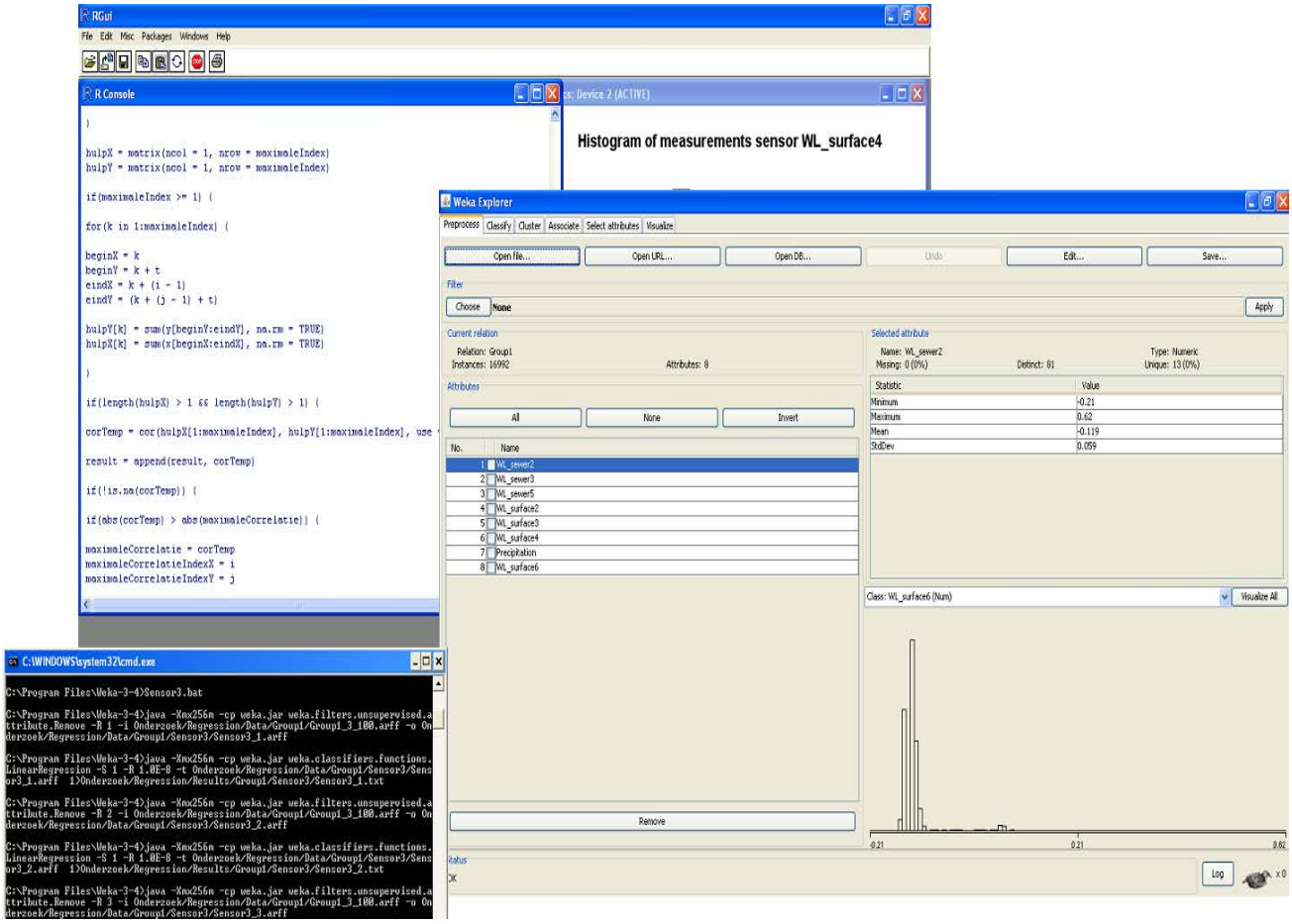
BMI Internship

Improving data quality control



Student name: Jaap de Rue
Student number: 1163493

Company: Witteveen+Bos
Department: Applied Statistics
Supervisor: dr. ir. J.L. Korving

Supervisor VU: dr. W.J. Kowalczyk
Second reader: dr. W. van Wieringen

Witteveen+Bos

Preface

The study Business Mathematics and Informatics at the Vrije Universiteit in Amsterdam has to be completed with an internship at a company, research facility, or institute. During this internship a real world problem is solved by the knowledge gained during the study Business Mathematics and Informatics.

For this internship I worked in the applied statistics department of Witteveen+Bos at their headquarters in Deventer. It consisted of the improvement of an existing validation model used for the analysis of measurement data collected in a sewer system.

I would hereby like to thank Hans Korving and Elke Ottenhoff of the applied statistics department of Witteveen+Bos. Without their time, help, and enthusiasm this report would not have been possible. I would like to thank Wojtek Kowalczyk for his most valuable guidance and comments, and his fast replies. Furthermore, I would like to thank all colleagues of Witteveen+Bos who have made these past six months a very pleasant and fun period.

Jaap de Rue

Deventer, August 2007

Management summary

(Confidential)

Table of contents

Part I Introduction

1 Introduction

1.1 Company

Witteveen+Bos is a company that provides advice and engineering services for projects in the sectors: water, infrastructure, environment, and construction. Typical for the method of working of Witteveen+Bos is the multidisciplinary project procedure: specialists from different sectors work together on the solution of complex assignments. The clients of Witteveen+Bos are governments, trade and industry, and different kinds of collaboration relationships. Witteveen+Bos serve them from eight offices in the Netherlands and from four offices abroad.

Via their work they give shape to society. Witteveen+Bos feel responsible for the delivery of reliable solutions for technical and social assignments. Clients and society may feel free to comment Witteveen+Bos on that. Employees (also Witteveen+Bos' stakeholders) are able to associate themselves with this responsibility awareness.

1.2 Background

(Confidential)

1.3 Relevance

(Confidential)

1.4 Value

The validation tool is used to answer questions like: “Is the sensor working?”, “Are the measurement data reliable?”, “Can the data be used in the reports?”, and “Are the data suitable for model calibration?”. The answers to these questions are then used by the sewer administrator to gain more certainty in locating and the effectivity of planning measures, and to test and increase the reliability of the sewer model. Next to that, a good quality control is of great importance for the quality of the so-called overflow-reports, as this is demanded by the water quality administrator. Furthermore, these overflow-reports are used to determine who is to blame in case of calamities.

(Confidential)

Lastly, the quality control of the measurement data can be used to test the operational control of the sewer system. For instance, tracing adjustments in the tuning of a consort. Although this is not the primary goal of the data control, it may be helpful in this way.

1.5 Goal

The main goal of this internship is to apply Data Mining techniques in order to improve the quality control.

To be able to determine whether a measurement is “good”, “questionable”, or “fault”, numerous mathematical models can be used. During the internship I will apply two Data Mining methods to examine the effect these have on the results of the quality control. Since the quality of these measurements is of such high importance to the clients of Witteveen+Bos (see paragraph 1.4), it is important to determine this quality with as high accuracy as possible.

(Confidential)

1.6 Approach

The bigger part of this internship will consist of applying the different techniques to the data. The first phase however will consist of an extensive data exploration. The measurement data will be summarized both numerical and graphical and the underlying probability distribution of the data will be studied in order to gain more insight into the behaviour of the measurement data.

After the data exploration is rounded up, an extensive analysis of the correlation between the measurement data of the sixteen sensors will be done. The goal of this analysis is to discover dependencies in the data. For instance, it is already shown that the correlation between the turbidity measurements and the precipitation on a four hour time interval is much higher than the correlation of the measurements based on the original five minute time interval. The necessary functions will be implemented in R (see paragraph 1.8) to calculate the correlation of all possible combinations. The goal of this correlation analysis is to divide the sixteen sensors into two groups. One group for which holds that the sensors have a high correlation, and one group for which holds that the sensors have less correlation. This way, comparison is possible between the results obtained by the M5P and robust regression.

When the correlation analysis is finished, the M5P and robust regression methods will be applied to try and improve the current quality control. After these methods are applied, a comparison with the current models will be done in order to study the effect these models have on the results of the quality control. At last a third technique will be devised, applied and tested to study whether faulty sensors can be identified.

1.7 Paper organization

This first part of the report gives the background of this internship. It is purely of an informative and introductory nature. In the second part the results of the data analysis are described. The data analysis consists of the numerical and graphical summary (chapter 2), the density study (chapter 3), and the correlation study (chapter 4). All steps taken, results, and conclusions will be given in the related chapters.

The third part of this report (Modelling, Evaluation, and Identification) will consist of the application of the two Data Mining techniques (chapters 5 and 6) and the evaluation of these techniques (chapter 7). All steps taken, all results, and the conclusions will be given in the corresponding chapters. Lastly, a technique will be described and applied to the measurement data in order to try to identify faulty sensors. The results of this technique will be given in chapter 8, where the conclusions will also be given.

The last part of this report (Conclusions) will describe all conclusions and recommendations that resulted out of this internship.

1.8 Programs

During this internship, two systems for data analysis will be used. For the first three phases, the open source program R will be used. R is a language and environment for statistical computing and graphics. It provides a wide variety of statistical and graphical techniques, and is highly extensible. One of R's strengths is the ease with which well-designed publication-quality plots can be produced, including mathematical symbols and formulae where needed.

The second program that will be used is Weka. Weka is also open source software that contains a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from one's own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes.

Two classification tasks will be performed to the measurement data in the Weka software. To be able to do so, the data will have to be converted to a so-called arff format. An arff (Attribute-Relation File Format) file is an ASCII text file that describes a list of instances sharing a set of attributes. ARFF files were developed by the Machine Learning Project at the University of Waikato for use with the Weka software.

ARFF files have two distinct sections. The first section is the Header information, which is followed by the Data information. The Header of the ARFF file contains the name of the relation, a list of the attributes (the columns in the data), and their types. An example of an ARFF file is as follows:

```
% 1. Title: Sewer network database
%
% 2. Sources:
%      (a) Creator:      J.M. de Rue
%      (b) Donor:        City of (Confidential)
%      (c) Date:         May, 2007
%

@RELATION Group1

@ATTRIBUTE WL_sewer2      Numeric
@ATTRIBUTE WL_sewer3      Numeric
@ATTRIBUTE WL_sewer5      Numeric
@ATTRIBUTE WL_surface2    Numeric
@ATTRIBUTE WL_surface3    Numeric
@ATTRIBUTE WL_surface4    Numeric
```

```
@ATTRIBUTE WL_surface6     Numeric
@ATTRIBUTE Precipitation   Numeric

@DATA
3.0,4.0,8.0,5.0,6.0,13.0,15.0,16.0
-0.14,-1.46,-1.03,?,?,?,0.6,0.0174
-0.14,-1.47,-1.03,0.57,0.55,0.55,0.59,.0174
-0.14,-1.46,-1.03,0.57,0.5413,0.55,0.61,0.0174
-0.14,-1.46,-1.03,0.57,0.54,0.55,0.61,0.0174
-0.13,-1.46,-0.98,0.57,0.5484,0.55,0.61,0.0174
-0.12,-1.47,-0.88,0.57,0.5418,0.5418,0.61,0.0174
-0.11,-1.45,-0.79,0.57,0.548,0.548,0.61,0.0174
-0.11,-1.46,-0.72,0.57,0.55,0.55,0.62,0.0174
```

All missing values, i.e. NaN's, are represented by a question mark.

Part II Data analysis

2 Data summary

In this chapter a numerical and graphical summary will be made to gain more insight into the measurement values. In the first paragraph the mean, median, standard deviation, minimum, maximum, skew ness, and kurtosis of the measurement values is calculated and the results will be discussed. In the second paragraph, histograms, box plots, and numerous different plots will be generated and discussed.

2.1 Data

(Confidential)

2.2 Choice of sensors

For this internship, sixteen sensors have been chosen. This group of sensors consists of three different types of sensors, namely: sensors that measure the water level in the sewer network (type W), sensors that measure the water level of the surface water (type Z), sensors that measure the discharge of the water (type D), and one sensor that measures the precipitation (N). This choice was made after taking a first glance at the data and was made such that multiple types of sensors, multiple types of faults in the data, and multiple locations are represented.

No turbidity sensors were chosen for this internship. The main reason for this is that quality labels still have to be developed for the turbidity measurements, i.e. the quality control of turbidity sensors is given as a different internship project.

The measurement data used during these three phases is the data of the months January and February. Since the data of the precipitation is largely missing of these two months, data of the precipitation in De Bilt is gathered from the website of the KNMI. This data was added because precipitation can be of great influence to the behaviour of the water levels and other measurements in the sewer network. (Confidential)

So to make sure that the vectors of measurement values are all of the same length, the data of the precipitation was adjusted such that the resulting vector contained the measurements of the precipitation every five minutes. This has as a side-effect that measurements of other sensors still can only be related to the precipitation on a daily basis, i.e. the precipitation is constant during the day, so small deviation or extreme values in the measurements of another sensor can not be related to the precipitation. This also means that the correlation study might lead to less good results wherever the measurement data of the precipitation is involved. Still, the precipitation is believed to be of great influence to the measurement data of the sewer network, even with measurement values on a daily basis instead of a five-minute basis.

Table 1 lists the names, locations, types, and comments of the sixteen sensors. Of these sixteen sensors, seven are of type W (water level in the sewer network), six are of type Z (water level of the surface water), two are of type D (discharge of the water), and one sensor is op type N (precipitation). Names are created for the sensors in order to gain clarity on what sensor is being discussed in this report. The seven sensors that measure the Water Level in the sewer network are named WL_sewer1 - 7, the six sensors that measure the Water Level in the surface water are named WL_surface1 - 6, the two sensors that measure the discharge of the water are named Discharge1 and 2, and the precipitation sensor is called Precipitation.

Name	Location	Type	Comments
WL_sewer1	02.0453	W	All values the same
WL_sewer2	21.1166	W	
WL_sewer3	01.1521	W	
WL_sewer4	05.0136	W	Many missing values
WL_sewer5	13.0208	W	
WL_sewer6	13.0222	W	
WL_sewer7	14.0037	W	
WL_surface1	02.0453	Z	All values the same
WL_surface2	05.0904	Z	
WL_surface3	06.0470	Z	
WL_surface4	06.0738	Z	
WL_surface5	20.0811	Z	
WL_surface6	03.0846	Z	
Discharge1	13.0222	D	
Discharge2	14.0037	D	
Precipitation	De Bilt	N	

Table 1: Types and locations of sensors

The sensors are spread out through the sewer network, their location is coded in its original sensor name, i.e., sensors WL_sewer1 and WL_surface1, WL_sewer6 and Discharge1, and WL_sewer7 and Dischare2 are placed at the same location. However, water flows through the sewer network, i.e., from the location of one sensor to the location of another sensor. It can thus be possible that despite different locations, the measurement values of two sensors have a high correlation, taking into account the time delay.

The comments in Table 1 mention the obvious faults in the data when taking a first glance at these. The first and eight sensor in the table contain measurement data of which the values are all the same, which could mean that the sensors are not operating properly, or they are not located at the proper location. Furthermore, the measurement data of sensor WL_sewer4 contains a lot (8792 out of 16992 measurements) of "NaN" values, i.e., missing values.

It should be noted here that the original measurement values of the sensors are used unless stated otherwise, i.e., in some cases the sums of the measurement values of one (or more) hour(s) is taken to gain clarity on the results. In this case, it will be stated clearly if and what sums were taken.

2.3 Numerical

To obtain a first understanding of the measurement data of the sixteen sensors, a numerical summary is made. The calculation of the mean, the standard error, the median, and the range can provide valuable information about the data. Table 2 lists the results of the calculations on the measurement data.

Sensor	Unit	Mean	Median	St. deviation	Min	Max
WL_sewer1	m. to NAP	0,05	0,05	0	0,05	0,05
WL_sewer2	m. to NAP	-0,119	-0,13	0,0587	-0,21	0,62
WL_sewer3	m. to NAP	-1,4681	-1,47	0,0072	-1,49	-1,43
WL_sewer4	m. to NAP	-1,4489	-1,47	0,1414	-1,47	0,2
WL_sewer5	m. to NAP	-0,9971	-1,03	0,0979	-1,04	0,04
WL_sewer6	m. to NAP	-1,4517	-1,8	0,4026	-1,8	0,41
WL_sewer7	m. to NAP	-1,9801	-1,64	0,569	-2,7	-1,08
WL_surface1	m. to NAP	0,05	0,05	0	0,05	0,05
WL_surface2	m. to NAP	0,554	0,55	0,0943	0,4726	0,83
WL_surface3	m. to NAP	0,5448	0,54	0,0752	0,48	0,66
WL_surface4	m. to NAP	0,5319	0,53	0,0254	0,4644	0,65

WL_surface5	m. to NAP	0,2684	0,46	0,3202	-0,33	0,5982
WL_surface6	m. to NAP	0,5962	0,6	0,0254	0,49	0,69
Discharge1	m3 per 5 min.	3,714	0	10,6973	0	70,7414
Discharge2	m3 per 5 min.	0,7351	0	3,3517	0	19,3139
Precipitation	mm. per 5 min.	0,1006	0,0278	0,1577	-0,0035	0,8611

Table 2: Characteristics of the data

The results for the sensors that measure the water level in the sewer network (WL_sewer1 – 7) show that the mean value is low with respect to the range, and that the median is close to the mean value. This could mean that there is a large number of measurements that are responsible for the distortion of the range. Furthermore, the median is somewhat smaller than the mean value, which could mean (with the exception of WL_sewer7) that there is a relatively small number of large measurement values, i.e., outliers or extreme values. This is emphasized by the positive skewness, which is a measure of the asymmetry of the probability distribution. A positive skew means that the mass of the measurement data is concentrated to the left. The skewness for the measurement values of all sensors is given by Table 3, together with the kurtosis which is a measure for the peaked ness of the measurement data. The higher the kurtosis, the more of the variance is due to infrequent deviations. The high skewness and kurtosis for sensors WL_sewer2, WL_sewer4, and WL_sewer5 means that mass of the measurement data is concentrated to the left and that it has a very high peak and thus very thick tails. The negative value for the skewness of WL_sewer7 means the mass of the measurement values is concentrated to the right and that there is a relatively small number of small measurement values.

Sensor	Skewness	Kurtosis
WL_sewer1	NaN	NaN
WL_sewer2	6,967438	62,73134
WL_sewer3	0,0602827	2,870637
WL_sewer4	8,766245	83,94596
WL_sewer5	6,183808	49,6689
WL_sewer6	0,3568109	1,260769
WL_sewer7	-0,4273429	1,271375
WL_surface1	NaN	NaN
WL_surface2	2,232511	25,70422
WL_surface3	0,5345028	4,188145
WL_surface4	0,3859615	3,675442
WL_surface5	-0,9075874	1,87412
WL_surface6	0,0627399	3,287606
Discharge1	3,186569	13,17393
Discharge2	4,659952	23,4081
Precipitation	2,463079	10,55633

Table 3: Skewness and kurtosis of the data

The results for the sensors that measure the water level of the surface water (WL_surface1 – 6) show that the range of measurement values of sensor WL_surface5 is about three to four times bigger than the other ranges and that it is the only sensor of this type that has a negative skewness. This could be an indication that the measurement values of this sensor contain extreme values.

Furthermore, the results for the discharge meters (Discharge1 - 2) show that the standard deviation and range of Discharge1 is far bigger than the standard deviation and range of sensor Discharge2, which could indicate that the measurement values of sensor Discharge1 contain a number of extreme values. In addition to this, the range of sensor Precipitation shows that the minimum value is negative, which is not possible. This could indicate that the measurement values are biased, or that the sensor is not working properly.

The measurement values of the bigger part of the sensors have a positive skewness, which means that the mass of the measurement data is concentrated to the left. This could be a result of the precipitation. When it does not rain, the sensors in the sewer network measure (very roughly spoken) the same. When it does rain, a lot of water flows into the network resulting in sensors measuring higher values. So the bigger part (or mass) of the data is concentrated to the left.

In addition to this, it can be concluded that the sensors WL_sewer1 and WL_surface1 measure a constant value. It is therefore decided to ignore these two sensors for the remainder of the internship. The technique in chapter eight will be used to try and identify faulty sensors. So for this chapter these two sensors will again be regarded.

2.4 Graphical

In this part of the report a graphical summary of the data will be given by providing different illustrations of the data. By means of this the numerical summary and its comments will be supported and accentuated.

2.4.1 Plots

To obtain a first impression on the behaviour of the measurement data, plots are made of the measurement values. Figure 1 gives the plots of the measurement data of the fourteen sensors. In this figure it can be seen that the measurement

values of sensors WL_sewer2 and WL_sewer5 have a number of extreme values. Furthermore, the plot of sensor WL_sewer4 shows a lot of gaps in the data, these are the "NaN's" or missing values that this sensor was seen to have in the numerical summary. Furthermore, the plot of sensor WL_sewer3 shows an extreme small range when compared to the other sensors that measure the water level in the sewer network.

The plots of sensors WL_sewer6 and WL_sewer7 show a gap in the data at the approximately the same dates as the plots of sensors Discharge1 and Discharge2 respectively. For all four sensors it holds that these are not gaps as data is missing, but gaps where the sensor measures a constant "low" value for the water level of the surface water and for the discharge of the water at that location. The plots of the sensors WL_surface2, WL_surface3, WL_surface4, and WL_surface6 seem to be in accordance with each other. They show the same behaviour and extreme values at the same dates, where the extreme values of sensor WL_surface2 are more extreme than those of the other three sensors. This can also be seen by the range of the y-axis. Due to these plots, it is made visible what happens to the measurement data of sensor WL_surface5. The measurement data makes a drop and seems to continue from this drop on. The measurements before this drop are in accordance with the measurements of the other sensors that measure the water level of the surface water. This drop in measurement data could be the result of one of the companies replacing the sensor, while its operation was not changed, i.e., it measures the water level to NAP, which of course cannot change like the plots suggest.

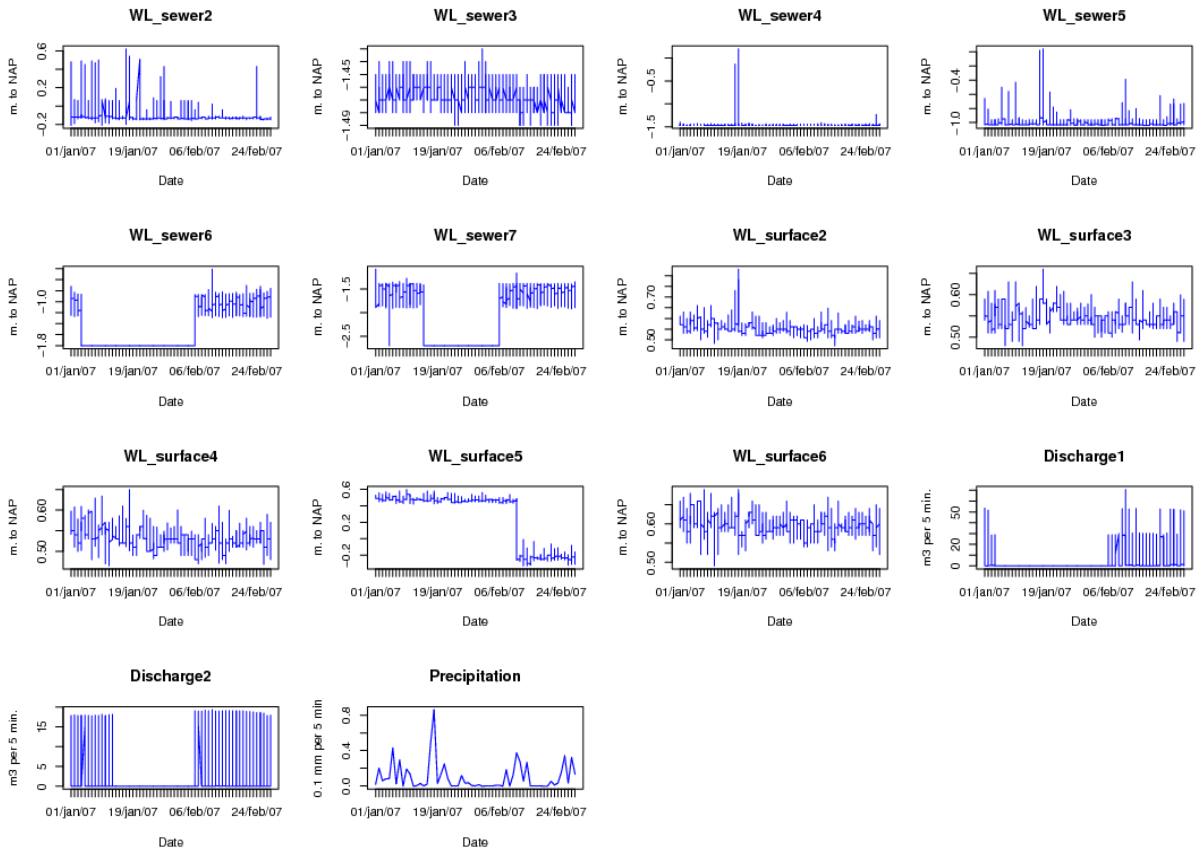


Figure 1: Plots of data of fourteen sensors

Next to this, the plots of sensors Discharge1 and Discharge2 show that the data of both sensors contain extreme values, where the extreme values of Discharge1 are more extreme than those of Discharge2.

Comparing the plots of the sensors that measure the water level in the sewer network with the plot of the precipitation it can be seen that most of the extreme values in the water level occur in cases of heavy downpour. An additional plot of these cases should provide more insight. Figure 2 gives a plot of all the data of all the sensors that measure the water level in the sewer network and the data of the precipitation. In this plot, the measurement data of the sensors are adjusted such that the data of a sensor is plotted above the data of the previous sensor. This way the plot is more clarifying and more information can be subtracted from it.

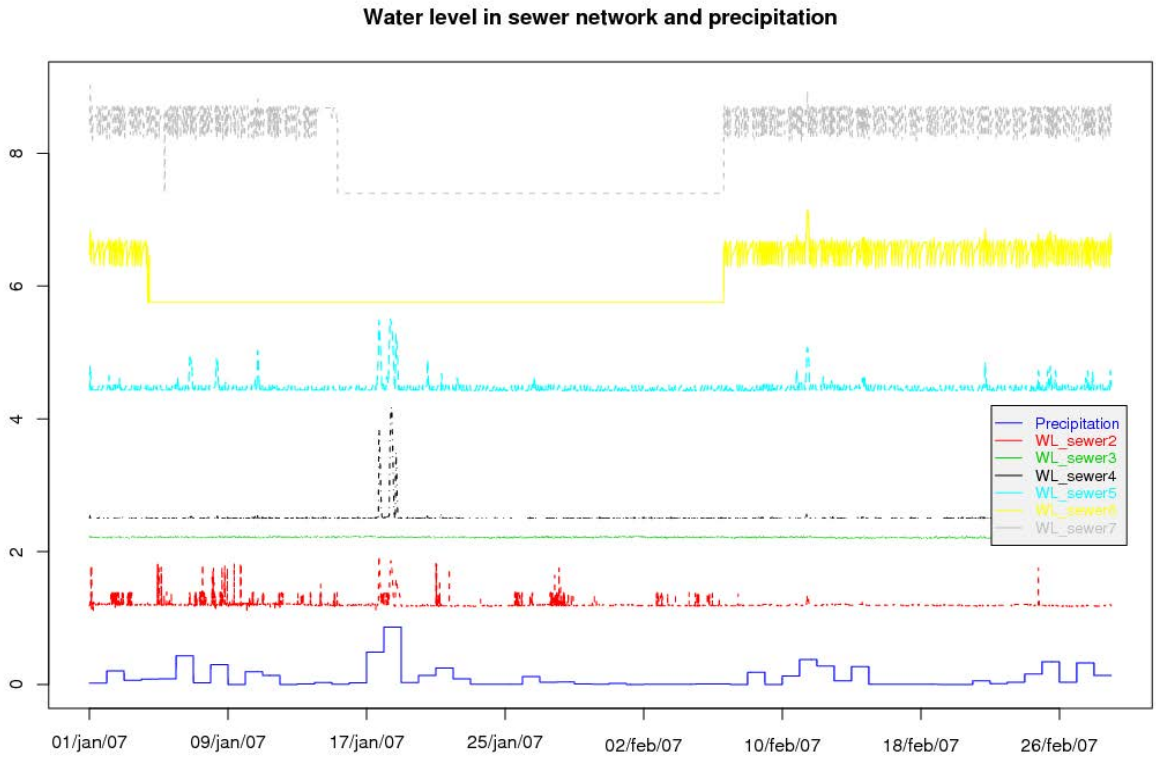


Figure 2: Plots of water level in sewer network and Precipitation

From this plot it can be seen that in most cases the extreme values of the water level in the sewer network occur in cases of (extreme) downpour (or precipitation). There are some exceptions to this rule, as the downpour seems to have no effect on the data of sensor WL_sewer3 whatsoever. The data of sensor WL_sewer4 seems to be effected only "once" by the downpour, but since a lot of measurement values are missing from the data of this sensor, it might be that the measurement values of this sensor are in fact being influenced by the downpour and that in a case of (extreme) downpour, this leads to a measurement error or a missing value.

In addition to this, the data of sensors WL_sewer6 and WL_sewer7 again seem not to be affected by the quantity of downpour. Except for the measurements around the twelfth of February, which results in an increase in water level at these two locations.

A same plot is made for the five sensors that measure the water level of the surface water together with the measurement data of the precipitation. For this

plot it also holds that the range of the data is adjusted such that the data of the sensors is plotted above each other.

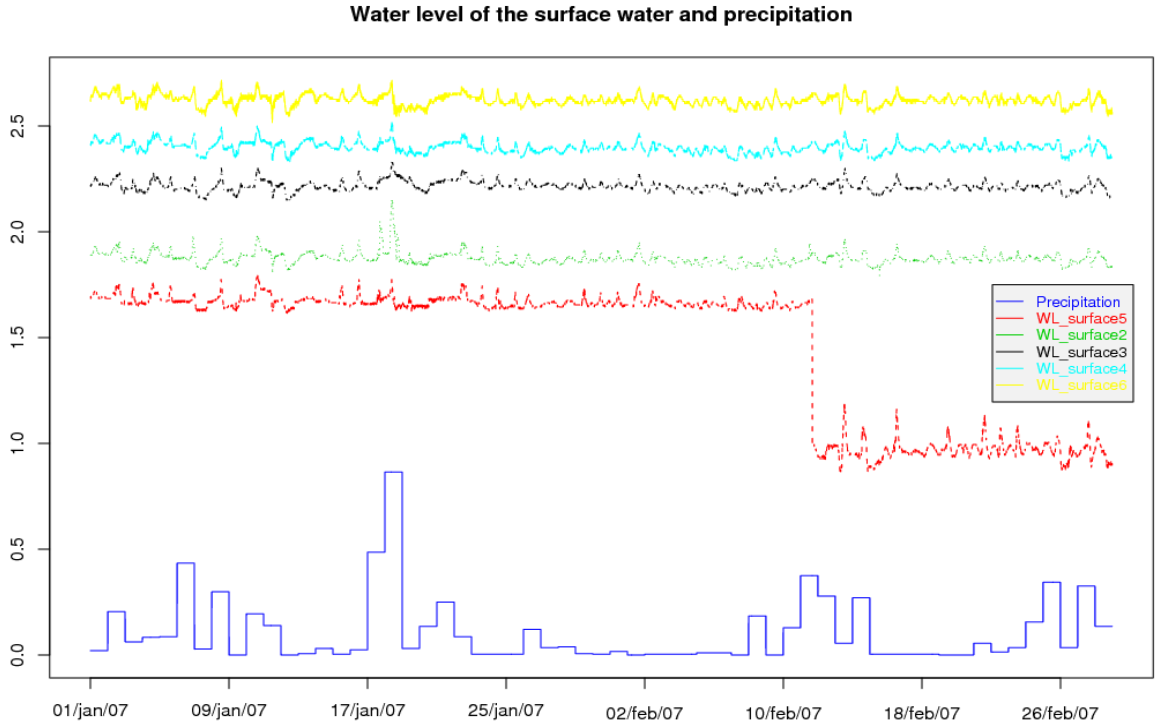


Figure 3: Plots of water level of surface water and precipitation

This plot shows that the measurement data of the sensors that measure the water level of the surface water follows the cases of (extreme) downpour. The data shows outliers where the precipitation shows outliers. Due to this plot it can be seen clearer that the measurement data of sensor WL_surface5 after the drop, is indeed from that point on in accordance with the other sensors of this type. Furthermore, the data of these sensors seem to be operating fine. With the exception of the drop in data of sensor WL_surface5 it seems that the measurement data “behaves” as one would expect. From this plot it can be seen that there is a strong correlation between the measurement values of these sensors and to some extend with the precipitation.

In addition to this, a third plot is made of the sensors that measure the discharge of the water. To this plot the data of the precipitation and the data of sensor WL_sewer6 and WL_sewer7 is added, since the sensors are placed at the same location as Discharge1 and Discharge2 respectively. The data of the discharge

sensors is rescaled to result in approximately the same range as the three sensors that were added, to keep the plot clarifying and informative.

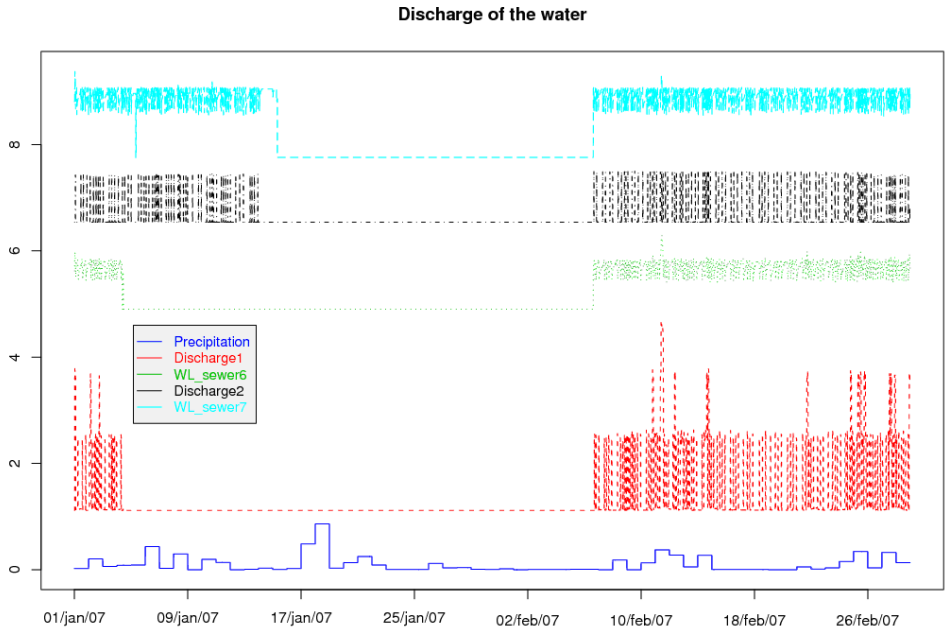


Figure 4: Plots of discharge and corr. water level in sewer netw. plus prec.

What can be seen from this plot is that, in contradiction with what was first concluded from the plot, the measurement data of sensors Discharge2 and WL_sewer7 do not drop at the same date. Furthermore, the precipitation does not seem to have any effect on the measurement values of Discharge2 and WL_sewer7. The measurement data of sensors Discharge1 and WL_sewer6 do seem to drop at the same date and with the exception of the drop it seems that the quantity of downpour does have any effect on the measurement data. In addition to this, it can be seen that for all four sensors it holds that the drop in values stops at the same date.

What has to be noted at these plots, is that the drop in the data is not the results of "NaN's" or missing values in the data, but is the result of a constant measurement. An explanation for this could be the location of the sensors, i.e. when these sensors are placed in a well where water is pumped in only when the water level at a nearby well has reached some specified level. This would result in water level that is constant over time, but increasing whenever the pump is started and water is pumped into this well. This can also explain why Discharge2 and WL_sewer7 seem to have no effect on the amount of precipitation, since the water level at this location (and thus the discharge of the

water) is only effected by a pump. The fact that Discharge1 and WL_sewer6 do seem to be effected by the amount of precipitation means that the pump that is responsible for the water level at that location is turned on faster.

2.4.2 Histograms

Next, histograms are made of the fourteen sensors. A histogram is a graph at which a scale for the measurements is presented at the horizontal axis. For every interval a rectangle is placed such that the surface of this rectangle is representative for the frequentation in that interval. This way a good impression of the spread of the measurements is obtained, which provides a first idea about the behaviour and distribution of the data.

For the histograms in this paragraph the density of the measurements is plotted, instead of the frequency of the measurements. This means that the probability densities are plotted, so the histogram has a total area of one. Since the measurements values of the sensors will also be studied for probability distributions, this was found more appropriate.

The histograms are given in Figure 5. The histograms of sensors WL_sewer2, WL_sewer4, and WL_sewer5 show that these measurement values contain extreme values that result in a very distorted range of the histogram. The histograms of sensors WL_sewer6 and WL_sewer7 show a distinction between measurement values to the left and to the right, which can be explained as a result of dry weather and rainy weather.

The histogram of sensor WL_sewer3 also shows a somewhat distorted range, which is the result of extreme values to the right of the mass of the measurement values.

Furthermore, the histograms of sensors WL_surface2 – 6 show that these measurement data might be normally distributed, with the exception of the histogram of sensor WL_surface5. The histogram of WL_surface2 also shows a distorted range, when compared to the other sensors of this type. Examining the range of the measurement data of WL_surface2, it can be seen that there are some measurement values greater than 0.7 where the other sensors of this type do not have such extreme values. The measurement data of WL_surface5 can be seen to be split into two parts. Of these two parts, the right part was earlier on in the report seen to be correct and the left part being biased measurements.

Looking at the range of this data and comparing this to the range of the other five sensors of this type, it can be seen that the right part of the measurement data is correct and the left part being biased measurements.

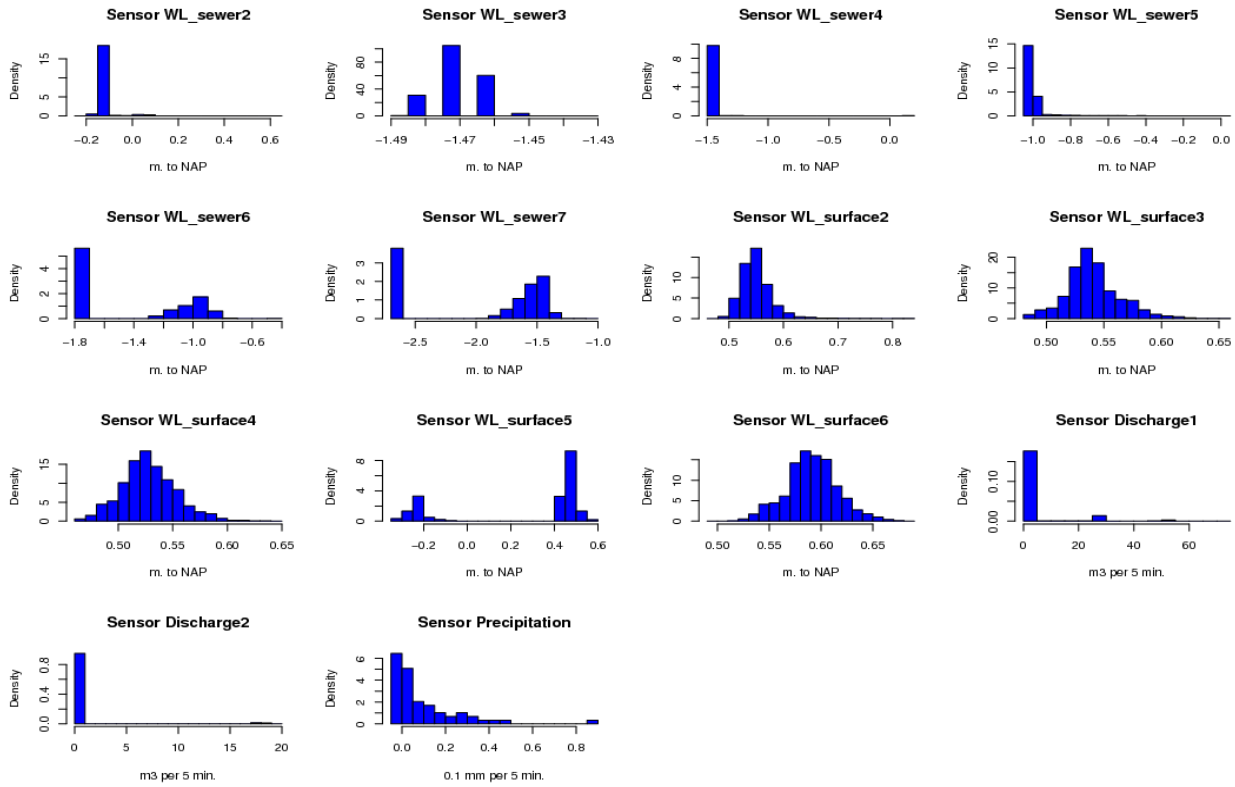


Figure 5: Histograms of all sensors

Next to this, the histograms (and in particular the range) of the two discharge sensors reveal that the measurement data of these sensors also contain extreme values. There is a huge amount of measurements to the left of the data and a small number of extreme values to the right. This is the result of a pump being switched on or off. Here, the extreme values of sensor Discharge1 are even bigger than those of sensor Discharge2. Furthermore, the histogram of sensor Precipitation might be exponentially distributed, except for the fact that there are (again) extreme values to the right of the mass of the measurement values.

2.4.3 Box plots

Where histograms provide a graphical image of the data, box plots provide a combination of a graphical and numerical summary. A box plot is a convenient

way of graphically depicting the minimum, the lower quartile (which cuts off the lowest 25% of the data), the median, the upper quartile (which cuts off the highest 25% of the data), and the maximum. A box plot thus provides information about possible extreme values in the data. Measurements that are smaller than the lower (or first) quartile plus one and a half times the inter quartile range (the upper quartile minus the lower quartile) or measurements that are larger than the upper (or third) quartile plus one and a half time the inter quartile range are marked as an extreme value and therefore marked separately. Furthermore, the median can be read out of the plot, with which the skewness of the measurement data again can be seen.

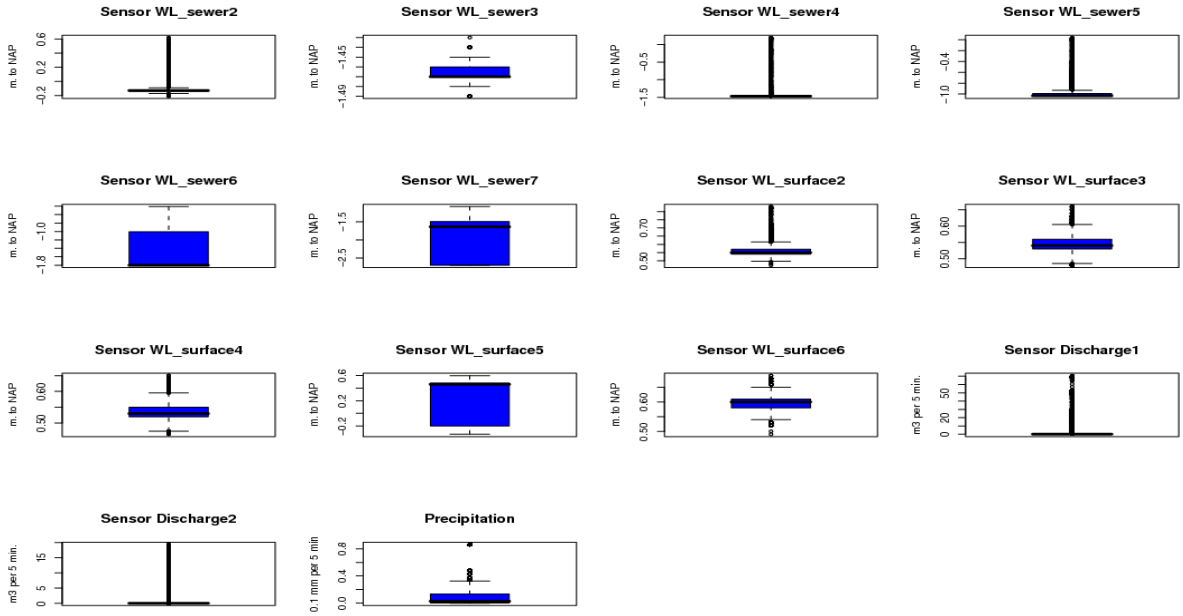


Figure 6: Box plots of the fourteen sensors

Figure 6 shows the box plots for all sensors. The box plots for sensors WL_sewer2, WL_sewer4, and WL_sewer5 show an enormous amount of extreme values, which was already noticed by means of the histograms. The box plot of sensor WL_sewer7 shows a few extreme values to the left and right, of which the extreme values to the left were not noticed by the histograms. Furthermore, the histograms of sensors WL_sewer6 and WL_sewer7 could be seen to be split up into two parts, of which it can now be seen that the median of WL_sewer6 is to the left of the data and the median of WL_sewer7 is to the right. This could already be concluded by the positive and negative (respectively) skewness, but is now made visible by these box plots.

Furthermore, the box plots of sensors WL_surface3, WL_surface4, and WL_surface6 can be seen to behave in a similar way; some extreme values to the left and right. The exception to the rule again is the box plot of sensor WL_surface5, which has no extreme values according to the box plot and for which holds that the mass of the measurement values is to the right. Furthermore, it can be seen that the measurement data of sensor WL_surface2 contains more extreme values than the measurement data of the other sensors of this type.

The results of the discharge sensors show a huge amount of extreme values, of which the extreme values of Discharge1 can be seen to be on a far greater scale than those of Discharge2. These extreme values are most interesting and will therefore be the subject of further analysis. The box plot of sensor Precipitation can be seen to have some extreme values to the right of the measurement data. This is logical because one can think of cases of extreme rainfall, which occur on a rare basis. These cases will result in extreme values to the right and since a negative precipitation is not possible, no extreme values will occur to the left of the measurement values.

2.5 Conclusions

Initial plots show that an extreme value for the precipitation does indeed in most cases lead to an extreme value in the data of sensors WL_sewer2, WL_sewer4, and WL_sewer5. Histograms of the data of these sensors showed that the data can be divided into two groups, one group of “stationary” measurements and one group of “real” measurements. With “real” measurements is meant that these measurements seem to be the result of precipitation.

The other three sensors of this type (water level in the sewer network) seem not to be effected by the quantity of downpour. Sensor WL_sewer3 is not effected at all, it can almost be said to measure a constant value. The data of sensors WL_sewer6 and WL_sewer7 both show a gap in the data which indicates that the measurement values of these sensors are heavily influenced by pumps, or water flowing over an overflow. This also holds for the measurement data of sensors Discharge1 and Discharge2, which are placed at the same locations respectively.

The behaviour of the measurement data of the sensors that measure the water level of the surface water is in accordance with each other. At least, when the measurement data of sensor WL_surface5 is disregarded from the twelfth of February on. The measurement values of this last sensor make a drop around

this date, but from that date on continue to operate completely in accordance with the other sensors of this type (water level of the surface water). One explanation for this drop could be that the sensor is replaced, which does not have any effect on the operation of the sensor. But, it does have an effect on the measurement values. In addition to this, the precipitation does have effect on the measurement data of these five sensors.

3 Density

In this chapter a study will be done to try and discover the underlying probability distribution function of the measurement data of the fourteen sensors. During this study, numerous plots will be used to obtain a first impression of the data. After that, multiple tests will be used to test the hypothesis that the measurement values of a certain sensor are distributed according to a certain well known probability distribution function.

When a probability distribution function can be discovered in the measurement data of a sensor, numerous calculations can be made on the data under the assumption that the data is distributed according to this probability distribution function. Furthermore, when a probability distribution can be assigned to the measurement data of, say, two vectors, it can also be said that the measurement data of these two sensors behaves accordingly to each other. This can be of great importance to the correlation study and thus to the choice of which sensors to place in the group that share a high correlation and which sensors in the other group.

3.1 Introduction

A first impression of the probability distribution function of measurement data can be obtained by means of histograms. Since these are already added to this report in paragraph 2.4.2, only the results will be given here.

What can be discovered by the histograms is that the measurement data of sensors WL_sewer2 – 7 cannot be assigned to any well known probability distribution.

Furthermore, the histograms of sensors WL_surface2 – 6 seems to be normally distributed with the exception of sensor WL_surface5. This histogram is split into two parts. These two parts were seen to be in accordance with the other sensors of this type, where the second (right) part has a much lower mean value. If this is correct, both parts of the measurement data of this last sensor should also be normally distributed. In addition to this, the histogram of the measurement data of sensor WL_surface2 seems to be normally distributed except for the measurements greater than 0.7, which distort the range.

Next to that, the only probability distribution that can be discovered from the histograms is that of an exponential distribution when looking at the histogram of the precipitation.

3.2 QQ-plots

To get a first impression of the probability distribution function of the measurement values of the fourteen sensors, QQ-plots are made. A QQ-plot is a method to judge whether a sample comes from a certain known distribution function. Suppose $x_1,....,x_n$ are independent replicas of a probability distribution *F*. De *i*-th order statistic $x_{(i)}$ will then approximately have a fraction of *i/(n+1)* of the measurements below. It is therefore approximately the *i/(n+1)*-quartile of the measurements. It is thus expected that the points

$$\left\{\left(F^{-1}\left(\frac{i}{n+1}\right), x_{(i)}\right) : i = 1,....,n\right\}$$

are approximately on a straight line. A QQ-plot is a plot of these *n* points. Therefore, if the plot approximately shows a straight line *y=x*, it can be assumed that this sample comes from the distribution function *F*.

The previous paragraph described that based on the histograms, the measurement data of sensors WL_surface2, WL_surface3, WL_surface4, WL_surface5, and WL_surface6 seem to be normally distributed. It should be noted that the measurements of sensor WL_surface5 need to be split up into two parts: all measurements smaller than 0.2 and the remaining part of the measurements. QQ-plots will be made according to this.

Figure 7 gives the QQ-plots for these seven sensors, of which the measurement data of sensor WL_surface5 is split into two parts, for which two QQ-plots are made. Next to that, only the measurement values of sensor WL_surface2 smaller than 0.7 are taken into account. All QQ-plots results in an approximate straight line, which indicates that the measurement values (or parts thereof) of these sensors indeed might be normally distributed. In addition to this, it can thus be assumed that these sensors generate similarly distributed data. This thus indicates a strong dependency between these sensors.

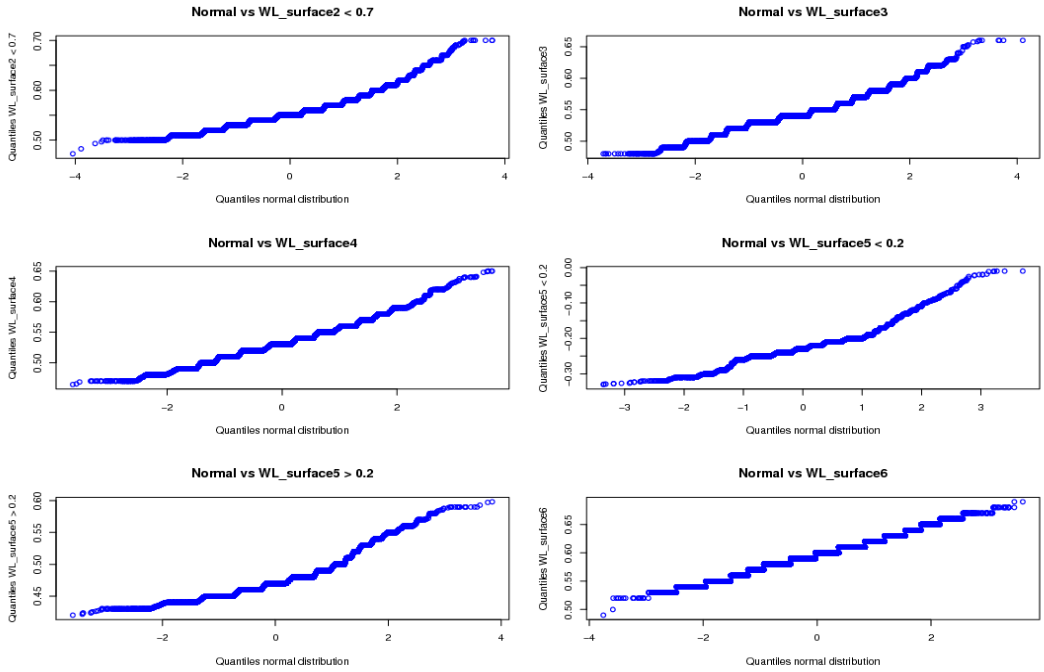


Figure 7: QQ-plots of the seven sensors

Of sensors WL_sewer2 – 7 and the discharge sensors, the measurement data was also split into two parts, again the so-called "stationary" measurement values and the "real" measurement values. This "real" part (or right part) of the measurement values was also tested for normality by means of a QQ-plot, but no convincing plots resulted from this. No normality of these measurements can thus not be assumed based on the QQ-plots.

In addition to this, a plot will be made of the quartiles of an exponential distribution against the quartiles of the data of the precipitation. First, the data is rescaled such that the minimum of the measurement data is equal to zero, thus assuming the measurement values of the Precipitation sensor are biased. Furthermore, the measurement values of this sensor that are greater than 0.6 are also removed.

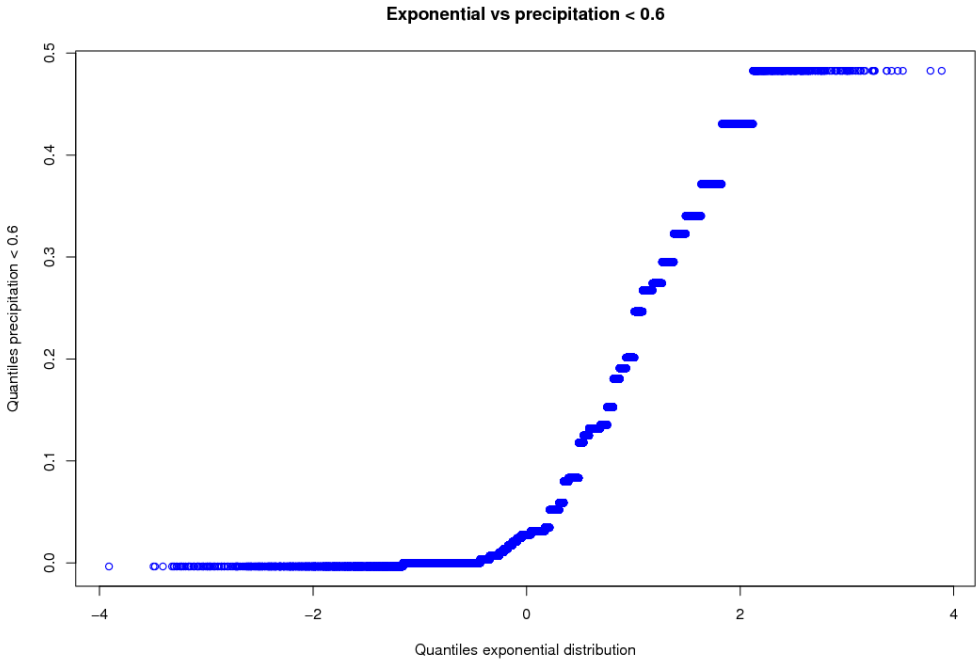


Figure 8: QQ-plot exp vs. Precipitation

Figure 8 shows the resulting QQ-plot, which cannot be said to result in a straight line. Based on the QQ-plot, it can thus be said that the measurement data of the sensor Precipitation are not exponentially distributed.

3.3 Pairs

The previous paragraphs showed that the measurement data of five sensors might be normally distributed. An additional graphical way to test this is by means of creating scatter plots of the quartiles of the measurement values of one sensor against the quartiles of the measurement values of another sensor. If the measurement values of the two sensors result approximately in the straight line $y = a + bx$, it can be assumed that these two sensors are from the same location-scale family. The function in the statistical program R that creates all combinations of QQ-plots given the measurement values of sensors is called Pairs, hence the name of this paragraph.

So, when the measurement values of the surface water level sensors are indeed normally distributed (and thus originate from the same location-scale family), a plot of the measurement values of one sensor against the measurement values of one of the other sensors, should approximately result in the straight line $y = a +$

bx. Since the measurement data of sensor WL_sewer6 contains a gap and thus a lot of measurements are missing, it is of no use to apply this function for this sensor.

Therefore, the function is only used for the four sensors that measure the water level of the surface water, since these are all of the same length. In addition to this, the behaviour of the measurement values of these four sensors seems to be in accordance with each other, i.e., they contain extreme values at approximately the same dates.

To take into account possible time delays between the measurement values of different sensors, the sums of the measurements are taken over four hours. This also means that the higher number of extreme values in the measurement values of sensors WL_surface2 are smoothed.

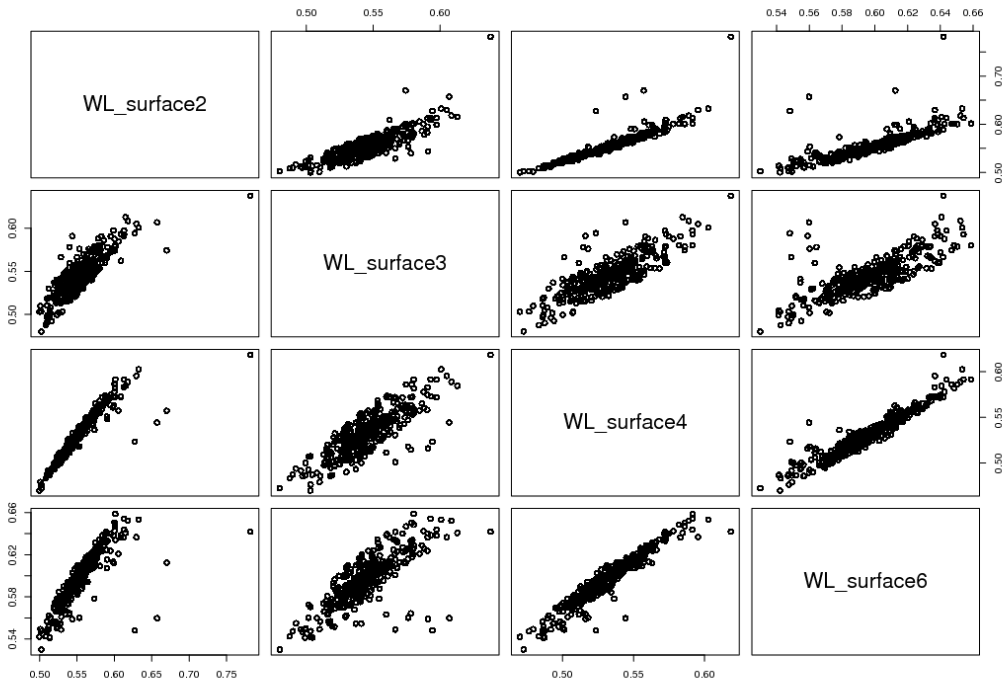


Figure 9: Plots of four surface water level sensors

Figure 9 shows that the plots of all different combinations results in approximately a straight line, which indicates that the (four hour sums of the) measurement values of these four sensors originate from the same location-scale family.

3.4 Tests

The previous paragraphs have shown some form of a distribution for some of the sensors. In this paragraph the measurement values of the sensors will be tested for normal or exponential distributions. To test this, some well known statistical tests will be performed.

For these tests, the following hypotheses are thus formulated:

H_0: the measurement values are normally distributed
H_1: the measurement values are not normally distributed

The null hypothesis is rejected at a p-value of $\alpha = 0.05$ and these test are performed on five of the fourteen sensors, namely: WL_surface2, WL_surface3, WL_surface4, WL_surface5 (measurement values < 0.2), WL_surface5 (measurement values > 0.2), and WL_surface6.

The Shapiro-Wilk test tests the null-hypothesis that a sample came from a normally distributed population. The test calculates a *W* statistic, of which small values are evidence of departure from normality. The results of the Shapiro-Wilk test showed that the data of none of the above mentioned sensors could be assumed to be normally distributed. The test resulted in very small p-values, but in some cases in very large values for the statistic. To gain some certainty on the rejection of the null-hypothesis, an additional test was done, namely the Kolmogorov-Smirnov test.

The Kolmogorov-Smirnov test is based on the biggest difference D_n between the empirical distribution function and the distribution function of the distribution that is tested for. The null-hypothesis is rejected for small values of D_n. This test proved not to be applicable since no correct p-values could be computed because of ties in the data.

Eventually, the Jarque-Bera test was used. This test is a goodness-of-fit measure of departure from normality, based on the sample kurtosis and skewness. The null hypothesis is a joint hypothesis of both the skewness and excess kurtosis being 0, since samples from a normal distribution have an expected skewness of 0 and an expected excess kurtosis of 0. Any deviations from this increases the statistic of this test.

After applying this last test to the data of the five sensors, no normality can be assumed of the data of any of the five sensors based on this test. Table 4 gives the results of the three tests, where the results of the Jarque-Bera test are given

by a value of 1 when no normality can be assumed based on this test, and a value of 0 when normality can be assumed based on this test.

	Shapiro-Wilk		Kolmogorov-Smirnov		Jarque-Bera
	Statistic	P-value	Statistic	P-value	Output
WL_surface2	0.8692	2.20E-16	0.9955	0.275	1
WL_surface3	0.9728	1.09E-15	1	0.2701	1
WL_surface4	0.9884	3.45E-09	1	0.2701	1
WL_surface5 (< 0.2)	0.9312	6.81E-13	1	0.2701	1
WL_surface5 (> 0.2)	0.9151	2.20E-16	1	0.2701	1
WL_surface6	0.9936	9.10E-06	1	0.2701	1

Table 4: Test statistics

3.5 Conclusions

Of the sensors that measure the water level of the surface water, the idea existed that these were normally distributed. Of one of these sensors (WL_surface5) the data was split into two parts, because previous plots learned that the measurements of this plot makes of drop and continues on measuring from that point on. Therefore, these two parts of the measurement data of this sensor were examined individually. QQ-plots that were made of the quantiles of the data against the quantiles of a normal distribution resulted in an approximated straight line, which indicates normality of the data. Eventually, the Jarque-Bera test pointed out that the measurement data of these sensors can not be assumed to be normally distributed.

Furthermore, the measurement data of some of the sensors that measure the water level in the sewer network (WL_sewer2, WL_sewer3, WL_sewer4, WL_sewer5, WL_sewer6, and WL_sewer7) and the sensors that measure the discharge of the water (Discharge1 and Discharge2), the measurement data was split into two parts. Namely, one “stationary” part of which the measurements are the result of dry weather, and one “real” part of which the measurements are the result of rainy weather. The “real” part was investigated for normality, but after examining plots of the quantiles of this data against the quantiles of a normal distribution, it became clear that normality could not be assumed.

Of the measurement data of the precipitation it was suggested by the histograms that these might be exponentially distributed. However, QQ-plots made it clear that this was not the case. Even after removing the extreme values, an exponential distribution could not be assumed. It has to be noted here, that the

measurement values of this sensor were first rescaled to have a minimum value of zero, assuming that the measurements are biased.

The measurement data of these fourteen sensors turns out to be too distorted to be approached by one of the known probability distributions. The numerical summary in paragraph 2.3 already learned that the values of the skewness and kurtosis of these measurement data are in many cases very high. Next to that, the number of extreme values in the data is also shown to be very high. This are most probably the reasons no known probability distribution can be assumed. Especially for the data that was believed to be normally distributed: for a normal distribution it holds that the skewness of a normal distribution is equal to zero and the kurtosis of a normal distribution is equal to three, of which the data of many sensors show a (huge) deviation.

Furthermore, considering the amount of measurements (16992 measurements of each sensor) it might not even be that peculiar that the statistical tests reject normality of the data. Since with such a huge amount, there will always be deviations from a normal distribution. However, what might be more important is the fact that a strong relation was shown between the measurements of the sensors that measure the water level.

4 Correlation

The correlation of two vectors $x = (x_1,....,x_n)$ and $y = (y_1,....,y_n)$ is classically calculated by Pearson correlation coefficient or Spearman rank coefficient. The Pearson correlation coefficient is a measure of the correlation of two variables *X* and *Y*, that is, a measure of the tendency of the variables to increase or decrease together. It is defined as the sum of the products of the standard scores of the two measures divided by the degrees of freedom:

$$r = \frac{\sum z_x z_y}{n-1}$$

Spearman's rank correlation coefficient is a non-parametric measure of correlation. Unlike the Pearson correlation coefficient, it does not require the assumption that the relationship between the variables is linear, nor does it require the variables to be measured on interval scales. In principle, ρ is simply a special case of the Pearson coefficient in which the data are converted to rankings before calculating the coefficient.

In this case, calculating the correlation between two vectors of measurements would not reveal all dependencies. Water can enter the sewer network by an enormous amount of locations. The water then flows through the network to so-called water purification plants. So from the point of entry to the water purification plant, the water comes by multiple sensors. This means that when the water level at the location of sensor one rises at some time point, the water level at the location of sensor two also rises, but with a given time delay.

To take into account a possible time delay between measurements of two sensors, the measurements on the basis of a certain time interval will be used. If there exist a time delay of, for instance, two hours between the measurements of two sensors, this should result in a higher correlation when this correlation is calculated based on the measurements on a two hour time interval. Therefore, the correlation will be calculated based on all possible time intervals of two sensors from a time interval of one hour up to a time interval of 48 hours. The following paragraph will describe all steps taken and all results of this study. In order to reduce the enormous calculation time in R, the measurement data used in this chapter is based on hourly values instead of the original five minute values.

4.1 Sums

The first step in this study is to examine all different time intervals, i.e. first the correlation of two vectors (with each the measurement data of one sensor) is calculated on a one hour basis, then on a two hours basis, up to a 48 hours basis (since it can be assumed that under normal conditions all water in the sewer network is flushed out in this time period). Furthermore, the correlation of these two vectors will also be calculated based on all different possible combinations of time intervals, i.e. the following table will provide more clarity on this last statement.

Vector 1	Vector 2
1 hr. t.i.	1 hr. t.i.
1 hr. t.i.	2 hrs. t.i.
1 hr. t.i.	3 hrs. t.i.
…	…
…	…
1 hr. t.i.	48 hrs. t.i.

The first row in this table states that the correlation of two vectors is calculated, where both vectors are based on measurements on a one hour base (or 1 hrs. t.i.). The second row states that the correlation of two vectors is calculated, where the first vector contains measurements on a one hour base and the second vector contains measurements on two hours base. This process is repeated up to the last row where the correlation is calculated of two vectors where the first vector contains measurements on a one hour base and the second vector contains measurements on a 48 hours base.

The next step is given by the following table, where this whole process is repeated but then the first vector is on a two hours base and the second is on a one hour base for the first calculation of the correlation and again up to a 24 hours base for the last correlation.

Vector 1	Vector 2
2 hrs. t.i.	1 hr. t.i.
2 hrs. t.i.	2 hrs. t.i.
2 hrs. t.i.	3 hrs. t.i.
…	…
…	…
2 hrs. t.i.	48 hrs. t.i.

This process is repeated over and over again until the first vector is on a 48 hours base and the second vector on a one up to 48 hours base. This way the correlation of two vectors based on all possible combinations of time intervals between one hour and 48 hours is calculated. This process is off course done for all possible combinations of vectors.

What has to be taken into account here, is that in the last step the correlation of two vectors is calculated where for both vectors the measurements are taken based on a 48 hours base. This means that the length of these vectors of measurement values is then 24 times shorter than the original vectors of measurement values based on one hour. The shorter the two vectors become, the higher the correlation will be, i.e. the correlation between two vectors both of length two will always be one. This means the behaviour of the correlation has to be examined while increasing the time intervals.

After all combinations have been generated and the correlation between these are calculated, three different cases come to light:

- The first case is the case where the maximum correlation is obtained due to true dependency between the vectors

- The second case is the case where the maximum correlation is obtained due to the (shorter) length of both vectors, i.e. the maximum correlation is obtained based on measurements on a 48 hours basis for both vectors

- The third case is the case where the maximum correlation is obtained due to a time interval of 48 hours for one of two vectors.

The cases of the first type need no further study, since a maximum correlation is obtained within a time interval of 48 hours. The cases of the second type need no further study, since it is evident that the correlation is increasing due to the decreasing length of the vectors with measurements.

The cases of the third type do need some additional calculation, since these cases fall right between the other two cases. Therefore, for these cases the whole process is repeated for time intervals up to 96 hours. The results show that some of the cases of the third type have now become cases of the first type.

The last step is to take into account a time shift. For this step, the same steps are taken as described above, i.e. all combinations of possible time intervals will be examined. But in addition to this, these combinations will be examined with a time shift between the two vectors of measurement values. To explain this, the table given above will be used. The first step is now to calculate the correlation of two vectors of which the first vector is on a one hour base and the second

vector is on a one hour base, but of this second vector, the measurements are taken of one hour later. For the second step, the measurements of the second vector are taken of two hours later, up to 24 hours later.

This process is repeated over and over again until the first vector is on a 48 hours base and the second vector on a one up to 48 hours base, with a time shift of one hour up to a time shift of 24 hours. This way the correlation of two vectors based on all possible combinations of time intervals between one hour and 48 hours is calculated and with a time shift between one and 24 hours. The results are given in Appendix A.

On the diagonal are the sensors that were studied. Above the diagonal is the increase in correlation that was obtained rounded of to two decimals. Under the diagonal the behaviour of the correlation is described in pictograms. When the correlation of two vectors increases after the first step (a case of type one), this is depicted by a green arrow. When the correlation of two vectors results in a case of type two, this is depicted by a red cross. Finally, when the correlation of two vectors results in a case of type three, this is depicted by a question mark. The latter cases were examined further for time intervals up to 96 hours. When the correlation of two vectors during this second step results in a case of type one, this is depicted as a green arrow (which is depicted after the earlier resulting question mark). When the correlation of two vectors during this second step results in a case of type two or three, this is depicted by a red cross.

It can be seen that in a lot of cases, the correlation of two vectors of measurement values can be increased. In one case, the correlation even increases with 0.36, which is the correlation of the measurement vectors of the two discharge sensors. The cases in which an increase in correlation of the measurements of two sensors can be obtained can indicate that these are cases where there is a physical connection between these two vectors. However not al obtained “winnings” in correlation are very high. The final correlations of the measurement values of all sensors can be found in Appendix B.

The conclusions on the previous chapter already learned that there is a strong dependency among four of the sensors that measure the water level of the surface water, i.e. WL_surface2, WL_surface3, WL_surface4, and WL_surface6. This strong dependency can no be seen from the correlations which are given in Appendix B. Furthermore, the measurement values of sensors WL_sewer2 and WL_sewer5 also share a high correlation. This was also seen in chapter 2, since the behaviour of the measurement values of these sensors was in accordance with each other. In addition to this, there is a high dependency between the measurement values of sensors WL_sewer6 and WL_sewer7, which already showed to behave in accordance with each other in chapter 2 where it was also shown that the measurement values of the two

discharge sensors is in accordance with each other. What went unnoticed before is the dependency between the measurement values of sensor WL_surface5 and WL_sewer3 and WL_sewer3 and WL_sewer5.

4.2 Spearman

For this part of the study, the rank correlation test of Spearman will be used to test the dependencies among the measurement vectors for significance. First the dependencies between measurement vectors will be tested as there are no time intervals or time shifts used. After that, the dependencies between the measurement vectors will be tested as they are calculated in the previous paragraph. The rank correlation test of Spearman is most interesting for this study, since this test is based on the ranks of the measurement values and it therefore does not require the assumption that the relationship between two vectors of measurement values is linear.

For this test, the following hypotheses are formulated:

H_0: X_i en Y_i are independent
H_1: X_i en Y_i are dependent

The null hypothesis is rejected when the bootstrap approach of the left p-value is smaller than $\alpha/2 = 0.025$. The results are given in Appendix C. It can be seen from the results that for three combinations of measurement vectors, the dependency was at first not significant, but after applying the time intervals and time shifts, the dependency is significant.

Furthermore, it is noticed that all sensors that measure the water level of the surface water share a dependency that is significant and have a high correlation. This could also be concluded from Figure 3. Next to that, the sensors at the same location (i.e. WL_sewer6 and Discharge1 and WL_Sewer7 and Discharge2) also share a dependency that is significant. Of the dependency between the sensors of type W nothing can be said, i.e. no hard conclusions can be made. The same holds for other combinations of types of sensors, with the exception of the discharge sensors, that seem to be have a significant dependency more often with sensors of type W, than of type Z.

4.3 Choice

The eventual choice for which sensors should belong in the two groups for the remaining part of this internship is based on the results obtained in the previous chapters. The goal is to create two groups of eight sensors where the measurement values of one group share a high correlation and one group where the sensors do not.

It is obvious that the four sensors that measure the water level of the surface water should end up in the group that share a high correlation, since all results point out that these are very dependent on each other. The same holds for the sensors WL_sewer2 and WL_sewer5, of which the strong dependency was also seen throughout this report. Furthermore, what was not seen throughout this report, but came to light only in the last chapter, is the strong dependency between sensors WL_sewer3 and WL_sewer5 and again for WL_sewer3 and WL_surface5. Next to that, the precipitation was seen to be responsible for most of the extreme values seen in the plots that were made in the second chapter. The extreme values that were measured were seen to be the result of precipitation in most of the cases. Since all three Data Mining tasks at hand in the next part of the internship are based on regression techniques, precipitation may very well be used as an explanatory variable. It is therefore decided that the sensor Precipitation is the eight, and therefore last, sensor to be added to the group of sensors that share a high correlation.

With this in mind, the following two groups are now created:

Group 1							
WL_sewer2							
0,13	**WL_sewer3**						
0,50	0,11	**WL_surface2**					
0,44	0,06	0,84	**WL_surface3**				
0,67	0,08	0,62	0,50	**WL_sewer5**			
0,24	0,11	0,90	0,83	0,30	**WL_surface4**		
0,13	0,08	0,79	0,77	0,12	0,95	**WL_surface6**	
0,48	0,03	0,57	0,49	0,51	0,33	-0,23	**Precipitation**

Group 2							
WL_sewer1							
-	**WL_surface1**						
-	-	**WL_sewer4**					
-	-	0,08	**WL_sewer6**				
-	-	0,12	0,52	**Discharge1**			
-	-	0,02	0,67	0,30	**WL_sewer7**		
-	-	0,18	0,20	0,64	0,28	**Discharge2**	
-	-	0,09	0,72	0,30	0,49	0,12	**WL_surface5**

These two groups will now be used for the three planned tasks for the remaining part of this internship and to study, among others, the differences in obtained results of these three tasks when sensors share a high(er) correlation and when sensors share a low(er) correlation.

Part III Modelling, Identification & Evaluation

5 M5P

The following two chapters will describe the application of two Data Mining tasks for numeric prediction. In numeric prediction the outcome of an attribute instance is written as a linear sum of the other attribute instances with appropriate weights. So for this internship, the outcome of a measurement of one sensor is written as a linear sum of the measurement values of the other sensors with appropriate weights. (Confidential)

In this chapter the M5P algorithm will be applied to the measurement data. The M5P algorithm is used for numeric prediction. It uses a decision tree, except that at each node it stores a linear regression model that predicts the class value of instances that reach the leaf. The model tree is constructed by first using a decision tree induction algorithm to build an initial tree. First, the technique of linear regression will be explained, followed by an explanation of the M5P algorithm.

5.1 Linear regression

In linear regression the relationship is modelled between a dependent variable Y and independent variables $X_0, X_1, ..., X_n$ [1]. The dependent variable is often called the response variable and the independent variables the explanatory variables. The idea is to express the response variable as a linear combination of the explanatory variables with predetermined weights:

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_n X_n$$

where Y is the response variable, $\beta_0, \beta_1, ..., \beta_n$ are the weights and $X_0, X_1, ..., X_n$ are the explanatory variables. The mostly used technique to calculate the weights is called the least-squares analysis, which is also used for the currently used models.

The weights are calculated from the training data. Suppose that the predicted value for the *i-th* instance of the response variable is given by the following notation:

$$\beta_0 + \beta_1 X_{i1} + ... + \beta_n X_{in}$$

Moreover, for notation purposes it is convenient to add an extra explanatory variable X_0 whose value is always 1. Then the predicted value for the *i-th* instance is given by:

$$\beta_0 X_{i0} + \beta_1 X_{i1} + ... + \beta_n X_{in} = \sum_{i=1}^{n} \beta_i X_i \, .$$

The method of least-squares analysis is to choose the weights $\beta_0, \beta_1, ..., \beta_n$ to minimize the sum of the squares of the differences between the predicted values and the actual values. The sum of the squares of these differences is defined by:

$$\sum_{j=1}^{k} \left(Y_j - \sum_{i=1}^{n} \beta_i X_{ji} \right)^2 .$$

This sum of squares is what is to be minimized. This results in a set of weights, based on the training data, which can be used for the prediction of a class of new instances.

Choosing the explanatory variables is the next difficulty. The goal is to obtain a good model with as few explanatory variables as possible. One way of building a model is to add explanatory variables one by one, where adding a variable has to meet certain criterions. Another way of building a model is to start with the model with all explanatory variables and to remove them one by one according to certain criterions.

By means of the correlation coefficient, different models can be compared and the quality of the model can be judged. The correlation coefficient is defined as follows:

$$\Re^2 = \frac{SS_{reg}}{SYY} = \frac{SYY - RSS}{SYY},$$

where

$$RSS = (Y - \beta X)^T (Y - \beta X)$$

and

$$SYY = \left(\sum_{j=1}^{k} Y_j - \bar{Y} \right).$$

The correlation coefficient is a measure of the overall quality of the model. It holds that $0 \leq \Re^2 \leq 1$. The closer $\Re^2$ is to 1, the better the model is.

5.2 M5P

The M5P algorithm uses a decision tree for numeric prediction. Trees used for numeric prediction are just like ordinary decision trees except that at each leaf they store either a class value that represents the average value of instances that reach the leaf, in which case the tree is called a regression tree, or a linear regression model that predicts the class value of instances that reach the leaf, in which case it is called a model tree [2].

The splitting criterion is used to determine which attribute is the best to split that portion T of the data that reaches a particular node. It is based on treating the standard deviation of the output of the response variable in T as a measure of the error at that node, and calculating the expected reduction in error as a result of testing each attribute at that node. The attribute which maximizes the expected error reduction is chosen for splitting at the node. So, for each attribute (explanatory variable) and for all possible split positions of this attribute, the standard deviation reduction (SDR) is calculated. The attribute for which the maximum SDR is obtained is split at this node and at the split position where this attribute obtained its maximum SDR.

The expected error reduction is calculated by:

$$SDR = sd(T) - \sum_i \frac{|T_i|}{|T|} \times sd(T_i),$$

where $T_1, T_2, \ldots$ are the sets that result from splitting the node according to the chosen attribute. The splitting process terminates when the class values of the instances that reach a node vary only slightly, that is, when their standard deviation is just a small fraction (e.g. less than 5%) of the standard deviation of the original instance set. Splitting also terminates when just a few instances remain, say, four or fewer. Experiments show that the results obtained are not very sensitive to the exact choice of these thresholds.

When a model tree is used to predict the value for a test instance, the tree is followed down a leaf in the normal way, using the instance's attribute values to make routing decisions at each node. The leaf will contain a linear model based

on some of the attribute values, and this is evaluated for the test instance to yield a raw predicted value. Instead of using this raw value directly, however, it turns out to be beneficial to use a smoothing process to compensate for the sharp discontinuities that will inevitably occur between adjacent linear models at the leaves of the pruned tree. This is a particular problem for models constructed from a small number of instances. Smoothing can be accomplished by producing linear models for each internal node, as well as for the leaves, at the time the tree is built. Then, once the leaf model has been used to obtain the raw predicted value for a test instance, that value is filtered along the path back to the root, smoothing it at each node by combining it with the value predicted by the linear model for that node. An appropriate smoothing calculation is:

$$p' = \frac{np + kq}{n + k},$$

Where *p'* is the prediction passed up to the next higher node, *p* is the prediction passed to this node from below, *q* is the value predicted by the model at this node, *n* is the number of instances that reach the node below, and *k* is a smoothing constant. Experiments have shown that smoothing substantially increases the accuracy of predictions.

As is mentioned, a linear model is needed for each interior node of the tree, not just at the leaves, for use in the smoothing process. Pruning is a method for reducing the error and complexity of induced trees. Prior to pruning, a model is calculated for each node of the un-pruned tree. The model takes the form:

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k ,$$

where $X_0, X_1, ..., X_n$ are attribute values. The weights $\beta_0, \beta_1, ..., \beta_n$ are calculated using standard regression. The pruning procedure makes use of an estimate, at each node, of the expected error for test data. First, the absolute difference between the predicted value and the actual class value is averaged over each of the training instances that reach that node. Because the tree has been built expressly for this dataset, this average will underestimate the expected error for unseen cases. To compensate, it is multiplied by the factor *(n+v)/(n-v)*, where *n* is the number of instances that reach the node and *v* is the number of parameters in the linear model that gives the class value at that node. The expected error for test data at a node is calculated as described above, using the linear model for prediction. Because of the compensation factor *(n+v)/(n-v)*, it may be that the linear model can be further simplified by dropping terms to minimize the estimated error. Dropping a term decreases the multiplication factor, which may be enough to offset the inevitable increase in average error

over the training instances. Terms are dropped one by one, so long as the error estimate decreases.
Once a linear model is in place for each interior node, the tree is pruned back from the leaves, as long as the expected estimated error decreases. If the expected error of a node is smaller than the expected error of the sub-tree below, the sub-tree is replaced by this single node.

```
+-------------------------------------------------------------------------------------+
|                                                                                     |
|       Attribute1 <= Criterion1                                                      |
|       |       Attribute1 <= Criterion2  :  LM1                                      |
|               |       Attribute1 >  Criterion2                                      |
|               |       |       Attribute2 <= Criterion3  :  LM2                      |
|       |       |       Attribute2 >  Criterion3  :  LM3                              |
|       Attribute1 > Criterion1                                                       |
|       |       Attribute3 <= Criterion4  :  LM4                                      |
|       |       Attribute3 >  Criterion4  :  LM5                                      |
|       .                                                                             |
|       LM1  :  Attribute4 = Constant + a*Attribute1                                  |
|       LM2  :  Attribute4 = Constant                                                 |
|       LM3  :  Attribute4 = -Constant + c*Attribute7                                 |
|       LM4  :  Attribute4 = Constant - d*Attribute6                                  |
|       LM5  :  Attribute4 = Constant + f*Attribute1                                  |
|                                                                                     |
+-------------------------------------------------------------------------------------+
```

Figure 10: Example M5P model tree

Figure 10 gives an example of an M5P model tree. Note here that attribute1 does not necessarily mean it is the first attribute of the data set.

To train and test the models, cross validation will be used. Cross validation is one of several approaches to estimating how well the model you've just learned from some training data is going to perform on future as-yet unseen data. There are several approaches of cross validation, among which test-set-validation, leave-one-out cross validation, and k-fold cross validation, of which the last will be used during the data mining tasks.

In k-fold cross validation, the data set of *N* samples is divided into *k* subsets ("folds") of equal size *N/k* and *k* models are built. Each time, one of the *k* subsets is used as the test set and the other *k-1* subsets are put together to form a training set. A common choice for the number of folds *k* is ten, which will also be used for the tasks at hand.

The results will eventually be compared to the results obtained with the current models. These results will be discussed in chapter 7. This chapter will describe the M5P algorithm in detail and next to that, it will discuss the results of this algorithm and all steps taken to achieve these results.

5.3 Results

(Confidential)

5.4 Summary

(Confidential)

6 Robust Regression

This chapter will first describe the algorithm of Robust Regression. This algorithm was applied to the measurement data of the fourteen sensors. The results of this application will be given in this chapter, together with all steps taken to obtain these results.

6.1 Algorithm

Noisy data is known to cause problems in linear regression. Therefore, statisticians often check data for outliers and remove them manually. In the case of linear regression, outliers can be identified visually, although it is never completely clear whether an outlier is an error or just a surprising, but correct, value. Outliers have a dramatic effect on the usual least-squares regression, because the squared distance measure accentuates the influence of points far away from the regression line.

Statistical methods that address the problem of outliers are called robust. One way of making regression more robust is to use an absolute-value distance measure instead of the usual squared one. This weakens the effect of outliers, but gross outliers can still have a considerable impact on the model. Another possibility is to try to identify outliers automatically and remove them from consideration. For example, one could form a regression line and then remove from consideration those 10% of points that lie farthest from the line. However, in chapter 5 these points were seen to be very important in modelling the data. A third possibility is to minimize the median, rather than the mean, of the squares of the divergences from the regression line, which will be applied to the measurement data.

Formally, the Least Median Squares (LMS) fit is determined by solving the following optimization problem:

$$\min_{b_0 b_1} medSR_i = Median\left\{(Y_1 - (\beta_o + \beta_1 X_1))^2, (Y_2 - (\beta_0 + \beta_2 X_2))^2, ..., (Y_n - (\beta_o + \beta_1 X_1))^2\right\}.$$

Since LS (Least Squares) is based on minimizing the sample mean and means are sensitive to extreme values, it makes sense that LMS, which replaces the mean by the much less sensitive median, will generate a more robust estimator.

Unlike LS, there is no closed form solution, or formula, with which to easily calculate the LMS line. Since the median is an order rank statistic, it is not amenable to calculation via derivatives or other calculations that rely on continuous functions. For each intercept and slope, the squared residuals have to be calculated and sorted in order to determine the middle, or median, value [4].

6.2 Results

(Confidential)

6.3 Summary

(Confidential)

7 Comparing results

Finally the outcomes of the M5P and LMS algorithm were provided with a quality label. These results were then compared to the results of the currently used models to study whether the two Data Mining tasks lead to better results. This chapter will describe these results and all steps taken in comparing these results. For the thirteen sensors, the quality labels on the basis of the currently used models have been provided by Elke Ottenhoff.

The data of the precipitation has been downloaded from the web-site of the KNMI, so this is not data of the city of (Confidential). Therefore, the quality labels of the precipitation data as assigned by the current models are not available. It takes a significant amount of time to apply the current models and the quality labels. Since my colleagues are busy enough with their own work, I did not ask them to apply the models to the precipitation data I used. Especially since the results are quite clear and conclusions can thus be drawn based on these.

7.1 Quality labels

(Confidential)

7.2 Labelling

(Confidential)

7.3 Conclusions

(Confidential)

8 Identifying faulty sensors

During this chapter a technique will be examined to try and identify faulty sensors. As has been concluded from the results of the previous chapters, there are two sensors that most probably are faulty. These sensors (WL_sewer1 and WL_surface1) have been seen to measure a constant water level. If the technique, which will be explained in the following paragraph, works in practice, these two sensors should be identified. This chapter will describe all steps taken and all results during the next paragraphs.

8.1 Technique

(Confidential)

8.2 Steps

(Confidential)

8.3 Results

(Confidential)

8.4 Conclusions

(Confidential)

Part IV Conclusions

9 Conclusions and recommendations

(Confidential)

10 Bibliography

Literature

[1] M.C.M. de Gunst and A.W. van der Vaart: Statistische Data Analyse. 2003.

[2] I.H. Witten and E. Frank: Data Mining. 2000.

[3] J. Oosterhoff and A.W. van der Vaart: Algemene Statistiek. 2000

[4] H. Barreto: An introduction to Least Median of Squares. 2001

[5] J. Erickson, S. Har-Peled, and D.M. Mount: On the Least Median Square Problem. 2003

Web-sites

[6] www.witteveenbos.nl

[7] www.riool.info

[8] www.knmi.nl

[9] www.nr.com

[10] www.mathworld.com

[11] www.wikipedia.org

[12] www.google.nl

[13] www.r-project.org

[14] www.cs.waikato.ac.nz/ml/weka

11 Appendices

11.1 Appendix A

(Confidential)

11.2 Appendix B

(Confidential)

11.3 Appendix C

(Confidential)

11.4 Appendix D

(Confidential)

11.5 Appendix E

(Confidential)