td>-199.2</td>
<td>270.6</td>
</tr>
<tr>
<td>departures</td>
<td>-123.58</td>
<td>238</td>
</tr>
</tbody>
</table>

Table 4: Estimated coefficients of reporting effects and average amounts per day.

As expected the estimated coefficients have all negative signs. A non-working day at the end of a month shifts the quantities into the next month and artificially reduces the quantities of the present month. However, the absolute amounts of our estimated coefficients seem to be a bit too small compared with the average amount per day because the reporting effect of a non-working day at the end of a month should be about as high if the present month would have 1 day less.

We adjust each month of the time series of births, deaths and influxes by the following formula:

$$\text{adjustedmonth}_i = \text{month}_i - \text{coefficient}_k \cdot \text{noworkdays}_i$$

The monthly value after the adjustment is $\text{adjustedmonth}_i$, whereas $i$ corresponds to the number of the month in the time series. The variable is the monthly value before the adjustment. $\text{coefficient}_k$ is the estimated coefficient depending on the indices $k$ for each time series. The variable $\text{noworkdays}_i$ describes how many more non-working days at the end of a month than non-working days there are than at the end of the previous month. So for each additional non-working day at end of the previous month in comparison to the end of this month, we subtract a certain amount
from this month (depending on the time series) to correct the reporting effect. The sum of the adjustments is zero because we add the same amount to a month to which we subtract from another.

2.5 Results

We adjusted the time series of births, deaths, influxes and departures for reporting effects resulting from the Christmas holidays, weekends and other holidays. The Panels A and B of Figure 13-16 present a before and after comparison of each time series.

Figure 13: Before and after comparison of the births-series.
Figure 14: Before and after comparison of the deaths-series.

Figure 15: Before and after comparison of influxes-series.
The graphical before and after comparison shows there are less outliers in the adjusted time series. In addition to this, the variance of each time series declined after the adjustment.

Second, we want to compare the auto-correlation function and the partial auto-correlation function of the initial time series and the adjusted time series of births and deaths. But according to Hamilton (1994) this requires stationary time series. The initial and adjusted time series of births, deaths do not satisfy the stationary conditions because the expected value varies over time (see Figure 13 and 14). The time series of births exhibit an upward trend, while the time series of deaths shows a downward tendency. In general, there are 2 types of trends: stochastic trends and deterministic trends. If it is a deterministic trend, shocks have permanent effects. In case of a stochastic trend shocks have no permanent effects (see Hamilton 1994). We run the Augmented Dickey-Fuller test to test whether the underlying trend is a stochastic trend. The p-value of the Dickey-Fuller test for both time series is smaller than 0.01 and we must reject the null hypothesis. Thus, the underlying trend is not a stochastic trend. Since both the time series exhibit an upward or downward tendency, it must be a deterministic trend. We estimate a deterministic trend by a polynomial of 2nd degree for the time series of births and deaths by the ordinary least squares method:

\[ y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + u_t \]
$y_t$ can either be the time series of births or deaths. $y_t$ consists of a deterministic function of time $\beta_0 + \beta_1 t + \beta_2 t^2$ and a stationary random variable $u_t$. In the second step we subtract the deterministic trend from both time series to get the detrended time series $x_t$.

$$
x_t = y_t - (\beta_0 + \beta_1 t + \beta_2 t^2)
$$

The detrended time series $x_t$ equals a stationary random variable and should be stationary with an expected value of zero over time. The detrended time series seem to exhibit stationary because the expected value and the variance remains fairly constant over time now. Figure 17 and 18 present a before and after comparison of the auto-correlation function and the partial auto-correlation function of the detrended time series of births and deaths.
Figure 17: Before and after comparison of the ACF PACF births.
Figure 18: Before and after comparison of the ACF PACF deaths.

The lag 12 of the ACF of both time series is remarkably reduced but remains significant as a result of the adjustment of the reporting effects. The ACFs of both time series now exhibit a pattern of damped sinusoids which could indicate seasonality. Additionally, the lag 12 of the PACF of both time series decreases. The adjustment intensifies the pattern of damped sinusoids of the PACF. The lag 12 remains still distortive to estimate an economically parameterized ARMA-Modell.
3 Theory

In this section we present and explain the methods for forecasting a time series we use. We start with some naive methods which will serve as benchmarks. Afterwards, autoregressive moving average models are explained, which are the main model class we use. Applying these models needs sometimes some data transforms which are also explained in the third subsection. The fourth subsection gives some details on the theory of forecasting ARIMA-models.

3.1 Naive Methods

Following Hyndman and Athanasopoulos (2013), who describe some forecasting methods as "very simple and surprisingly effective" which can serve as benchmarks for more complicated models, we want to introduce some of these concepts. The first idea is simply to propose that all future values are equal to the mean of the historical data, hence proposing that

\[ \hat{y}_{T+h|T} = \frac{1}{T} \sum_{t=1}^{T} y_t. \]

We call this method the average method. It is clear that this method makes only sense if we do not find any trend behaviour.

The second method closely connected to the average method which was described in Hyndman and Athanasopoulos (2013) is to set each forecast to be equal to the last observed value. This would overcome the poor performance of the average method if a time series exhibits a trend but would miss any possible cyclical behaviour of a series.

To overcome this disadvantage someone can modify the naive method when setting the forecast to be equal to the last observed value from the same season. For example, if we want to forecast the value for august 2013 we set this value equal to the value of august 2012.

Again, these methods only serve as benchmarks here.

3.2 (S)ARIMA-Models

Autoregressive integrated moving average (ARIMA) models build up on a little more sophisticated but not much more complicated ideas compared to the ideas introduced in the previous section. First introduced in 1970 by Box and Jenkins, ARIMA-models appear to be perhaps the most prominent method in time series analysis for forecast-
ing. To start with, we introduce the concept of weak stationarity: We call a time series stationary if the mean \(\mu\) and the variance \(\sigma^2\) are constant over time and the autocovariance \(\text{cov}(y_t, y_s)\) only depends on the time lag \(k = |t - s|\) between two observations. Illustratively, the properties of a weakly stationary time series do not depend on the time of an observation but only on the time difference between two observations.

The following paragraphs build closely on Shumway and Stoffer (2006, ch. 3). The first component of an ARIMA-model is the autoregressive part which is simply an observation \(y_t\) in \(t\) as a weighted sum of past observations plus an error component \(\omega_t\), hence

\[
y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + \omega_t,
\]

where \(\omega_t \sim N(0, \sigma^2)\). Using \(p\) past observations leads to an AR\((p)\) model. For notational simplicity it is convenient to use a backshift operator \(B\) in a sense such that

\[
B^k y_t = y_{t-k}.
\]

Using a backshift operator we can write an AR\((p)\)-model as

\[
(1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p) y_t = \omega_t
\]

or simply as

\[
\phi(B) y_t = \omega_t,
\]

where \(\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p\) is called the autoregressive operator which determines whether an AR\((p)\) process is stationary.

The second component of an ARIMA-model is the moving average part. A moving average expresses the current observation \(y_t\) as the weighted sum of past noise terms, explicitly as

\[
y_t = \omega_t + \theta_1 \omega_{t-1} + \ldots + \theta_q \omega_{t-q}
\]

\[
\Leftrightarrow y_t = \theta(B) \omega_t,
\]

where \(\theta(B) = 1 + \theta_1 B + \ldots + \theta_q B^q\).

With this two components we define an ARMA\((p, q)\)-process close to Shumway and Stoffer (2006, Definition 3.5): A time series \(\{y_t\}_{t=1}^T\) is ARMA\((p, q)\) if it is stationary
and for $\phi_p, \theta_q \neq 0$ and $\sigma^2 > 0$ we have

$$y_t = \phi_1 y_{t-1} + \ldots + \phi_p y_{t-p} + \omega_t + \theta_1 \omega_{t-1} + \ldots + \theta_q \omega_{t-1}$$

$$\Leftrightarrow \phi(B) y_t = \theta(B) \omega_t.$$ 

The plots of the autocorrelation function (ACF) as well as the partial autocorrelation function (PACF) are often used to determine the optimal number of parameters $p$ and $q$ (for example using table 3.1 in Shumway and Stoffer, 2006). We do not follow this approach and use the Akaike Information Criterion (AIC, see Akaike 1973) instead. This measure builds directly on the likelihood function and takes also the number of model parameters into account.\(^1\)

Unfortunately, many time series do not have constant expectations - especially the time series we deal with here. For example, $\mu_t$ of the main residence time series depends clearly on $t$, thus exhibits a clear trend behaviour. A simple solution for this problem is taking the $d$th difference of a time series. An ARIMA($p, d, q$)-model is therefore simply an ARMA($p,q$)-model of the $d$-times differenced original time series. More formally, if we write $\nabla^d y_t = y_t \sum_{k=1}^{d} (\binom{d}{k}(-B)^k(1-B)^d y_t$ we can say that $y_t$ is an ARIMA($p,d,q$)-process if $\nabla^d y_t = (1-B)^d y_t$ is an ARMA($p,q$)-process. Hence, such a model can be written as

$$\phi(B)(1-B)^d y_t = \theta(B) \omega_t.$$ 

Unfortunately, simple ARIMA-models can not cover seasonal effects directly but many time series show some (short term) seasonal patterns. Fortunately, extending ARIMA-models to capture seasonal patterns is rather easy. If we consider the time series $\text{influx}$ in figure 1, we observe some sort of cyclical behaviour, which is also easy to explain: Many new jobs start in august or september, universities start in october. This structure has to be covered in an appropriate way. For this we can introduce AR- and MA-polynomials that identify with the seasonal lags, thus

$$\Phi_P(B^s) y_t = \Theta_Q(B^s) \omega_t.$$ 

\(^1\)Unfortunately the common approach to identify the number of parameters by the (P)ACF is in our opinion inferior to the AIC for our task for several reasons: Firstly, the pathway of the (P)ACF can be interpreted differently and is therefore ambiguous. Secondly, this approach does not account for the number of model parameters and if so, this choice is not based on statistical model selection criteria.
which yields an ARMA($P,Q$)-model with a seasonal period of $s$.\(^2\) The operators $\Phi_P(B^s)$ and $\Theta_Q(B^s)$ are defined similar to their general counterparts. A seasonal ARMA($p,q$,$(P,Q)_s$) - model can be written as

$$
\Phi_P(B^s)\phi(B)y_t = \Theta_Q(B^s)\theta(B)\omega_t
$$

As an example (see Shumway and Stoffer, 2011, example 3.42), a SARIMA($0,1,1$) $\times (0,1,1)_{12}$ can be expressed in the differenced form as

$$
y_t = y_{t-1} + y_{t-12} - x_{t-13} + \omega_t + \theta\omega_{t-1} + \Theta\omega_{t-12} + \Theta\theta\omega_{t-13}.
$$

We use this models especially with a season of 12 months. This means, for example with the same model as above, that we model the value for september 2012 as the linear combination of not only as the weighted value of the previous month, but also with the weighted value of september 2011 and august 2011.

To find a suitable SARIMA-model one can use the ACF and PACF according to table 3.2 in Shumway and Stoffer (2011). Again, we do not use this table but try to find the model with the smallest AIC.

### 3.3 Applying (S)ARIMA-Models

One important property of a time series is stationarity. This means that the time series must exhibit neither a trend nor changes in variance. Both issues can be tackled using (i) methods for detrending, e.g. filters, and for changes in variance the (ii) Box-Cox-transformation (Box and Cox, 1964). The Box-Cox-transformation is defined as

$$
\tilde{y}_t^\lambda = \begin{cases} 
\frac{y_t^{\lambda}-1}{\lambda}, & \text{if } \lambda \neq 0 \\
\ln(y_t), & \text{if } \lambda = 0
\end{cases}
$$

where $\lambda \in \mathbb{R}^+$, which is preferably estimated when estimating the ARIMA-coefficients. This transformation can be used to stabilize a time series’ variance and is used in this survey if necessary.

For detrending a time series there exist many approaches. The first is simply to take differences in all observations of a time series, which is of course directly related to ARIMA-models. When combining the Box-Cox-transformation for $\lambda = 0$, we can also represent a time series in terms of their approximated growth rates if we stabilize the

\(^2\)Note that we use capitals for the seasonal parts and small letter for the general parts.
variance of the differenced time series (which for example is done for the main residence time series), hence modeling \( \tilde{y}_t = \ln(y_t - y_{t-1}) \). Note that this approximation is only sufficiently close to the geometric mean if the (true) growth rates are less than 0.05. But if the growth rates\(^3\) or any other differenced Box-Cox transformation still exhibit a trend, we have to distinguish whether this trend is either stochastic or deterministic. To test for a stochastic trend, we use the augmented Dickey-Fuller-Test (ADF-test). We do not cover this topic in detail. This test tests whether the AR-part of a series has a unit root. We use this test in the sense that if we can dismiss this hypothesis, we assume that the series does not have a stochastic but a deterministic trend instead. For details, see Said and Dickey (1984).

When estimating with maximum likelihood (ML), which is necessary for estimating MA-terms, one has to specify a distribution at first, in this case a normal distribution. To identify whether the data is normally distributed, someone can use either quantile to quantile plots (qq-plot) or the Shapiro-Wilk test. When using a qq-plot, we compare the observed value against the values of a (corresponding) normal distribution. Intuitively, if these values are sufficiently close to each other, we assume that the data is normally distributed. The Shapiro-Wilk test tests the hypothesis if the data is normally distributed or if this hypothesis can not be supported.

\(^3\)If we speak of growth rates in the following, we mean their logarithmic approximation
4 Forecasts

In this section we present our forecasts for the time series we deal with here. We follow the same scheme for every series: At first, we identify the best SARIMA-models using data until december 2012. We forecast the first six months of 2013 and calculate the MSE of the forecasted and the true values. We compare this MSE to a benchmark-forecast, which is simply the same values of the previous season (or more complicated, a SARIMA(0,0,0)×(0,1,0)-model). All data transforms will be motivated where they are needed. In the last subsection we summarize all results and discuss them.

4.1 Births

We do not model the number of births directly because the more people live in a town the more childs are delivered (at least it appears to be more likely). For this, we model births per capita by dividing the number of births in a month by the current number of main residents in the same month. In addition to this, we multiply this series by 100 which gives us the number of births per 100 capita. The series and forecast is shown in figure 19, the corresponding (P)ACFs are shown in figure 24 in appendix 6.1.

At first, we only look at the series until december 2012. We assume this time series to be stationary, but the (P)ACFs show again cyclical patterns. Hence we model this series directly. A SARIMA(0,1,2)×(3,0,0)_{12} was identified as best model by the AIC, which fulfills the relevant model assumptions, illustrated by the diagnosis illustrated in \[ \text{Figure 19: Forecast, births per 100 capita.} \]
The residuals seem to be normally distributed, according to a Shapiro-Wilk test ($p = 0.2825$), the ACF does not show a significant cyclical pattern and all dependencies seem to be covered, as indicated by the Ljung-Box-test. But most importantly, the MSE for the model is smaller compared to the MSE for the benchmark-forecast, since $0.000119 < 0.000271$. Hence we choose the ARIMA-approach to forecast this series.

Our final model for this time series is a SARIMA(2,1,1)$\times$(2,0,0)$_{12}$ with estimates

$$ARIMA(2,1,1)(2,0,0)[12]$$

Coefficients:

<table>
<thead>
<tr>
<th></th>
<th>ar1</th>
<th>ar2</th>
<th>ma1</th>
<th>sar1</th>
<th>sar2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$-0.4771$</td>
<td>$-0.2592$</td>
<td>$-0.9664$</td>
<td>$0.3511$</td>
<td>$0.2851$</td>
</tr>
</tbody>
</table>

s.e. 0.0774 0.0799 0.0174 0.0793 0.0854

sigma^2 estimated as 6.701e-05: log likelihood=557.87

AIC=-1103.75  AICc=-1103.22  BIC=-1085.08

and point-forecasts as well as confidence intervals (95%) in table 5. Again, all model assumptions are fulfilled: The data as well as the residuals seem to be normally distributed ($p = 0.3593$ and $p = 0.1652$, respectively), the ACF does not show any cyclical behaviour and all dependencies seem to be modeled, according to the Ljung-Box-plot in figure 26.

<table>
<thead>
<tr>
<th></th>
<th>Point Forecast</th>
<th>Low 95</th>
<th>High 95</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan 2013</td>
<td>0.098317</td>
<td>0.082426</td>
<td>0.114207</td>
</tr>
<tr>
<td>Feb 2013</td>
<td>0.086392</td>
<td>0.069261</td>
<td>0.103523</td>
</tr>
<tr>
<td>Mar 2013</td>
<td>0.094734</td>
<td>0.077601</td>
<td>0.111868</td>
</tr>
<tr>
<td>Apr 2013</td>
<td>0.081787</td>
<td>0.064652</td>
<td>0.098923</td>
</tr>
<tr>
<td>May 2013</td>
<td>0.086461</td>
<td>0.069323</td>
<td>0.103599</td>
</tr>
<tr>
<td>Jun 2013</td>
<td>0.088070</td>
<td>0.070930</td>
<td>0.105211</td>
</tr>
<tr>
<td>Jul 2013</td>
<td>0.092859</td>
<td>0.075716</td>
<td>0.110002</td>
</tr>
<tr>
<td>Aug 2013</td>
<td>0.095016</td>
<td>0.077871</td>
<td>0.112161</td>
</tr>
<tr>
<td>Sep 2013</td>
<td>0.089794</td>
<td>0.072647</td>
<td>0.106941</td>
</tr>
<tr>
<td>Oct 2013</td>
<td>0.096721</td>
<td>0.079571</td>
<td>0.113870</td>
</tr>
<tr>
<td>Nov 2013</td>
<td>0.086947</td>
<td>0.069795</td>
<td>0.104099</td>
</tr>
<tr>
<td>Dec 2013</td>
<td>0.077867</td>
<td>0.060712</td>
<td>0.095021</td>
</tr>
</tbody>
</table>

Table 5: Births per 100 capita: Forecast and CIs.


4.2 Deaths

As shown in figure 20, it seems that the number of deaths per month seems to decrease, although more people are living in Munich. This trend is not stochastic but deterministic, as indicated by an ADF-test ($p < 0.01$). We estimated this trend by simple linear regression where its negative slope coefficient is highly significant ($p < 0.01$). Furthermore, the logarithm of this series appears to be normally distributed (Shapiro-Wilk test, $p = 0.2386$), therefore not only the t-test is valid but also the distributional assumption for ML-estimation of possible MA-terms are fulfilled. In the following we speak of this series as the logarithm of deaths per 100 capita per month. A plot of this series with the ARIMA-forecasts is shown in figure 20, the ACF and PACF in figure 27 in appendix 6.2 as well as the diagnostic plots.

The first step is to use only the data until December 2012 to identify a suitable SARIMA-model with the smallest AIC and to forecast the six following months. If the MSE of this forecast is better compared to the MSE of the benchmark-forecast, then we choose the SARIMA-model. We identified a SARIMA$(2,0,2) \times (1,0,0)_{12}$ - model as suitable. The MSE of its forecast is also smaller compared to the MSE of the benchmark-forecast ($0.000208 < 0.000331$). The diagnostic plot is given in figure 28 in appendix 6.2.

Now using the six data points from 2013, we identify a SARIMA$(2,0,2) \times (1,0,0)_{12}$
on basis of the AIC (with apparently normally distributed residuals, as indicated by a Shapiro-Wilk test with \( p = 0.5409 \)) with estimates

ARIMA\((3,0,0)(2,0,0)[12]\) with non-zero mean

Coefficients:

<table>
<thead>
<tr>
<th></th>
<th>ar1</th>
<th>ar2</th>
<th>ar3</th>
<th>sar1</th>
<th>sar2</th>
<th>intercept</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-0.344</td>
<td>0.0328</td>
<td>0.2177</td>
<td>0.2697</td>
<td>0.1794</td>
<td>0.0039</td>
</tr>
<tr>
<td>s.e.</td>
<td>0.077</td>
<td>0.0807</td>
<td>0.0771</td>
<td>0.0810</td>
<td>0.0842</td>
<td>0.0144</td>
</tr>
</tbody>
</table>

\(\sigma^2\) estimated as 0.01463: log likelihood=114.58

AIC=-215.16 AICc=-214.46 BIC=-193.34

and point-forecasts as well as confidence intervals (95\%) in table 6. These values are transformed back to the series of deaths per 100 capita. The diagnostic plot is given in figure 29 in appendix 6.2.

<table>
<thead>
<tr>
<th></th>
<th>Point Forecast</th>
<th>Low 95</th>
<th>High 95</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jul 2013</td>
<td>0.083249</td>
<td>0.064499</td>
<td>0.101767</td>
</tr>
<tr>
<td>Aug 2013</td>
<td>0.076122</td>
<td>0.058178</td>
<td>0.094332</td>
</tr>
<tr>
<td>Sep 2013</td>
<td>0.074466</td>
<td>0.056768</td>
<td>0.092516</td>
</tr>
<tr>
<td>Oct 2013</td>
<td>0.078269</td>
<td>0.059511</td>
<td>0.097497</td>
</tr>
<tr>
<td>Nov 2013</td>
<td>0.080464</td>
<td>0.061078</td>
<td>0.100397</td>
</tr>
<tr>
<td>Dec 2013</td>
<td>0.074339</td>
<td>0.056389</td>
<td>0.092820</td>
</tr>
<tr>
<td>Jan 2014</td>
<td>0.087476</td>
<td>0.066354</td>
<td>0.109224</td>
</tr>
<tr>
<td>Feb 2014</td>
<td>0.078263</td>
<td>0.059362</td>
<td>0.097727</td>
</tr>
<tr>
<td>Mar 2014</td>
<td>0.080957</td>
<td>0.061400</td>
<td>0.101098</td>
</tr>
<tr>
<td>Apr 2014</td>
<td>0.079733</td>
<td>0.060472</td>
<td>0.099571</td>
</tr>
<tr>
<td>May 2014</td>
<td>0.077181</td>
<td>0.058536</td>
<td>0.096384</td>
</tr>
<tr>
<td>Jun 2014</td>
<td>0.073714</td>
<td>0.055907</td>
<td>0.092055</td>
</tr>
</tbody>
</table>

Table 6: Deaths per 100 capita: Forecast and CIs.
4.3 Immigrants

In contrast to the previous subsection, we do not observe a negative but a (determinis-
tic) positive trend in the number of immigrations (again indicated by an ADF-test and
$p < 0.01$). This trend is highly significant as indicated by the significant slope coeffi-
cient. As shown in figure 21 and the ACF and PACF in figure 30 in appendix 6.3, we
observe also a strong seasonal pattern in this series. Because of the seemingly changed
pattern in january 2006, we only use data beginning in january 2006 to carry out our
estimations. One has to keep in mind that the data for this series is not normally
distributed (Shapiro-Wilk test, $p < 0.01$).

In the first step, where we only utilize the data until december 2012, a SARIMA$(3, 0, 0) \times (1, 1, 0)_12$
was identified as the best ARIMA-model. Unfortunately, the MSE of the this models
forecast is higher than the forecast of our benchmark forecast ($9088520 > 859093$). The
diagnostic plot is given in 31 in appendix 6.3.

Now using the six data points from 2013, we identify a SARIMA$(3, 0, 0) \times (0, 1, 1)_12$
on basis of the AIC with estimates

ARIMA$(3, 0, 0)(0,1,1)[12]$

Coefficients:
\begin{verbatim}
 ar1  ar2  ar3  sma1
 -0.1483  0.0926  0.5924  -0.3429
 s.e.  0.0912  0.0922  0.0959  0.1370

sigma^2 estimated as 276805:  log likelihood=-600.72
AIC=1211.44  AICc=1212.28  BIC=1223.23

and point-forecasts as well as confidence intervals (95%) in table 7. Again, the detrended values were added. The diagnostic plot of this model is given in 32 in appendix 6.3. One has to keep in mind that our model until december 2012 did not beat the benchmark forecast.

\begin{table}[h]
\centering
\begin{tabular}{lccc}
\hline
      & Point Forecast & Low 95 & High 95 \\
\hline
Jul 2013 & 9391.736585 & 8694.587849 & 11090.986222 \\
Aug 2013 & 8296.993109 & 7588.568517 & 10007.518603 \\
Sep 2013 & 10359.993490 & 9644.898253 & 12077.189628 \\
Oct 2013 & 11851.155252 & 10986.805621 & 13717.605785 \\
Nov 2013 & 6619.335963 & 5743.607588 & 8497.165239 \\
Dec 2013 & 5111.424542 & 4226.661416 & 6998.288569 \\
Jan 2014  & 8084.759517 & 7162.216982 & 10009.402953 \\
Feb 2014  & 6076.667386 & 5147.449336 & 8007.986337 \\
Mar 2014  & 7251.113324 & 6314.671975 & 9189.655574 \\
Apr 2014  & 8671.485245 & 7726.326152 & 10618.745240 \\
May 2014  & 6530.281820 & 5582.207149 & 8480.457392 \\
Jun 2014  & 6668.628492 & 5716.117146 & 8623.240741 \\
\hline
\end{tabular}
\caption{Immigrants per month: Forecast and CIs.}
\end{table}
\end{verbatim}
Figure 22: Forecast, number of migrants per month.

4.4 Departures

As we can see in figure 7.4, the number of migrants drops at the end of the 80’s. Because these values would shift the mean downwards, we analyze this series from January 1990 onwards. All outliers above 15000 are replaced by their mean values for the month in which the outliers occur. We also assume that this time series does not exhibit neither a deterministic nor a stochastic trend (as supported by an ADF-test with $p < 0.01$ at least for the stochastic case). As indicated by the ACF- and PACF-plots in figure 33 in appendix 6.4, we again observe a yearly seasonal pattern. Unfortunately, this series is not normally distributed (as indicated by a Shapiro-Wilk test with $p < 0.01$).

In the first step, where we only utilize the data until December 2012, a SARIMA$(0,1,1) \times (4,0,0)_{12}$ was identified as the best ARIMA-model. Here, the MSE of the this model’s forecast is smaller than the forecast of our benchmark forecast (1083720 < 6284204). The diagnostic plot is given in 34 in appendix 6.4. The residuals are not normally distributed, but it seems that the model covers all absolute as well as all (identifiable) seasonal dependencies.

Now using the six data points from 2013, we identify the same model as above on basis of the AIC with estimates

$$\text{ARIMA}(0,1,1) (4,0,0) [12]$$
Coefficients:

<table>
<thead>
<tr>
<th></th>
<th>ma1</th>
<th>sar1</th>
<th>sar2</th>
<th>sar3</th>
<th>sar4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-0.7943</td>
<td>0.1822</td>
<td>0.1448</td>
<td>0.0862</td>
<td>0.2407</td>
</tr>
</tbody>
</table>

\[
s.e.\quad 0.0415 \quad 0.0586 \quad 0.0612 \quad 0.0659 \quad 0.0649
\]

\[\sigma^2 \text{ estimated as 1637966: log likelihood=-2412.66}\]

AIC=4837.32 AICc=4837.62 BIC=4859.15

and point-forecasts as well as confidence intervals (95%) in table 8. The diagnostic plot of this model is given in 35 in appendix 6.4. Again, the residuals are not normally distributed, but - as indicated by the nonsignificant ACF - seasonal dependencies are covered.

<table>
<thead>
<tr>
<th>Point.Forecast</th>
<th>Lo.95</th>
<th>Hi.95</th>
<th>True</th>
<th>Benchmark</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan 2013</td>
<td>7579</td>
<td>5064</td>
<td>10093</td>
<td>9526</td>
</tr>
<tr>
<td>Feb 2013</td>
<td>8490</td>
<td>5922</td>
<td>11058</td>
<td>8102</td>
</tr>
<tr>
<td>Mar 2013</td>
<td>8156</td>
<td>5536</td>
<td>10777</td>
<td>8208</td>
</tr>
<tr>
<td>Apr 2013</td>
<td>7965</td>
<td>5294</td>
<td>10636</td>
<td>8706</td>
</tr>
<tr>
<td>May 2013</td>
<td>6875</td>
<td>4153</td>
<td>9596</td>
<td>7883</td>
</tr>
<tr>
<td>Jun 2013</td>
<td>8303</td>
<td>5532</td>
<td>11074</td>
<td>7307</td>
</tr>
<tr>
<td>Jul 2013</td>
<td>8282</td>
<td>5462</td>
<td>11101</td>
<td>0</td>
</tr>
<tr>
<td>Aug 2013</td>
<td>9209</td>
<td>6342</td>
<td>12076</td>
<td>0</td>
</tr>
<tr>
<td>Sep 2013</td>
<td>9225</td>
<td>6311</td>
<td>12139</td>
<td>0</td>
</tr>
<tr>
<td>Oct 2013</td>
<td>8513</td>
<td>5553</td>
<td>11473</td>
<td>0</td>
</tr>
<tr>
<td>Nov 2013</td>
<td>8488</td>
<td>5483</td>
<td>11493</td>
<td>0</td>
</tr>
<tr>
<td>Dec 2013</td>
<td>7573</td>
<td>4523</td>
<td>10623</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 8: Migrants: Forecast and CIs.

### 4.5 Main residences growth rate

We model the main residence series in terms of growth rates, which is simply the first differences of the logarithmized values. We replaced all outliers (growth rates larger than \(|0.005|\)) by the mean of the monthly values in which the outlier occurred. The plot (again with the final forecast) is shown in figure 23.

As we already mentioned many times, we find a strong cyclical pattern, as shown by the ACF and PACF in figure 36. The identified model, using the data only until december 2012, is a SARIMA(0,1,1)×(3,0,0)_{12}. Although the data is not normally distributed, as indicated by a Shapiro-Wilk-Test \((p < 0.01)\), all dependencies seem to
be covered, as shown in figure 37. Furthermore, the MSE of the ARIMA-forecast is smaller than the MSE of the benchmark-forecast ($7.078 \cdot 10^{-7} < 2.173 \cdot 10^{-3}$).

Using the data until June 2013, we also identified a SARIMA($0, 1, 1) \times (3, 0, 0)_12$ as suitable model. The diagnostic plot is shown in figure 38. All dependencies seem to be covered. The forecasts and the corresponding 95% confidence intervals are given in table 9.
### Table 9: Main residences growth rate: Forecast and CIs.

<table>
<thead>
<tr>
<th></th>
<th>Point Forecast</th>
<th>Low 95</th>
<th>High 95</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jul 2013</td>
<td>0.00178</td>
<td>-0.00023</td>
<td>0.00379</td>
</tr>
<tr>
<td>Aug 2013</td>
<td>0.00133</td>
<td>-0.00071</td>
<td>0.00338</td>
</tr>
<tr>
<td>Sep 2013</td>
<td>0.00212</td>
<td>0.00003</td>
<td>0.00420</td>
</tr>
<tr>
<td>Oct 2013</td>
<td>0.00306</td>
<td>0.00094</td>
<td>0.00518</td>
</tr>
<tr>
<td>Nov 2013</td>
<td>0.00146</td>
<td>-0.00070</td>
<td>0.00362</td>
</tr>
<tr>
<td>Dec 2013</td>
<td>0.00088</td>
<td>-0.00131</td>
<td>0.00308</td>
</tr>
<tr>
<td>Jan 2014</td>
<td>0.00142</td>
<td>-0.00081</td>
<td>0.00364</td>
</tr>
<tr>
<td>Feb 2014</td>
<td>0.00126</td>
<td>-0.00100</td>
<td>0.00352</td>
</tr>
<tr>
<td>Mar 2014</td>
<td>0.00139</td>
<td>-0.00091</td>
<td>0.00368</td>
</tr>
<tr>
<td>Apr 2014</td>
<td>0.00182</td>
<td>-0.00051</td>
<td>0.00415</td>
</tr>
<tr>
<td>May 2014</td>
<td>0.00195</td>
<td>-0.00041</td>
<td>0.00432</td>
</tr>
<tr>
<td>Jun 2014</td>
<td>0.00113</td>
<td>-0.00126</td>
<td>0.00353</td>
</tr>
</tbody>
</table>

### 5 Summary

In this thesis we took a look on five time series regarding population data, namely the births, deaths, influxes, departures and main residence population in Munich from 1987 until 2013. Reporting effects resulting from the Christmas holidays shift the date of data entries from December to January. We show that the strength of this reporting effect depends on how the 24th of December falls each year and adjust for this effect with regards to the weekday of the 24th of December.

Furthermore, we find other reporting effects resulting from weekends and other holidays at the end of a month and adjust for them. The adjustment for reporting effects reduced the number of outliers and the variance of the time series. In addition, the adjustment improved the ACF and PACF of the time series of births and deaths by a reduction of the lag 12. Furthermore, we also modeled the data using a seasonal-ARIMA-approach which at least performed better than a chosen benchmark-model.

It is sometimes crucial to improve the quality of the data and the resulting analysis by simply being aware of the problems in collecting the data. This perhaps has a larger impact than the correct implementation of a model and seems to be underestimated in statistical courses or even practical data analysis.

Nevertheless, as our seasonal-ARIMA-analysis shows, a sophisticated choice of a linear model may cover many structures in the data. But also this method can not straighten out bad data quality.
References


6 Appendix

6.1 Birth plots

Figure 24: Births per capita, (partial) autocovariance function.

Figure 25: Births per capita -12/2012, diagnostic plot.
Figure 26: Births per capita -6/2013, diagnostic plot.
6.2 Death plots

Figure 27: log deaths per capita, (partial) autocovariance function.

Figure 28: log deaths per capita -12/2012, diagnostic plot.
Figure 29: log deaths per capita -6/2013, diagnostic plot.
6.3 Immigrant plots

Figure 30: Detrended immigrants, autocovariance function.

Figure 31: Detrended immigrants -12/2012, diagnostic plot.
Figure 32: Detrended immigrants -6/2013, diagnostic plot.
6.4 Migrates plots

Figure 33: Detrended immigrants, autocovariance function.

Figure 34: Detrended immigrants -12/2012, diagnostic plot.
Figure 35: Detrended immigrants -6/2013, diagnostic plot.
6.5 Main residence growth rate plots

Figure 36: Main residences growth rate, autocovariance function.

Figure 37: Main residences growth rate -12/2012, diagnostic plot.
Figure 38: Main residences growth rate -6/2013, diagnostic plot.
6.6 Cleaned data

![Normal Q–Q Plot (births)](image1)

![Normal Q–Q Plot (deaths)](image2)

Figure 39: QQ-plot, cleaned births- and deaths-series.

![Standardized Residuals](image3)

![ACF of Residuals](image4)

![p values for Ljung–Box statistic](image5)

Figure 40: Standardized Residuals ACF of Residuals and Ljung-Box statistic of an ARMA(6,7) model time series of births.
Figure 41: Standardized Residuals ACF of Residuals and Ljung-Box statistic of an ARMA(6,3) model time series of deaths.