Uniwersytet Wrocławski
Wydział Matematyki i Informatyki
Instytut Matematyczny
specjalność: Ogólna

Nela Tomaszewicz

Statistical Analysis of Climate Data

Praca licencjacka
napisana pod kierunkiem
doktor Liudmily Zaitsevy

Wrocław, wrzesień 2020

Contents

1 Introduction

A famous American essayist Charles Dudley Warner once said

> Everybody complains about the weather, but nobody does anything about it.

He referred to the climate of the New England region in the USA. The quote is easily applicable to today's world. We all complain about the weather, regardless if it's hot Summer, freezing Winter or rainy Autumn. The temperature plays a huge part in our daily conversations with peers and strangers. It's the most commonly noticed physical quantity impacting our lives every day. Everybody understands that in January in Poland no one would expect a 20-degree Spring weather. What is not commonly known is that mathematics, as always, provides us with models explaining the reality and even predicting the future.

The goal of the thesis is to investigate real climate data regarding temperature measures.

In the first chapter of the thesis, we will take a closer look at daily temperature data measured in the last 35 years in six cities around the world. We will analyse its properties and introduce the concept of time series approach for data analysis. We will explain why the data does not meet the assumptions of stationarity and why it needs to. We will examine the relationship between observations using the autocorrelation and partial autocorrelation functions. In the next chapter we start the preparation of the data for future modelling. The aim is to investigate a long term trend and seasonality. In chapter 4 we apply a proper model for the data, which is the autoregressive model with the order p. We check its correctness in the next and last chapter, by performing 90-day predictions. What we find out is, that this simple model gives generally correct predictions, however, it does not perform well in the presence of weather anomalies.

2 Exploring the data

The first step of any data examination is getting to know the kind of information we are about to work on. In this section we will explore the data properties and we will choose the right mathematical approach for the next step of the analysis.

2.1 About NASA Power

NASA **P**rediction **O**f **W**orldwide **E**nergy **R**esources, in short NASA POWER [6] is a project containing over 200 satellite-derived meteorology and solar energy Analysis Ready Data (ARD), at three temporal levels: daily, inter-annual, climatology. The POWER Data is updated all the time. There are three "communities" that the POWER Project targets: Surface meteorology and Solar Energy (SSE), Sustainable Buildings (SB), Agroclimatology (AG). In this paper we will consider the data of daily temperature measures in six cities around the world using the resources collected for the agroclimatology community.

2.2 The choice of cities

We will analyse data from six different cities around the world:

1. Wroclaw (Poland),
2. Reykjavik (Iceland),
3. Melbourne (Australia),
4. Rio de Janeiro (Brasil),
5. New York City (United States of America),
6. New Delhi (India)

Each of these cities lays in different part of the world and in a different climate zone. Wroclaw and New York City both lay in the temperate zone, however they differ drastically in terms of population number, traffic and access to the ocean. Reykjavik belongs to the subpolar zone and even though it's the smallest one in terms of population, there's the common belief that Iceland, and especially its glaciers, has a very serious melting issue because of the temperature getting higher. Rio de Janeiro's climate is considered a "tropical savanna" meaning its winters are warm and summers are hot and humid. It is also an enormous city with huge traffic. It lays in the subtropical climate zone as well as Melbourne. The climate of New Delhi is characterised by strong monsoon in the June and July as well as a significant difference between summer and winter months. It is also influenced by the nearby Himalayas causing cold winds especially in the winter.

2.3 Time series approach

We will inspect daily temperature measured from 1st of January 1985 to 31st of December 2019, so the last 35 years. The temperature is computed as a daily average. Because of the way it's measured we will use the *time series* approach to model and examine the results.

Definition 1. *A time series is a set of observations $x_1, x_2 \ldots, x_n$ taken at successive periods of time.*

We will consider a *discrete-time series*, where the time set T_0 of times at which observations take place is discrete.

A crucial part in time series analysis is to define a model to describe the data. Because we cannot really predict the future observations, we need to assume a certain probability attached to them. This leads us to suppose that each observation x_t is a realized value of a random variable X_t. Then, the time series $\{x_t, t \in T_0\}$ is a realization of of the family of random variables $\{X_t, t \in T_0\}$. So, we can model the data as a realization of a stochastic process $\{X_t, t \in T\}$, where $T_0 \subseteq T$.

Definition 2 (Stochastic process)**.** *A stochastic process is a family of random variables $\{X_t, t \in T\}$ defined on a probability space $(\Omega, \mathscr{F}, \mathbb{P})$.*

In our case the set T is a set of time points being every day from 01.01.1985 until 31.12.2019.

Definition 3 (Realizations of a Stochastic Process)**.** *The functions $\{X(\omega), \omega \in \Omega\}$ on T are called the realizations of the process $\{X_t, t \in T\}$.*

We will use the term time series to describe both the data and the stochastic process of which it is a realization.

Definition 4 (Time series model)**.** *A time series model of the observed data $\{x_t\}$ is a specification of the joint distributions (or only the means and covariances) of a stochastic process $\{X_t\}$ of which $\{x_t\}$ is a realization.*

In the case of daily temperature data we will consider the *classical decomposition model*:

$$X_t = m_t + s_t + Y_t, \quad t \in T, \tag{1}$$

where $\{m_t\}$ is the slowly changing function called a *trend component*, $\{s_t\}$ is the function with a perdiod d known as a *seasonal component* and $\{Y_t\}$ is a *random noise component.*

2.4 Time plots

Let's take a look at the daily temperature measures. For that we will use *time plots* which show the observations in the following time moments.

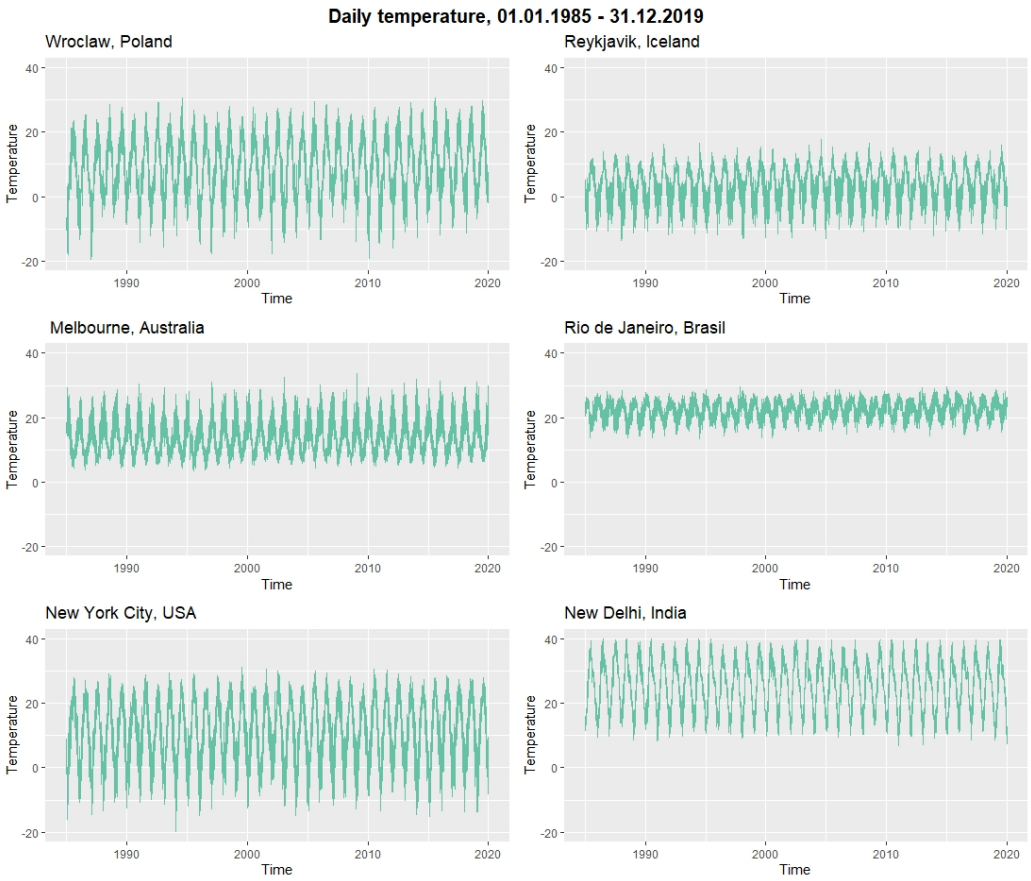

Figure 1: Time plots of the daily temperature measures.

The plots are put on the same Y axis to compare the amplitude between the measures and also to see the differences between temperatures in different cities. The daily temperature in New Delhi is far different than the daily temperature in Wroclaw. We can also observe a noticeable seasonality in the data by the way the plots are shaped. For now, by just looking at these simple plots, we can't really notice a trend. Because the seasonality is visible, we can make an assumption that the data is strongly autocorrelated. We will explore this property later on.

2.5 Seasonal plots

In the previous subsection we have observed that the data has a visible seasonal factor. Indeed, the seasons are changing and so does the weather and temperature. We can explore it more with the use of *seasonal plots*. Each line of the plot represents a year. Daily temperature was aggregated into months and the lines were smoothed to observe clearer seasonal tendency.

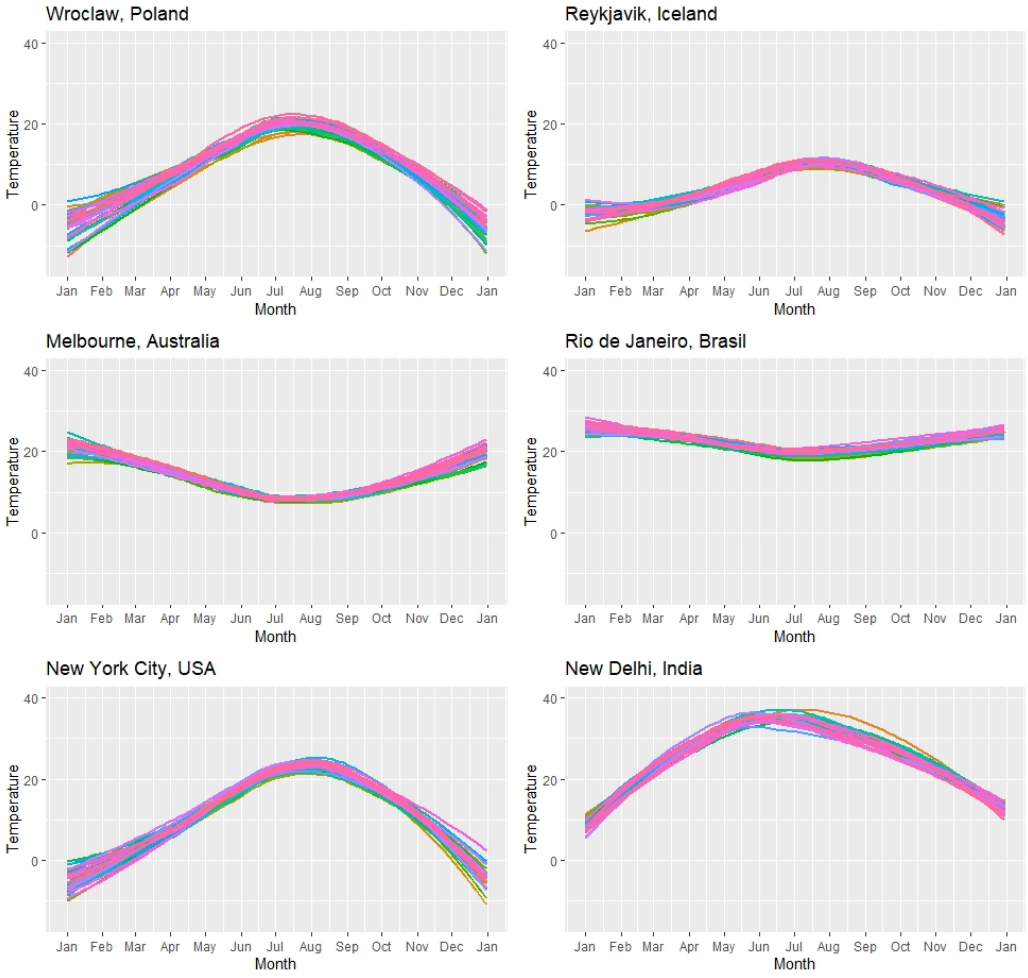

Figure 2: Seasonal plots of aggregated daily temperature.

With help of the seasonal plot, we can notice the hottest and the coldest months in each city. Also, there is a significant difference between Melbourne and Rio de Janeiro, being on the other half of the globe than the rest of the cities. The charts of these two areas are exactly opposite. However, from

this descriptive analysis we can't really tell if there is a increasing tendency of temperature in recent years comparing to the past. There are some values that do not fit to the overall line chart especially in the winter and summer months. What we can notice is that in some of the cities the temperature amplitude is higher than in other. For example, the New Yorkers can experience winters with temperatures quite below 0°C , but the citizens of Rio de Janeiro would have to deal with mean temperatures between 20°C and 30°C.

2.6 Stationarity and the Autocorrelation Function

One of the most important issues we need to tackle while analysing time series data is asking a question on what impact do the previous observations have on the current value and how much they are dependent on each other. We examine if there exists a time correlation between observed values, called the *autocorrelation* and also, regarding the dependence, we compute the autocovariance function (sometimes just called covariance function).

Definition 5 (The Autocovariance Function)**.** *If* $\{X_t, t \in \mathbb{Z}\}$ *is a process such that* $Var(X_t) < \infty$ *for each* $t \in \mathbb{Z}$*, then the autocovariance function* γ_X *of* $\{X_t\}$ *is defined by*

$$\gamma_X(r,s) = Cov(X_r, X_s) = \mathbb{E}[(X_r - \mathbb{E}X_r)(X_s - \mathbb{E}X_s)],\ r,s \in \mathbb{Z} \qquad (2)$$

The autocovariance function definition leads us to the definition of an important time series feature called *stationarity.*

Definition 6 (Stationarity)**.** *The time series* $\{X_t, t \in \mathbb{Z}\}$ *is stationary if:*

(i) $\mathbb{E}X_t^2 < \infty$ *for all* $t \in \mathbb{Z}$

(ii) $\mathbb{E}(X_t) = m$ *for all* $t \in \mathbb{Z}$

(iii) $\gamma_X(r,s) = \gamma_X(r+t, s+t)$ *for all* $r,s,t \in \mathbb{Z}$

This stationarity we consider in the above definition is often called *weak* sationarity or *wide–sense* stationarity. It will be extremely important in the next steps of the analysis, especially fitting a model for the data.

Remark. *If* $\{X_t, t \in \mathbb{Z}\}$ *is stationary then* $\gamma_X(r,s) = \gamma_X(r-s, 0)$ *for all* $r,s \in \mathbb{Z}$*. So, it we can redefine the autocovariance function of the stationary process as the function of one variable:*

$$\gamma_X(h) \equiv \gamma_X(h,0) = Cov(X_{t+h}, X_t),\ t,h \in \mathbb{Z}. \qquad (3)$$

Now, the function $\gamma_X(\cdot)$ *will be reffered to as the autocovariance function and* $\gamma_X(h)$ *is its value at lag* h*.*

To examine the autocorrelation we will be using the *Autocorrelation Function (ACF)*. It is the measure of the linear correlation between observations of the time series separated by h time stamps, so at lag h.

Definition 7 (The Autocorrelation Function)**.** *The autocorrelation function of* $\{X_t, t \in \mathbb{Z}\}$ *at lag* h *is:*

$$ACF(h) = \frac{\gamma(h)}{\gamma(0)} = Cor(X_{t+h}, X_t). \tag{4}$$

The autocorrelation function definition that was presented is more of a theoretical approach. In practice and in the case of temperature data, we use the *sample autocorrelation function*. If the data are realized values of a stationary time series $\{X_t\}$, then the sample ACF will give us an estimate of the ACF of $\{X_t\}$. Like before, to define the sample autocorrelation function, we need to define the sample autocovariance function.

Definition 8 (The Sample Autocovariance Function)**.** *Let* $x_1, x_2, \ldots, x_n$ *be observations of a time series. We define the sample mean of* $x_1, x_2, \ldots, x_n$ *as*

$$\bar{x} = \frac{1}{n} \sum_{t=1}^{n} x_t. \tag{5}$$

The sample autocovariance function is:

$$\hat{\gamma}(h) = \frac{1}{n} \sum_{t=1}^{n-|h|} (x_{t+|h|} - \bar{x})(x_t - \bar{x}), \ -n < h < n. \tag{6}$$

Definition 9 (The Sample Autocorrelation Function)**.** *The sample autocorrelation function is defined as:*

$$\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}, \ -n < h < n. \tag{7}$$

Let's take a look at the charts of the sample autocorrelation function. Here we care about the relationship between observations, so we will examine only charts related to daily temperature measures in Wroclaw, because the other cities are very similar. The data is measured the same way in every region, so we don't have to consider all of the places to see if the autocorrelation exists.

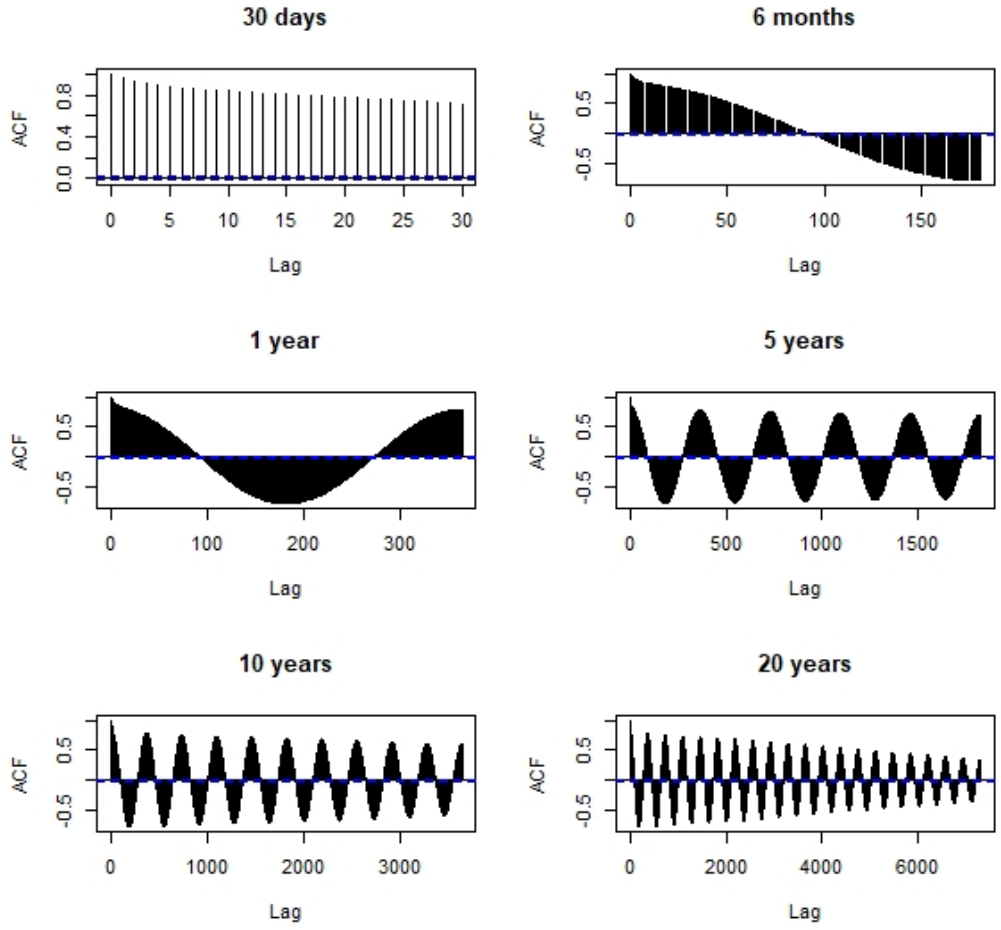

Figure 3: Autocorrelation charts of the daily temperature in Wroclaw, Poland.

We analyse data autocorrelation in 30 days, 6 months, 1 year, 5 years, 10 and 20 years lags. Very strong cyclical oscillations that are slowly decreasing, indicate that there is a significant seasonal component in the data. The oscillation period is about 4 months, which is actually true to the seasons changing and the seasonal plots analysed in the section 2.5. It also indicates a strong dependence between observations.

Another very important function describing the correlation between observations in a time series is the *partial autocorrelation function*. The partial autocorrelation $\alpha(h)$ at lag h is the correlation between X_1 and X_{h+1} adjusted for the intervening observations $X_2, \ldots, X_h$.

Definition 10 (The Partial Autocorrelation Function)**.** *The partial auto-*

correlation function $(PACF)$ $\alpha(\cdot)$ *of a stationary time series is defined by*

$$\alpha(1) = Corr(X_2, X_1) = \rho(1) \tag{8}$$

and

$$\alpha(h) = Corr(X_{h+1} - P_{\bar{sp}\{1,X_2,\ldots,X_h\}}X_{h+1}, X_1 - P_{\bar{sp}\{1,X_2,\ldots,X_h\}}X_1),\ h \geq 2, \tag{9}$$

where $P_{\{1,X_2,\ldots,X_h\}}$ *is an orthogonal projection on a linear subspace given by* $1, X_2, \ldots, X_h$.

Let's see the results after using the PACF function with different maximum lags h.

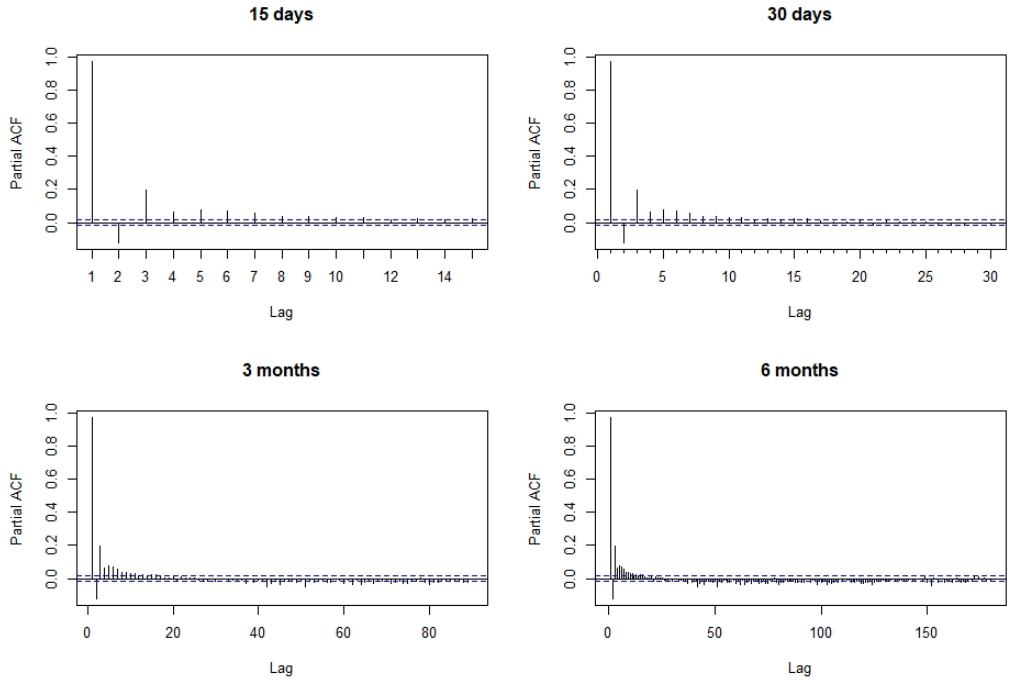

Figure 4: Partial autocorrelation charts of the daily temperature in Wroclaw, Poland.

The smaller the maximum lag h, the stronger is the partial autocorrelation. After about 15 days, the spikes of the plot are becoming statistically insignificant by appearing between blue dotted lines indicating a 95% confidence intervals. For lag h larger than 15 we can observe that the spikes appear randomly outside the confidence intervals.

3 Time series decomposition

3.1 Data adjustment

It is worth mentioning that in order to conduct time series decomposition in terms of seasonality removal, we need to make every year 365 days long. We will make this happen by removing 8 observations made on 29^{th} of February of eight different years. Because of this transformation, we can successfully perform the decomposition of the time series.

3.2 Theory

To perform a reasonable time series analysis, we need to get to know the data better. The technique that will help us is the time series decomposition. Time series decomposition relies on extracting the trend and seasonal component from the time series. Because of this we can see if there is a long-term increasing tendency in the temperature data and we can also take a better look at the seasonal factor. Time series decomposition will allow us to extract random noise component and later on choose and fit the best time series model for the most accurate predictions. The aim here is to make non stationary time series stationary, because that's the assumption that model, we're about to apply, has. Decomposition's goal is also to reduce the dependence between observations.

We're dealing with an additive time series, so the decomposition will be additive as well. This means simply subtracting trend and/or seasonal component from the data.

3.3 Trend component

We shall start with identifying the trend in our data. For that we will use the *Simple Moving Average (MA)* algorithm. The idea of this method is to "smoothen" the time series by taking mean of the values next to each other. It's a non–parametric method for trend estimation.

Let's consider a nonnegative integer q and a two-sided moving average

$$W_t = \frac{1}{2q+1} \sum_{j=-q}^{q} X_{t-j} \tag{10}$$

of the process $X_t, t \in \mathbb{Z}$ defined by the classic decomposition model (1). Then

for $q+1 \leq t \leq n-q$,

$$W_t = \frac{1}{2q+1}\left(\sum_{j=-q}^{q} m_{t-j} + \sum_{j=-q}^{q} Y_{t-j}\right) \approx m_t. \tag{11}$$

We assume that m_t is approximately linear over the interval $[t-q, t+q]$ and that the average of the error terms over this interval is close to zero. The moving average provides us with a trend estimator:

$$\hat{m}_t = \frac{1}{2q+1}\sum_{j=-q}^{q} X_{t-j},\ q+1 \leq t \leq n-q. \tag{12}$$

To determine the estimate value of the trend in the moment t we take a mean of q previous and q following observations of the one appearing in the moment t. In result there are $2q+1$ values of which we take the mean. The choice of q is in charge of the control of the degree of smoothing.

Remark. *$\{X_t\}$ is not observed for $t < 1$ and $t > n$, therefore in `R` we replace those values with `NA` meaning 'Not Available" which is a code used to mark that there is no value. Another approach is to replace X_t with X_1 for $t < 1$ and X_t with X_n for $t > n$.*

Let's take a look at the results of a simple moving average filter on the daily temperature data. We'll start with Wrocław.

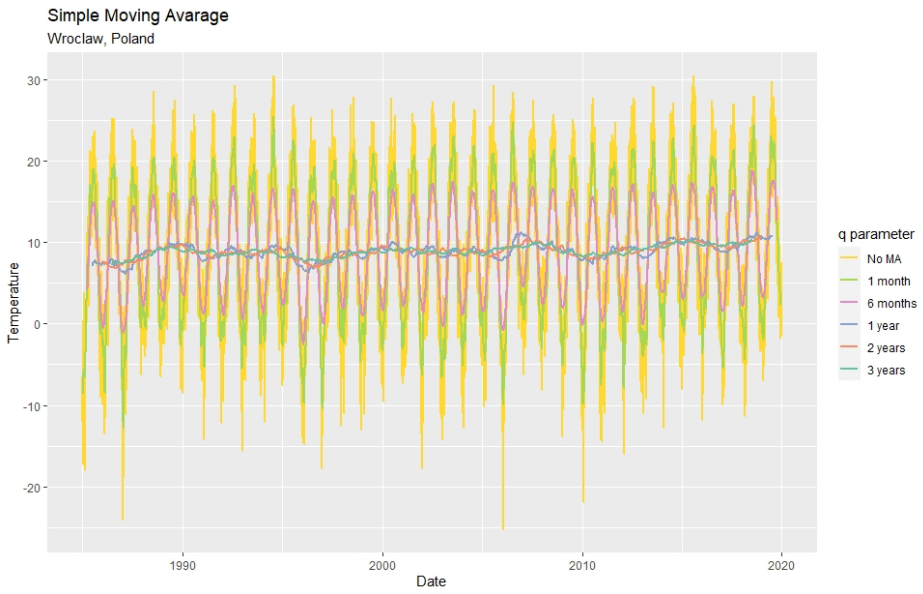

Figure 5: Moving Average method with different q parameter values on daily temperature data in Wrocław, Poland.

We can observe that when we're increasing the q parameter, the daily temperature plot gets smoother. We will zoom in and see the structure of the lines closer.

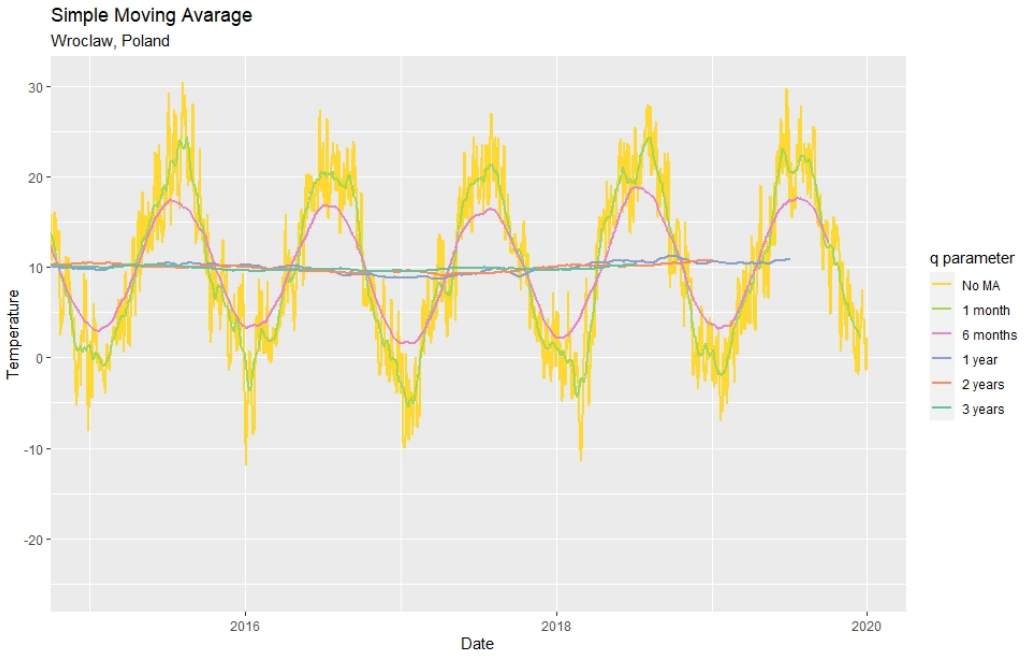

Figure 6: Zoom on the smoothing effect of the simple moving average procedure.

We can see that in this scale, the lines are not as smooth as we thought they would be. However, what we're interested in is not the effect of flattening the chart, but the trend of the data. Let's analyse the the trend factor in the daily temperature data in Wrocław after applying the simple moving average procedure.

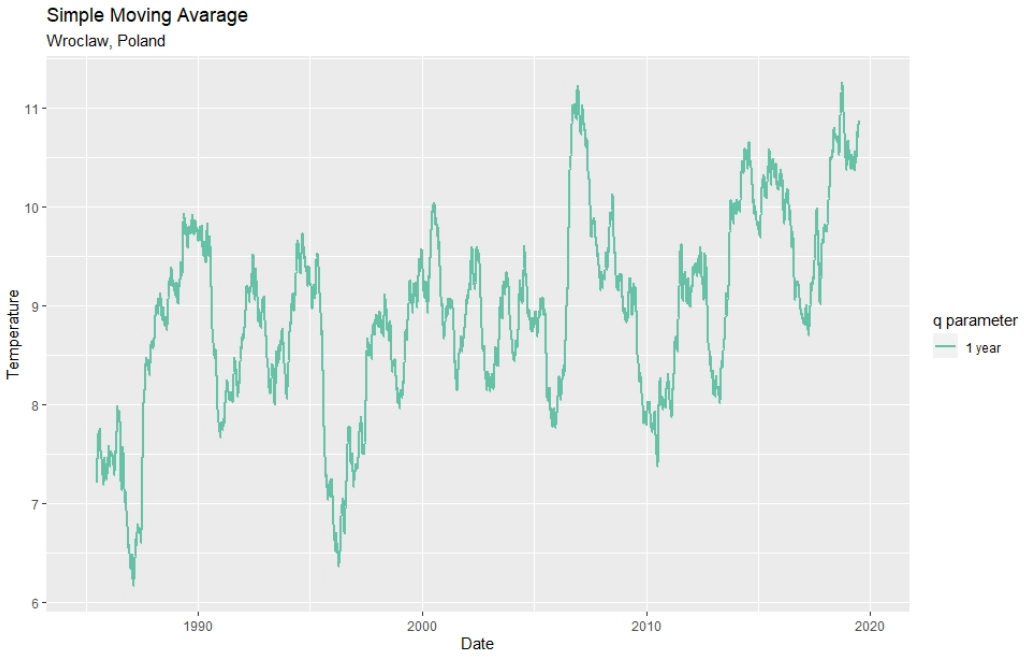

Figure 7: Plot of the temperature data after applying the simple moving average with q parameter equal to 1 year.

We can observe a slightly increasing tendency. However, to say that there exists a clear trend factor in the analysed time series, we shall see it on the Figure (5). The moving average procedure gives us a first insight into the data.

Now, we will extract trend factor from two different time series in two different cities in climate zones far different from Wrocław – Melbourne and Reykjavik, to see if there can be observed an increasing tendency.

The moving average of Melbourne's (Australia) time data is the following:

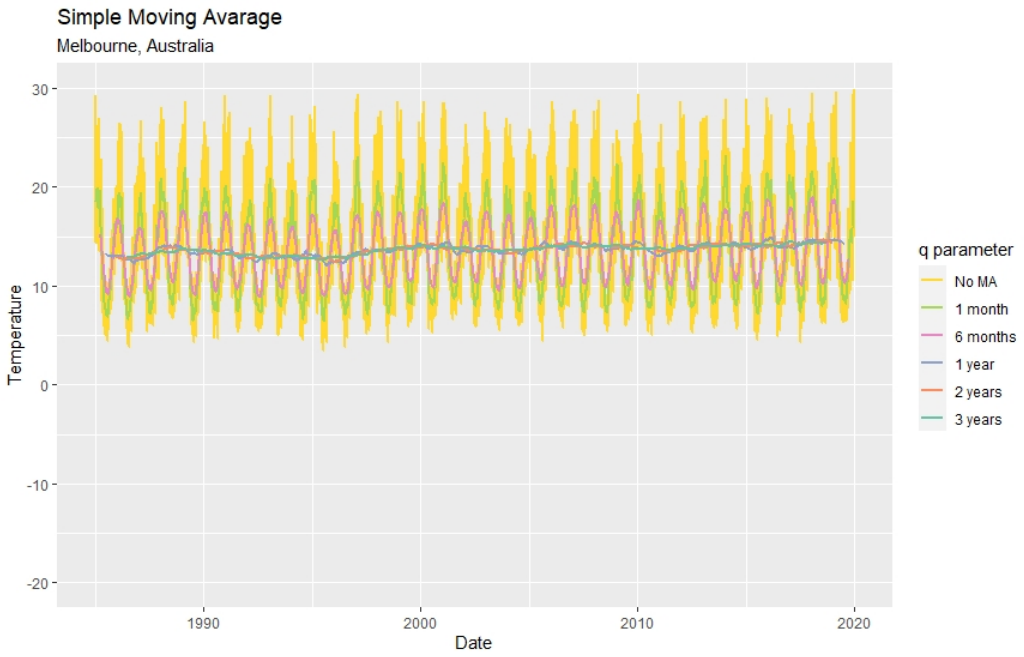

Figure 8: Moving Average method with different q parameter values on daily temperature data in Melbourne, Australia.

The Y axis was fit to the one of Wroclaw's chart for comparison. Obviously, the daily temperature in Melbourne is higher than the daily temperature in Wroclaw. We can also see the smoothing aspect of the moving average procedure. Once again, there is a slight increasing tendency, but not very significant.

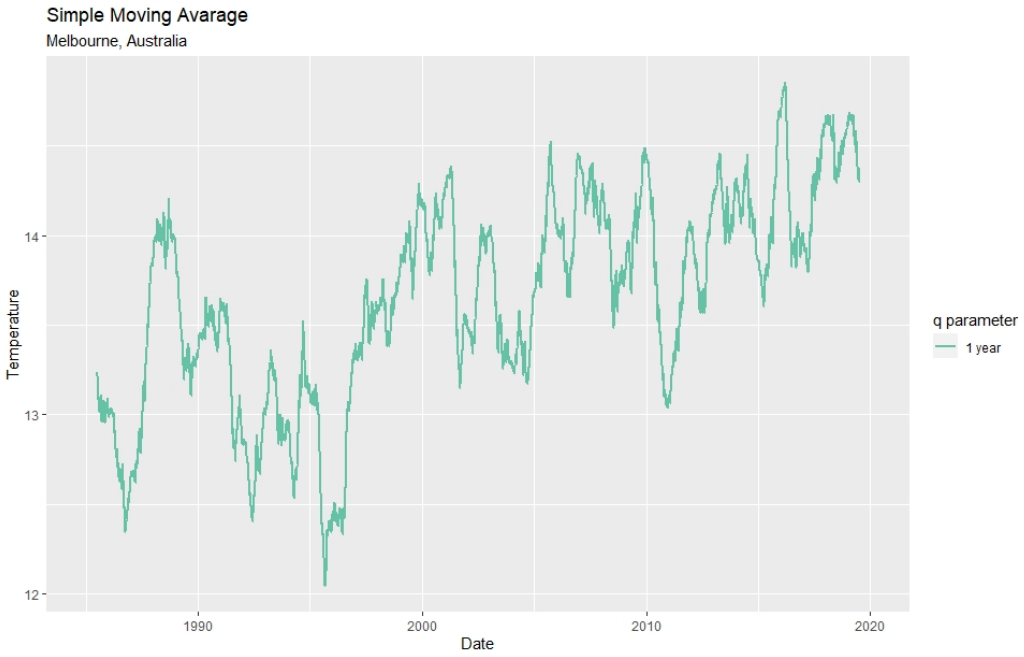

Figure 9: Plot of the temperature data after applying the simple moving average with q parameter equal to 1 year.

The difference between January 1985 and December 2019 is about 1°C. The average level at which the plot is running is from about 12°C to almost 15°C. For Wroclaw's temperature data the highest moving average point is just above 11°C. We can see a very significant temperature difference between these two places.

Let's compare it with Reykjavik's chart after applying the simple moving average model with 1–year q parameter.

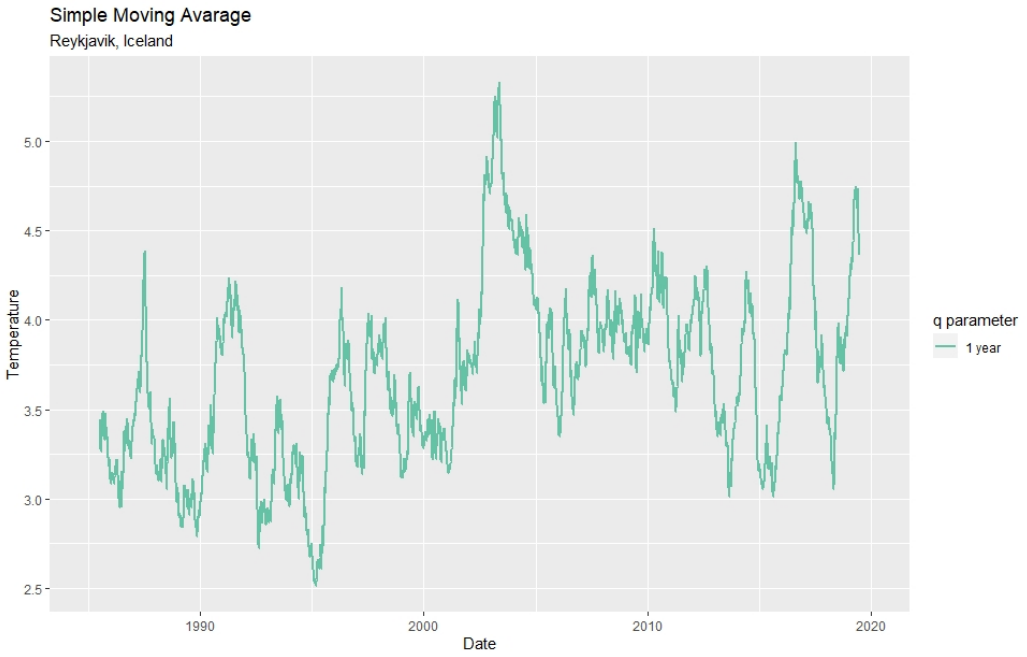

Figure 10: Plot of the temperature data after applying the simple moving average with q parameter equal to 1 year.

In terms of Reykjavik, the increasing trend is even less visible. However, its moving average model shows some very interesting observations. We can see a significant rise in temperature around 2003. According to [3] years 2003 and 2004 were the warmest ones in the recent climatic history of Iceland. Another anomaly can be observed around 2015. As it is stated in [8], 2015 was the coldest year since 2000 which can be seen on the plot above.

Summing up, the moving average algorithm didn't give any significant results in terms of long-term trend identification. To eliminate the trend component it should have a visible regularity in a long-term run. We will not include it in the further decomposition. However, as we've seen in the example of Reykjavik's temperature data, it can be quite useful for spotting some historical dependencies and facts that happened in the past.

3.4 Seasonal component

Next step in time series decomposition is identification of the seasonal component. We have already observed that the data we're dealing with is strongly seasonal due to changing seasons.

We transform the data $\{X_t\}$ into a matrix with 35 columns and 365 rows. That's why we needed the transformation mentioned in 3.1. Later we will calculate mean value of every observations in each row. That will be the estimate of seasonal factor.

The matrix that we are constructing:

$$\begin{pmatrix} X_{1,1} & X_{1,2} & \cdots & X_{1,35} \\ X_{2,1} & X_{2,2} & \cdots & X_{2,35} \\ \vdots & \vdots & \ddots & \vdots \\ X_{365,1} & X_{365,2} & \cdots & X_{365,35} \end{pmatrix}. \tag{13}$$

The seasonal factor estimator is calculated as

$$\hat{s_i} = \frac{1}{n} \sum_{j=1}^{n} X_{i,j}, \tag{14}$$

where $i = 1, \ldots, 365$ and $n = 35$. Because all of the analysed data is seasonal, plots for different cities would only differ in case of Y axis and its values. Therefore, we will only show the seasonal plot for Wrocław.

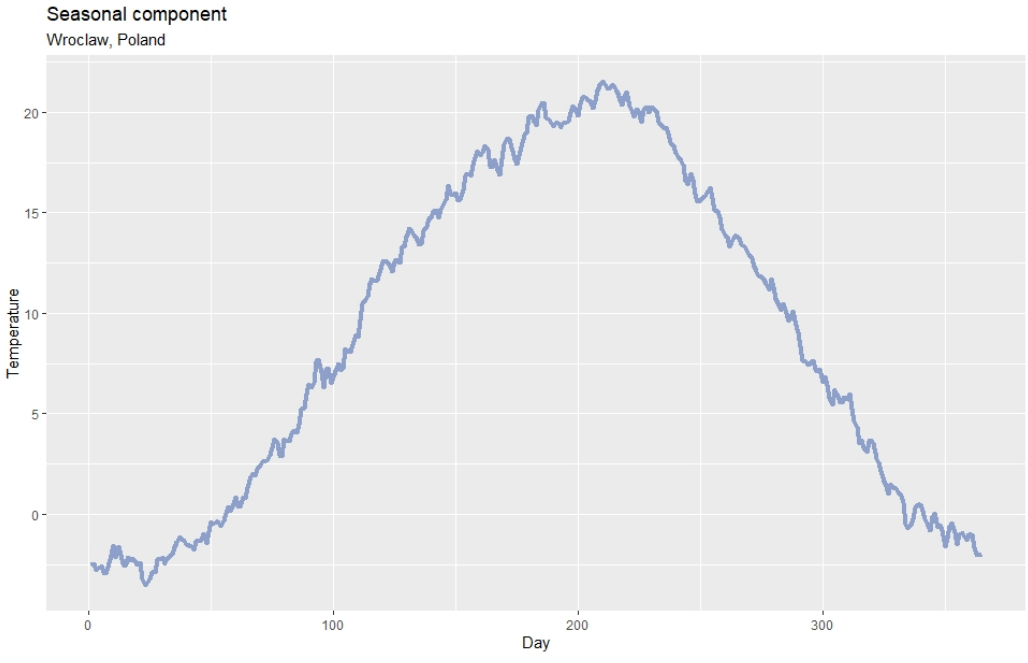

Figure 11: Seasonal factor of daily temperature data in Wroclaw, Poland.

The plot is increasing up until roughly 200^{th} day of the year, so until July. It is comparable with the seasonal plot on the Figure 2. Then the line on the plot is decreasing, indicating the lower temperature during the colder months. Seasonality has a significant impact on the observations, determining their values. This happens cyclically every year.

3.5 Random noise component

The objective of this step is to produce a stationary time series. Because we don't deal with trend factor, but have strongly seasonal data, we need to remove the identified seasonality from the original temperature measures. We will get a random noise component, which will be further analysed and used to fit an appropriate model, because for that we need the analysed time series to be stationary. The process of eliminating the seasonal component and extracting random noise concludes in one simple equation:

$$\hat{Y}_t = X_t - \hat{s}_t,$$

where $\hat{Y}_t$ is the estimated value of random noise component.

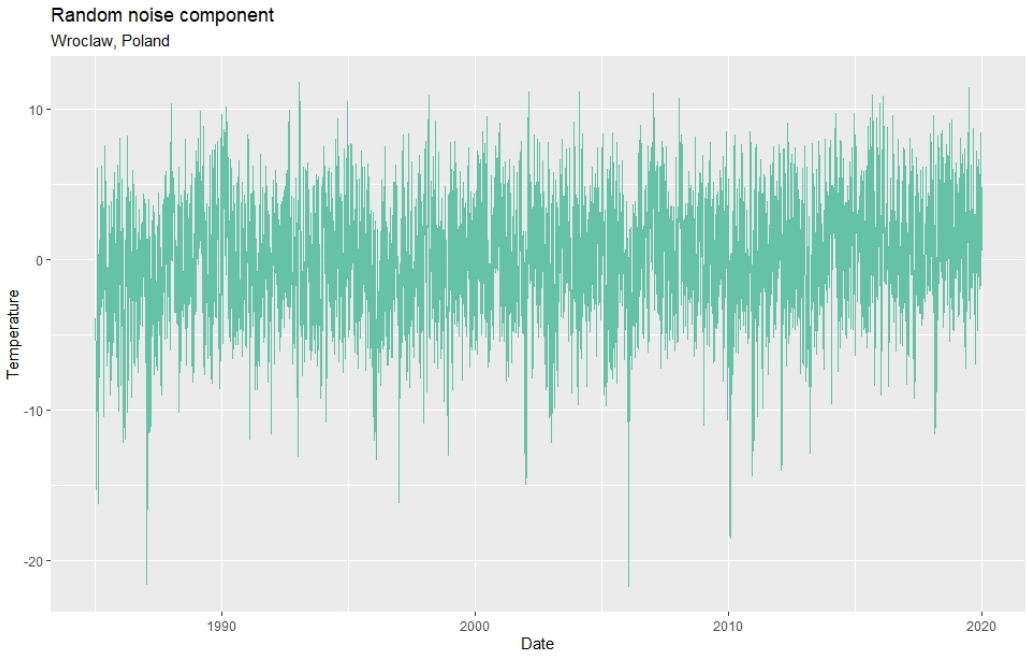

Figure 12: Random noise component of daily temperature in Wrocław, Poland.

The data doesn't have any specific trends. We cannot observe any regularities as we did at the beginning, by analysing Figure 1. The time series has become stationary, it's variance is approximately the same with the time changing. The objective to make the input time series stationary has been achieved.

To confirm if we have successfully removed the seasonality, we need to examine ACF and PACF charts.

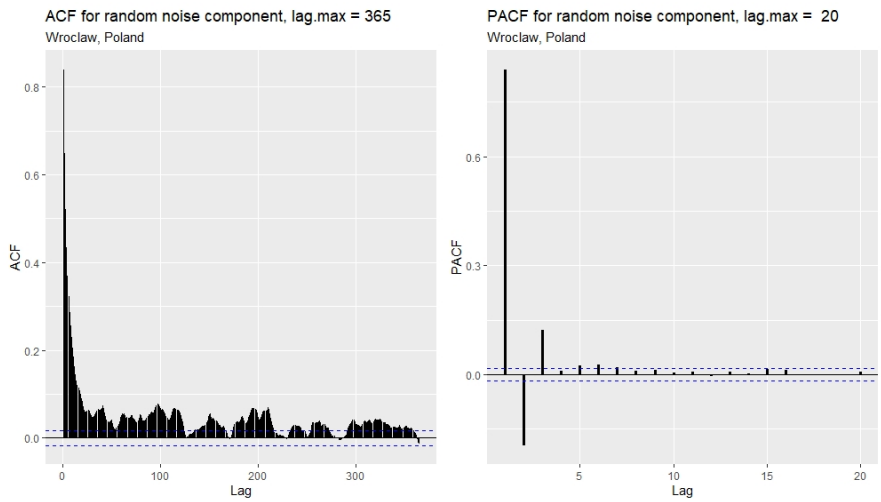

Figure 13: Autocorrelation and partial autocorrelation plots of random noise component of time series data in Wrocław, Poland.

For us to say, that the autocorrelation between data doesn't exist, the lags of the ACF plot should be between blue dotted lines indicating a 95% confidence interval $\pm 1.96\frac{1}{\sqrt{n}}$. They are not, so the observations are still dependent on the previous ones. This is actually an advantage, because we can use that to predict future values of daily temperature. The ACF plot also confirms that we have successfully removed seasonal component, because by comparing it with Figure 1, we can see that the autocorrelation lags no longer have an almost cosine shape for the maximum lag equal to 365. It indicates a significant improvement in terms of increasing the dependence between data.

The partial autocorrelation plot shows far less spikes outside the confidence interval. It also gives us an evidence of decreased dependence on the previous observations. Comparing it with Figure 4, we can observe that the maximum significant lag is reduced from about 15 to 7. This plot is a good indication of the model we should apply for the data. If the partial autocorrelation function $PACF(h)$ is between the confidence intervals $\pm 1.96\frac{1}{\sqrt{n}}$ for

lags $h > p$ we can make an assumption, that the time series data are a realization of the *autoregressive model* AR with the order p. Here $p = 7$, because after 7^{th} lag the spikes of the plot are between the confidence intervals. We will introduce the $AR(p)$ model in the next section.

We will construct ACF and PACF plots for random noise component for other cities.

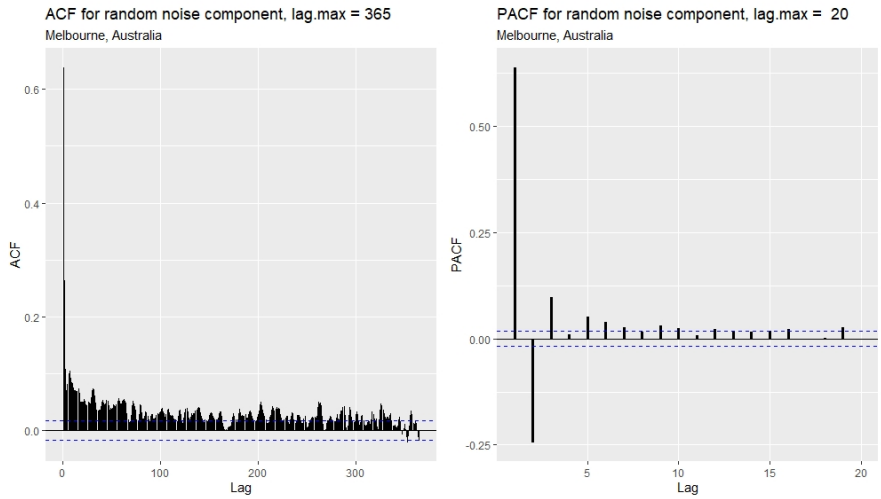

Figure 14: Autocorrelation and partial autocorrelation plots of random noise component of time series data in Melbourne, Australia.

Melbourne's partial autocorrelation plot shows that there are some lags after lag 7 that are statistically significant. It means that the order p of the autoregression model will be probably greater than 7 and, by looking at the plot, even greater than 18. We can conduct that the daily temperature in Melbourne is actually significantly dependent on a longer period of previous observations. Analyzing Figure 2 (seasonal plots) we can make an observation, that the change in monthly mean temperature is smaller than in Wroclaw. The temperature in winter months in Australia is about 10°C lower than in the summer, whereas in Wrocław the difference is twice as high. This leads us to the conclusion that the greater number of statistically significant values of PACF in Melbourne's data random noise could be related to the more stable temperature throughout the year.

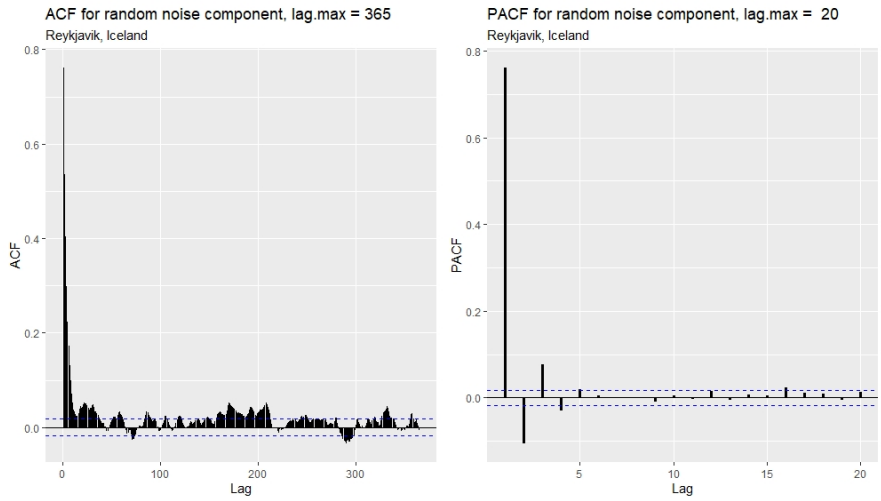

Figure 15: Autocorrelation and partial autocorrelation plots of random noise component of time series data in Reykjavik, Iceland.

The spikes of ACF plot of Reykjavik's decomposed time series drop faster to confidence intervals than plots of Wrocław and Melbourne. The observations are less correlated which is also visible at PACF plot, where almost all of the spikes after lag 5 are between the intervals. This could mean that the Icelandic temperature is prone to rapid changes. Quoting [7]: 'Iceland, especially inland and during winter, is frequently subject to abrupt and dramatic changes in weather that can sharply reduce visibility, as well as rapidly increasing wind speed and precipitation, and shift temperature.'

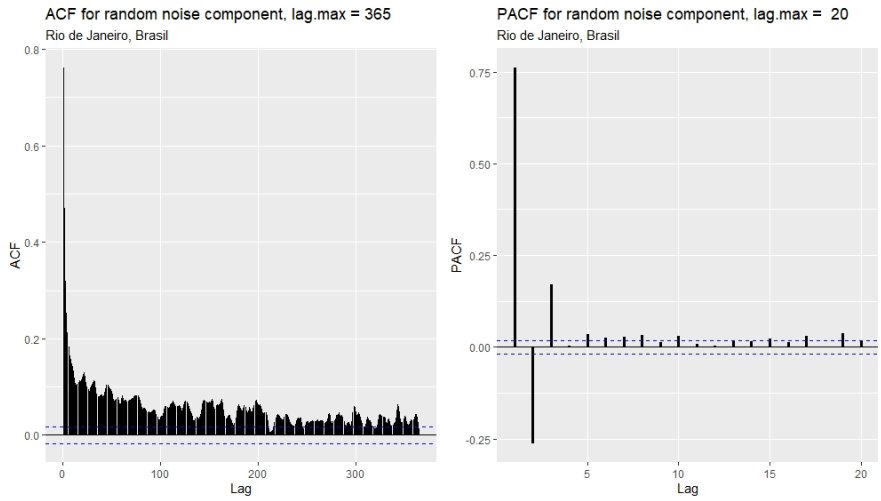

Figure 16: Autocorrelation and partial autocorrelation plots of random noise component of time series data in Rio de Janeiro, Brasil.

Rio de Janeiro's random noise component's ACF plot is quite similar to Melbourne's but with even higher lags. The spikes of PACF plot are also statistically significant up until 20^{th} lag. Once again, comparing the results with seasonal plot, we can see that the temperature throughout the months is changing even less than in Australia, so it could be a reason for higher dependence on previous days.

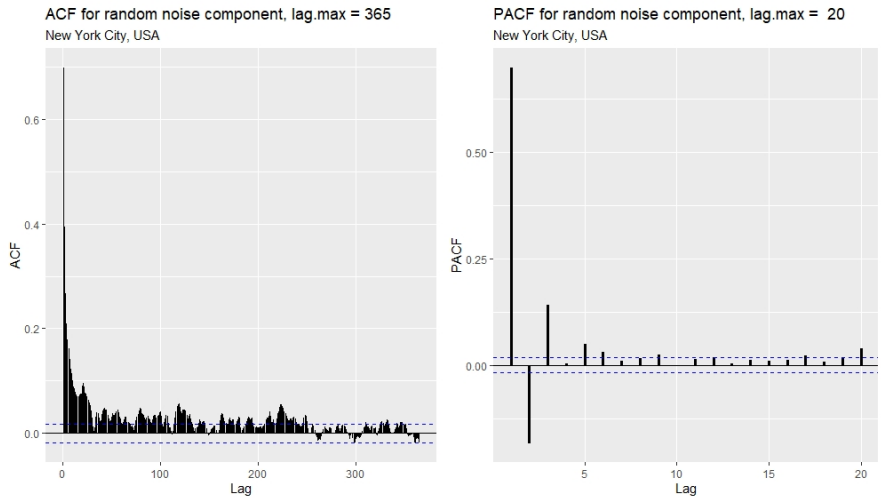

Figure 17: Autocorrelation and partial autocorrelation plots of random noise component of time series data in New York City, USA.

ACF plot for New York City random noise component is quite similar to Wrocław. However, the PACF plot shows some significant spikes that are at lag greater than 7. This indicates that the current observations are also dependent on more than a week-long time period.

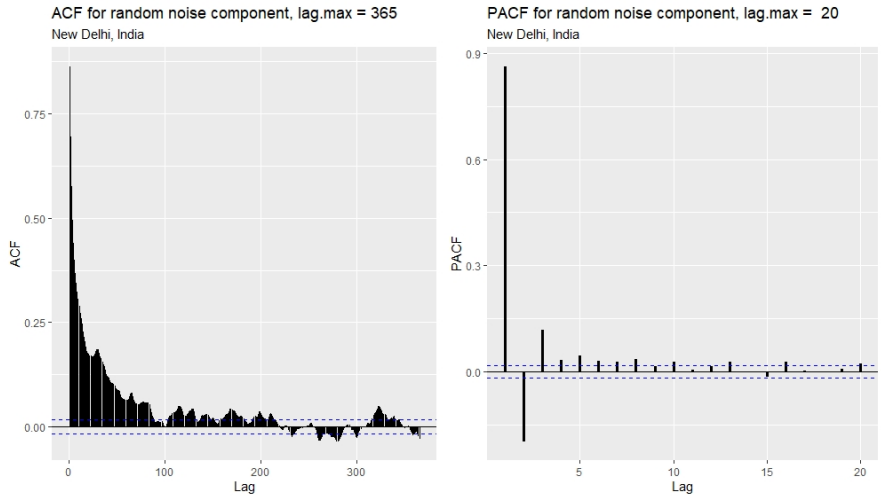

Figure 18: Autocorrelation and partial autocorrelation plots of random noise component of time series data in New Delhi, India.

New Delhi's random noise component has ACF function values dropping quite slowly. The partial autocorrelation spikes are also significant up until 20^{th} lag. New Delhi's temperature shown on the seasonal plot shows some significant difference between summer and winter months. Summer in New Delhi is extremely hot and winter is quite moderate.

4 Modelling

Choosing the right model for the given data is a critical step in time series analysis. It lets us do the most accurate predictions in the future. Most of the models considered for time series are stationary and for that reason we have transformed the temperature data time series into a stationary one.

4.1 Theory

The main property of the autoregressive (AR) model is that it works under the assumption that past observations influence the current values. In terms of temperature data, we have already proved that by using autocorrelation and partial autocorrelation functions.

Before we present a definition of the autoregressive process, we shall review a stationary process that is a component of the AR process and also many other processes.

Definition 11 (White Noise $WN(0, \sigma^2)$)**.** *The process $\{Z_t\}$ is called white noise with mean 0 and variance σ^2 written $\{Z_t\} \sim WN(0, \sigma^2)$ if and only if $\{Z_t\}$ has 0 mean and covariance function defined as:*

$$\gamma(h) = \begin{cases} \sigma^2 & \text{if } h = 0 \\ 0 & \text{if } h \neq 0 \end{cases} \tag{15}$$

Remark. *If the random variables Z_t are independently and identically distributed with mean 0 and variance σ^2 then we write $\{Z_t\} \sim IID(0, \sigma^2)$*

The white noise process is an important building block for other stationary models.

The autoregressive process is a part of more general process called the autoregressive–moving average model, in short $ARMA$. The investigation of the $ARMA$ process will let us understand the AR process better.

Definition 12 (The Autoregressive–Moving Average Process, $ARMA(p, q)$)**.** *The process $\{X_t, t \in \mathbb{Z}\}$ is an ARMA(p, q) process if $\{X_t\}$ is stationary and if for every t:*

$$X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q}, \tag{16}$$

where $Z_t \sim WN(0, \sigma^2)$.

We can simplify the equation (16):

$$\phi(B) X_t = \theta(B) Z_t, \ t \in \mathbb{Z}, \tag{17}$$

where ϕ and θ are the p^{th} and q^{th} degree polynomials called the autoregsessive and moving average polynomials. The autoregressive polynomial:

$$\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p. \tag{18}$$

The moving average polynomial:

$$\theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^q \tag{19}$$

B is the backward shift operator defined by

$$B^j X_t = X_{t-j}, \ j \in \mathbb{Z} \tag{20}$$

The ARMA process consists of the moving average process and the autoregressive process. We will focus on the definition of the second one because it applies to the data we're analysing.

Definition 13 (The Autoregressive Process, $AR(p)$)**.** *If* $\theta(z) \equiv 1$ *then*

$$\phi(B)X_t = Z_t \tag{21}$$

and the process is called the autoregressive process of order p *(AR(p)).*

4.2 Application of the Autoregressive Model

Now we will conduct the application of the $AR(p)$ model to the temperature data. In section 3.5 we have discussed the partial autocorrelation plot of the random noise component. From each PACF plot we could deduct the order p of autoregressive model. It's obviously a flawed method and we will use `ar()` function from `R` to estimate the autoregression polynomial coefficients ϕ_i, $i \in \{1, \ldots, p\}$.

The `ar()` method uses the Akaike Information Criterion to estimate the model's order. Coefficients are estimated using Yule–Walker method. Here are the fitted models plotted in orange, compared with the original daily temperature data plotted in gray. Each plot contains information about the order p of the autoregressive model.

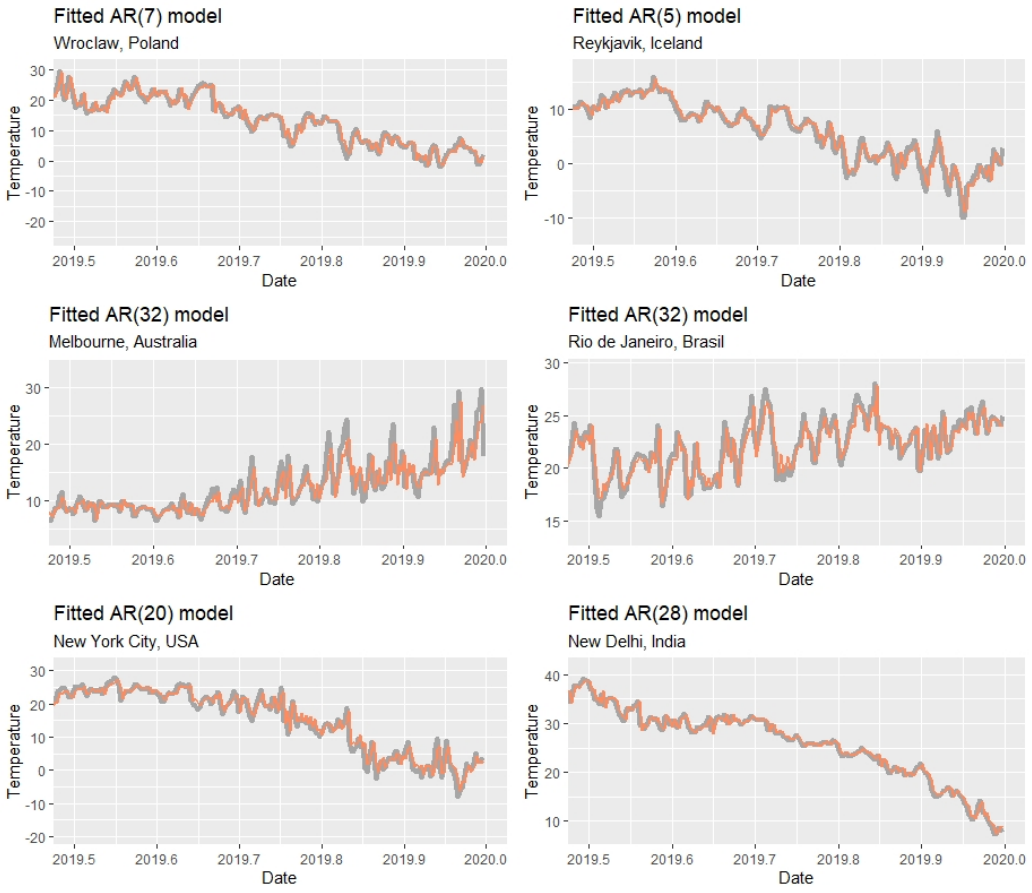

Figure 19: Comparison of fitted AR(p) models for daily temperature data of different cities.

The plots show last half of the year 2019 for each city. We can see that the model looks similar to the data, however the bigger the order p, the more of the original data is far from the points estimated by the model. For example, Melbourne's and Rio's plots show the biggest differences at the highest "spikes". This gives us the first insight into the model and its ability of dealing with extreme values.

4.3 Stationary solution of the Autoregressive Process

We will check the stationarity of the processes calculated in the previous section. To do that we shall examine the inverse roots of the characteristic polynomial

$$\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p. \tag{22}$$

If the inverse roots are inside the complex unit circle, then the $AR(p)$ process is stationary and causal [see Theorems 3.1.1 and 3.1.3, pages 85, 88 in [1]].

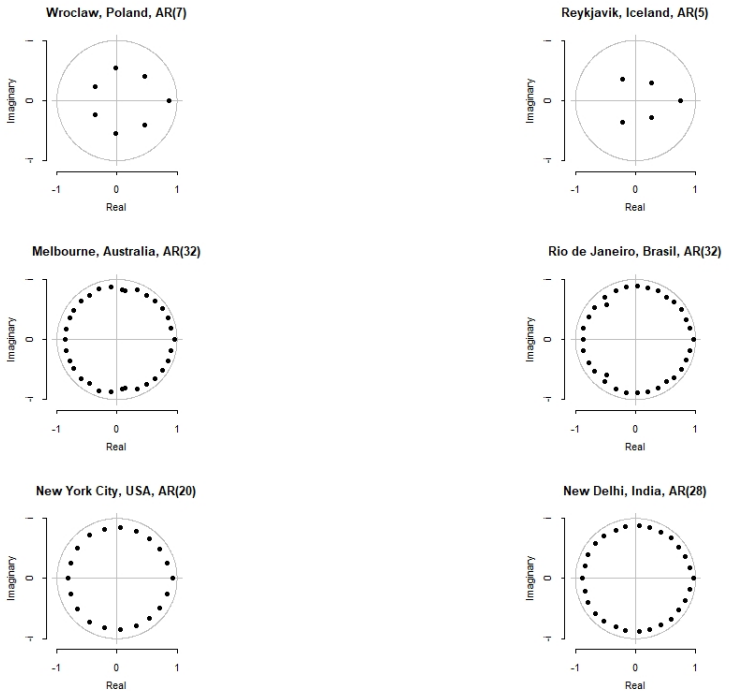

Figure 20: Inverse roots of the polynomial $\phi(z)$ for every city.

The inverse roots are indeed inside the complex unit circle. Thus, the $AR(p)$ processes are stationary and we can move on to forecasting.

5 Forecasting

In this section we will perform the last but not least part of time series analysis. Predicting future data will give us foundation for a discussion about modelling information about daily temperature.

5.1 Theory

The problem we want to tackle is predicting the value of X_{n+1} observation of a stationary time series with known mean μ and autocovariance function γ_X based on the values $\{X_n, \ldots, X_1\}$. For the purpose of this paper and later prediction of the temperature data we will restrict the group of stationary

processes to the ones that are zero-mean. In this analysis we will not conduct the X_{n+h}, $h > 1$ prediction, because even empirically, we know that forecasting temperature a few days ahead is not accurate. The forecasting we will be doing is called *one-step ahead*. One-step ahead prediction means prediction of just one observation based on p previous ones, where p is the autoregression order.

The goal is quite simple – we want to find a linear combination of $X_n, X_{n-1}, \ldots, X_1$ that forecasts X_{n+1} with minimum mean squared error. The best linear predictor in terms of $X_n, X_{n-1}, \ldots, X_1$ will be the following:

$$\hat{X}_{n+1} = a_1 X_n + \cdots + a_n X_1 = \sum_{j=1}^{n} a_{nj} X_{n+1-j} \tag{23}$$

for some a_n. It is supposed to minimize:

$$v_n = \mathbb{E}[(X_{n+1} - \hat{X}_{n+1})^2]. \tag{24}$$

By solving (24) using (23), we get

$$v_n = \mathbb{E}\left[\left(X_{n+1} - \sum_{j=1}^{n} a_{nj} X_{n+1-j}\right)^2\right] = \mathbb{E}[X_{n+1}^2] + a_n' A_n a_n - 2a_n' b_n, \tag{25}$$

where $a_{n,ij} = \gamma_X(i-j)$, $b_{n,j} = \gamma_X(j)$. We find the optimal a_n by solving:

$$A_n a_n = b_n. \tag{26}$$

This includes inverting the matrix A_n, which can be very computationally inefficient for large values of n. However, there are two algorithms that give the value of $\hat{X}_{n+1}$ without the need of inverting the matrix. The first algorithm is called the Durbin-Levinson algorithm and the second one is referred to as the innovations algorithm. We will use the innovations algorithm to compute one-step ahead predictions for $AR(p)$ process.

The innovations algorithm is a recursive algorithm that is applicable to all time series, regardless if they are stationary. It relies on the fact that what we are interested in is the prediction of the observations themselves, not the values a_n. It produces forecasts that are linear combinations of prediction errors, therefore for $j = 1, \ldots, p$ we have

$$\hat{X}_{n+1} = \sum_{j=1}^{n} d_{nj}(X_j - \hat{X}_j). \tag{27}$$

Now we will want to find the coefficients d_n of forecast errors. Applying the innovations algorithm to $AR(p)$ process leads us the the final form of one-step ahead prediction that we will be using:

$$\hat{X}_{n+1} = \phi_1 X_n + \phi_2 X_{n-1} + \cdots + \phi_p X_{n+1-p},\ n \geq p, \tag{28}$$

where the coefficients $\{\phi_1, \ldots, \phi_p\}$ are the coefficients of the autoregressive model computed in the section 4.2. The thing about the calculated coefficients is that ϕ_1 is close to 1. For example, the $AR(7)$ prediction model for the temperature of Wroclaw, Poland is the following:

$$\begin{aligned}\hat{X}_{n+1} = 1.0257X_n - 0.3174X_{n-1} + 0.1173X_{n-2} - 0.0102X_{n-3} + \\ + 0.0028X_{n-4} + 0.0083X_{n-5} + 0.0189X_{n-6}.\end{aligned} \tag{29}$$

As we can see the first coefficient is close to 1, whereas the rest of the coefficients are not very impactful. This will be the reason for the predictions' one-day shift that we will see in the next subsection.

5.2 Daily temperature forecasting

We perform one-step ahead forecasting for time period from 01.01.2020 to 31.03.2020. We compare the predicted values with actual data that has become a test set for the model. Actual data from NASA Power database is marked as turquoise dashed line and the predicted values are shown as orange solid line. We will also analyse forecast errors plots. The blue dashed lines indicate $[-\sigma, \sigma]$ and $[-3\sigma, 3\sigma]$ intervals.

Figure 21: Comparison of 90–day forecasting and errors for temperature data in Wroclaw, Poland.

For temperature data in Wroclaw the prediction doesn't do very well with weather anomaly like almost 10°C at the beginning of February. It is also clear on the forecast errors (residuals) plot. The point standing for residual for that observation and its fitted value lays far outside the $[-\sigma, \sigma]$ interval. Another outstanding prediction error happens at the end of February, when the model doesn't fit to a sudden temperature drop and rise. Generally, the prediction points for some rapid weather changes are badly fitted and therefore their residuals indicate difference greater than the model's standard deviation σ.

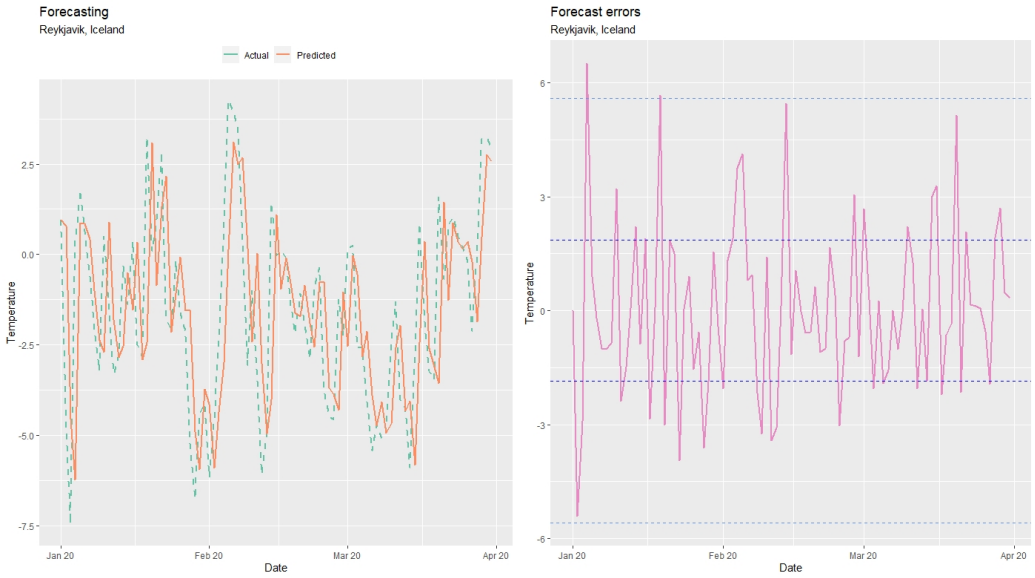

Figure 22: Comparison of 90–day forecasting and errors for temperature data in Reykjavik, Iceland.

Reykjavik's prediction reflects the trends and fluctuations of the data. However, like in the previous example, it doesn't do well with sudden temperature changes like the ones in February or at the beginning of January. February had both temperature measures below $-5°$C and above $4°$C in a very short time stamp. The highest temperature in this time interval was measured on the 5^{th} of February and was $4.25°$C high. The fitted value for this observation has a forecast error equal to almost $4°$C. Another interesting conclusion is the fact, that the lowest temperature was $-7.47°$C and it was measured on 3^{rd} of January. The forecast error for this observation is quite high – it is equal to almost $-3°$C. However, it is the observation happening on the following day that has the highest residual value and the least accurate prediction. On the 4^{th} of January 2020 the daily mean temperature was about $0°$C. For the prediction model this is more of an anomaly than the extremely low daily temperature on the previous day.

Figure 23: Comparison of 90–day forecasting and errors for temperature data in Melbourne, Australia.

Melbourne's model has one of the worst–looking fitted values and errors plots. The reason for that could be the infamous Australian "Black Summer" that happened at the beginning of 2020. The term refers to unusually intense bushfires. Before the 9^{th} March they burnt over 18.6 million hectares of land. In the state of Victoria, where Melbourne is, the bushfires destroyed about 1.5 million hectares and burnt about 400 homes. It is important to know, that the bushfires in Australia are part of the summer season, however in the summer of 2020 they were very severe mainly due to the drought and very high temperatures, which can be seen on the plot. By the end of January, the citizens of Melbourne experienced particularly hot days with the mean daily temperature of about 30°C. Especially last two days of January brought 30–degree heat. This are also observations that the model had trouble predicting very accurately.

Figure 24: Comparison of 90–day forecasting and errors for temperature data in Rio de Janeiro, Brasil.

Most of the residuals of Rio de Janeiro's predicted data is between the $[-\sigma, \sigma]$ interval with a few exceptions. There was a drastic drop of temperature in the mid–January. In the second part of January 2020, the citizens of Rio de Janeiro experienced a subtropical storm called *Kurumi*. The storm has lead to a severe flooding and heavy rains. This could be the reason of the quite low temperature measured in that time. The model has made some actually good predictions. The reason is probably the overall monsoon characteristic of Rio's climate. The long heavy rains happen there all the time between December and March, so the model should have included that.

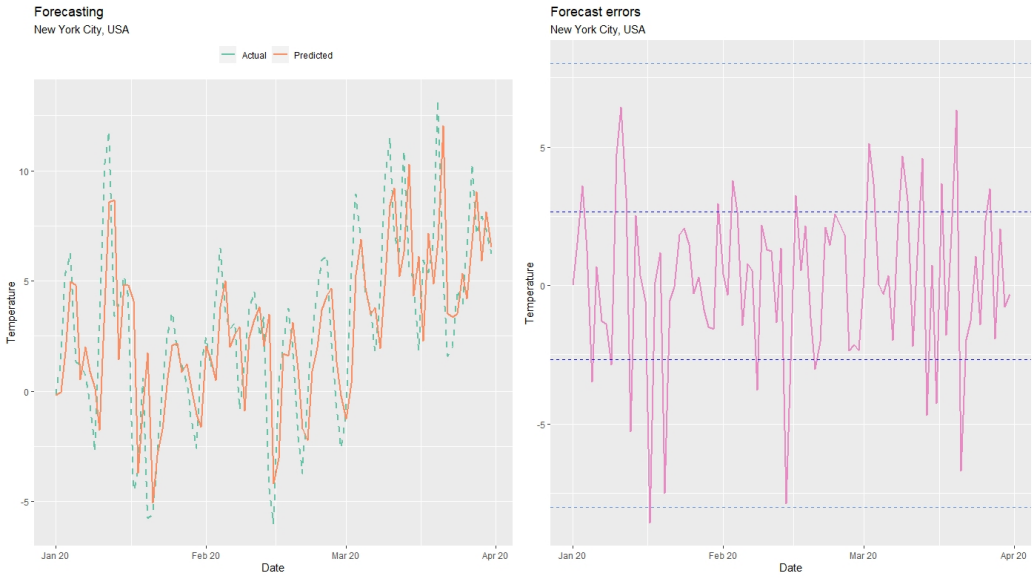

Figure 25: Comparison of 90–day forecasting and errors for temperature data in New York City, USA.

January 2020 in the New York City had some pretty varying temperatures, which once again were predicted by the model, but not very accurately with some rather big forecast error values. Especially big differences happened in the first half of the month. The model didn't perform well with the sudden rise and drop of the temperature resulting in a residual value even lower than -3σ.

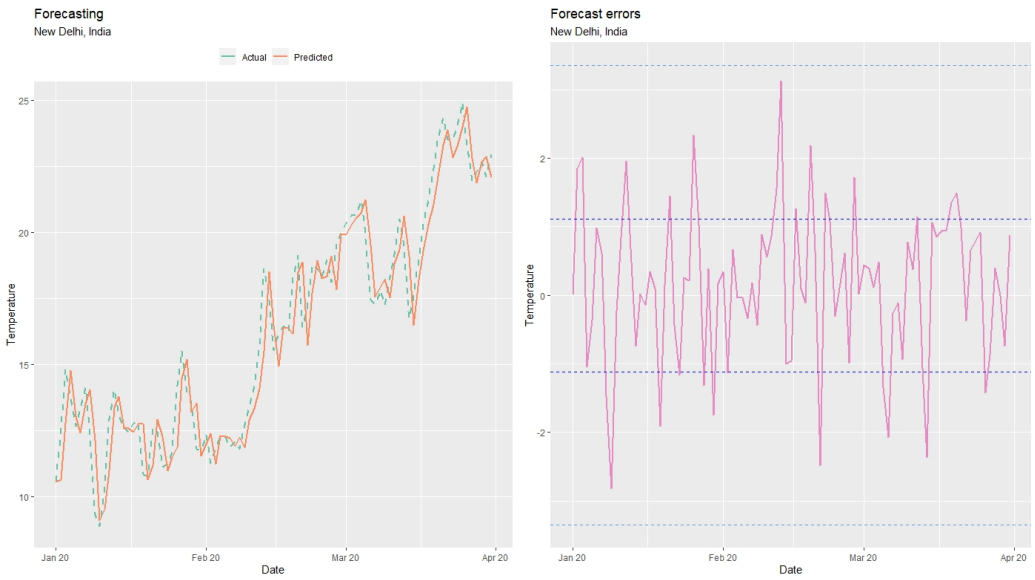

Figure 26: Comparison of 90–day forecasting and errors for temperature data in New Delhi, India.

Even though the forecast plot for New Delhi's data looks quite like the actual temperature measures, there are a few forecasting errors outside the $[-\sigma, \sigma]$ interval. Especially high residual values are connected to predictions of a significant temperature drop at the beginning of January and also a drastic temperature rise at the end of March. Overall, the graph shows some increasing trend tendency which is characteristic to the Spring time in Delhi.

6 Conclusion

We need to remember, that the temperature is extremely dependent on many other factors related to different weather properties as well as geographic and social characteristics. The wind, sudden weather anomalies like tornadoes and floods or unusually intense bushfires are somewhat predictable for us, but not really for a mathematical model as simple as the one presented. The models used in serious meteorological predictions take many factors into consideration and can produce very accurate predictions.

I started my thesis by quoting a famous writer, I will end it by referring to George Box – "one of the greatest statistical minds of 20^{th} century":

> All models are wrong, but some are useful.

7 Code

The code for analysis was implemented using `R` programming language and is available here: `https://github.com/nelanz/BSc-Thesis`.

References

[1] P. J. Brockwell, R. A. Davis, *Time Series: Theory and Methods, Second Edition*, Springer Science + Business Media New York, 1991.

[2] P. J. Brockwell, R. A. Davis, *Introduction to Time Series and Forecasting, Second Edition*, Springer-Verlag New York Inc. 2002.

[3] E. Hanna, T. Jónsson, J. E. Box, *Recent changes in Icelandic climate*, Royal Meteorological Society, 2006, Vol. 61, No. 1, s. 1-9.

[4] W. Szczotka, *Szeregi czasowe*, Uniwersytet Wrocławski, Wrocław 2012.

[5] A. Zagdański, A. Suchwałko, *Analiza i prognozowanie szeregów czasowych. Praktyczne wprowadzenie na podstawie środowiska R.*, Wydawnictwo Naukowe PWN SA, Warszawa 2016.

[6] NASA Power

[7] Climate of Iceland, Wikipedia Article [access: 01.09.2020].

[8] The Icelandic Meteorogical Office, *Climate summary 2015*.