Urban Heat Effect in Ghent: a Time Series Analysis

Kumar Kshitij Patel & Sébastien Roels

September 6, 2018

Abstract

In this paper, we investigate the temperature evolution at fixed locations and verify the hypothesis that there is a difference between the temperature inside and outside the city. We find this difference by comparing the temperature trends (fitted local polynomials) and variations (best describing models) between both locations. First, we analyzed the initial data and transformed it into a stationary form. It was done by removing (in chronological order) yearly seasonality, daily seasonality and doing minute-differencing. We follow this with an initial analysis of the ACF/PACF of the differenced data, to facilitate ARMA modeling. After that, we do GARCH modeling for capturing volatility in the data, since the residuals obtained after using appropriate ARMA models are not normal (i.e., Gaussian), which means that even if they are uncorrelated, they are not independent and therefore not white noise. Finally, we investigate whether those models indeed give a good description and we look for clues for which model could describe the data even better. For getting a physical perspective, we compare the local fits and the best fitting models between the two locations to give a conclusion about the temperature difference between city and countryside and hence the urban-heat effect based on the methods at hand.

Contents

1 Initial data analysis

1.1 About the data

The datasets contain year-long temperature measurements with one measurement every minute (for one year). The initial data (of one out of six locations) is shown in Figure 1. One should mention that some locations have missing data or even data with measured temperatures while their ventilation was not active, this results in a 'shift' of temperatures and thus locations containing these issues were avoided in order to be able to focus on the main idea behind this project, which is time series analysis.

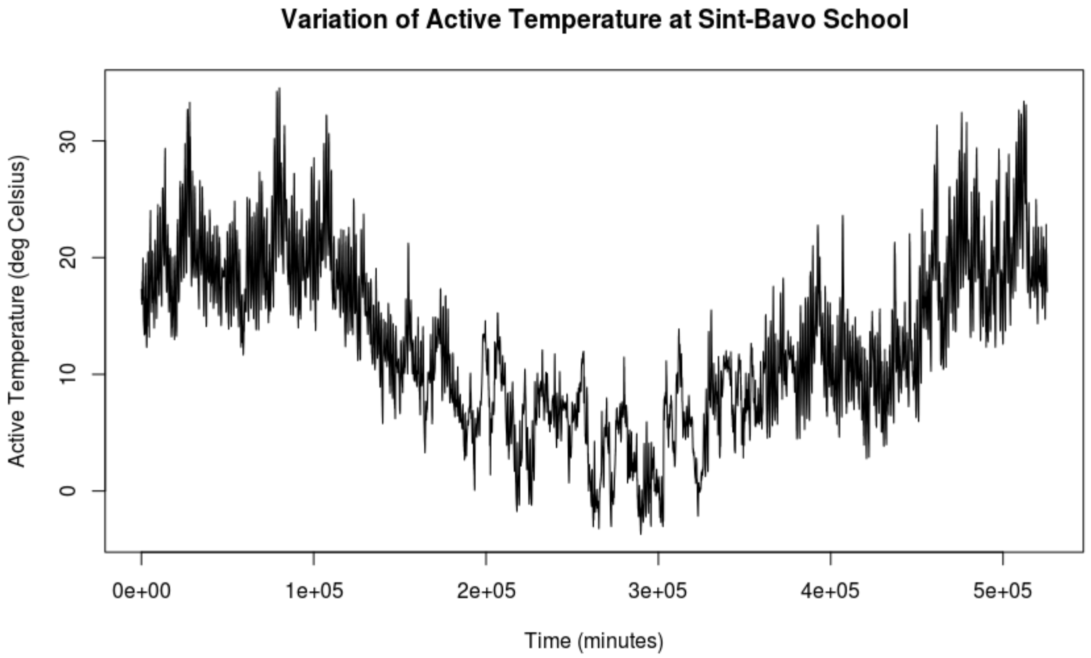


Figure 1: Initial data of location: Sint-Bavo school

1.2 Seasonality removal

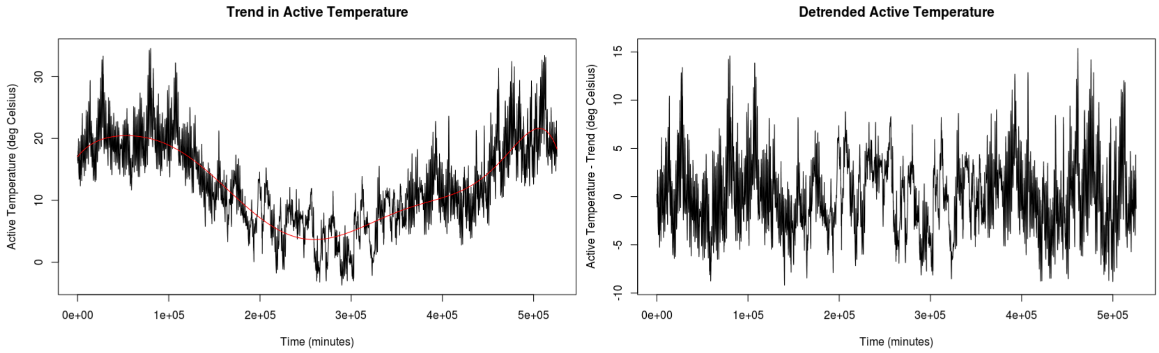
Seasonality removal is a key aspect in this analysis since the data has to be made stationary in such way that we have a minimal information loss, i.e. in a 'least restricting' way. In our data set, we have two seasonalities; daily(DS) and yearly(YS). We will remove these seasonalities from high to low scale, meaning we start by removing YS and then remove the DS. However, the order doesn't really matter since differencing operators commute, i.e. it doesn't matter in which order they are applied.

1.2.1 Yearly Seasonality

YS can be removed in two ways, by an ordinary polynomial regression or local polynomial regression. We use the former, because YS in temperature is well observed and known to be higher during the summer and low during the winter, in a rather predictable manner. Thus, using a local polynomial fit is clearly an overkill for this simple task! Also, differencing ($D^{365}_{days} = D^{525600}$) can not be used since we have only one year of data. The data acquired by removing YS is shown in Figure 2 and now becomes:

$$Y_t \leftarrow Y_t - \mu(t) \tag{1}$$

with $\mu(t)$ a local polynomial fit of order 8 to the data.



(a) Fitted trend of order 8 to data.

(b) Data after (yearly) trend removal

Figure 2: Removal of YR in data.

1.2.2 Daily Seasonality

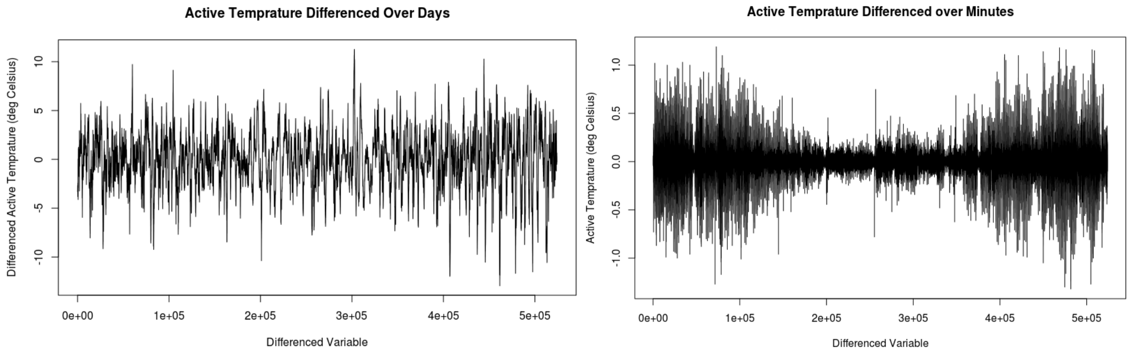
DS can be removed by using ordinary differencing. We difference over minutes on consecutive days, because physically there isn't any other simpler periodicity. The rather obscure side of this decision is the fact that uncertainty increases with differencing which results in the fact that forecasting will be pointless (i.e. will have enormous error bands). The acquired data is shown in the Figure 3a and the new data is given by:

$$Y_t' \leftarrow (I - B^{1440})Y_t, \quad 1440 = 24(hours) \cdot 60(minutes) \tag{2}$$

1.3 Remarks About the Minute Data

In this project we chose not to work with hourly- or even daily averaged time series which is much simpler to deal with owing to a smaller dependence in consecutive data-points. The reason for this is simply that we naively started working this way. Later on. It is uncommon to be able to model minute temperature data very well but after a short discussion with prof. Thibault we decided to continue to work in this direction and see where it would lead us.
It didn't take long before we bumped into the problem of high dependence in the data, even when the new data "survived" multiple stationarity tests (KPSS and PACF plot). The ACF remained almost constant and didn't decay exponentially, which was a clear problem for modeling with ARMA processes. In order to resolve this problem we finished by differencing the data. Thus, the (final) data became:

$$Y_t'' \leftarrow (I - B)Y_t' \tag{3}$$



(a) The acquired data after removal of its daily seasonality.

(b) The acquired data after differencing Equation 2

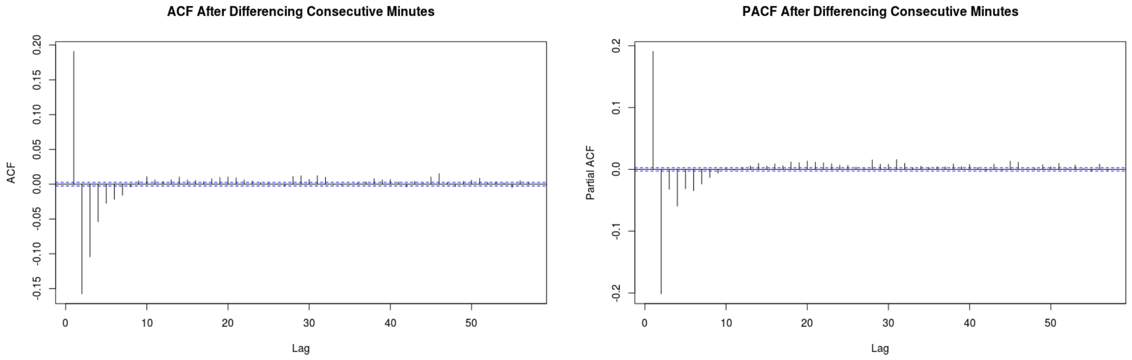
Figure 3: This figure shows the stepwise stationarization of the initial data.

From Figure 3b one clearly sees that the variance is time dependent. Also, when one knows that the data runs over one year and starts and stops in the summer, one sees that for high temperatures

the variance is high and for low temperatures the corresponding variance is low.
One final remark concerning the structure of the report should be made. In the following analysis, we keep the **chronological order** in which we actually worked. Thus, at some point we will make a wrong assumption and only later on we will correct this and finally find the optimal model. The reason for this is that this seems, for us, **the most natural way of reporting what we learned and how we learned it**.

2 Stationarity

After the previous data processing steps have been applied, one has to check if the acquired data is indeed stationary. For this we look at the ACF (Figure 4a), PACF (Figure 4b) and perform a KPSS test.

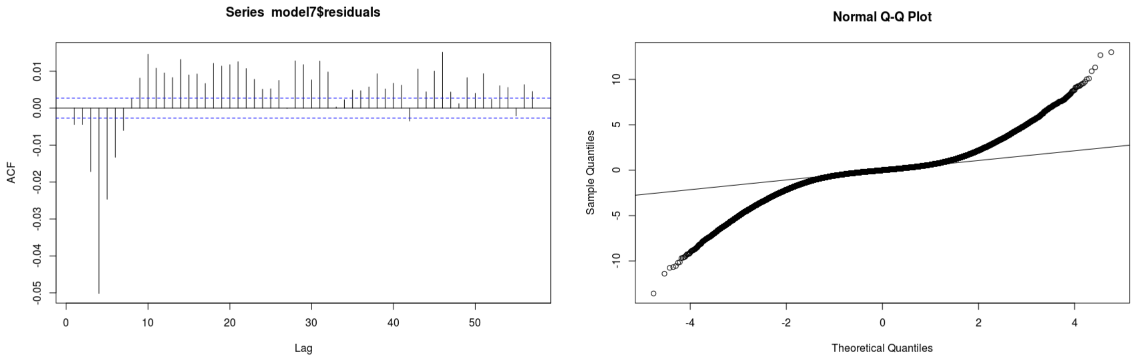


(a) ACF

(b) PACF

Figure 4: ACF and PACF of the acquired (stationary) data.

The ACF and PACF look good and the KPSS test results in a value of 0.1 (>0.05), thus we conclude that the data is stationary and that we can start searching for models which describe the data well. At this point, through domain knowledge and otherwise we try to model the volatility in the temperature data in the following ways:



(a) ACF

(b) QQ-Plot

Figure 5: ACF and QQ-Plot of the temperature dependent variance model.

- **Temperature dependent variations**
  In a physical system, several physical theories state that temperature fluctuations are temperature dependent [1], more precisely that variations will increase when temperature increases, in a rather linear fashion. In order to resolve this we divided the data by the square root of the temperature, to normalize the standard deviation. However, as one sees in Figures

[1] Can for example be proven using 'fluctuation dissipation theory in non-equilibrium statistical physics'.

5b and 5a, the resulting QQ-plot and ACF of squared standardized residuals are not good enough, post the ARMA modeling.

- **AR(8) with GARCH(1,1) sampled with normal distribution**
  From the PACF of the stationary data (Figure 4b) one could argue that AR(8) is a good model.[2] However, also taking into account the time dependent variance, we tried to fit an AR(8)-GARCH(1,1) model fitted with a normal distribution. The acquired ACF of the squared standardized residuals looks fine, however the QQ-Plot is not good, as can be seen in Figure 6.

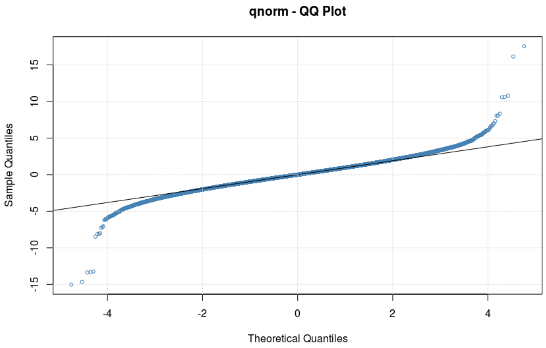
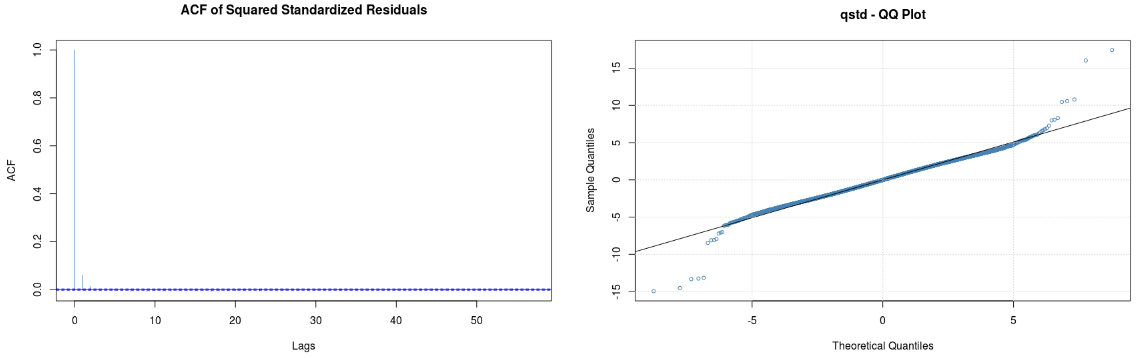


Figure 6: QQ-Plot of the AR(8)-GARCH(1,1) model with normal residuals.

The intuitive reason why we say that the QQ-Plot is not good enough is because the residuals diverge at the ends, which hints that the underlying noise distribution has a non zero sampling probability at the extremities.

- **AR(8) with GARCH(1,1) sampled with t-distribution**
  The ACF of squared standardized residuals and the QQ-Plot of this model look very good, as can be seen in Figure 7.



(a) ACF

(b) QQ-Plot

Figure 7: ACF and QQ-Plot of the residuals from AR(8)-GARCH(1,1) model with t-distribution

3 Model Fitting

In the previous section we concluded that an AR(8)-GARCH(1,1) model sampled with a t-distribution looks rather good. In this section, we will investigate this idea even deeper and analyze whether we can find an even better model. We'd look at the Log-likelihood and the AIC of the model. The major intuitive difference between the two is that the latter besides maximizing the posterior probability of the model given the data also penalizes over-parameterization and thus helps avoid

[2] Here we assumed that the PACF cuts off, later on we will see that this was not the case and thus that the assumption made here is incorrect.

over-fitting. it must be noted that BIC, penalizes over-fitting even more, but the obtained results for AIC and BIC were similar in our case and thus we choose to report only the former.

3.1 Comparison with other models

Now, the natural question is if the used parameters for AR and GARCH are actually optimal. For this we assume, that the optimality of the parameters of the AR model does not depend on the ones of the GARCH model, and vice versa. Thus, we begin by fixing GARCH(1,1) and varying the degree of the AR process. For each of these, we calculate the corresponding log likelihood. The acquired results are shown in Figure 8[3]. In order to find the optimal parameter we search for the maximal gradient. The reason for this is that if we increase the number of parameters, the likelihood will continue to increase but more slowly because we are over-parameterizing. Thus, using this reasoning one finds that AR(10) describes our data best[4].

.

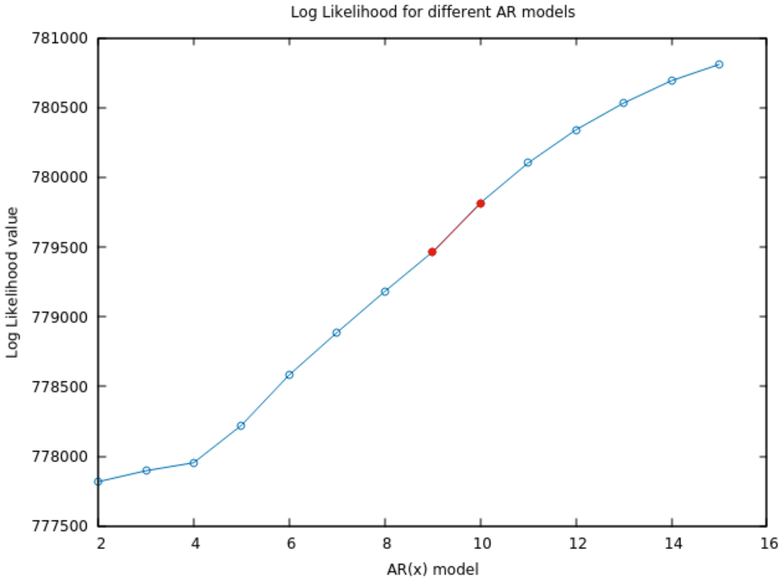


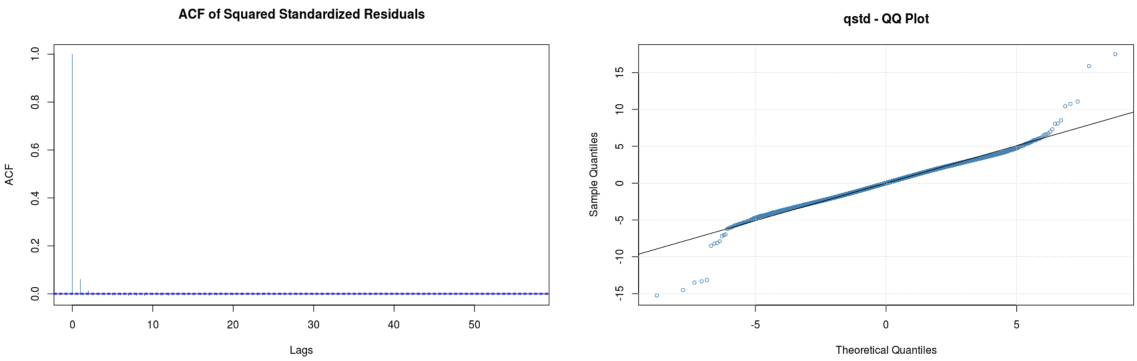
Figure 8: The Log Likelihood values for different AR models fitted together with GARCH(1,1) to the data.

Next, we keep AR(10) fixed and start varying the parameters of the GARCH model. We tried two different models, GARCH(1,0) and GARCH(1,1) with their corresponding log likelihoods being 690916.6 and 779816.2 respectively. From this, one concludes that GARCH(1,1) describes the data best. In-fact literature[1][2] and practice both seem to point at the fact that GARCH(1,1) is usually the most appropriate model for temperature volatility.

In conclusion, one finds that AR(10)-GARCH(1,1) is the best model to describe our data. To support our claim that this is the the optimal model, different aspects of this model are given in Figure 9. Again, one should note that this conclusion is still with the **wrong** assumption that the PACF of the stationary data cuts off.

[3] Note, that we realize that AIC is a better evaluation metric than log likelihood but we chose this method to highlight another parameter estimation technique often used in learning, called the **"elbow-method"**.

[4] One should also mention that the method we used in R didn't return AIC but only the log likelihood and one thus had to be 'creative' to come op with an optimal interpretation



(a) ACF

(b) QQ-Plot of SS-residuals

Figure 9: Results for the AR(10) GARCH(1,1) model sampled with a t-distribution.

3.2 ARMA(4,5)-GARCH(1,1) Model

Thus so far, we made one big mistake. We took the for granted the thumb-rule: ACF: tails- off, PACF: cuts- off ⇒ AR, MA for the inverse as a dogmatic rule. However, we learned the hard way that this is only a heuristic rule and that the PACF in Figure 4b could also be interpreted as tailing off. But this would mean that we have to work with an ARMA instead of an AR model (again together with GARCH(1,1). That is what we did, we computed the corresponding log likelihoods and most importantly AIC's. We omit the elbow-method this time. The AIC's are given in table 1, note that all of them were computed however only the most important ones are shown[5].

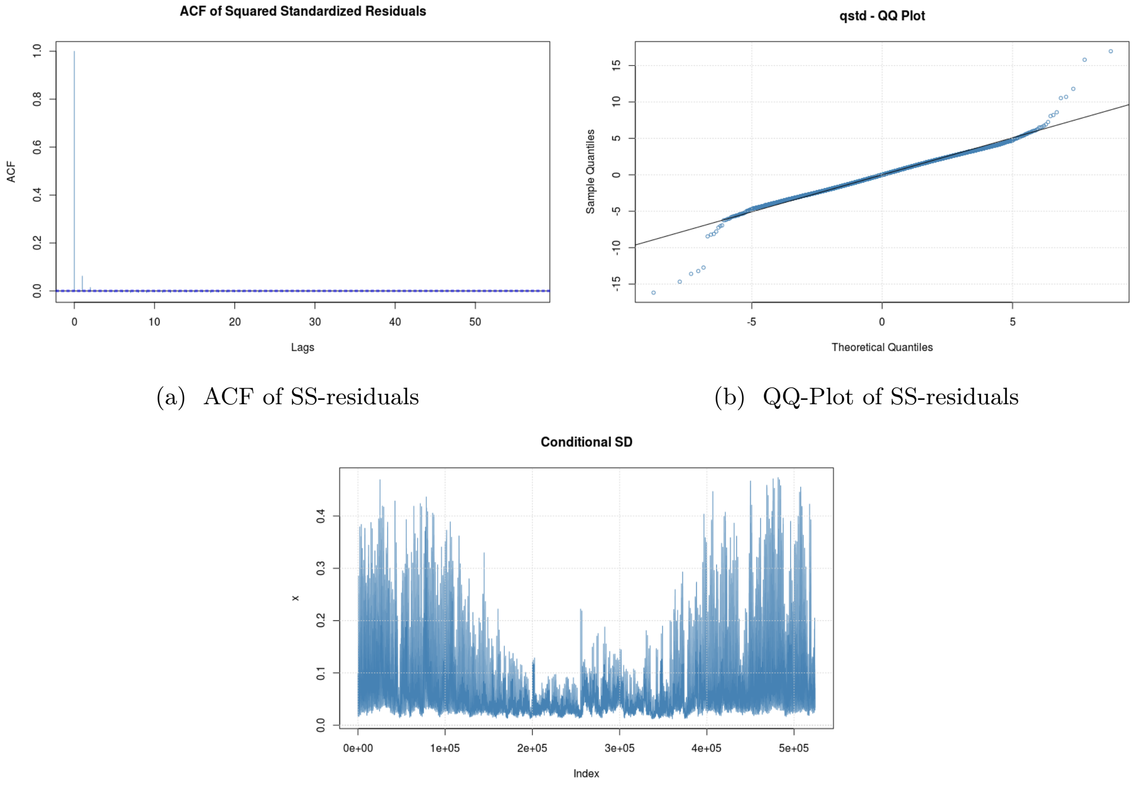
ARMA(x,y)GARCH(1,1)- model	AIC
1,1	-2.955951
1,2	-2.957138
...	...
3,5	-2.971601
4,5	**-2.971638**
5,5	-2.971635
...	...
4,4	-2.971465
...	...
10,0	-2.964169
13,0	-2.966696
16,0	-2.968186

Table 1: ARMA models with their corresponding AIC

From this, one can draw the conclusion that ARMA(4,5)GARCH(1,1) is the model which describes our data best. It balances the trade-off between degrees of freedom and likelihood and works like a soft-regularization. The corresponding results are shown in Figure 10.

One sees that indeed, the QQ plot is quit good. Also, the estimated standard deviation looks promising since it shows the same 'physical' variances as our data: low variances for low temperatures and vice versa. The ACF of residuals has very little dependence at lag 1, which we agree too was difficult to model after multiple strenuous attempts at remodeling, and seems to be the consequence of highly correlated data.

[5] Note that all their digits are shown since they are important!



(a) ACF of SS-residuals

(b) QQ-Plot of SS-residuals

(c) Estimated Standard Deviation

Figure 10: Results for the AR(4,5)-GARCH(1,1) model sampled with a t-distribution.

4 Forecasting

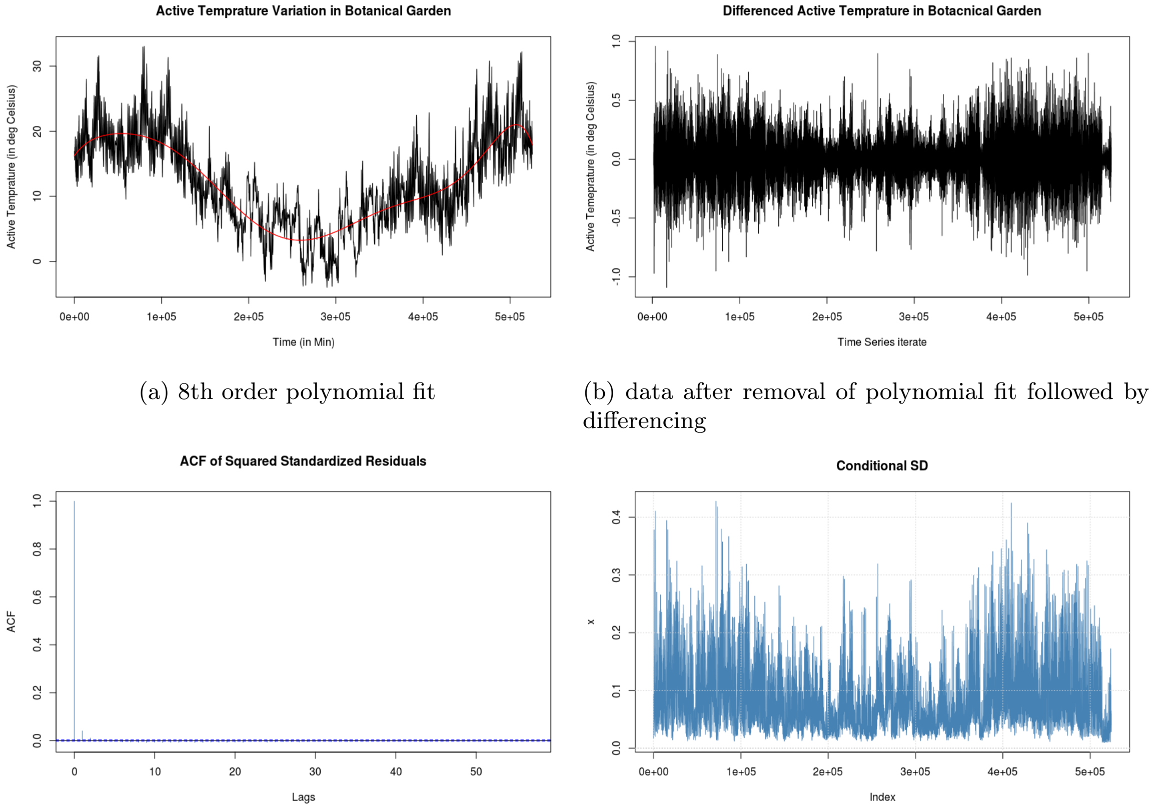
In time series analysis, forecasting is a very important and in some cases useful tool. In our analysis however, we argue that forecasting is pointless. The reason for this is the following. First we made our initial data stationary by:

- fitting a polynomial fit to remove the yearly seasonality
- removing daily seasonality by differencing
- differencing once more to make the data stationary

Differencing twice makes forecasting already poor, because it adds the uncertainty of the previous time-step while forecasting, that too recursively. Further when the data was stationary, we found that AR(10)GARCH(1,1) describes the data best, but this again means we added two models each with a number of parameters to describe our data. Above that, we are working with minute data and knowing that 'all' the uncertainty in forecasting comes from our noise which we 'add' some every minute, on a daily scale our prediction will be flooded with uncertainty, thus useless. Most importantly our data is limited to a single year which means in order to even predict we need to at-least replicate our data, which doesn't seem to be the right thing to do and hence we avoid it totally. Moreover, since our aim is to understand the differential heating between city and countryside we further proceed in that direction.

5 Outside the city

With the help of the cues we obtained we obtained above for inside the city modeling, we began modeling an outside location: The botanical garden in Ghent, using a similar chronology which we avoid mentioning here for brevity. The model we obtained was (again) ARMA(4,5)-GARCH(1,1) and the corresponding figures are shown in Figure 11.



(a) 8th order polynomial fit

(b) data after removal of polynomial fit followed by differencing

(c) ACF of squared standardized residuals

(d) conditional SD

Figure 11: Important figures concerning countryside location: Botanical Garden

One sees that the general picture with the previous location is the same except that variations (and the corresponding conditional SD) look different. One sees that temperature fluctuations in the botanical garden are smaller and show less difference between summer and winter in the botanical garden compared to the previous city location.[6]

6 Discussion

The work required for this project was divided in three large parts. The first part was making the data properly stationary without having to 'lose' the minute data by averaging out. The second was to try to find the best model for this data. In this part we made a mistake in the assumptions we made, which at the end led us to the final part; the correct model description.
On reflection we are glad that we sticked to minute data since it 'obliged' us to go through different aspects of time series which we think would not have been necessary if we had averaged out to days. However, this is just a feeling and knows no solid proof (except if we would **also** redo the complete analysis for averaged data).

Regarding the urban heat effect[3][4] we find ourselves in a rather 'inconclusive' spot. There are two ways in which we could 'compare' the temperature behavior between city and countryside [7]. One way is to look if there is a significant difference between the polynomial fit's for both locations. One find that their difference is not significant and one can thus not conclude that the 'average' temperature behavior modeled by a polynomial fit of order 8 between both locations is different. The second way in which we could look for difference is in the model which describes each locations data best. We found that they are best described by the same model. In conclusion one can thus say that no significant difference between both locations (city and countryside) was found subject

[6] Sint Bavo School

[7] i.e. these are the ways possible with the time series analysis techniques learned during this course[2]

to our time series analysis. It doesn't falsify the existence of such an effect, but rather points out at the fact that ARMA-GARCH style modeling is not sufficient for capturing it.

References

[1] Terence C Mills. Time series modelling of temperatures: an example from kefalonia. *Meteorological Applications*, 21(3):578–584, 2014.

[2] Emeric Thibaud. Time series analysis. *MATH-342*, 2017-18/II.

[3] M Santamouris. *Heat-island effect*, volume 402. James & James: London, 2001.

[4] Brian Stone Jr and Michael O Rodgers. Urban form and thermal efficiency: how the design of cities influences the urban heat island effect. *American Planning Association. Journal of the American Planning Association*, 67(2):186, 2001.