Radboud University

Faculty of Social Sciences

Spatio-temporal analysis and augmentation of consumer weather station data

Evaluation and enhancement of crowdsourced weather data quality

Lisa Tostrams

Committee:
Prof. dr. T. Heskes
Dr. L. Vuurpijl
T. de Ruijter, Msc.

First version

Nijmegen, January 2017

Abstract

Currently there are thousands of Consumer Weather Stations (CWS) uploading data in the Netherlands. This data could potentially be used in a range of applications. However, there are some challenges facing the CWS that need to be adressed before this data is usable. In this thesis, data is first preprocessed to discard obviously incorrect data. Radiation error is removed using Kalman filter. Station covariance is analysed with Principal Component Analysis. Spatial correlation is examined by Kriging to determine validity of local measurements. From each method, a quality measure is computed to evaluate individual station quality.

Contents

Chapter 1

Introduction

Consumer Weather Stations

Over the past few years, there has been a steady growth in the use of automated weather stations by meteorology hobbyists. Besides being able to monitor your local weather state, these stations can notify users of interesting measurements, predict local weather, and can even be integrated with other home automation systems. Many of the most common Consumer Weather Stations (CWS) have the possibility to upload the weather data to the internet to share and compare measurements with fellow meteorology enthusiasts. Some popular websites are the Netatmo site (www.netatmo.com), Wunderground (www.wunderground.com), WOW-UK (wow.metoffice.gov.uk) and WOW-NL (wow.knmi.nl).

There are currently thousands of CWS in The Netherlands regularly uploading weather data, compared to 31 official automated weather stations from the KNMI (www.knmi.nl/nederland-nu/weer/waarnemingen). In Amsterdam alone there are around 400 CWS produced by Netatmo (see fig. 1.1). All this new data could potentially be used in a range of applications, including high spatial resolution weather forecasting and modelling of the urban heat-island effect [Steeneveld et al., 2011]. Temperatures in urban areas can be expected to be higher than in rural areas, due to daytime heat storage and subsequent heat release after sunset. Steeneveld et al. used CWS data to asses the urban heat island effect in cities in the Netherlands.

Challenges of Consumer Weather Station data

Before CWS data can be used on a large scale there are some issues regarding the data that need to be addressed. There are many challenges facing the CWS, leading to a high uncertainty as to whether the data matches reality [Bell, 2014]. These challenges include calibration issues of individual sensors leading to bias or drift, design flaws, or sensor quality. Most stations do not have metadata available about the upkeep and location of the modules. Another often occurring problem is software and communication errors, leading to long gaps in measurements. In general, more data is available during the day than during the night, probably because some users switch of Wi-FI at night. The CWS may also be placed in a way that is not optimal for weather monitoring, in the full sun or in an area without ventilation, under umbrellas, or even inside the house. In their user manuals, Netatmo itself lists a number of reasons why measurements may be missing, including low battery power, inference in signal by walls or wind, the devices being located too close to each other, unpowered inside module, missing Wi-Fi signal, or network change detection.

[Meier et al., 2015] developed a 5-level quality measure for Netatmo stations, based on available station-specific metadata, data availablity and a comparison with UCON atmospheric data. Reasons for different levels are listed as Netatmo API and server limits, user-specific operating errors (especially in communication metadata), failure of wireless network, loss of battery power, user-specific installation error, device being set up indoors, and device not having been set up in the shade. Part of the problems with the stations is attributed to the fact that there are no standard guidelines on how to use the devices, leading to incorrect installment and use of sensors.

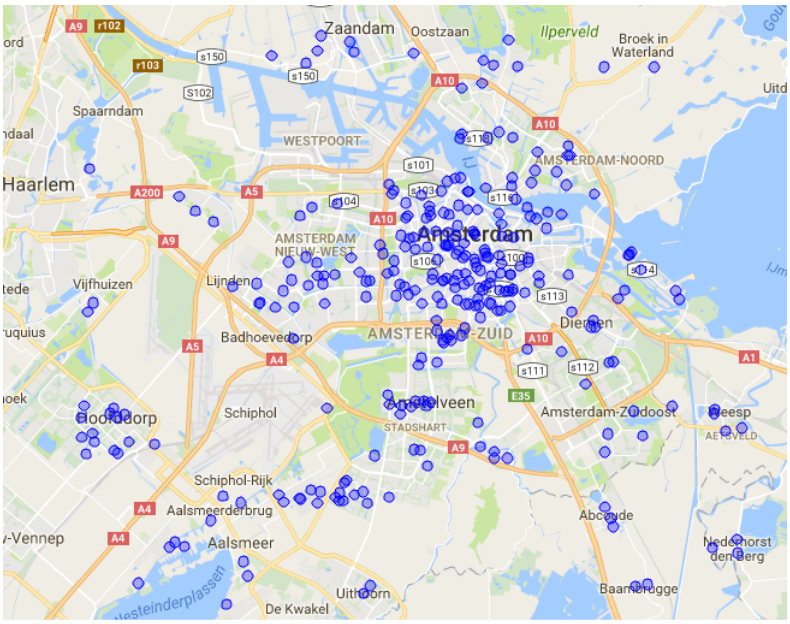

Figure 1.1: Locations of stations in Amsterdam, the Netherlands.

At the highest quality level, the daily number of available stations varied between 350 and 831, out of 1100 total stations.

Research question

These problems can make CWS data difficult to process. There is research employing meteorology methods to estimate the quality of crowd sourced weather data [Bell, 2014, Meier et al., 2015]. However, there are many other fields of study that deal with high uncertainty measurements, such as time series analysis, robotic motion planning and control, and trajectory optimization, among others. These fields of study have developed mathematical add statistical methods for the processing of possibly unreliable data. Which leads us to the main research question of this thesis: How can we apply existing spatial and temporal analysis to correct for measurement errors and bias in consumer weather station temperature data?

Chapter 2

Methods

2.1 Data

The data used for analysis came from CWS linked to the NetAtmo website. Over the months of April and November, 2016, the air temperature measurements in the area of Amsterdam were considered (see figure 2.1). The CWS data used here came from Netatmo weather stations [man, 2012]. Besides the outside module, the setup includes an inside device that transfers the data automatically to the Netatmo website. The popularity of these devices can be attributes to their cost effectiveness, ease of setup, and visualisation of the data via smartphone. The stations use AA batteries for power, and transfer their data over Wi-Fi. Since it is not known exactly how reliable the CWS data is, and the actual temperatures at the measurement locations are unknown, the measurements from the nearby KNMI station at Schiphol are used as a means of comparison.

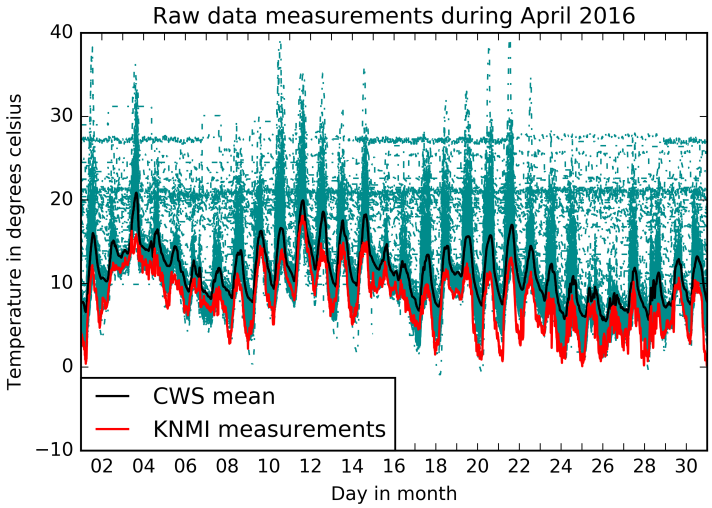

Figure 2.1: Temperature measurements of NetAtmo stations in Amsterdam, April 2016. Plot includes KNMI measurements at Schiphol for comparison.

2.2 Data Pre-processing

Before analysing the data, some pre-processing was done.

Resampling

The different stations offered measurements of varying intervals, ranging from a measurement every minute to a few measurements per day. This variation did not only occur between stations, but also within stations, as many stations were affected by gaps in measurements. These varying intervals and gaps make certain computations difficult, and comparisons impossible. The data was re-sampled at an interval of 10 minutes. New measurements we computed using the mean of surrounding measurements, the non-numerical values occurring in the data were padded with the nearest non-NaN value. Stations having over 75% of missing data were considered unreliable and therefore excluded.

Outlier removal

After re-sampling, the most unlikely measurements ($z - score > 3$) were removed from the data. Z-score of a measurement x_i at a certain time point t is computed as

$$Z = \frac{x_i - m_t(x)}{std_t(x)}$$

with $m_t(x)$ being the mean and $std_t(x)$ being the standard deviation of all measurements at time step t. The data is not normally distributed but has a right skewed distribution, so generally a z-score would not be considered an applicable quantifier for outliers. However, considering the nature of the data, a temperature measurement being more than 3 standard deviations away from the mean is very unlikely.

Per day, the correlation of each station with measurements from the KNMI were computed. The Pearson product-moment correlation coefficient r_s on day d of a station's measurements s_d with the KNMI measurements k_d is computed as

$$r_s = \frac{covariance(s_d, k_d)}{std(s_d) * std(k_d)}$$

with $std(s_d)$ and $std(k_d)$ being the standard deviations of the measurements on day d. The covariance of the two timeseries s_d and k_d, each including $n = 144$ measurements, is

$$\frac{1}{n}\sum_{i=1}^{n}(s_{d_i} - m(s_d))(k_{d_i} - m(k_d))$$

with $m(s_d)$ and $m(k_d)$ being the means of the measurements on day d.

Since a certain type of variation in the data is highly expected, namely the 'warm during the day, cold during the night' variation or diurnal cycle, a minimal correlation with high quality temperature measurements from the KNMI is likely. To determine a reasonable minimal correlation, measurements of multiple CWS were plotted against KNMI measurements. A visual comparison of 'good' and 'bad' stations (for example, see figure 2.6) determined that a minimal correlation coefficient of 0.3 resulted in the filtering of obviously incorrect data.

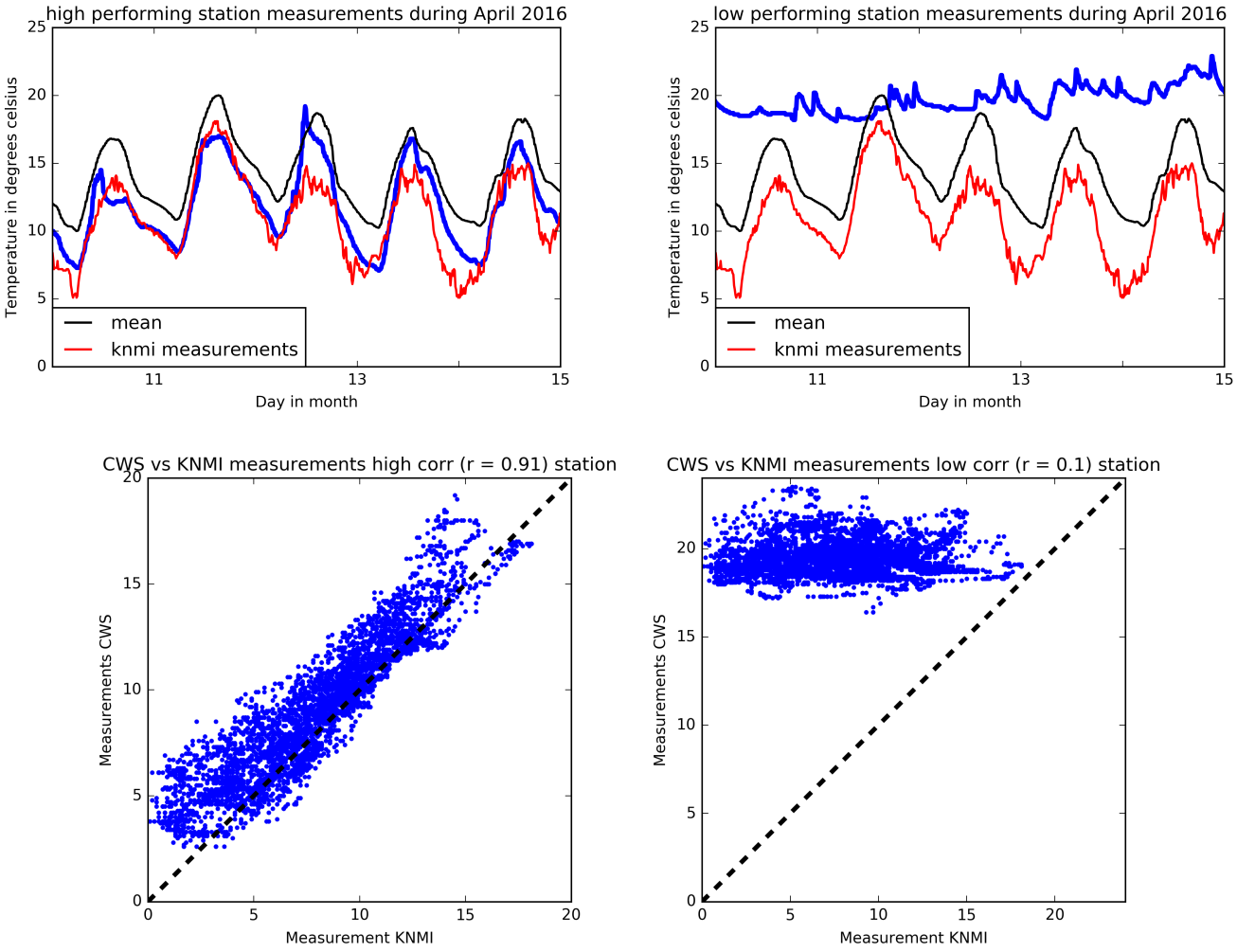

Figure 2.2: Visual comparison of a 'good' (high correlating) station vs a 'bad' (low correlating) station.

2.3 Temporal Data Processing

2.3.1 Individual station variance: Kalman filter

The Kalman filter is a set equations that provides a means to estimate the state of a process, in a way that minimizes the mean of the squared error [Welch and Bishop, 2006]. The Kalman filter addresses the general problem of trying the estimate the true state x_k of a time controlled system from a measurement z_k at time step k with

$$z_k = x_k + R$$

where the variable R represents the measurement noise. R may vary with time or may remain constant. In this application, R is estimated to vary during the day.

The *a priori* state estimate at step k is defined as $\hat{x}_k^-$, and the *a posteriori* state estimate is defined as $\hat{x}_k$, at time step k given measurement z_k. The goal of the Kalman filter is to compute the *a posteriori* state estimate $\hat{x}_k$ as a linear combination of an *a priori* estimate $\hat{x}_k^-$ and a weighted difference between an actual measurement z_k and measurement prediction related to the previous estimate (in this application, simply the previous estimate) $\hat{x}_k^-$, given by

$$\hat{x}_k = \hat{x}_k^- + K(z_k - \hat{x}_k^-)$$

The difference $(z_k - \hat{x}_k^-)$ is called the measurement *residual*. The residual reflects the difference between the predicted measurement and the actual measurement. The n by m matrix K is chosen to be the *gain factor* that minimizes the *a posteriori* error estimate P_k

$$P_k = \frac{1}{k}\sum_{i=1}^{k}(x_i - \hat{x}_i)$$

One from of K that minimized P_k is given by

$$K = \frac{P_k^-}{P_k^- + R}$$

When the measurement error R approaches zero, the gain K weights the residual more heavily. When the estimate error P_k approaches zero, the gain K weights the residual less heavily. After each time and measurement update, the process is repeated with the previous *a posteriori* estimates used to project or predict the new *a priori* estimates.

Determining the measurement variance R is done by comparing the measurements to the mean or the nearby KNMI station. Determining the process variance Q is more difficult because, in theory, the process that is estimated cannot directly be observed. A more detailed look into the setting of the parameters of the Kalman filter can be found in appendix A.

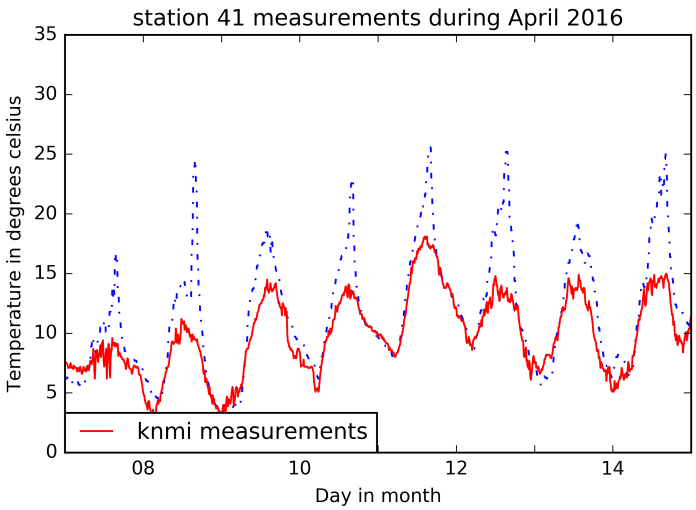

Figure 2.3: Temperature measurements of a single station. Plot includes KNMI measurements at Schiphol for comparison.

The Kalman filter relates the predicted state $\hat{x}_k$ to the previous state estimate $\hat{x}_k^-$ and the actual measurement $z(x)$. If the expected variance is low, say 0.1 degrees Celsius, but the difference between the previous estimate and the actual measurement is high, say 2.5 degrees Celsius, then the actual measurement becomes more unlikely. In this case, the current state estimate is more strongly related to the previous estimate than to the actual measurement.

CWS seem to measure much higher temperatures during the day than the KNMI station, as seen in figure 2.3. This station records temperature measurements that are likely during most of the night and mornings. In the afternoon the measurements become more unlikely, indicating

that the temperature at the station is heating up at a much higher rate than at the KNMI station. This might happen because air temperature measurements are susceptible to solar radiation error [Nakamura and Mahrt, 2005], where the temperature is overestimated due to overheating in the sun. The Kalman filter is expected to work especially well here, when the variation from measurement to measurement far exceeds the expected variation rate.

2.3.2 Global station covariance: Principal Component Analysis

Principal component analysis (PCA) is a statistical method that can be used to capture the variability of data in less attributes. PCA uses orthogonal transformation to represent the data in statistical uncorrelated dimensions. From these dimensions, the data can be reconstructed, losing a certain amount of variation in the process [Tan et al., 2013]. PCA tends to identify the strongest patterns in the data. Since patterns caused by unlikely measurements is hopefully weaker than patterns caused by unlikely measurements, reduction of dimensionality can eliminate some of the measurement noise. When the data is reconstructed, this noise is left behind.

The variability of an m by n data set D can be summarized in an m by m covariance matrix C, which has entries c_{ij} defined as:

$$c_{ij} = covariance(d_{*i}, d_{*j})$$

where d_{*i} and d_{*j} are the i^{th} and j^{th} attribute (column) of the data set D respectively. The covariance of any two vectors x and y with mean vales $m(x)$ and $m(y)$ and length n is defined as

$$covariance(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - m(x))(y_i - m(y))$$

On the diagonal of the covariance matrix you find the variance in each attribute, or dimension, of the data. The direction with the largest variance is the first Principal Component. The orthogonal direction with the second largest variance is the second Principal Component, and so on. These directions correspond with the eigenvectors V of the covariance matrix C, ordered by their corresponding eigenvalues $s_1, ..., s_n$. The relative size of the eigenvalues also corresponds to the variance explained by each Principal Component.

The data can be represented in these new dimensions by projecting the data onto the first i eigenvectors V_1 to V_i, resulting in m by i projection Z_i:

$$Z_i = [V_1, .., V_i] * [D - m(D)]$$

where vector $m(D)$ contains the means of the attributes of D. The percentage of variance explained p_i by this projection can be computed from the eigenvalues $s_1, .., s_i$:

$$p_i = \frac{1}{i^2}\sum_{j=1}^{i} s_j^2$$

. The data can be reconstructed from this projection to m by n reconstruction $\hat{D}$:

$$\hat{D} = Z_i * [V_1, .., V_i]^T + m(D)$$

with $[V_1, .., V_i]^T$ being the transpose of $[V_1, ..., V_i]$. In this reconstruction, p_i percent of variance is captured.

2.4 Spatial Data Processing

2.4.1 Spatial variance: Kriging

Kriging is a method of interpolation, originally stemming from geostatistics to predict spatially correlated ore grades in gold mines. The goal of Kriging is to provide best linear unbiased predictions for the value of a function at a given point by computing a weighted average of known

values in the neighborhood [Oliver and Webster, 2014]. The estimate is based on assumptions of spatial covariance, for which a function is computed beforehand.

When the same degree of variation can be expected from place to place, and the mean is constant, the spatial covariance as a function of distance h, called the variogram, is given by

$$C(h) = \frac{1}{n}\sum_{i=1}^{n}(Z(x_i) * Z(x_i + h) - \mu^2)$$

Here, function $Z(x)$ is the value of function Z at two dimensional location $x = x1, x2$, and μ is the mean of the process. When the mean is not constant, the covariance is replaced by half the variance of the differences, the semivariance:

$$\gamma(h) = \frac{1}{2n}\sum_{i=1}^{n}(Z(x_i) + Z(x + h))^2$$

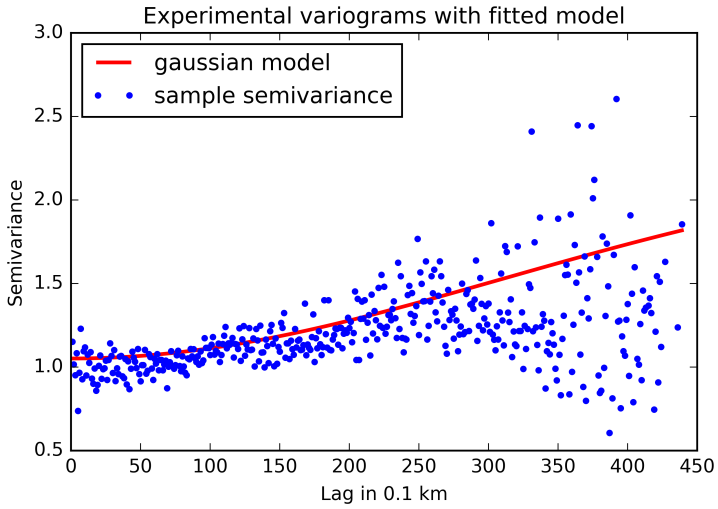

Figure 2.4: The average experimental variogram over the month April, with the Gaussian model fitted.

The experimental variogram is the variogram calculated from the data (see figure 2.4).

$$\hat{\gamma}(h) = \frac{1}{2m(h)}\sum_{j=1}^{m(h)}(z(x_j) - z(x_j + h))^2$$

Here $m(h)$ is the number of observations that are at distance h from each other (see figure 2.5 for histogram of lags). The next step is to fit a smooth curve on the experimental variogram. Popular models are the exponential, spherical, and Gaussian models. From these three, the Gaussian had the best fit (See appendix B for other models).

The Gaussian model is defined as

$$\gamma(h) = c_0 + c\left(1 - exp\left(-\frac{h^2}{a^2}\right)\right)$$

The nugget variance c_0 depicts the uncorrelated variance, c is the correlated component of the variation. See figure 2.4 for the fitted Gaussian model on the averaged experimental variogram.

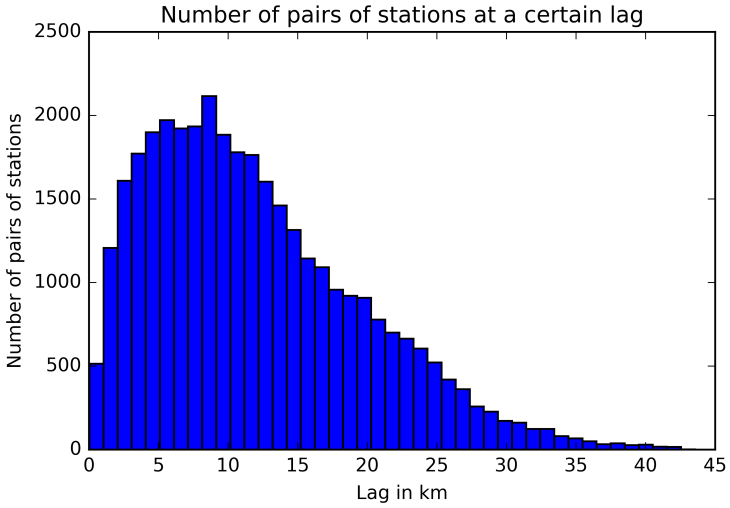

Figure 2.5: Histogram of lag between pairs of stations. For each lag h, the number of pairs of stations at that lag is m(h) in the experimental variogram.

Spatial estimation of the value Z at a location x_0, is calculated as a linear combination of the n closest observed values $z_i = Z(x_i)$ and corresponding weights for x_0 $w(x_0) = [w_1..w_n]$:

$$\hat{Z}(X_0) = \sum_{i=1}^{n} w_i * z_i$$

The weights for x_0 are calculated as follows

$$\begin{pmatrix} w_1 \\ \vdots \\ w_n \end{pmatrix} = \begin{pmatrix} \gamma(x_1 - x_1) & \cdots & \gamma(x_1 - x_n) \\ \vdots & \ddots & \vdots \\ \gamma(x_n - x_1) & \cdots & \gamma(x_n - x_n) \end{pmatrix}^{-1} * \begin{pmatrix} \gamma(x_1 - x_0) \\ \vdots \\ \gamma(x_n - x_0) \end{pmatrix}$$

The Kriging error variation is given by

$$\sigma^2(x_0) = w(x_0) * (\hat{Z}(x_0) - Z(x_i))$$

Since the Kriging prediction is the mean of a normal distributed range of predictions, the Kriging error variation provides the possibility to calculate a 99% confidence interval for the predicted value:

$$conf_{99} = (\hat{Z}(x_0) - 2.58 * \sigma(x_0), \hat{Z}(x_0) + 2.58 * \sigma(x_0))$$

The number of closest observed values n to consider in these computations was determined by comparing the runtime with the mean squared error MSE,

$$MSE = \forall x_0 \sum_{i=1} (\hat{Z}(x_0) - Z(x_0))^2$$

for $n = 1, .., 99$. Both $n = 10$ and $n = 38$ had the lowest MSE for all n with runtime < 50 seconds. $n = 10$ was used for all computations.

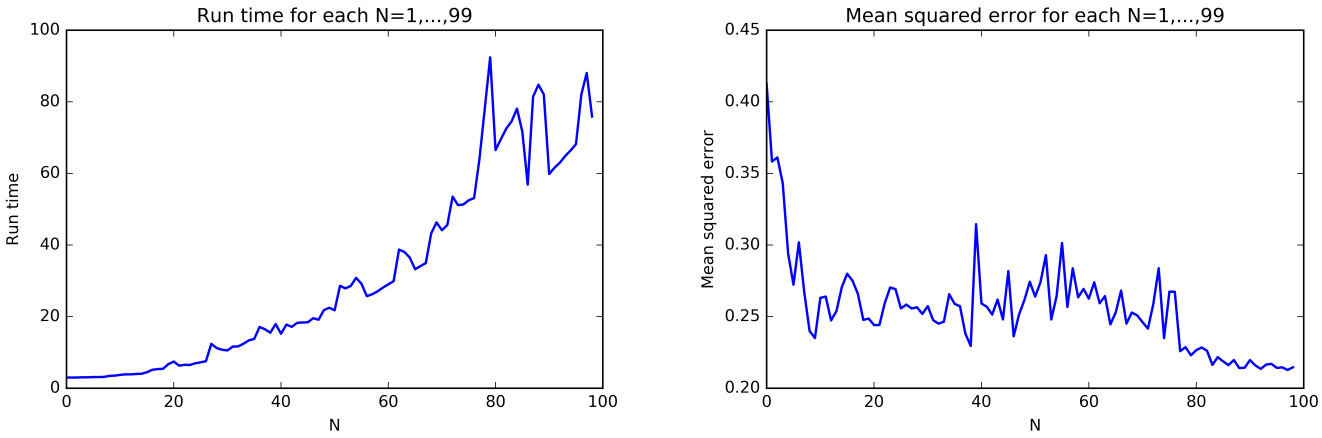

Figure 2.6: Runtime and mean squared error for each number of considered closest observed values.

Chapter 3

Results

From each applied method, a quality evaluation was calculated to estimate how reliable the measurements from individual stations are. By plotting the 10 best and worst performing stations according to each quality evaluation, it is possible to see exactly what each method contributes to the overall analysis. The original measurements are shown in figures 3.1 and 3.2. The dataset started with 354 stations in April and 419 stations in November.

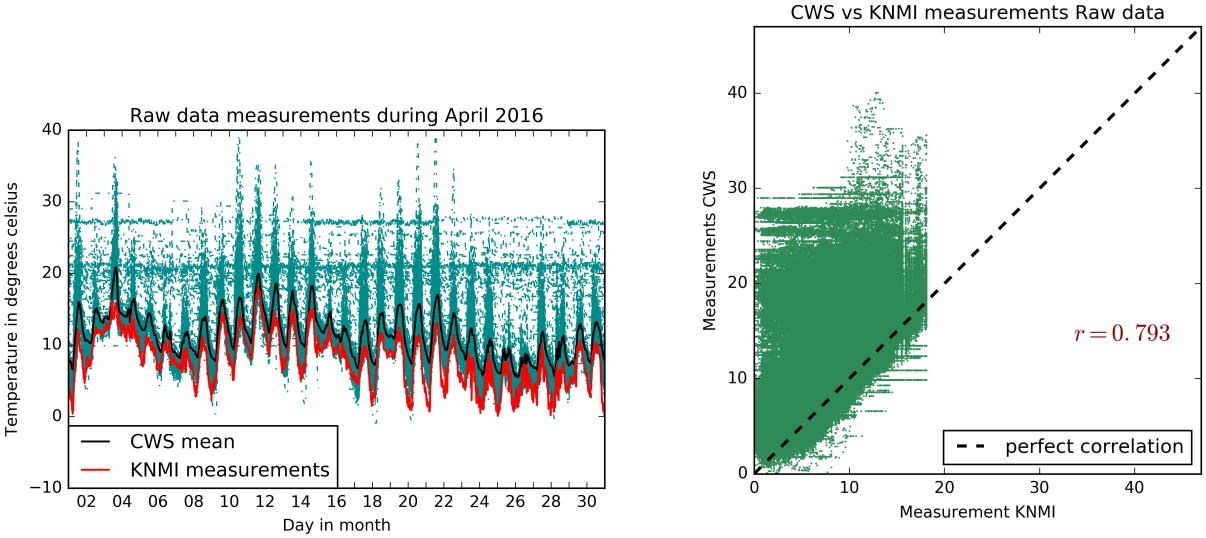

Figure 3.1: Unprocessed temperature measurements during April

The unprocessed CWS measurements compared to the KNMI measurements show three major issues:

- A few stations do not show the diurnal cycle
- Some CWS measure extremely high temperatures during the day
- Some station measurements do not correlate with KNMI measurements

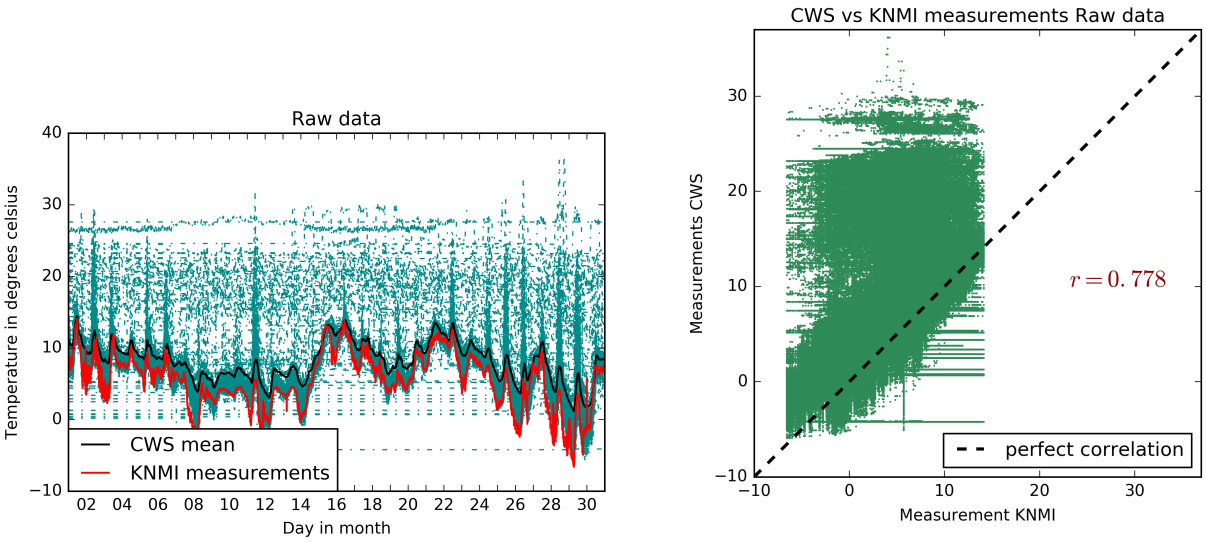

Figure 3.2: Unprocessed temperature measurements during November

3.1 Pre-processing

After removing stations that missed 75% of data, 297 and 330 stations were left to be resampled in April and November. For all of these stations, the daily correlation coefficient with the KNMI station was computed. By removing the measurements during days with a correlation coefficient lower than 0.3, most of the obviously non-cyclic data is removed (see figure 3.4 and **??**). For 47 and 70 stations in April and November, all days were marked as not correlating, and thus these stations were disregarded completely. The measurements that are marked as outlier by their z-score are considered unreliable and thrown out as well. This results in a much higher average correlation with the KNMI measurements.

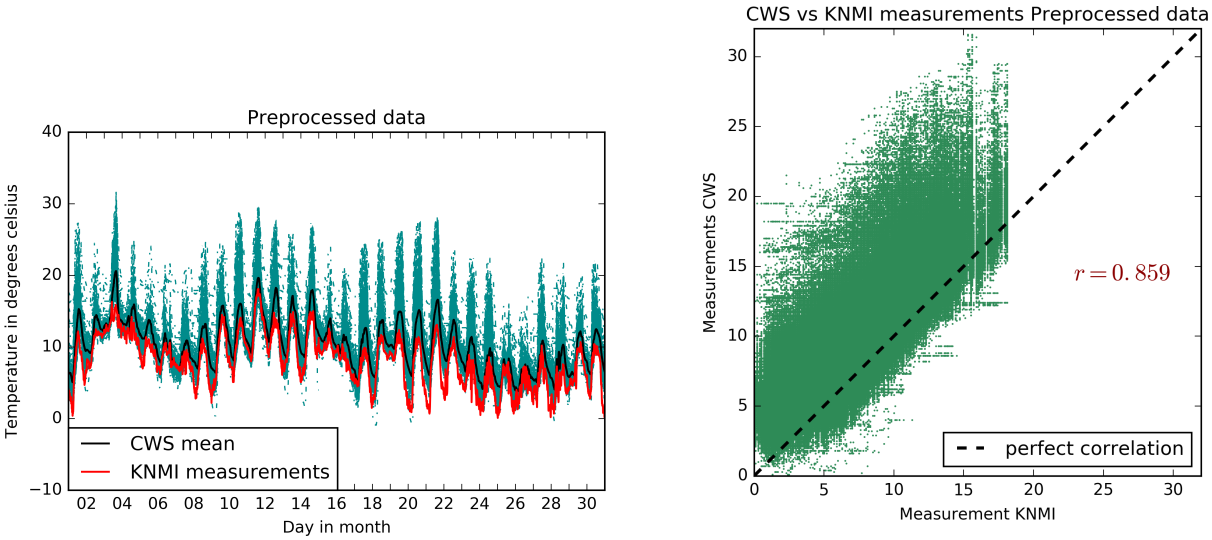

Figure 3.3: Preprocessed temperature measurements during April

Most stations correlate highly with the KNMI station, with a median coefficient of 0.91 and

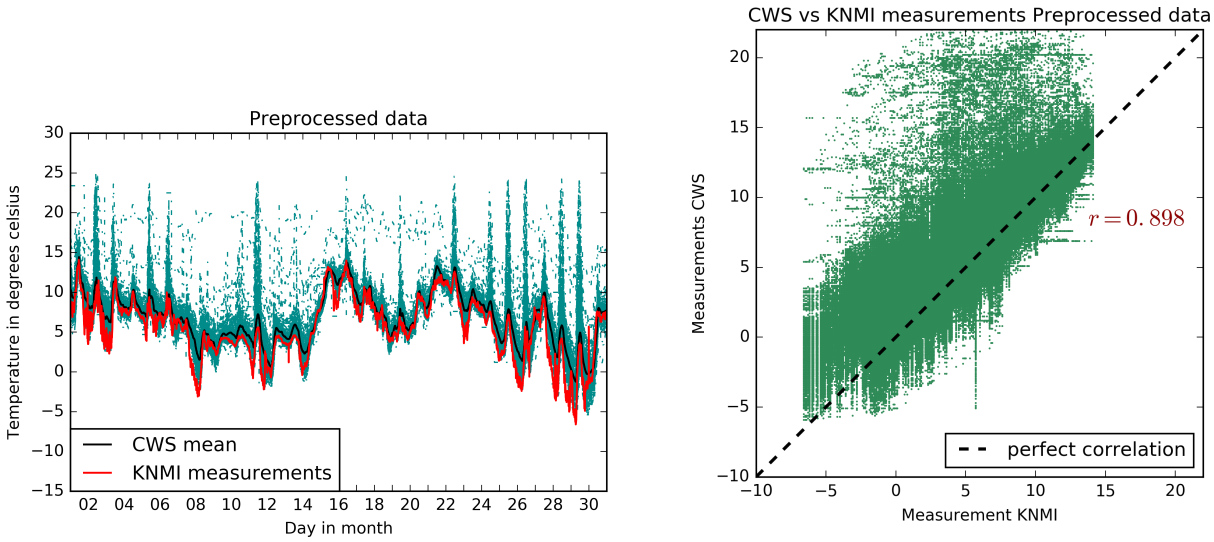

Figure 3.4: Preprocessed temperature measurements during November

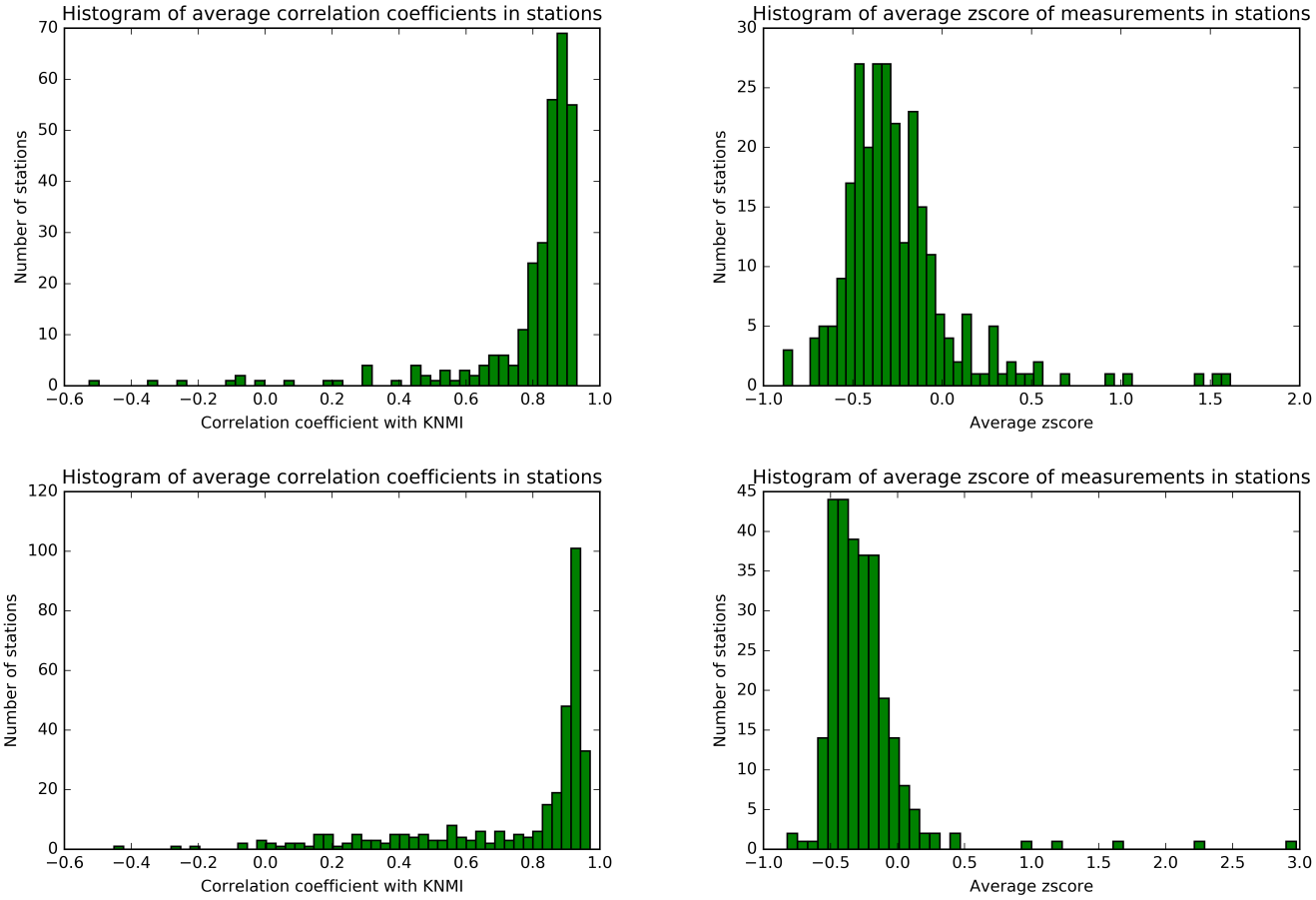

Figure 3.5: Quality evaluation scores of stations during April (top) and November (bottom)

0.93 in April and November respectively. The stations with a coefficient of around 0 do not vary with the KNMI at all. The negative coefficients might be caused by having lagged data uploads, reporting measurements from hours ago as if recorded now.

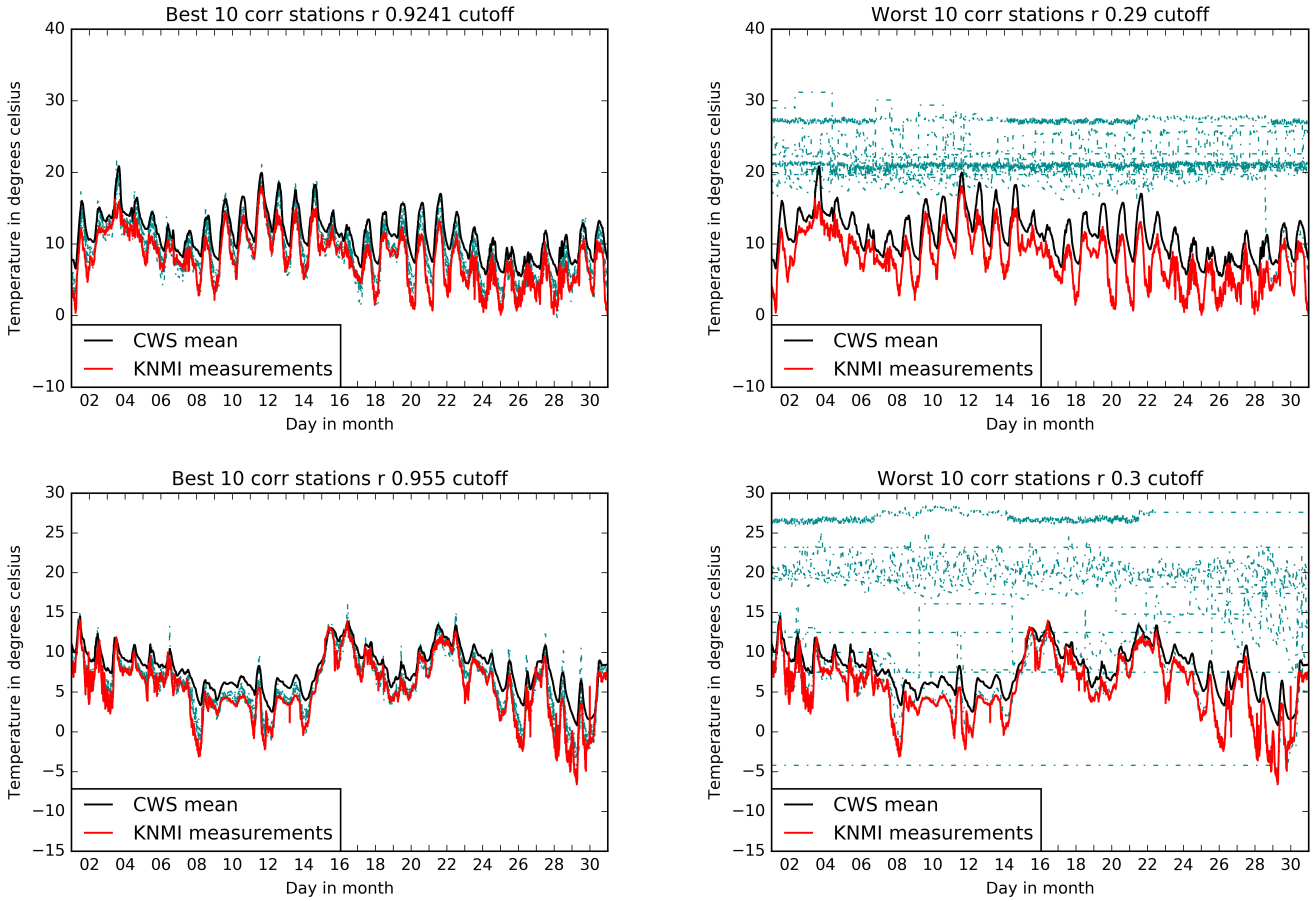

Figure 3.6: Best and worst performing stations during April (top) and November (bottom)

Figure 3.6 shows the best and worst performing stations according to their correlation with the KNMI station. This gives an indication of which stations were removed from the dataset during preprocessing, and the type of station that remained.

3.2 Temporal processing

3.2.1 Kalman filter

The effect of the Kalman filter on a station that records rapidly increasing temperatures can be seen in figure 3.7. In this example, the filter handles the high temperature increases well. When applied to all stations, the Kalman filter results in a decrease in overestimation of temperature during the day.

The slight decrease in correlation (see figures 3.8 and 3.9) might be caused by the fact that the Kalman filter mostly influences the measurements when the temperatures are high. This could potentially de-linearize the relationship with the KNMI station slightly.

The difference between the Kalman filter estimate and the original measurement is expressed as the mean squared error (MSE). A high MSE indicates that the Kalman filter generally applied a high correction to the measurements. In figure 3.10 the MSE of all stations is plotted in a

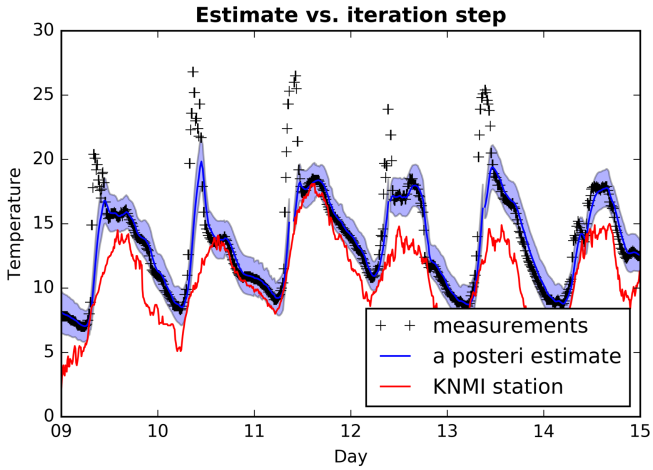

Figure 3.7: Kalman filter effect on station measurements. The blue area represents the 95% confidence interval of the Kalman filter estimate (blue line).

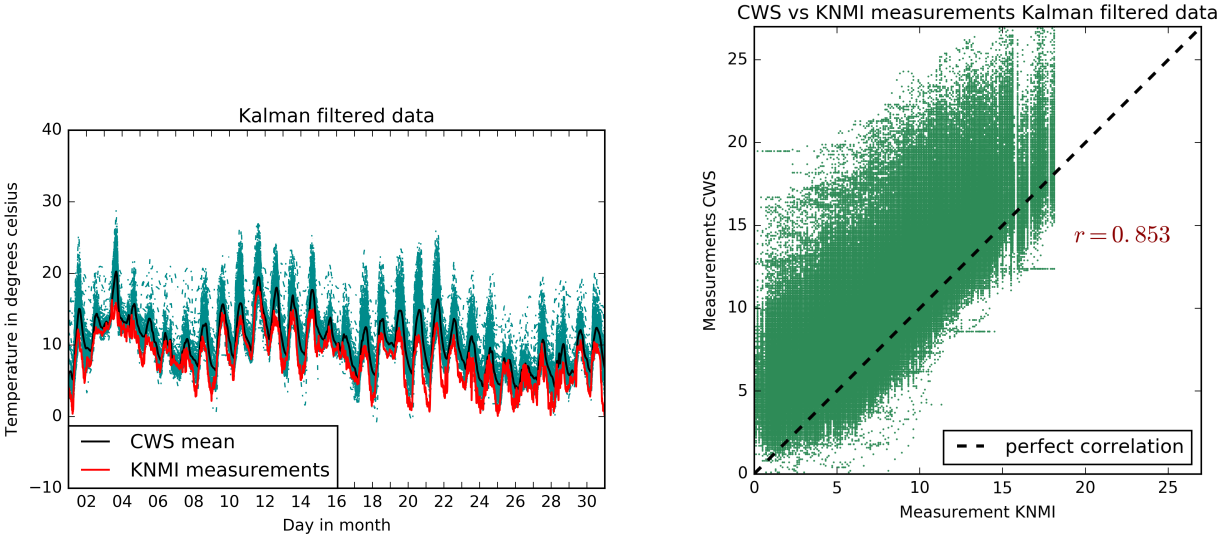

Figure 3.8: Kalman filtered temperature measurements during April

histogram. The MSE scored in November are generally much lower than the MSE scores in April, indicating that either the Kalman filter was dealing with a much higher uncertainty in estimating the measurement error, or that the measurements in November were in less need of correction by the filter than the measurements in April.

Plotting the best and worst performing stations according to the MSE (figure 3.11) shows what type of measurements are corrected most by the filter. As expected, these are measurements from stations that record rapid temperature increasements.

3.2.2 Principal component analysis

During PCA the measurements were projected onto the principal components that explained 99.99% of the variance in the complete dataset. The remaining unexplained variance was computed per station as a measure of how much the measurements diverged from the most important patterns

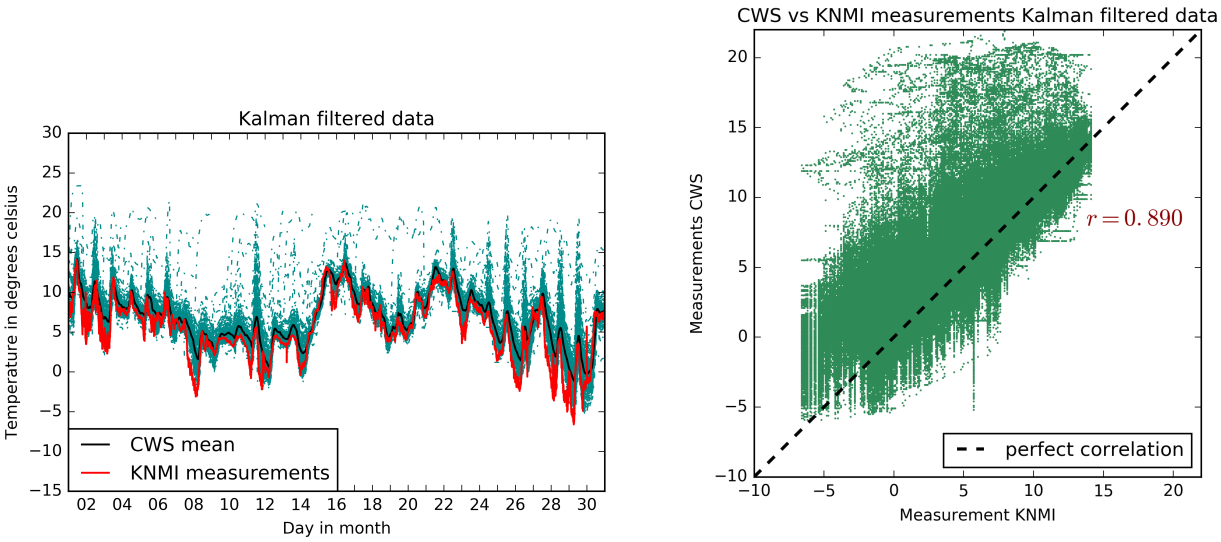

Figure 3.9: Kalman filtered temperature measurements during November

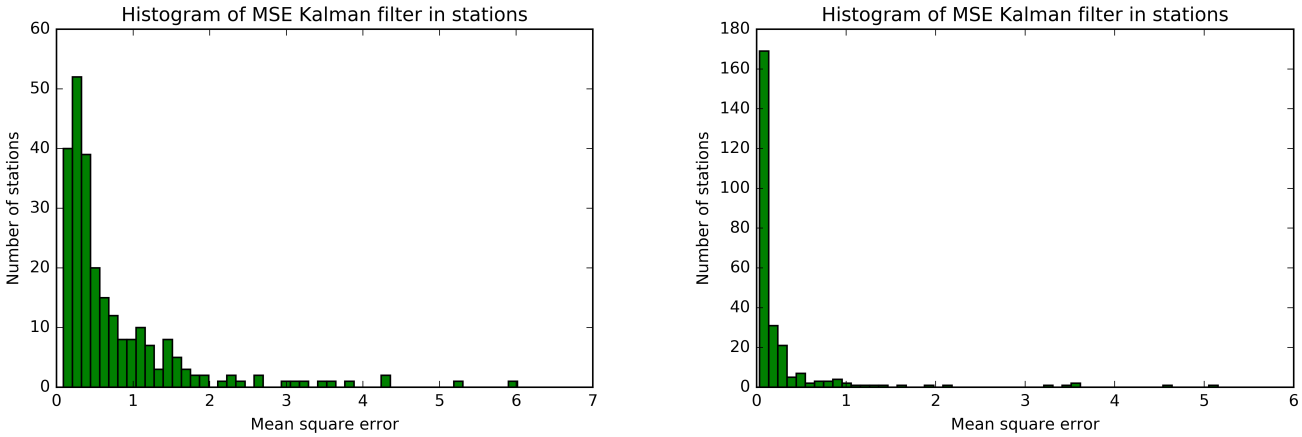

Figure 3.10: Quality evaluation scores of stations based on mean squared error April data (left) and November data (right)

in the data (see figure 3.14).

Projecting the data onto the first few principal components did not modify the measurements considerably, but computing the unexplained variance per station did identify some seemingly 'noisy' stations, as can be seen in figure 3.15. Contrary to the correlation coefficient with the KNMI, this quality evaluation score does not necessarily identify stations that do not behave like the KNMI station behaves, but it identifies stations that do not behave like the stations in the rest of the city does. This is shown in figure 3.15 as well, as the best performing stations do tend to overestimate the temperature compared to the KNMI, but this slight overestimation is apparently a pattern that can be validated by most stations.

Using the unexplained variance as a quality measure, the 'worst' 15% of the dataset was removed, resulting in the remaining 212 and 221 station measurements depicted in figures 3.16 and 3.17.

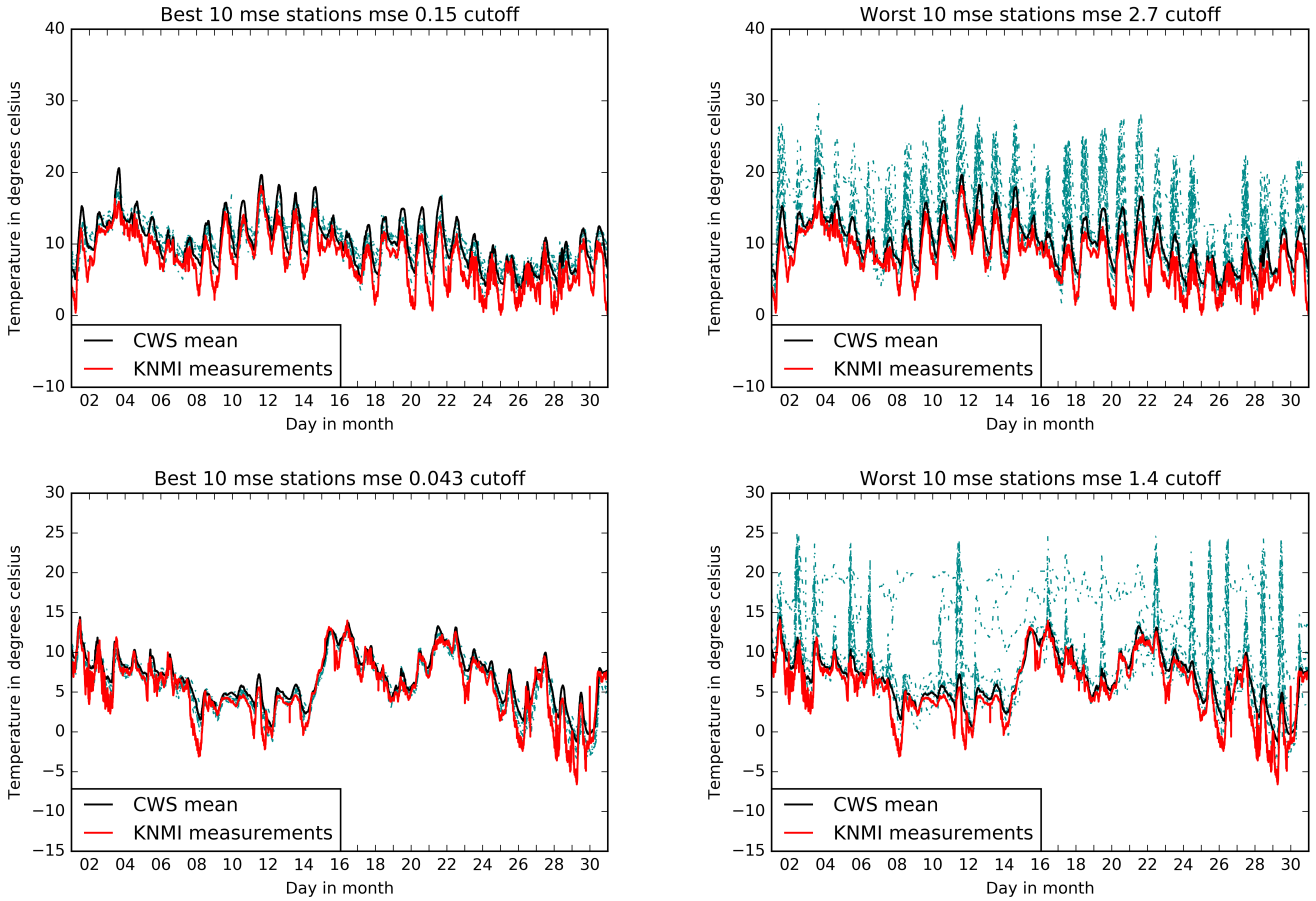

Figure 3.11: Best and worst performing stations in April (top) and November (bottom)

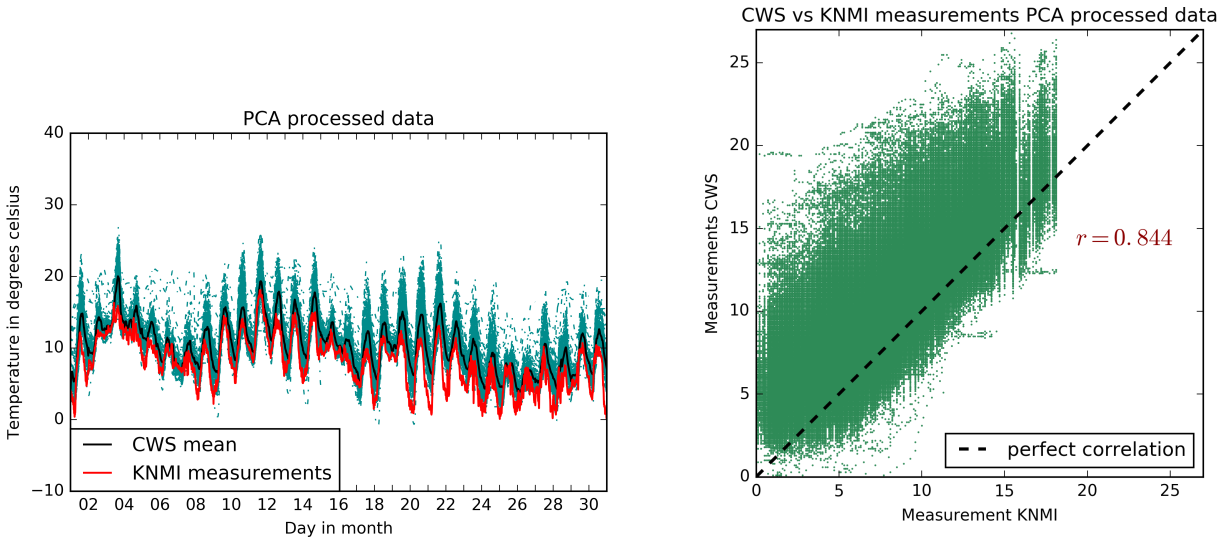

Figure 3.12: PCA processed temperature measurements during April

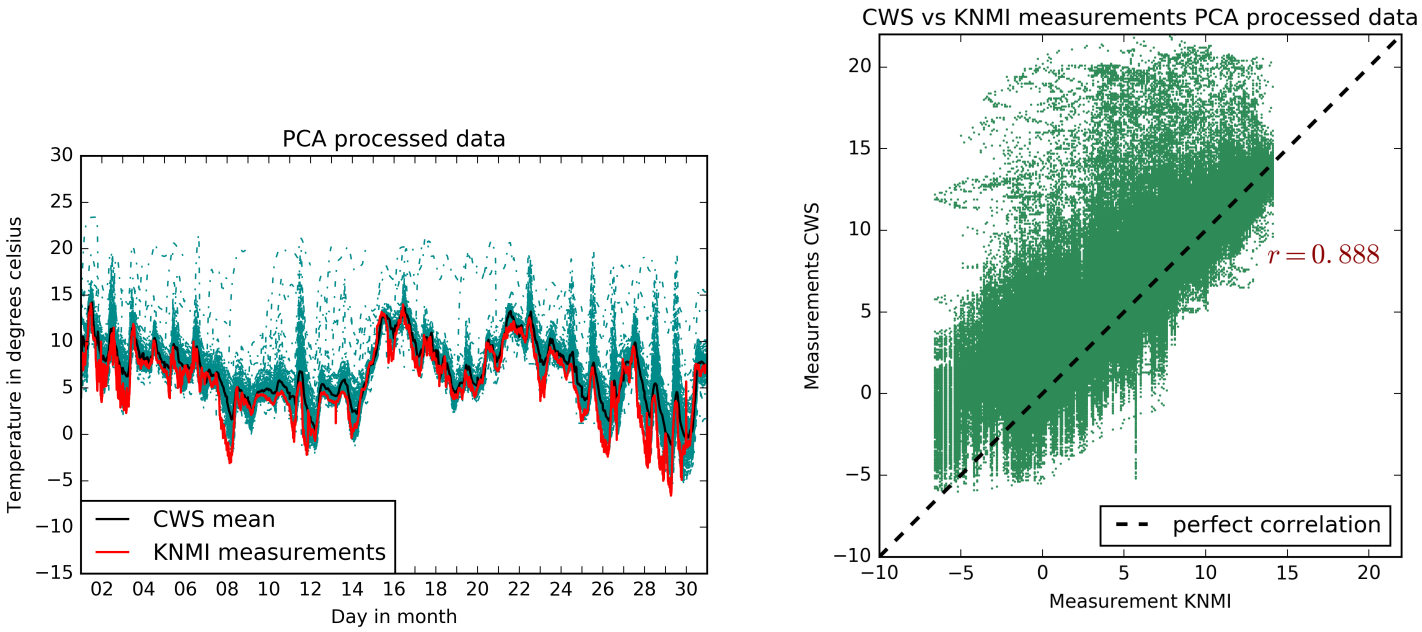

Figure 3.13: PCA processed temperature measurements during April

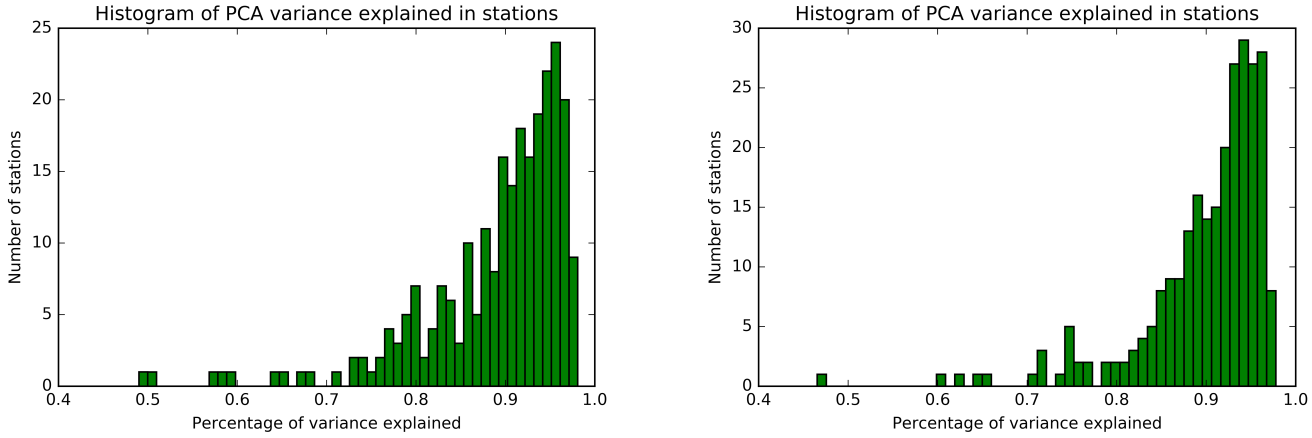

Figure 3.14: Quality evaluation scores of stations based on percentage of variance explained by PCA in April (left) and November (right)

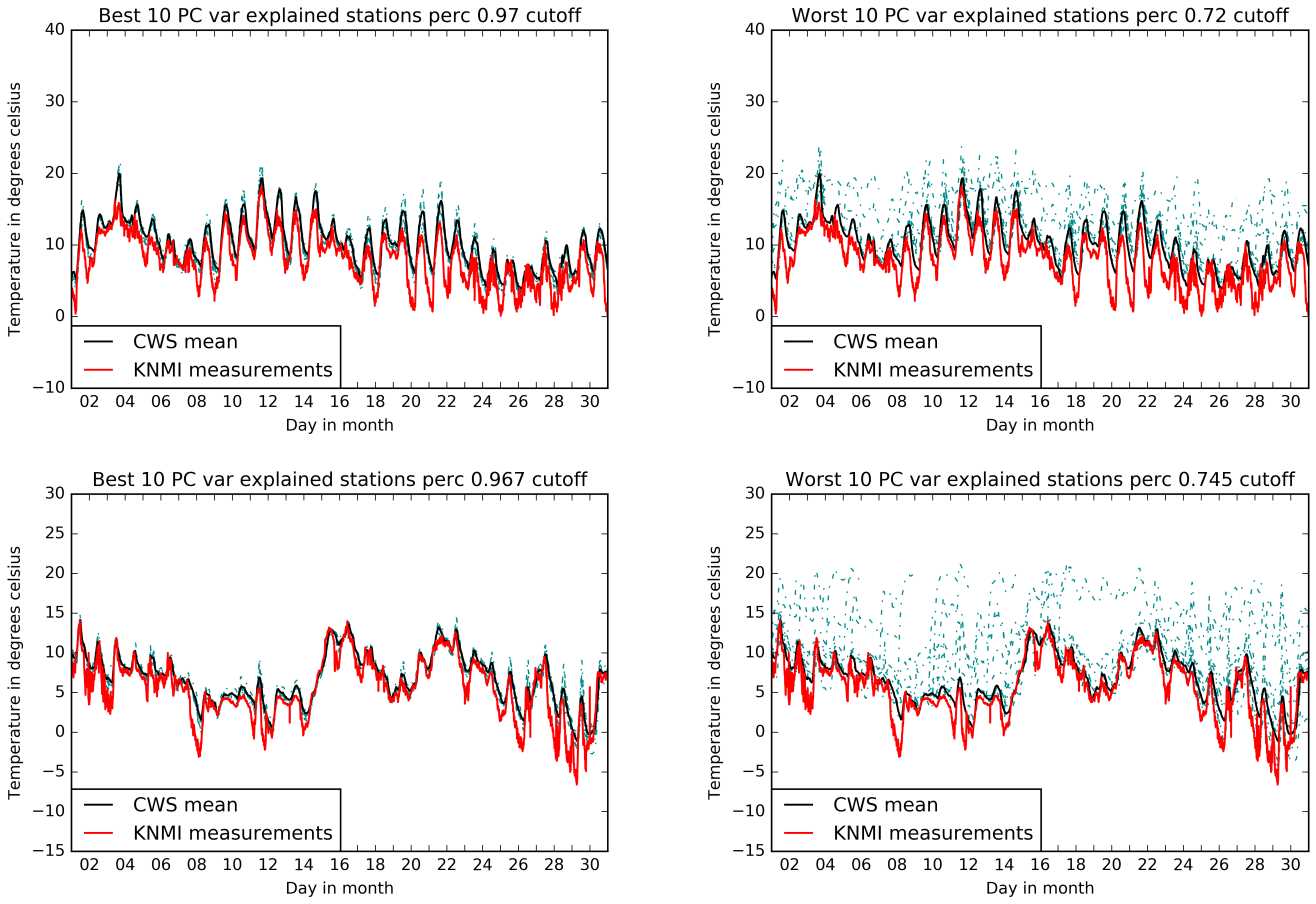

Figure 3.15: Best and worst performing stations

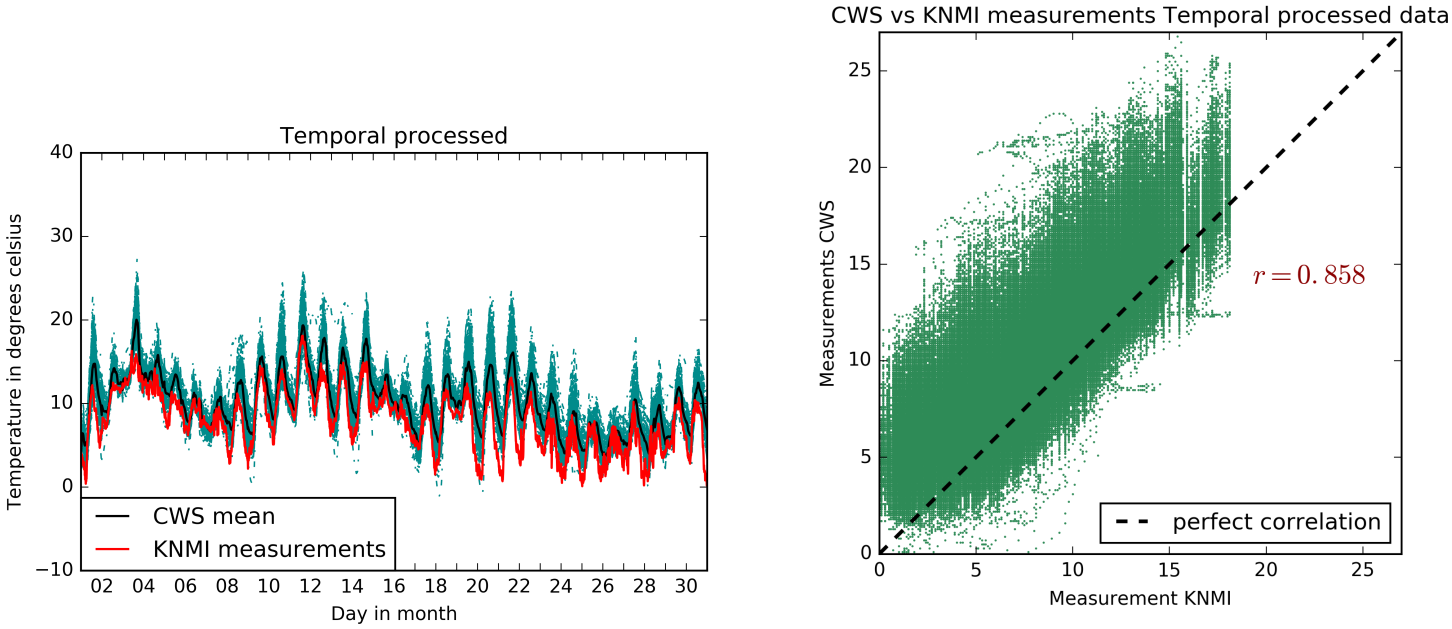

Figure 3.16: Temporally processed data without worst 15% performing stations, according to PCA explained variance (April)

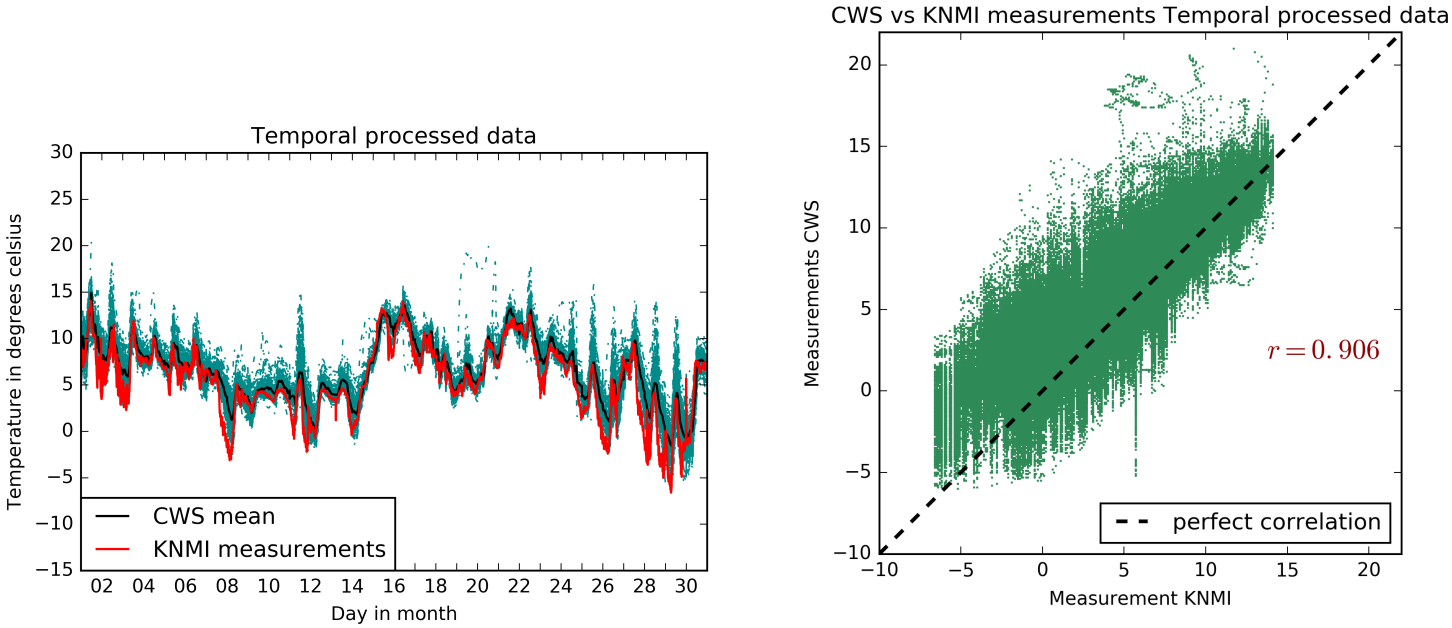

Figure 3.17: Temporally processed data without worst 15% performing stations, according to PCA explained variance (November)

3.3 Spatial processing

3.3.1 Kriging

Kriging was done after temporal processing was finished, since spatial correlation between stations was non-existent or even negative before outlier removal and temporal error corrections (see Appendix C). Using the Kriging error variation, a confidence score can be computed for each actual measurement, compared to the Kriging estimate at that location. A confidence score of below 2.58 indicates that the actual measurement falls within the 99% confidence interval of the prediction. To allow the possibility for the urban climate to cause strong spatial variations without those variations being disregarded, a measurement was not marked an outlier until a confidence score of over 6. Outlier measurements were replaced by the Kriging estimate for their locations.

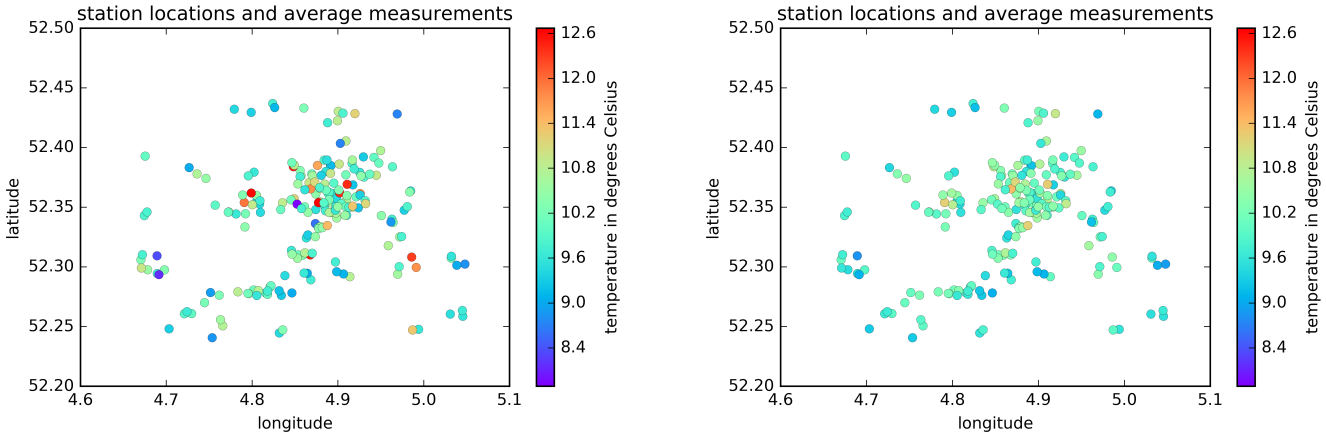

Figure 3.18: Locations and average measurements during April of stations before and after Kriging correction

Figures 3.18 and 3.19 show the locations and average measurements of stations before and after Kriging. The model of spatial covariance seemingly allows for warmer temperature measurements towards the center of the city.

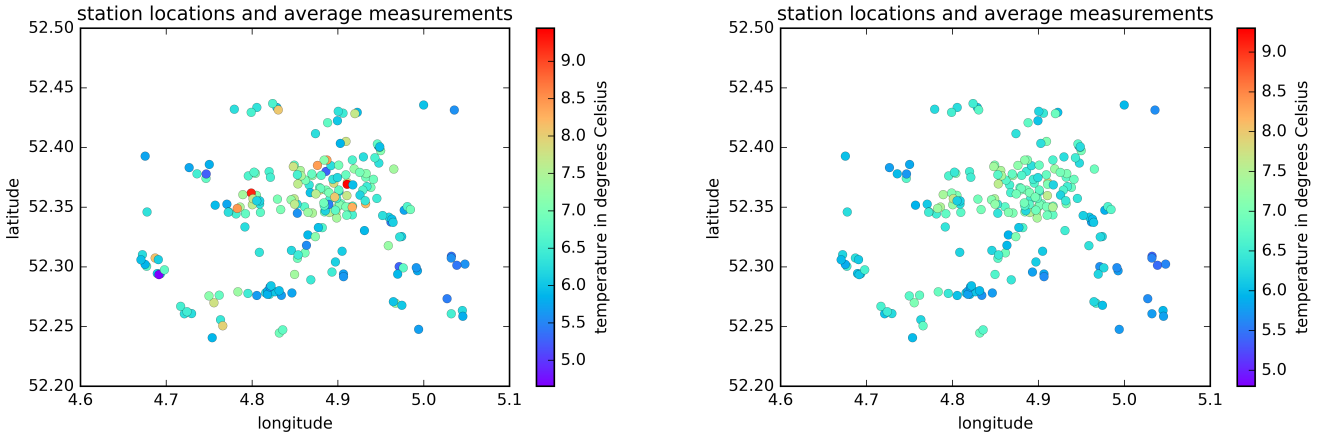

Figure 3.19: Locations and average measurements during November of stations before and after Kriging correction

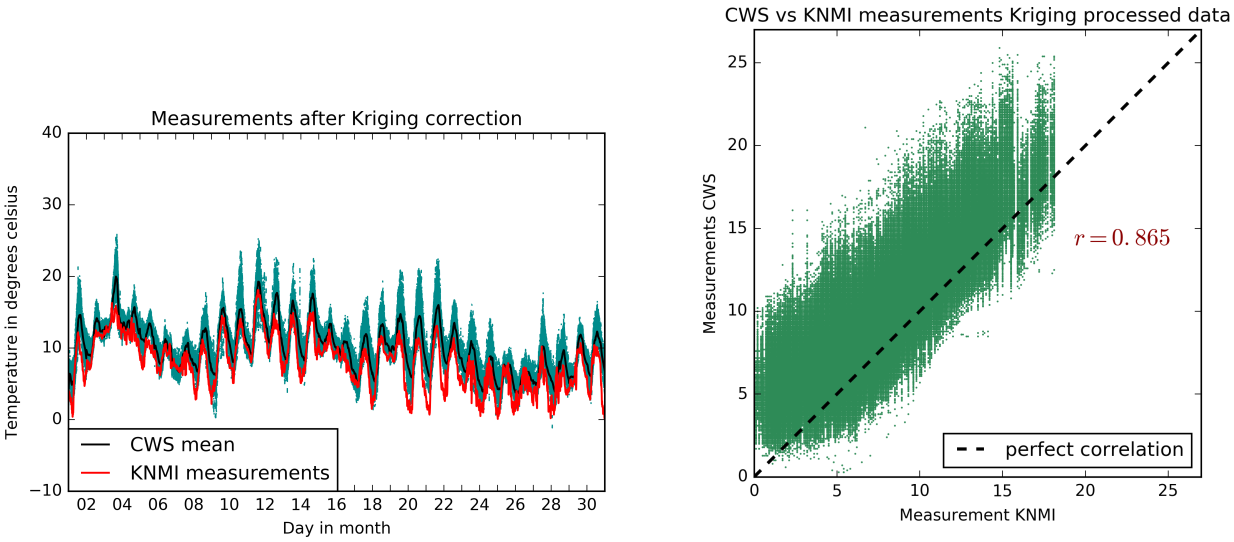

Figure 3.20: Measurements after Kriging correction

Contrary to earlier processing, Kriging identifies measurements that are unlikely based on nearby measurements at the time. The resulting measurements are plotted in figures 3.20 and 3.21.

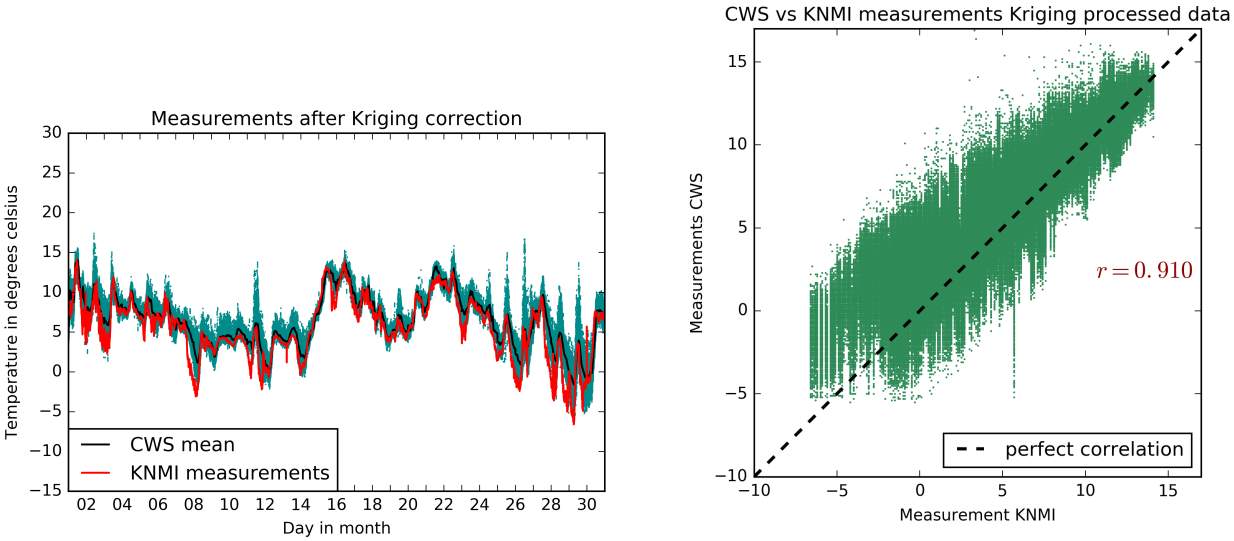

Figure 3.21: Measurements after Kriging correction

For each station, the difference between measurements and estimate is expressed in the mean squared error. Again, the measurements during November require fewer corrections than the measurements during April, resulting in lower average MSE scores (figure 3.22). The average MSE of all stations during the day changes between April and November as well, with a generally lower MSE score, a smaller difference between MSE during the day and during the night, and a shift from maximum MSE occurring around 15:00 to occurring around 12:30. It can be concluded that even spatial outliers might be caused by some stations being in the sun while others are

not, especially since the higher MSE scores are seemingly strongly influenced by the longer days in April (higher MSE between 07:00 and 19:00) and the shorter days in November (higher MSE between 09:00 and 17:00).

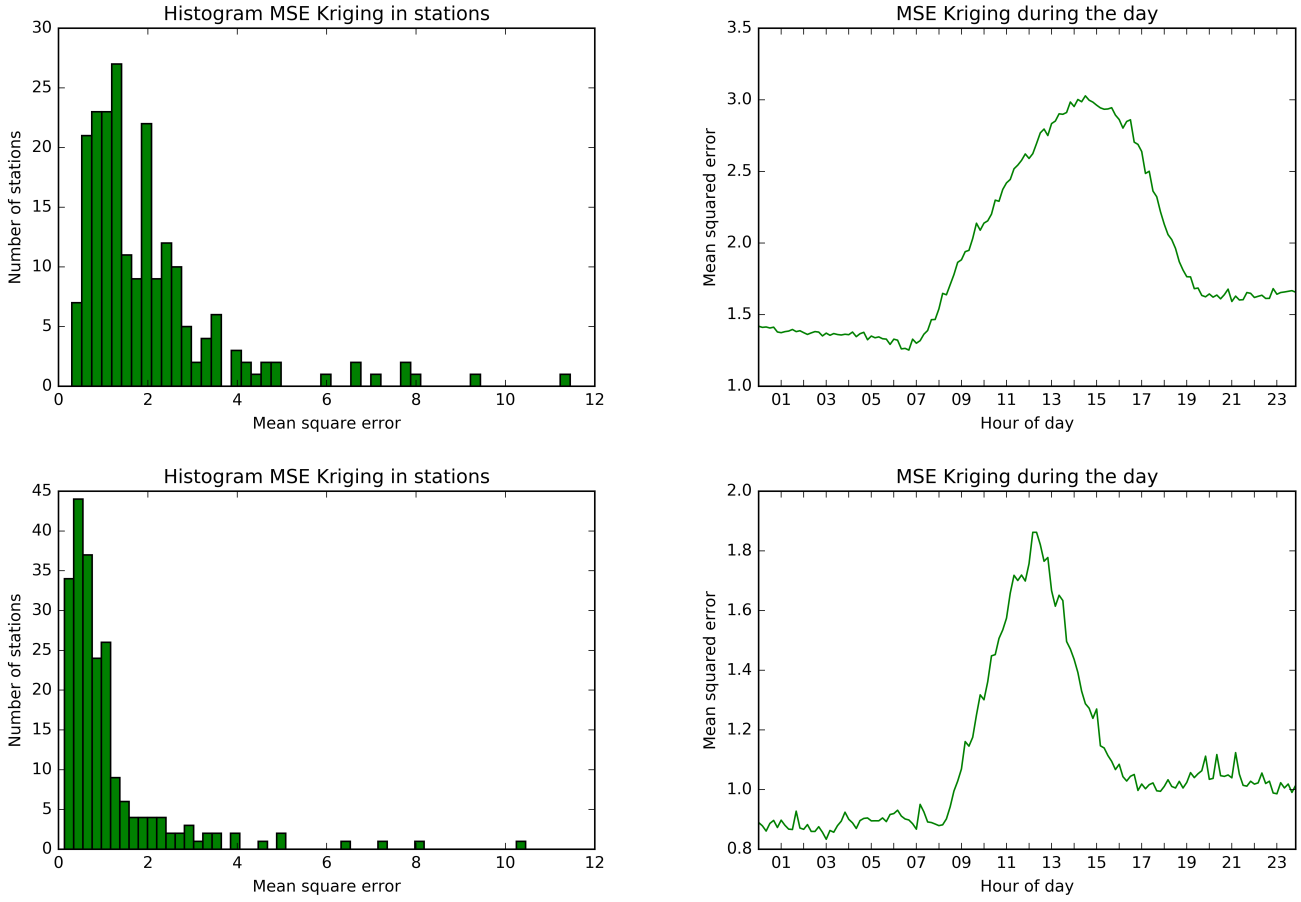

Figure 3.22: Quality evaluation scores of stations based on Kriging mean squared error, and average mean squared error during the day, for April (top) and November (bottom)

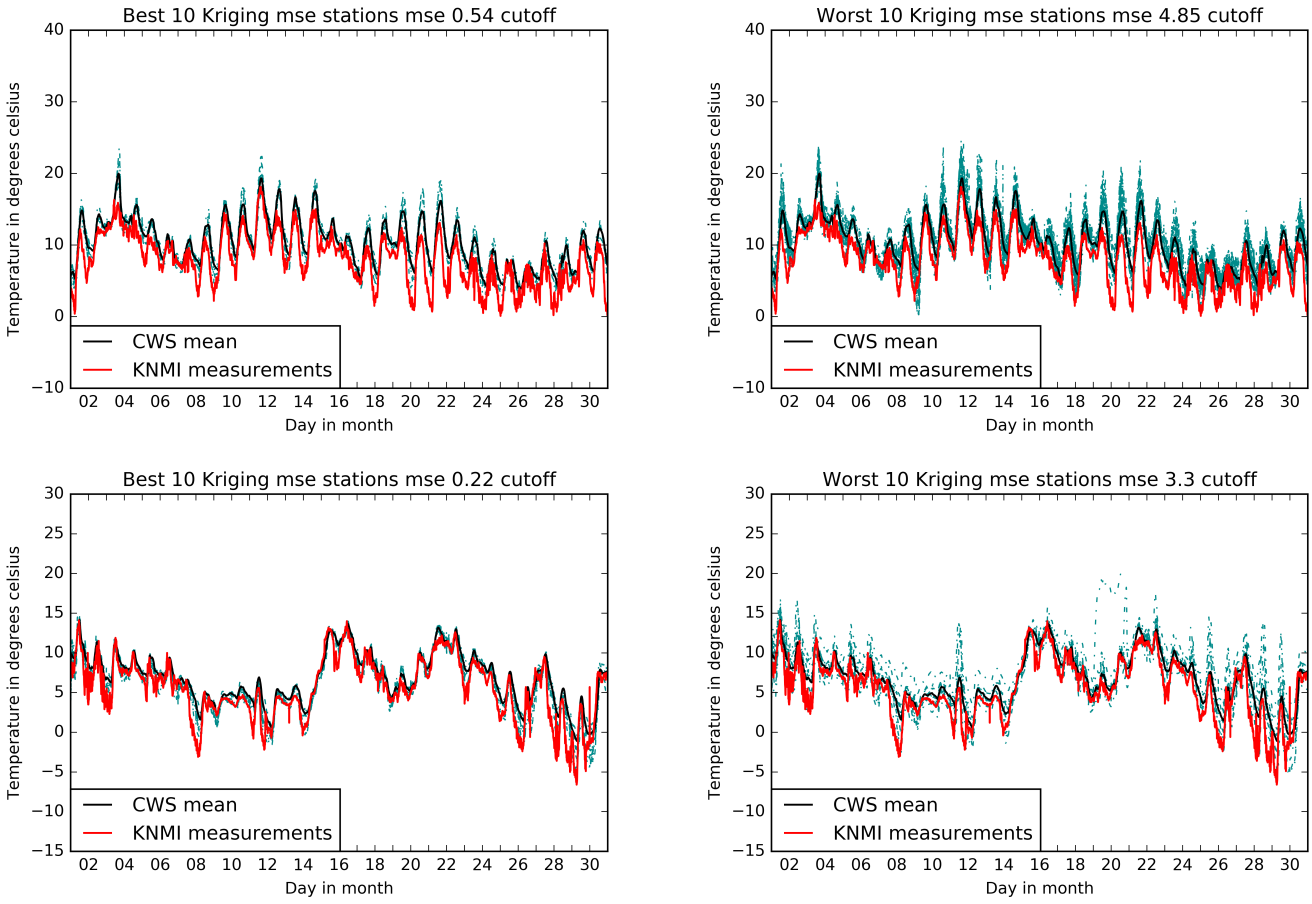

Figure 3.23: Best and worst peforming stations according to Kriging mean squared error

Chapter 4

Discussion

4.1 Conclusions

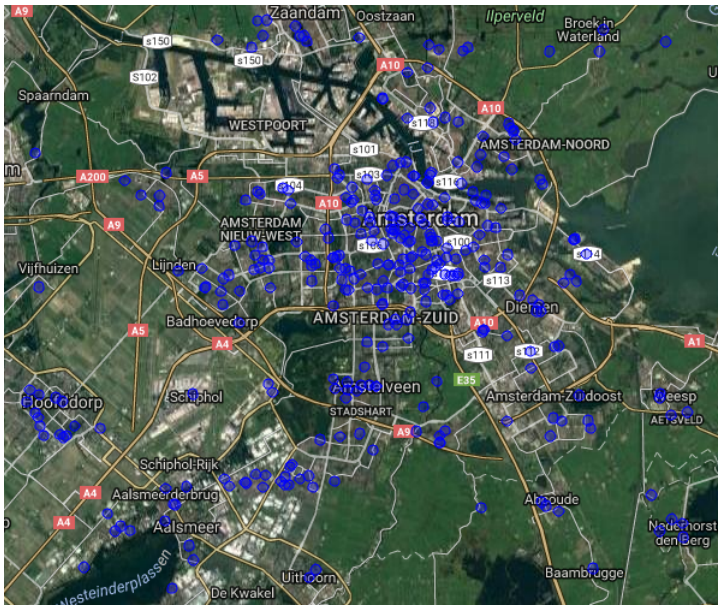

Figure 4.1: Satalite image of Amsterdam with station locations

4.2 Discussion

Z-score gebruiken als outlier measure omdat het onlogisch is dat een meeting 3x de std van het gemiddelde afligt. RJ: "Hoewel ik snap dat er keuzes gemaakt moeten worden blijf ik dit toch een zwak argument vinden. Als je meer tijd gehad zou hebben (bv. een Msc thesis) had je ook per tijdsmoment statistiek kunnen gebruiken om te bepalen of een sample afwijkt (iets met sample/re-sample met en zonder terugleggen ofzo...)."

Keuze van modelleren van spatiele variantie is vaag. Stations gaan na een bepaalde afstand weer op elkaar lijken. Nu geen last van omdat alleen de dichstbijzijnde 10 worden gebruikt voor Kriging, maar een slimmer persoon met domein kennis zou er wat beters van kunnen maken.

Resamplen kan beter.

Sommige QM werken beter voor bepaalde toepassingen: als je de maximum temperatuur op een dag wilt weten, zal een fase verschil je niet zo dwars zitten en wil je liever stations hebben met

een lage Kalman MSE dan met een hoge correlatie met KNMI of een hoge PCA explained variance. Als je het effect van opwarming in de zon wilt modelleren, kan je beter de stations gebruiken met een hoge correlatie met het knmi en een hoge kalman MSE, aangezien die waarschijnlijk in de zon staan. Als je het gedrag van temperatuur in de stad wilt weten, heb je meer aan de stations die sterk met elkaar covarieren, en wil je de stations die veel PCA explained var hebben. Stations met een hoge Kriging MSE liggen of a) in de zon, of b) hebben een unique locatie ten opzichte van de stations eromheen.

Als je wilt weten hoe de temperatuur in de stad zich gedraagt, heb je niet per se veel aan het weggooien van data punten die niet correleren met het KNMI. Als je een fenoneem wilt beschrijven dat zich alleen voordoet in de stad, dan zal dit sowieso niet correleren met het KNMI. Maar als alle stations in de stad een bepaald gedrag vertonen, dan zal dit zich wel uitdrukken in de PCA projectie. Dus dan kan je beter niet stations weggooien die niet correleren met het KNMI, maar stations die niet covarieren met de rest van de stad.

Niet correleren met het KNMI is niet per se fout. It is good to keep in mind that stations that don't behave like the KNMI are not necessarily wrong, as a intented purpose of CWS is to model behaviour that, with current KNMI station placement, cannot be modelled by KNMI in the first place.

De rest van de spatiele variantie kan zeer waarschijnlijk verklaard worden door hoe dense de bebouwing in die area is. In delen van Amsterdam met veel parken (duidelijk: amsterdamse bos, gaasperplas, oosterpark etc) zie je veel meer afkoeling dan de meest drukke delen (duidelijk: amstelveen, centrum, west).

Bibliography

Netatmo User Manual, 2012. 3

S.J. Bell. *Quantifying uncertaintry in Citizen Weather Data*. PhD thesis, Aston University, 2014. 1, 2

F. Meier, D. Fenner, T. Grassman, B. Jnicke, M. Otto, and D. Scherer. Challenges and benefits from crowdsourced atmospheric data for urban climate research using berlin, germany, as testbed. In *ICUC9 - 9th International Conference on Urban Climate jointly with 12th Symposium on the Urban Environment*. Department of Ecology, Technische Universitt Berlin, Germany,, 2015. 1, 2

R Nakamura and L. Mahrt. Air temperature measurement errors in naturally ventilated radiation shields. *Journal of atmospheric and oceanic technology*, 22(6):1046–1058, 2005. 7

M.A Oliver and R. Webster. A tutorial guide to geostatistics: Computing and modelling variograms and kriging. *CATENA*, 133:56–69, 2014. 8

G.J. Steeneveld, S. Koopmans, B.G. Heusinkveld, L.W.A van Hove, and A.A.M. Holtslag. Quantifying urban heat island effects and human comfort for cities of variable size and urban morphology in the netherlands. *Journal of Geophysical Research*, 116(D20129): doi:10.1029/2011JD015988, 2011. 1

P. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining*. Pearson Education Limited, 2013. 7

G Welch and G. Bishop. *An introduction to the Kalman Filter*. Department of Computer Science, University of North Carolina at Chapel Hill, 2006. 5

Appendix A

Kalman Filter parameters

In general, the Kalman filter estimates the state x of a discrete-time controlled processes that is governed by a *linear* difference equation. Since temperature measurements are not considered to be linear, technical term for the application on this dataset is the *extended Kalman filter*. Here, the estimate uses the derivative of the process. In principle, the underlying process cannot be observed, but here the assumption is made that the KNMI measurements from the nearby station at Schiphol are representative for the process. Therefore, parameters for expected measurement error and process variation are calculated from KNMI measurement characteristics. For the derivative of the KNMI measurements, see figure A.1.

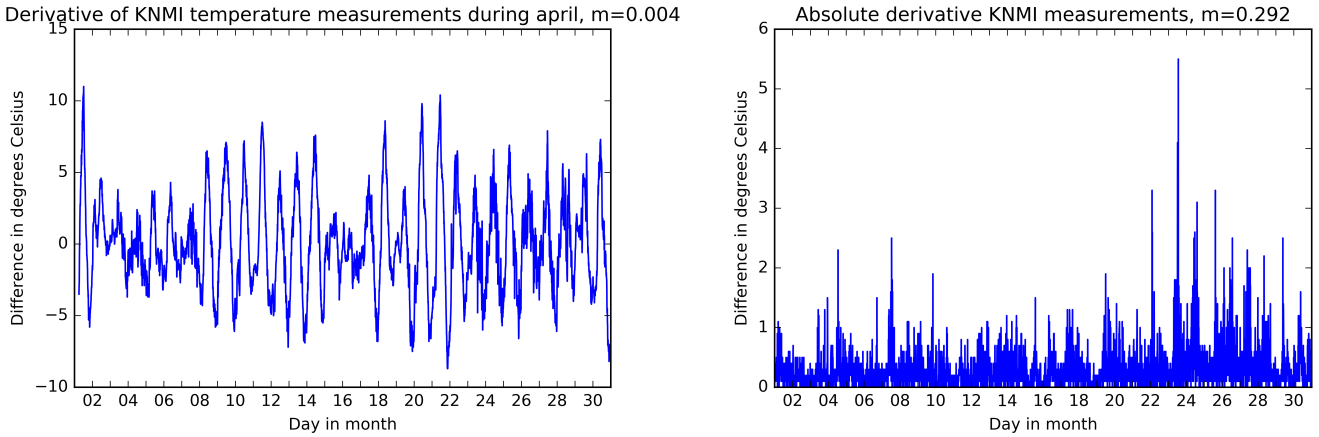

Figure A.1: Derivative of KNMI measurements during April (smoothed over hour) and absolute derivative of KNMI measurements (unsmoothed)

To determine the process noise parameter Q, the absolute derivative of KNMI measurements is considered, see plot A.1. Here the variance from timestep to timestep can be seen. The average m represents the average difference between to concurrent timesteps. This can be seen as the allowed process noise, since some variation between timesteps is expected. In this application, the parameter Q was assumed to be constant and set to the mean m.

The next parameter to be estimated is the measurement error R. In all but a few applications, the measurement error is not expected to remain constant. Again, in lack of direct observation of the underlying process, the KNMI measurements at Schiphol are assumed to be a valid representation of the process. The absolute mean difference between CWS measurements and KNMI measurements at each timestep can be seen in figure A.2. The difference in general seems to be highest during the late afternoons, and lowest (near zero at some days) during the early morning. In figure A.2 the average difference between KNMI and CWS measurements during the day is also

depicted.

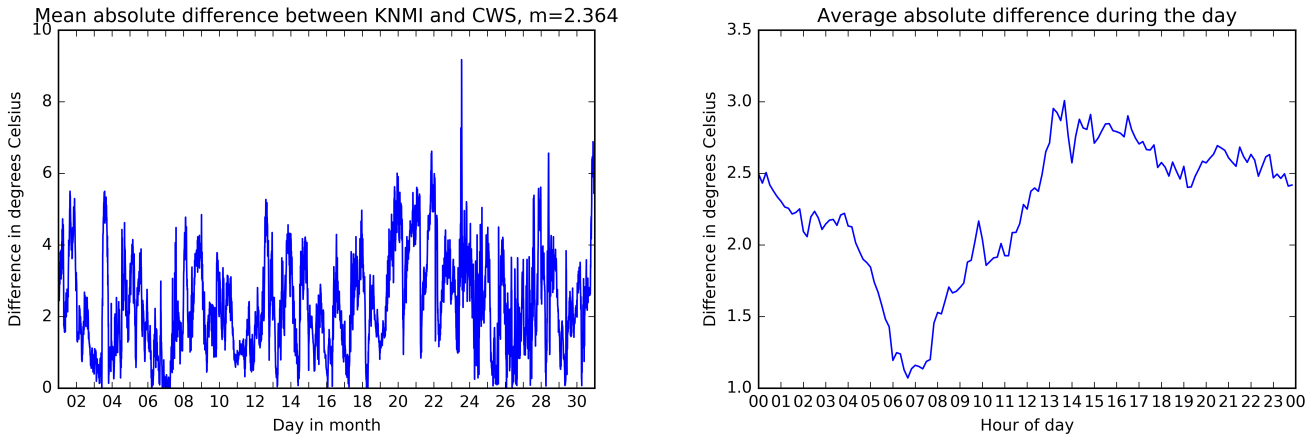

Figure A.2: Average difference between KNMI and CWS measurements at each timestep, and average difference during the day

Seeing the relationship between difference in measurements and time of day, the expected measurement error can probably be modelled as a direct function of time. However, due to both the variance of quality of CWS stations (should not 'punish' the well performing station measurements for the behaviour of the badly performing stations) and need for simplicity, the expected measurement error R is simply the difference between the measurement at the CWS station and the measurement at KNMI at each timestep. This relfects both differences in expected measurement error during the day, and difference in quality between CWs stations.

Appendix B

Variogram models

Common functions for variogram modelling are the spherical, exponential and Gaussian models. In all functions, h is lag in distance, c is the correlated and c_0 is the uncorrelated component of the variance.

The spherical model is defined as

$$\gamma(h) = c_0 + c\left(\frac{3h}{2r} - \frac{1}{2}\left(\frac{h}{r}\right)^3\right)$$

r is the range (maximum distance between stations) of the function. The quantity $c + c_0$ is known as the 'sill'.

The spherical model fitted to the average variogram can be seen in figure B.1. The c_0 is estimated to be low, compared to the other models. This does not seem to reflect the variance at the zero lag.

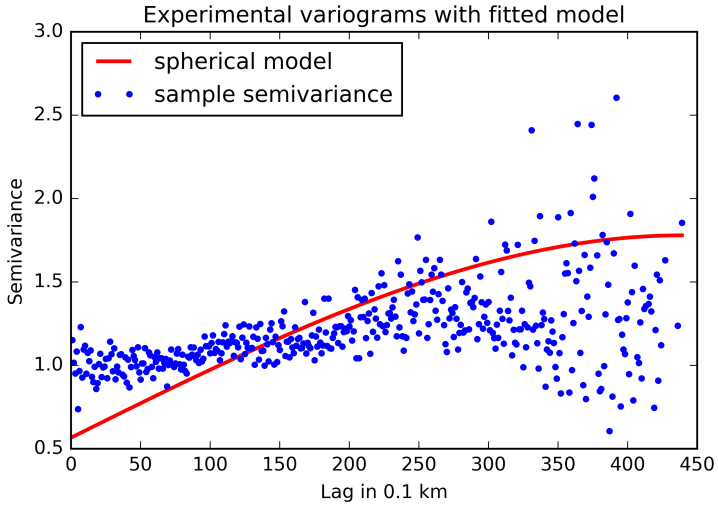

Figure B.1: Temperature measurements of a single station. Plot includes KNMI measurements at Schiphol for comparison.

The exponential model is the function

$$\gamma(h) = c_0 + c\left(1 - exp\left(-\frac{h}{r}\right)\right)$$

The exponential model fitted on the average variogram can be seen in figure B.2. The increase in variance over the first few kilometers of lag seems to be overestimated.

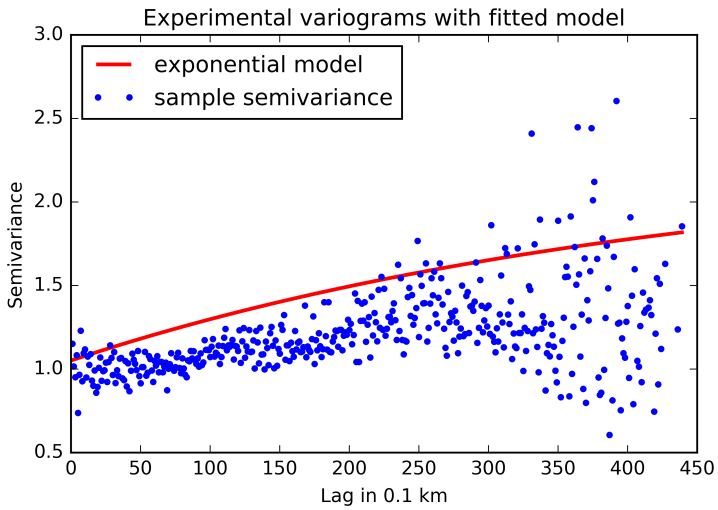

Figure B.2: Temperature measurements of a single station. Plot includes KNMI measurements at Schiphol for comparison.

Since the Gaussian model had the best fit, this model was chosen as variogram model for Kriging (see Methods chapter, figure 2.4).

Appendix C

Validity of assumption of spatial correlation

In order to perform Kriging, there has to be an assumption that the data is positively spatially correlated: semivariance between stations is low when stations are close, and high when stations are far apart. In order to determine if this is a valid assumption, the variogram that is calculated from the samples has to be considered. The experimental variogram was modelled for the raw dataset, the pre-processed dataset and the temporally processed dataset, and visually inspected.

C.1 Spatial correlation in raw data

Figure C.1 shows the average measurements of stations and their locations. Some stations averages are extremely high compared to others. This can be seen in both the experimental variogram in and the fitted model in figure C.2 where the variance at the zero lag is estimated to be very high at over 10 degrees Celsius.

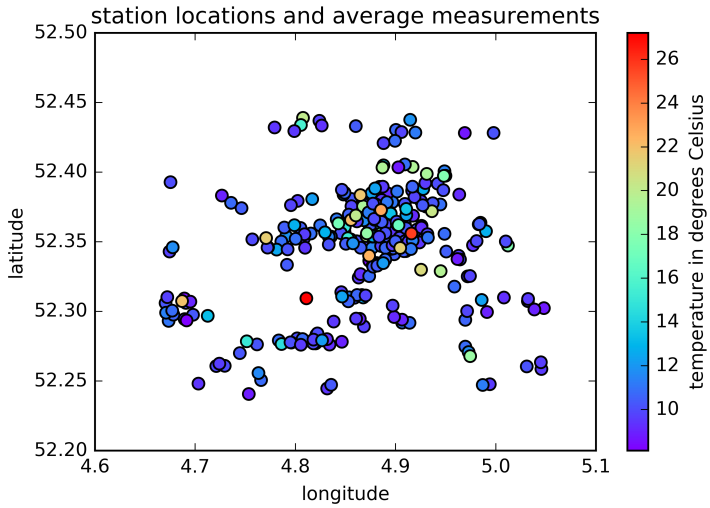

Figure C.1: Location and average measurements of unprocessed stations.

In the sample variogram, semivariance and distance are negatively correlated (with a correlation coeffiient of -0.57), meaning that the further away a pair of stations is, the more similar they are. This can perhaps be explained by the observation that the stations that report the highest average

temperatures seem to be located towards the center, and thus being close to the rest of the stations. In any case, the assumption of positive correlation between distance and semivariance does not hold.

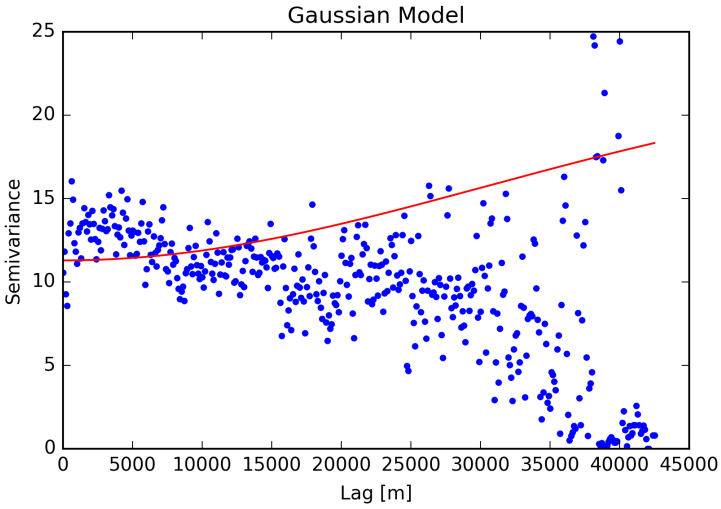

Figure C.2: Gaussian model fitted on unprocessed data.

C.2 Spatial correlation in preprocessed data

Figure C.3 shows the average measurements of stations after preprocessing. Besides the one red dot, most stations have similar average measurements. This is reflected in the experimental variogram and the Gaussian model, with an uncorrelated component at the zero lag at around 1.5 degrees.

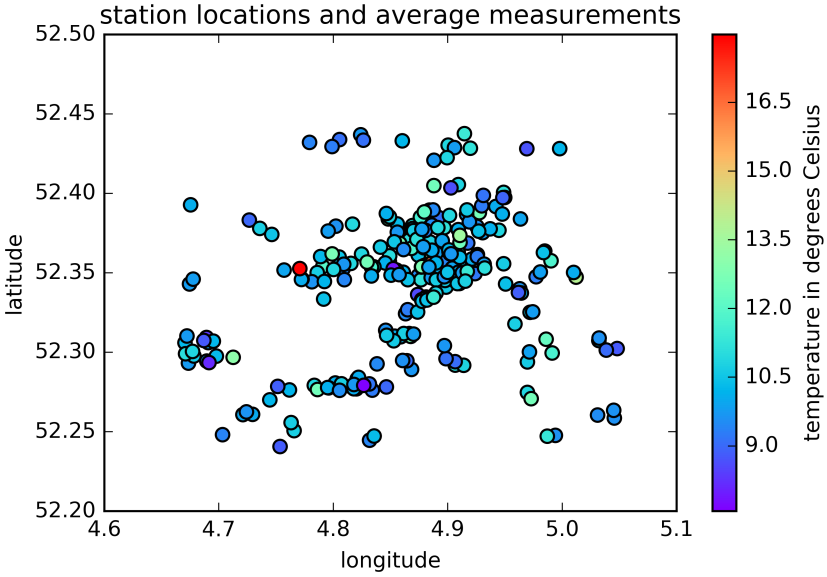

Figure C.3: Location and average measurements of preprocessed stations.

However, there is still a negative trend in the experimental variogram. The correlation between distance and semivariance is negative with a correlation coefficient of -0.33.

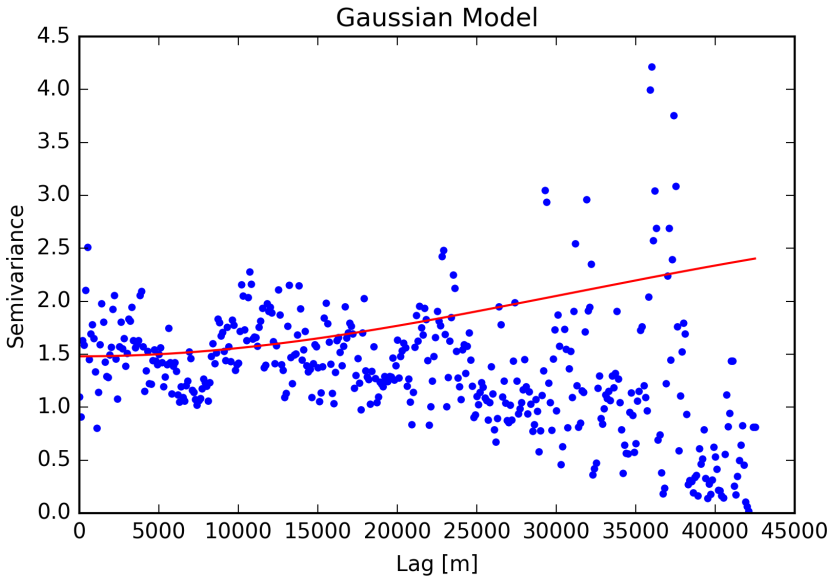

Figure C.4: Gaussian model fitted on preprocessed data.

C.3 Temporally processed data

After data has been processed by the Kalman filter and PCA, the resulting averages are as shown in figure C.5. Most stations have very similar averages, which is reflected by the experimental variogram and the Gaussian model, where the variance at the zero lag is at 0.2 degrees Celsius.

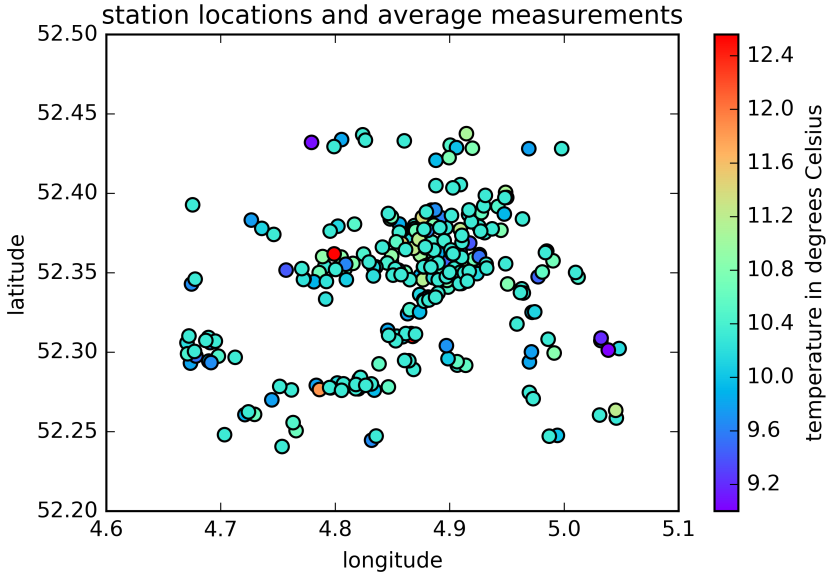

Figure C.5: Location and average measurements of processed stations.

The correlation between distance and semivariance is positive, with a correlation coefficient of 0.17. This indicates that after temporal processing, the assumption that the remaining variance in the data can partially be spatially explained, can be a valid assumption.

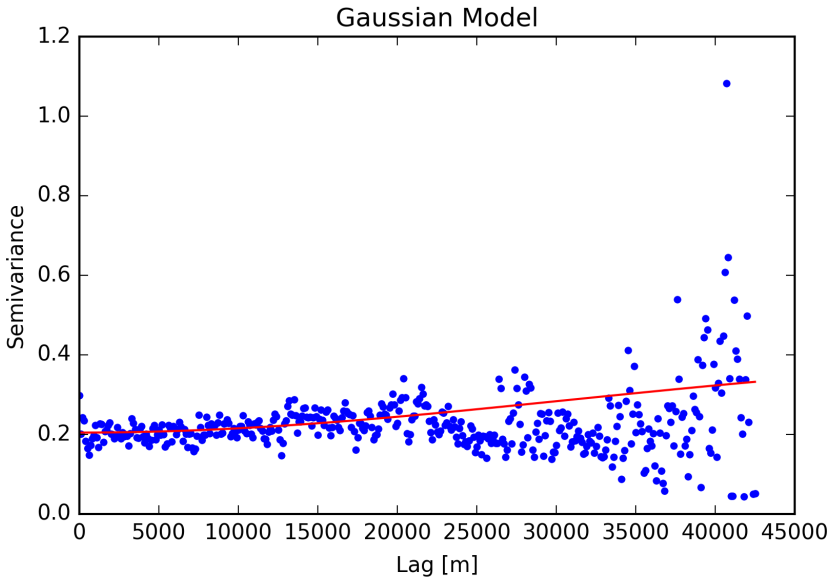

Figure C.6: Gaussian model fitted on processed data.