Modeling the relationships towards the performance of non-revenue water

Using structural models to link the performance of non-revenue water and its components with the obtainable variables

In partial fulfillment of the requirements for the degree of Master of Science in Construction Management and Engineering

Graduation report 'Modeling the relationships towards the performance of non-revenue water'
March 14th 2013
Construction Management and Urban Development 2012-2013
Eindhoven University of Technology

Author: Ing. C.B. (Bertine) Korevaar
Student nr.: 0742057

Graduation committee:
Prof. dr. ir. W.F. Schaefer
Dr. ir. B. Glumac
Ir. B. van Weenen
Ir. M. Riemersma (Royal HaskoningDHV)

Royal HaskoningDHV
Enhancing Society Together

TU/e Technische Universiteit Eindhoven University of Technology

COLOPHON

Title	Modeling the relationships towards the performance of non-revenue water
Subtitle	Using structural models to link the performance of non-revenue water and its components with the obtainable variables
Keywords	Non-revenue water, multiple regression analysis, structural models
Organization	Eindhoven University of Technology Architecture Building and Planning Construction Management & Urban development Studio process engineering for urban development
Commission	Eindhoven University of Technology Prof. dr. ir. W.F. Schaefer Dr. ir. B. Glumac Ir. B. van Weenen
Company	Royal HaskoningDHV Consultant Ir. M. Riemersma
Author	Ing. C.B. (Bertine) Korevaar
Student number	0742057
E-mail	c.b.korevaar@student.tue.nl, bertinekorevaar@gmail.com
Telephone	06 142 25 176
Report	Graduation thesis
Status	Definitive
Date	March 2013
Course code	7CC30
Graduation date	14 March 2013
Contact	Eindhoven University of Technology Faculty Architecture Building and Planning Den Dolech 2 5612 AZ Eindhoven Postbus 513 5600 MB Eindhoven Tel: 040 247 91 11

PREFACE

This thesis is the last chapter of the masters program Construction Management and Engineering at the TU/e, faculty of Architecture. During the graduation period I was employed to Royal HaskoningDHV. The performed research is a part of the KENWIB initiative; KENWIB deals with issues arising from the ambition of the municipality of Eindhoven to be an energy neutral municipality in the year 2040.

Before I began the masters program at the TU/e, I finished my bachelor degree in Civil Engineering. The world of water, whether it was storm water management, drinking water management or wastewater management, has always fascinated me. Royal HaskoningDHV provided me an opportunity to share thoughts about the problems encountered in the drinking water management field; one of the rising problems and an interesting topic is non-revenue water. An interesting and relevant research could be performed regarding this subject, combining the knowledge gained from the masters program and the knowledge and experience from Royal HaskoningDHV.

This report could not have been possible without the support, advice and cooperation of a lot of people. First of all I want to thank Michel Riemersma (from Royal HaskoningDHV) for his guidance and knowledge, for the discussions and the feedback on the subject and the report. Great thanks to the supervisors from the TU/e, Brano and Bart. I would like to thank Brano for his many comments, which made my arguments clearer and Bart for his support throughout the research.

I would also like to thank my family and friends. I would like to thank my boyfriend for the many discussions we had about the topic. I would like to thank my rugby teammates for their support, especially Paulien who has given me a lot of support and tips. And of course a special thanks to my family which gave me support throughout the research.

I hope you enjoy reading my master's thesis,

Bertine Korevaar
March 2013

MANAGEMENT SUMMARY

Currently the world population is growing, the increase in world population leads to an increase in water demand, putting a growing pressure on water as a resource. Water must be managed efficiently since there are already regions where water is a scarce resource. Water can be managed more efficiently by reducing the non-revenue water (NRW). "*Non-revenue water is the difference between the volume of water put into a water distribution systems and the volume that is billed to customers*" (Kingdom, et al., 2006). NRW can be divided into three components, of which the components apparent losses and the real losses are responsible for the main part of the volume of NRW. Real losses are caused by leakages in all parts of the systems and overflows at the utility reservoirs, while the apparent losses are often the result of influences on the company, which are usually beyond the scope of the day-to-day operation practices. An effective NRW reduction plan can be developed when the performance of NRW and its components (apparent losses and real losses) is analyzed based upon the performed measurements in the system. The current process of analyzing the performance of NRW and its components should be more efficient in time, so no time gets lost. To take a step forward to a more time sufficient approach it is proposed to examine the relationships between the performance of NRW and its components and the variables which influence of are influenced by the performance. Royal HaskoningDHV recognizes this as an opportunity to link one of their respected products with NRW to create a new business opportunity. The product that will be linked to the performance of NRW is OPIR (Optimal Production through Intelligent Control). Some of the variables will be obtained from OPIR; the other variables are variables which can be obtained at the beginning of a project.

There are several performance indicators that can be used to analyze the performance of NRW. For this research is chosen to express the performance of NRW and its components in liters/connection/day. This performance indicator is used by many water companies and it can be used to analyze NRW, as well as the components of NRW.

The variables obtained from OPIR are the pressure; the day peak factor and the night peak factor. The day peak factor and night peak factor can be estimated using the water demand pattern obtained from OPIR. The other variables which are included are the connection density, the number of connections and the sector type. The performance of NRW and its components is influenced by the connection density and the pressure, while the day peak factor and night peak factor are influenced by the performance of NRW or one of its components, the number of connections and the sector type. The day peak factor is not used in the analysis, since the data did not provide sufficient data points including the day peak factor.

Structural models are developed to model the relationships between the performance of NRW and its components, the variables obtained by OPIR and the variables that can be obtained at the beginning of a project. Structural models are developed for (1) the performance of NRW, (2) the performance of the real losses and (3) the performance of the apparent losses. The structural models are developed using the path analysis technique. The structural model is analyzed using the regression analysis, the regression analysis is chosen over the machine learning technique since the data set dispose of a hundred cases. The structural equations

evolving from the structural models are analyzed using the multiple regression backward elimination analysis. When the assumptions of the multiple regression analysis are violated, robust techniques are used to be able to perform the regression analysis.

The results show that the pressure and the connection density can be used to predict the performance of NRW and the apparent losses and that only the variable connection density can be used to predict the performance of the real losses. This is peculiar since this relationship is defined in the literature and thus expected. The found relationship between the performance of the apparent losses is not substantiated by the literature and thus it is unknown whether this is a causal relationship. Since there is no relationship between the performance of the real losses and the pressure, the found relationship between the performance of NRW and the pressure is based upon the relationship between the performance of the apparent losses and the pressure. It is unknown whether this relationship is causal, which means that it is also unknown whether the relationship between the performance of NRW and the pressure is causal.
A relationship is expected between the NPF and the performance of the real losses. The analysis shows that a structural equation is developed for the NPF using the structural model of the performance of the real losses, as is expected. This proven relationships shows that OPIR can be linked to NRW, based upon the use of the water demand patterns.

From the broad 95% confident boundaries, the small proportion of the variance in the dependent variable that can be explained using the independent variables, the even smaller shrinkage levels and the validation plots, where the predicted value mismatches the observed values can be concluded that the structural equations resulting from the models are too inaccurate to be used for predictive purposes. The inaccuracy may arise from the size of the used dataset as well as the reliability of the used dataset. It is recommended that a bigger verified dataset will be used in future research.

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

LIST OF ABBREVATIONS

ALC	Active Leakage Control
ALI	Apparent Loss Index
CAAL	Current Annual Apparent Loss
CARL	Current Annual Real Loss
DMA	District Meter Area
DPF	Day Peak Factor
ELL	Economic Level of Leakage
ICF	Infrastructure Correction Factor
ILI	Infrastructure Leakage Index
IWA	International Water Association
MNF	Minimum Night Flow
MVPA	Measured Variable Path Analysis
NDF	Night-Day Factor
NPF	Night Peak Factor
NRR	Natural Rate of Rise
NRW	Non-Revenue Water
OLS	Ordinary Least-Squares
OPIR	Optimal Production through Intelligent Control
PI	Performance Indicator
RHDHV	Royal HaskoningDHV
SEM	Structural Equation Model
UAAL	Unavoidable Annual Apparent Loss
UARL	Unavoidable Annual Real Loss
VIF	Variance Inflation Factor
WLTF	Water Loss Task Force

1 RESEARCH LAY-OUT

1.1 Context

Ismail Serageldin (then vice-president of the World Bank) has warned the world in August 1995 about the approach used to manage water:

"If the wars of this century were fought over oil, the wars of the next century will be fought over water – unless we change our approach to managing this precious and vital resource" (Seralgeldin, -).

Currently the world population is growing, while people are living in area's that experience water stress (World Water Assessment Programme, 2009). This increase in world population will lead to an increase in water demand, putting a growing pressure on water as a resource. It is crucial that water is managed efficiently, especially in regions where water is scarce (Gonzalez-Gomez, et al., 2011).

One of the solutions to the water demand problem is reducing non-revenue water (NRW); this reduction contributes to the development of sustainable cities and regions (Gonzalez-Gomez, et al., 2011). "*Non-revenue water is the difference between the volume of water put into a water distribution systems and the volume that is billed to customers*" (Kingdom, et al., 2006). The NRW (%) in developed countries differs from 5% in the Netherlands till 29% in Italy. NRW (%) in developing countries are more concerning, the values differ from 5% in Saldanha Bay, South Africa till 70% in LWSC, Liberia (Wikipedia, 2012).

It is not feasible for developing countries to eliminate all NRW in an area, but half of the current level of NRW appears to be a realistic target. The reduction in NRW generates cash and service without new investments in production facilities or extracting more water from scarce water resources (Kingdom, et al., 2006).

The first step in reducing NRW is performed by measuring the water loss in the system. Subsequently the performance of the system needs to be analyzed with the help of a performance indicator (PI); the PI can be determined using the water loss measurements. A PI can be used to understand NRW better, to define and set targets for improvements and to compare performances between countries and utilities. The last step is carried out by implementing a NRW reduction plan; this plan describes the actions that will be taken to reduce NRW (Farley & Trow, 2003). The interactions between these three steps are modeled in Figure 1.

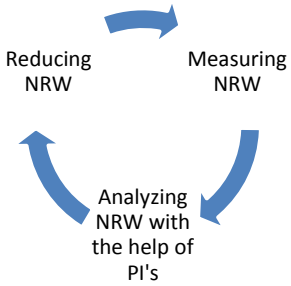


Figure 1: Process of reducing non-revenue water

1.2 Research approach

1.2.1 Problem

Before a NRW reduction plan can be made, the PI needs to be determined based upon the performed measurements in the system. It takes time to collect the proper measurements, because meters need to be installed in the system. The current process of analyzing the performance of NRW should be more efficient in time.

To take a step forward to develop a more time sufficient approach to analyze NRW it is proposed, in cooperation with Royal HaskoningDHV (RHDHV), to analyze the relationships between the performance of NRW or its components and the variables that are influenced or are influencing this performance. RHDHV recognizes this as an opportunity to connect one of their respected products with NRW to create a new business opportunity. The product that will be linked to the performance of NRW is OPIR (Optimal Production through Intelligent Control). OPIR is a system that optimizes the operation of drink water systems. *"The OPIR system determines the optimal operation based on its accurate demand forecaster. The demand forecaster is an intelligent self-learning system, which functions by comparing distribution curves from the past, with its current measurements"* (RoyalHaskoning DHV, -). *"OPIR plans the operation for each pumping station, valve or treatment plant of the water supply system for 48 hours ahead. The operation is planned to meet the lowest (electrical and other) cost, within the given operational boundaries"* (Royal HaskoningDHV, -). The variables will not only be obtained from OPIR, the other variables that will be used are variables that can be obtained at the beginning of a project. From these examined relationships a model will be built, subsequently the model will be analyzed and validated to determine whether the model can be used for predictive purposes.

The problem definition for the research, resulting from the context and the problems, is:

The current method used to analyze non-revenue water in a system is too time consuming, another approach is proposed but the nature of the relationships between the performance of non-revenue water or its components, the variables obtained by OPIR and the variables obtained at the beginning of a project are unknown, it is also unknown if the models developed using these relationships can be used for predictive purposes.

1.2.2 Phrasing of the question

What is the nature of the relationships between the performance of NRW or its components, the variables obtained from OPIR and the variables obtained at the beginning of the project and what is the predictive value of the models developed using these relationships?

To answer this central question research questions are stated. These questions are used to create a roadmap to answer the central question.

1. Why is non-revenue water becoming an important topic around the world?
2. What is non-revenue water?
3. What is the current way to measure and analyze non-revenue water, which performance indicator can be used in the research?
4. Which variables can be obtained by OPIR and which variables can be obtained at the beginning of a project?
5. What is the nature of the relationships between the performance of NRW or its components, the variables obtained by OPIR and the variables obtained at the beginning of a project?
6. What kind of model can be developed using the nature of the relationships and can this model predict the performance of NRW or its components?

1.2.3 Research limitations

There are some limitations that may occur during the research; these limitations have an influence on the outcome of the research.

The first limitation can occur in the data that needs to be collected. First of all only systems can be used which have a continuous supply and where the consumers do not rely on the use of roof tanks, only then the performance of NRW can be analyzed. For a reliable result the dataset should be verified, though this is difficult to achieve for this topic. The size of the dataset also has an influence on the reliability of the results, but it is difficult to collect a lot of data points since it is relatively new topic.

The second limitation that has an influence on the outcome of the research is the effect of the cultural aspects in some of the variables. It is important to use data that is collected from the same region. The outcome can be validated for other regions using an external validation.

1.3 Research model

1.3.1 Research model

Figure 2 shows the research model that is used for the research.

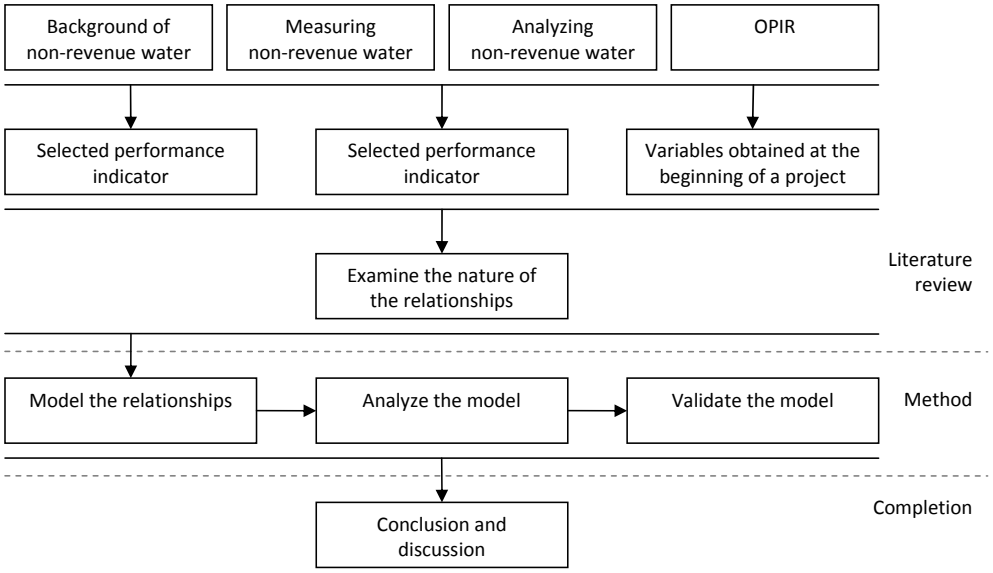


Figure 2: Research model

First a literature study will be performed to review the topic NRW. The literature study discusses the methods used to measure NRW, the PI's that can be used to analyze NRW concluding with the PI that will be used to express the performance of NRW. Since variables will be obtained by OPIR, the purpose and operation of OPIR will be reviewed, as well as the variables that can be obtained from OPIR. The variables that can be obtained at the beginning of a project will be discussed as well. Finally the nature of the relationships between all the variables is reviewed.

The method that is chosen must be able to model the relationship between the performance of NRW or its components, the variables obtained by OPIR and the variables obtained at the beginning of a project. The method must also be able to analyze the model, as well as validate the model.

The last step is to draw conclusions from the provided results and the validation of the model. The used PI, variables, research method and results will be discussed, to determine whether this model and relationships can be used to predict the performance of NRW.

1.3.2 Research method

First a desk research will be performed; this desk research provides the theoretical input for the research. The desk research will review the current method to analyze NRW explaining the

different PI's; subsequently a PI will be chosen which will be used in the remainder of the research. The desk research will also provide the variables that will be used and will discuss the relationships between the variables. The desk research will be performed by reviewing literature and by interviewing experts regarding the topics of NRW and OPIR.

The relationships will be modeled using the path analysis method. The model will be analyzed using the regression method. *"Regression analysis is a statistical technique that attempts to predict the values of one variable using the values of one or more other variables."* (Allen, 1997). The predicted variable is the dependent variable; the variables that are used to predict the dependent variable are called the independent variables. Regression methods determine the relationships between the dependent and independent variables. The method can be used to describe the relationships in a mathematical form, which makes it possible to predict the dependent variable. When the relationship between these variables is not statistically significant the independent variable cannot be used to predict the dependent variable. The regression method can also be used to analyze how accurate the variables can predict the dependent variable; this feature can be used to validate the model.

The statistical package SPSS will be used to perform the multiple regression analysis. SPSS is one of the most used programs in the statistical analysis.

1.4 Reading guide

The contextual orientation goes into depth about NRW and it also discusses OPIR, the system that will be linked to the performance of NRW. The second chapter starts the contextual orientation by discussing the importance of reducing NRW. The basic components of NRW are described in chapter three, as well as how these can be measured. The fourth chapter discusses the presence of NRW in the world. The fifth chapter reviews the performance indicators that are used to analyze the performance of NRW; the chapter also reviews which PI is used during the analysis. In chapter six OPIR is discussed as well as the link between OPIR and the used variables. The seventh chapter discusses the variables that will be used in the analysis and the nature of their relationships.

When the contextual orientation is performed the research methodology is discussed. Chapter eight discusses the method and the used validation techniques. In chapter nine the data is reviewed. Chapter ten presents the results of the regression analysis. Chapter eleven discusses these results, while in chapter twelve the conclusions are drawn as well as that some recommendations are given for future research.

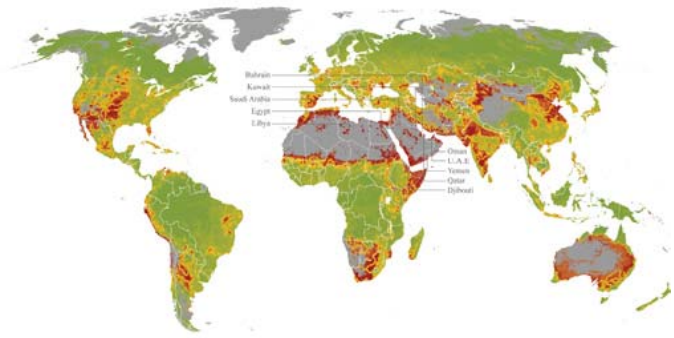
2 THE IMPORTANCE OF REDUCING NON-REVENUE WATER

William Hope has stated in the 19th century:
''There is no water supply in which some unnecessary waste does not exist and there are few supplies, if any, in which the saving of a substantial proportion of that waste would not bring pecuniary advantage to the Water Authority'' (Page, 2005)
This chapter will explain which problems exists because of NRW and it will also discuss the vicious cycle the water utilities have to break regarding NRW.

2.1 Problem 1: Drinking water scarcity

People experience drinking water scarcity, even though the amount of drinking water on the planet is enough to supply six billion people. Drinking water scarcity exists because the water is unevenly distributed, polluted, wasted and inefficiently managed. It is the first and main reason to reduce NRW (Page, 2005).

Figure 3 shows the water stress in the world in 2012, in the figure the top ten countries are ranked based upon the level of water stress. The disturbing fact of this figure is that there is no data available for some countries, while those countries are close to the countries experiencing water stress.



Legend
Extreme risk
High risk
Medium risk
Low risk
No Data

Rank	Country	Rating
1	Bahrain	Extreme
2	Qatar	Extreme
3	Kuwait	Extreme
4	Libya	Extreme
5	Djibouti	Extreme

Rank	Country	Rating
6	U.A.E.	Extreme
7	Yemen	Extreme
8	Saudi Arabia	Extreme
9	Oman	Extreme
10	Egypt	Extreme

Figure 3: Water stress index 2012 (Mc Geown, 2012)

The world population has tripled in the past century, while the use of water has increased with a factor six. Currently a total of 9 trillion m^3 of water is used per year, but it is projected that with a growth of 80 million people a year, the demand for fresh water will increase with 64 billion m^3 of water per year. Of the 3 billion people who are expected to be added until 2050, 2.7 billion people will be born in developing countries especially in regions where water stress is experienced (World Water Assessment Programme, 2009).

Figure 4 projects the forecast of the pressure on water resources. Currently the biggest consumer of water is the agriculture sector, with the use of 70% of the fresh water (Niemczynowicz, 1999). In the future the water demand will increase for urban uses and it will transcend the water use in the agriculture sector (Gonzalez-Gomez, et al., 2011).

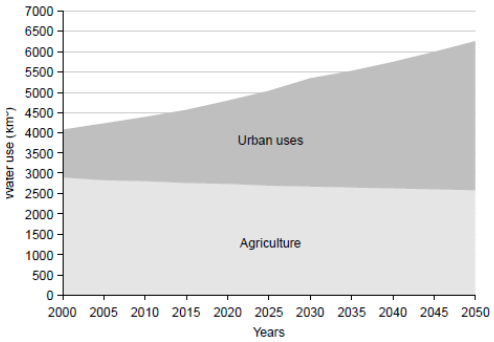


Figure 4: World water use by sector, 2000 – 2050 (Gonzalez-Gomez, et al., 2011)

Climate change also influences the water resources and the use of water. It affects the average weather patterns, which are used for present and future planning. These changing weather patterns have an effect on the water resources (Charalambous, 2012), for example the global warming can decrease the water availability (ERM, 2007).

The growth in world population will cause the water demand to grow and due to the climate change the water resources will decrease, the result of these two trends is that water will become an even scarcer resource. Water must be managed efficiently, especially in regions currently experiencing water stress. Managing water efficiently will lead to an increase in the sustainable use of water (Gonzalez-Gomez, et al., 2011).

2.2 Problem 2: Use of energy

One of the major causes in energy waste experienced in water supply systems is caused by water leaks and inefficient use of water (Feldman, 2009). Energy is needed during the treatment process and to supply the water to the consumers. Reducing the water losses will lead to a more efficient use of the energy in the utility, this result in a reduction of the emissions caused by the energy use. Utilities will find it more interesting to experience that the reduction of energy will lead to a reduction in energy costs, affecting them in a positive manner (Vouk, et al., 2012).

2.3 Problem 3: Health and safety issue

A high pressure in a water supply system tends to lead to a high flow rate in the system, while a low pressure tends to lead to a low flow rate. Consequently a high pressure will cause more water to flow out of the leaks, than a low pressure. However, lowering the pressure must be carried out carefully since it can have consequences as well. A low flow rate may cause the consumers to receive water under a low pressure or to receive no water at all; depending on the consumers' location in the system (Tabesh, et al., 2009). Another consequence of a low flow

rate is the possibility of the degradation of the water quality. The system will not be thoroughly filled with water, which gives (polluted) groundwater a chance to infiltrate the system (Tabesh, et al., 2009).

Pressure changes cause a lot of problems as well. When there are large variations in the pressure of the system, turbulence could occur, this is called the hammer effect. The hammer effect tears down the system, which leads to more breaks in the system. These breaks in combination with a low pressure could lead to (polluted) groundwater infiltrating the system (Tabesh, et al., 2009).

2.4 Vicious cycle and virtuous cycle of non-revenue water

Water utilities can end up in a vicious cycle trying to reduce NRW; this will only enlarge the problems described in the previous sub-chapters. The vicious cycle describes that when water gets lost, substantial capital expenditure programs are often promoted to meet the ever increasing demand. These programs will cause the treatment and distribution costs to increase and the total water sales to decrease, because of the increased losses of water. To turn this vicious cycle into a virtuous cycle, utilities need to have the political will and the full support of the utilities management. The virtuous cycle shows that the investment in water loss management leads to maintaining the water loss level or even decreasing it, which leads to a more efficient operation with less produced water. The capital saved by the efficient operation, can be invested in the NRW reduction program. The virtuous cycle will lead to a decrease of NRW. The vicious and virtuous NRW cycles are shown in Figure 5 (Frauendorfer & Liemberger, 2010).

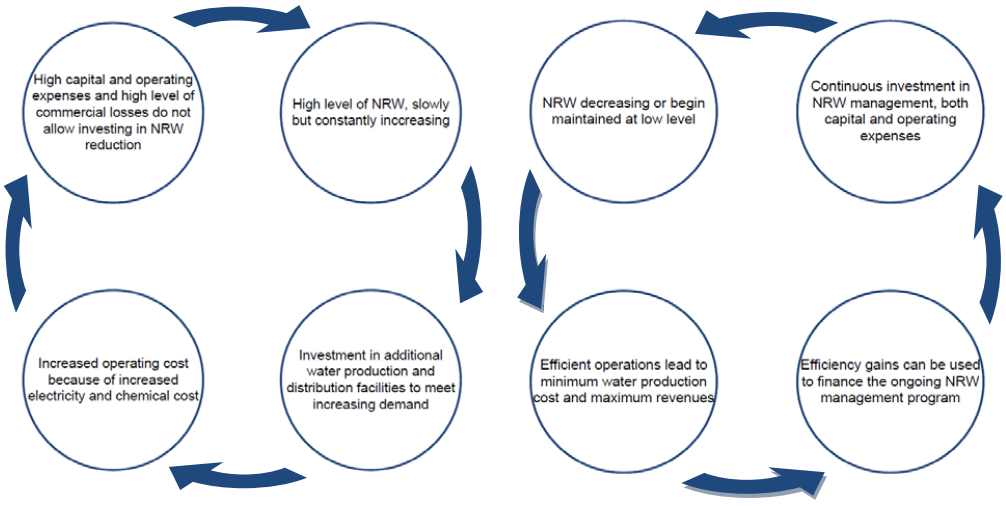


Figure 5: Vicious and Virtuous non-revenue cycle (Frauendorfer & Liemberger, 2010)

3 NON-REVENUE WATER AND THE COMPONENTS

In the early 90's there was no standard term to express and asses the water losses in the distribution system. The international water association (IWA) has acknowleged this problem and established the water loss task force (WLTF). The WLFT examined the international best practices and developed a standardized terminology for NRW (Frauendorfer & Liemberger, 2010).

"Non-revenue water is the difference between the volume of water put into a water distribution system and the volume that is billed to customers" (Kingdom, et al., 2006).

In addition the WLTF developed the international water balance to establish one uniform method to measure and analyze NRW, this method is called the top-down approach. The water balance is shown in Table 1. The components in the water balance are measured in kl/day or m^3/year etc. The system input volume can be split up in four components; the billed authorized consumption, the unbilled authorized consumption, the apparent losses and the real losses. NRW consists of the components; unbilled authorized consumption, the apparent losses and the real losses. The sub-chapters will discuss the system input volume and the four major components and how to measure or determine their volumes. The volumes are needed to analyze the performance of NRW (Frauendorfer & Liemberger, 2010). Appendix I describes how NRW can be reduced.

3.1 System input volume

The system input volume is the water from the reservoirs which is entering the system.

<table>
<tr><td rowspan="4">System input volume</td><td rowspan="2">Auth. Cons.</td><td>Bill. Auth. Cons.</td><td>Rev. water</td></tr>
<tr><td>Unbill. Auth. Cons.</td><td rowspan="3">Non – rev. water</td></tr>
<tr><td rowspan="2">Water loss</td><td>Ap. losses</td></tr>
<tr><td>Real losses</td></tr>
</table>

Figure 6: System input volume

The system input volume is measured with a meter at the beginning of the system. The values can be inaccurate due to the use of old, unfit, broken or tempered meters. The input meters can be verified using portable flow measuring devices. When discrepancies show up between the data measured by the normal meters and by the portable meters, the discrepancies should be examined (Farley & Liemberger, -).

When the system is not metered, the system input volume can be determined using (1) temporary flow measurements with portable devices, (2) reservoir drop tests or (3) analysis of pump curves, analysis of the pressure of the pumps and the average pumping hours (Farley & Liemberger, -).

3.2 Billed authorized consumption

The billed authorized consumption consists of billed metered consumption and billed non-metered consumption as is shown in Table 1.

<table>
<tr><td rowspan="4">System input volume</td><td rowspan="2">Auth. Cons.</td><td>Bill. Auth. Cons.</td><td>Rev. water</td></tr>
<tr><td>Unbill. Auth. Cons.</td><td rowspan="3">Non – rev. water</td></tr>
<tr><td rowspan="2">Water loss</td><td>Ap. losses</td></tr>
<tr><td>Real losses</td></tr>
</table>

Figure 7: Billed authorized consumption

The billed metered consumption can be determined by summing up all the bills. The billed metered consumption is

Table 1: IWA standard international water balance and terminology

System input volume	Authorized consumption	Billed authorized consumption	Billed metered consumption (including water exported)			Revenue water
			Billed non-metered consumption			
		Unbilled authorized consumption	Unbilled metered consumption			Non-revenue water
			Unbilled non-metered consumption			
	Water loss	Apparent losses	Unauthorized consumption		Illegal connections	
					Tampering with meters	
					Unrecorded consumers	
			Metering inaccuracies	Inaccurate meters	Slow running meter	
					Broken meter	
				Meter reading errors	Reading errors	
				Data handling and billing errors	Data entry errors	
					Administration losses	
					Delays	
					Loss of records	
		Real losses	Leakage on transmission and/or distribution mains			
			Leakage and overflows at utilities storage tanks			
			Leakage on connections up to the Consumers' meters			

greatly influenced by the large consumers of a distribution area, the utilities should give special attention to these consumers. Utilities must also know the total number of connections of the different consumer types (domestic, commercial or industrial); these different consumer types influence the billed metered consumption as well (Farley & Liemberger, -).

Billed non-metered consumption can occur when the consumers' meter is not working and the billing is based upon an estimation of the water use. The billed non-metered consumption can be obtained from the metering output and the billing output. To validate the accuracy of the estimates the unmetered consumers should be monitored, this can be carried out trough individually monitoring a consumer or by monitoring an area that supplies unmetered consumers (Farley & Liemberger, -).

3.3 Unbilled authorized consumption

Unbilled authorized consumption is water that is used by utilities for operational purposes. The unbilled authorized consumption should be a small portion of the water balance (less than 1% of the system input volume). The first time the unbilled authorized consumption is measured, the amount can be unnecessarily high. As long as a utility does not measure the unbilled authorized consumption, the consumption cannot be managed resulting in exceeding the normal portion of less than 1% of the system input volume. When the measurements are known, the utilities should be able to manage the unbilled authorized consumption and decrease it to the normal proportion (Farley & Trow, 2003). Unbilled authorized consumption consists of unbilled metered consumption and unbilled non-metered consumption, as is described in Table 1.

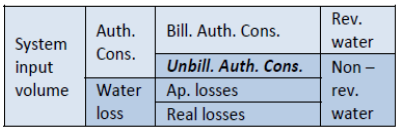


Figure 8: Unbilled authorized consumption

Unbilled metered consumption can be estimated using the metering and billing output. The process to validate the values is the same as for the billed non-metered consumption described in sub-chapter 3.2. An example of unbilled metered consumption is water that is free for certain villages or a population group, this water is often metered, but it is not billed to the consumer (Farley & Liemberger, -).

Unbilled non-metered consumption needs to be individually identified and estimated. An example of unbilled non-metered consumption is water that is used to flush the mains after fixing a break. Other examples of unbilled non-metered consumption is water that is used for street cleaning, fire fighting and fire flow tests (Farley & Liemberger, -).

3.4 Apparent losses

The apparent losses are often the result of influences on the company, which are usually beyond the scope of the day-to-day operational practice. Most developed countries use rules of thumb to estimate the apparent losses, though the rules of thumb are only applicable for the best performing systems (Mutikanga & Sharma, 2012). Table 1 shows the divisions of components in the apparent losses.

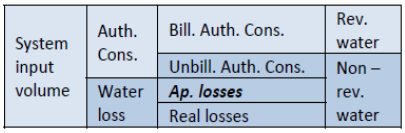


Figure 9: Apparent losses

The upcoming paragraph will discuss the unauthorized consumptions and the metering inaccuracies consisting of the use of inaccurate meters, the meter reading errors and the data handling and billing errors.

Unauthorized consumption can occur when there are tampered or demolished meters present in the system or it is caused by the presence of illegal connections in the system. Reducing unauthorized consumption supports greater fairness between consumers; the money saved can be used for new investments in the system (Roger & Bettin, 2012). Unauthorized consumption can be measured by carrying out individual site inspections. Historical records need to be used to quantify the total volume of unauthorized use. The unauthorized use (q) can be split up in domestic illegal use (q_d), commercial illegal use (q_c), illegal use at government institutions (q_g) and illegal use at public standpipes (q_p) (Mutikanga & Sharma, 2012). When the volumes of the unauthorized use are unknown, equation 3-1 can be used to estimate the unauthorized consumption. The volumes used in equation 3-1 are the losses expected per type of illegal connection.

$$\begin{gathered} q = q_d + q_c + q_g + q_p \\ q_d = number\ of\ properties * 20m^3\ per\ month \\ q_c = number\ of\ properties * 500m^3\ per\ month \\ q_g = number\ of\ government\ institutions * 500m^3\ per\ month \\ q_p = number\ of\ standpipes * 50m^3\ per\ month \end{gathered} \quad \textbf{3-1}$$

The first cause of metering inaccuracies is the use of inaccurate meters. The causes for the metering inaccuracy are:

- The use of an inaccurate meter size for the actual demand pattern;
- The used meter type is not correct for the operating range;
- The used service line size is not optimal for the operating range.

To assess the accuracy of the meters, the weighted meter accuracy needs be determined by analyzing the accuracy of the meters at a high flow, medium flow and low flow. The weighted meter accuracy is the parameter that defines the meter performance when measuring the water consumption of consumers. The meters that will be tested need to come from different parts of the system, to create a homogeneous and well mixed population. The number of samples depends on the total number of meters in the system and should give a good representation of the system. The accuracy of the existing meters can be double checked with the weekly readings of the master meter (Mutikanga & Sharma, 2012).

The second cause of metering inaccuracies is the meter reading errors. Meter reading errors can be measured by carrying out audits to verify the accuracy of the meter readers. The readings of the auditors will be compared with the readings submitted by the meter readers. The readings with large variances need to be regarded as erroneous readings. The erroneous meter readings will be summed up to determine the value of the meter-reading error for that part of the system. With this figure the total meter-reading error for the whole system can be estimated (Mutikanga & Sharma, 2012).

Data handling and billing errors are the third cause for metering inaccuracies. Data handling and billing errors can be captured by carrying out an audit. The audit will compare the input data used for the billings and the values on the meter reading sheets submitted by the meter readers. The readings that are wrongly captured in the billing system will be summed up to estimate a total volume of that part of the system. With the volume for that part of the system, the data handling error for the whole system can be estimated (Mutikanga & Sharma, 2012).

3.5 Real losses

The real losses are caused by leakages in all parts of the systems and overflows at the utility reservoirs. The real losses mainly occur due to poor operations and maintenance, the poor quality of the underground assets and the lack of active leakage control. Effective active leakage control is applied when a utility has staff members which are hired to find leakages that are not reported by consumers or other means (Farley & Trow, 2003). The real losses can be split up in leakage on transmissions and/or distribution mains, leakage and overflows at utilities storage tanks and leakage on connections up to the consumers' meters as is described in Table 1.

<table>
<tr><td rowspan="4">System input volume</td><td rowspan="2">Auth. Cons.</td><td>Bill. Auth. Cons.</td><td>Rev. water</td></tr>
<tr><td>Unbill. Auth. Cons.</td><td rowspan="3">Non – rev. water</td></tr>
<tr><td rowspan="2">Water loss</td><td>Ap. losses</td></tr>
<tr><td>Real losses</td></tr>
</table>

Figure 10: Real losses

There are three methods which can be used to determine the real losses; (1) the component analysis, (2) the 24 hour zone measurement and (3) the minimum night flow analysis; these methods are all described in the following sub-chapters. The outcome of the methods should be compared with each other to determine the reliable real loss value (Farley & Liemberger, -).

3.5.1 Component analysis

The component analysis does not only split the real losses up in the components described in Table 1, but it is also split up in three categories:

- *"Background (undetectable) leakage: Small flow rate, runs continuously;*
- *Reported breaks: High flow rate, relatively short duration;*
- *Unreported breaks: Moderate flow rates, duration depends on intervention policy"* (Thornton, et al., 2008).

The component analysis focuses on analyzing the volume of each component for every category. Appendix 2 describes how the component analysis should be performed.

3.5.2 24 Hour Zone Measurement

During a 24 hour zone measurement the inflow and pressure in the system will be logged. An accurate value of the real loss component can be determined using the logged measurements. The distribution area should be isolated to perform a 24 hour zone measurement. The area can only be supplied via one or two inflow points during the measurement (Puust, et al., 2010).

3.5.3 Minimum night flow analysis

A minimum night flow (MNF) analysis can only be performed in a district meter area (DMA). *"A DMA is a hydraulically discrete part of the distribution system that is isolated from the rest of the*

distribution system. It is normally supplied through a single metered line so that the total inflow to the area is measured" (Thornton, et al., 2008). Other conditions that are needed to perform a MNF analysis are a continuous flow and the consumers in the area should not fill their roof tanks during the night. Roof tanks in continuous supplied areas are used when the pressure in the system is too low, which leads to a low flow rate out of the orifice (Trifunovic & Veenstra, 2012).

The MNF indicates the lowest water consumption level associated with the reduction of people's activities. The MNF period normally occurs in the early morning, usually between 02:00 and 04:00. The water losses during this time of day are predominantly leaks, therefore it can be assumed that the minimum flow rate minus consumption rate equals the maximum real losses (Water Loss Task Force, 2007).

Since the pressure during the night is usually the largest and the pressure varies throughout the day, the real loss value of the whole day cannot be determined by extrapolating the real loss value measured during the MNF. Extrapolating the real loss value would lead to the overestimation of the daily leakage, because of the lower pressure during the day (Water Loss Task Force, 2007). To determine the real loss value for the whole day, the night-day factor (NDF) should be estimated, this factor can vary between 18 to 24. When the NDF is estimated, the real loss value can be determined using equation 3-2.

$$Real\ losses = NDF * Real\ losses\ during\ MNF \quad \textbf{3-2}$$

Appendix 3 describes how to perform an accurate MNF analysis.

4 THE PRESENCE OF NON-REVENUE WATER IN THE WORLD

The global NRW (%) is around the 35%. According to DAI (an economic development consultancy) there is a volume of 32 billion m^3 of real loss water and an additional volume of 16 billion m^3 of apparent loss water every year. The costs for these losses are estimated at US$14 billion per year (DAI, 2010), while Kingdom, et al. (2006) has estimated the costs of water loss at approximately US$141 billion per year. Even though these values do not correspond with each other, it can be concluded that action is needed to diminish the water and money loss.

The trend is that the NRW (%) in developed countries is small, though it can incur to 30%. The NRW (%) in developing countries can vary extremely, with the lowest level at 4,4% incurring to 70% (Wikipedia, 2013). The NRW (%) in developing countries depends on the will to reduce NRW, which can vary per country and per city. Figure 11 shows the NRW (%) values for several countries and cities.

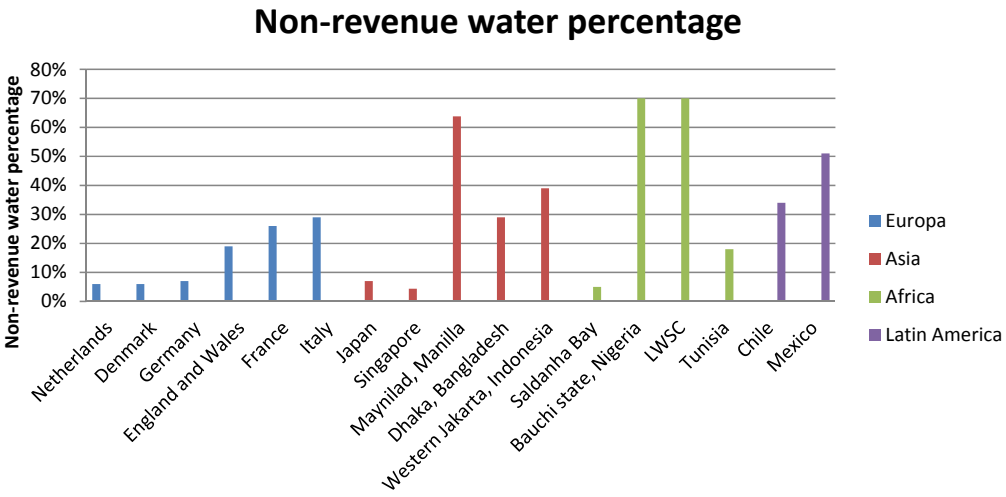


Figure 11: Non-revenue water percentages over the world (Wikipedia, 2013)

The real loss component is usually the largest component (around the 80% of the total NRW) in developed countries (Kingdom, et al., 2006). This can be explained due to the good administration in these countries (so low apparent losses) and the deteriorating system. The water distribution systems in the developed countries are built decennia ago. The systems are currently deteriorating because of their age. Often the system is already leaking water before utilities start with the renovation or the replacement of the system (Gonzalez-Gomez, et al., 2011).

The apparent loss component in developing countries is almost as high as the real loss component, it is around 40% of the total NRW (Kingdom, et al., 2006). This can be explained by institutional and management causes, like fraudulent activities and corruption (Gonzalez-Gomez, et al., 2011). Figure 12 shows the division of the real and apparent losses in the developed and developing countries.

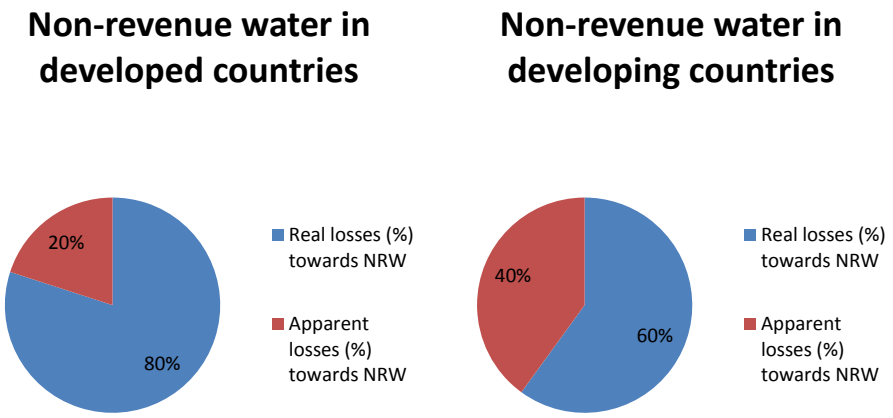


Figure 12: Division of the components of non-revenue water in developed and developing countries (Kingdom, et al., 2006)

5 PERFORMANCE INDICATORS

The performance of NRW and its components can be analyzed, when the volumes of the components described in chapter 3 are determined. This analysis is performed using a performance indicator (PI); a PI helps the water utilities in the following way:

- Understanding the water losses better;
- Defining and setting targets for improvement;
- Measuring and comparing performance;
- Developing standards;
- Monitoring compliance;
- Prioritizing investments (Farley, et al., 2008).

There are several PI's available to asses NRW, these PI's are described in Table 2. The table shows the accuracy level of the PI's and which volume the PI's can analyze (Farley, et al., 2008). The sub-chapters will describe the PI's shown in Table 2.

Table 2: Performance indicators (Farley, et al., 2008)

Accuracy level	Performance indicator	Used to analyze
1 (basic)	Volume of NRW in percentages (% of system input volume)	Non-revenue water, real losses and apparent losses
1 (basic)	Liters/connection/day	Non-revenue water, real losses and apparent losses
	m^3/km mains/day	
2 (medium)	Liters/connection/day/mWc	Non-revenue water, real losses and apparent losses
	m^3/km mains/day/mWc	
3 (detailed)	Infrastructure leakage index (ILI)	Real losses
3 (detailed	Apparent loss index (ALI)	Apparent losses

5.1 Non-revenue water in percentages

One of the most basic PI's is the NRW (%). The NRW (%) can be estimated using equation 5-1.

$$NRW\ (\%) = \frac{(System\ input\ volume - Billed\ authorized\ consumption)}{System\ input\ volume} \quad \text{5-1}$$

NRW (%) was formerly the standard PI. Currently it has been accepted that expressing NRW (%) is not a good technical measure, but NRW (%) is still used by most water utilities and countries (Department of Water Affairs, 2010). Using NRW (%) favors the utilities with a high consumption, a low pressure and operating with an intermittent supply. A utility that has a high consumption has a lower NRW (%) than utilities with the same NRW but with a lower consumption. More leakage is experienced by utilities that need a high pressure to pump the water to a high location than by utilities pumping the water to a valley using a low pressure. Utilities operating with an intermittent supply benefit from the reduction of time that the system is pressurized and thus can leak (Farley & Trow, 2003). The NRW (%) cannot be used to compare systems with each other and it cannot be used to compare figures of different months of the same system with each other (Farley, et al., 2008).

5.2 Liters/connection/day and m^3/km mains/day

To estimate the PI's liters/connection/day and m^3/km mains/day the volume of NRW is divided by the size of the area (number of connections or the length of the mains). Which of the two PI's is chosen is based upon where the greater portion of the losses is associated with, the mains or the connections. When there are more than 20 connections per km main, the greater portion of the losses is associated with the connections and thus liters/connection/day should be used. When there are less than 20 connections per km main, the greater portion of the losses is associated with the length of the mains and thus m^3/km mains/day should be used. Table 3 describes the explained rules (Lambert, 2001).

Table 3: Use of the performance indicator (Farley, et al., 2008)

Liters/connection/day	m^3/km mains/day
X > 20 connections/km mains	X < 20 Connections/km mains

The outcome of the PI is strongly influenced by the density of the connections, the consumers' meter location and the average pressure. The outcome of the PI can be used to compare the values of different months of the same system with each other, when the pressure values of these months do not differ much (Farley & Trow, 2003).

5.3 Liters/connection/day/mWc and m^3/km mains/day/mWc

These PI's are not only estimated by dividing the volume of NRW by the size of the area, but also by dividing it by the average pressure (mWc) in the area. The same rules apply for the choice of PI as is discussed in sub-chapter 5.2. The PI's can be used to compare systems with each other (Farley, et al., 2008).

5.4 Infrastructure leakage index (ILI)

The infrastructure leakage index (ILI) is *"A measure of how well a distribution system is managed for control or real losses, at the current operating pressure"* (Yeboah, 2008).

IWA has developed ILI as a quantitative method to compare performances between systems based upon the real losses (Pearson & Trow, 2012). ILI is a ratio estimated by dividing the current annual real loss (CARL) (the component real losses in the water balance) by the unavoidable annual real loss (UARL) for a system of the same size as the assessed system.

$$ILI = \frac{Current\ annual\ real\ loss\ (CARL)}{Unavoidable\ annual\ real\ loss\ (UARL)} \quad \text{5-2}$$

The UARL is based on a wide survey of the 'best' achieved systems across the world. The value of UARL is determined by the length of the mains, the number of connections, the length of the supply pipe and the average operating pressure (Farley & Liemberger, -).

A system is a 'world-class' system when it has an ILI value close to 1, the real losses are then close to the minimum technical value at the operating pressure. A low ILI value is only likely to be economically justified when the marginal costs of the water supply is relatively high or the water is scarce (Lambert & McKenzie, 2002).

5.5 Apparent Losses Index (ALI)

The performance of the system can also be determined based upon the apparent losses. The apparent loss index (ALI) is the ratio between the current annual losses (CAAL) and the unavoidable annual apparent losses (UAAL) (Thornton, et al., 2008).

$$ALI = \frac{Current\ annual\ apparent\ loss\ (CAAL)}{Unavoidable\ annual\ apparent\ loss\ (UAAL)} \quad \text{5-3}$$

Currently the water loss task force (WLTF) is engaged in developing an equation which can be used to estimate the UAAL. As long as this UAAL equation is not developed, the UAAL is estimated by using 5% of the metered consumption value. This value is high for water utilities in developed countries, but is reasonable for developing countries (Thornton, et al., 2008).

The ALI has the same index as the ILI, which is described in sub-chapter 5.4.

5.6 Conclusion

The performance in the system will be analyzed using one of the described PI's. The analysis will focus on systems with a continuous supply and where consumers totally depend on the system (the consumers do not depended on their roof tanks). RHDHV has stated some requirements the PI should meet:

- The PI is not obliged to compare systems with each other;
- The PI should be able to determine performance of NRW, the real losses and the apparent losses (Riemersma, 2013).

The PI's ILI and ALI will not be used, since the ILI and ALI cannot be used to determine the performance of NRW, the real losses and the apparent losses and RHDHV did not demand that the chosen PI should be able to compare systems with each other. All the other PI's meet the requirements RHDHV has set, though it is useful to select a PI that is well-known and often used by water companies. The PI's liters/connection/day/mWc and m^3/km mains/day/mWc are not often used, thus will be dropped. The PI's liters/connection/day and m^3/km mains/day are preferred over the PI NRW (%) since these PI's can easily be transformed (if needed) to a more detailed PI by dividing them by the average pressure (mWc) in the system.

The choice between the PI's liters/connection/day and m^3/km mains/day is based upon where the greater portion of the real losses occurs, in the connections or in the mains (Lambert, 2001). The projects are mainly introduced in urban areas, where the greater portion of the losses is associated with the connections; this is why the PI liters/connection/day is chosen.

6 OPIR

To take a step forward to a more time sufficient approach the relationships between the performance of NRW and its components and the variables that influence or are influenced by the performance will be analyzed. Some of these variables can be obtained from OPIR. Currently 70% of the Dutch water companies are using OPIR; OPIR is also used by water companies from Belgium, Canada, Portugal and Poland. NRW is and will become an important topic in the world; linking NRW with OPIR will create a new business opportunity for OPIR.

"OPIR is the real-time software solution which optimizes the operation of water supply systems. Water treatment plants and pumping stations consume large amounts of energy to abstract, treat and transport water to their consumers. Intelligent control of the water system by OPIR can lead to a significant reduction of the total energy consumption" (RoyalHaskoning DHV, -).

A level based control (a distribution system without using OPIR) consists of the following steps:

- Drinking water is supplied from a drinking water storage to meet the drinking water demand of the consumers;
- When the water level drops below a certain threshold, the pumps will start and the storage is filled with drinking water;
- When the water level has reached a certain threshold, the pumps will turn off.

The disadvantages of a level based control system are that the space of the basement is not totally used and every time the pumps start up and turn off extra energy is used (Bakker, 2012). Figure 13 shows a system where OPIR is not used, the production level changes during the day.

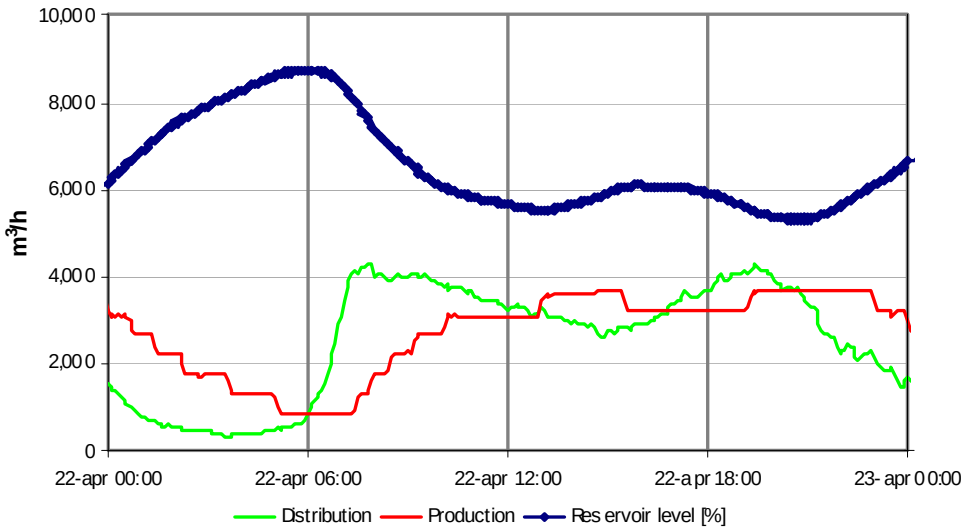


Figure 13: Storage and flow output of a level based control

The optimal operation determined by OPIR is based on the water demand forecaster. It forecasts the water demand and the optimal operation two days ahead. *"The demand forecaster is an intelligent self-learning system, which functions by comparing distribution curves from the past, with its current measurements. The energy costs can be reduced even more, by shifting pump operation to low energy tariff hours, or by buying energy on the on-line energy market at the lowest cost. Operating at a constant flow rate will provide optimal process conditions for a water treatment plant. Optimal process conditions will lead to optimal water quality"* (RoyalHaskoning DHV, -).

Figure 14 shows the predictive control using OPIR. The figure shows a constant production level during the day (RoyalHaskoningDHV, -).

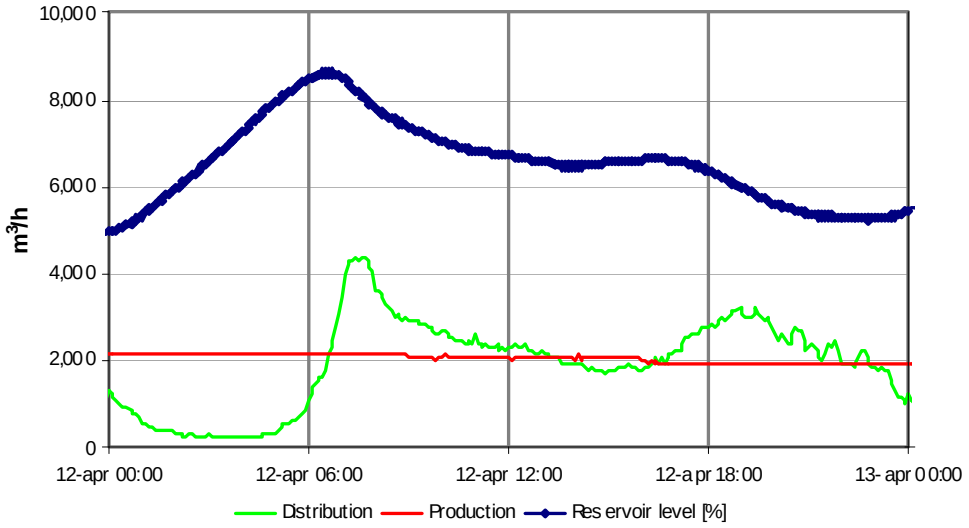


Figure 14: Storage and flow output of a predictive control based on OPIR

OPIR can be extended with a pressure module, in the module OPIR controls the pressure of the distribution system and thus in the distribution area. OPIR controls the pressure for the weakest point in the system; the pressure in this point should never be lower than the minimum allowable pressure. Most of the time this results in a lower average pressure in the whole system (RoyalHaskoning DHV, -).

6.1 Linking OPIR with the performance of non-revenue water

There are two elements of OPIR which can be used to link OPIR with NRW:

- The first element is the measured distribution curves, also called the daily water demand patterns, which are used in the demand forecaster of OPIR;
- The second element is the pressure, measured and controlled using the pressure model, an extension of OPIR.

The daily water demand pattern can be used to determine two variables, the day peak factor (DPF) and the night peak factor (NPF). These two variables can be estimated by dividing the maximum day consumption or the minimum night consumption by the average consumption of the day (Trifunovic, 2006). The NPF and the DPF are shown in Figure 15.

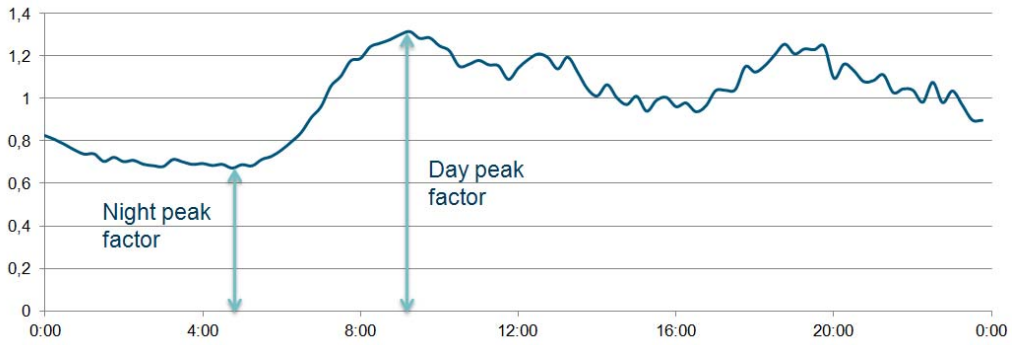


Figure 15: Day peak factor and night peak factor

The element pressure can be used directly as a variable.

7 VARIABLES FOR NEW APPROACH

The nature of the relationship between the performance of NRW or its components, the variables obtained by OPIR and the variables obtained at the beginning of a project will be discussed in this chapter. Chapter 6 concludes that three variables can be obtained from OPIR; the NPF, the DPF and the pressure. Boomsma (2013) stated that the variables that can be obtained at the beginning of a project are the number of connections, the length of the mains and the sector type. The connection density can be used as well, since this variable can be estimated using the number of connections and the length of the mains. Table 4 shows which variable can be obtained from OPIR and which variables at the beginning of a project.

Table 4: Variables

Variables obtainable from OPIR	Variables obtainable at the beginning of a project
Day peak factor	Connection density (# con/km)
Night peak factor	Number of connections
Pressure (mWc)	Length of the mains
	Sector type

The sub-chapters will discuss the variables and their relationship with each other. The performance of NRW, the real losses and the apparent losses will not be discussed, since these are discussed in the previous chapters.

7.1 Day peak factor (DPF)

The DPF is the ratio of the maximum peak flow and the average consumption as is described in sub-chapter 6.1.

Zhang (2005) describes in his research that there is a relationship between the DPF and the performance of NRW. An equation is developed to estimate the DPF for areas with domestic use only. Table 5 shows the DPF equation that is determined for different NRW (%) values, the P expresses the population in 1000s (Zhang, 2005). The last row in the table shows the estimated DPF values using P = 2000.

Table 5: Day peak factor expression (Zhang, 2005)

Day peak factor expression			
NRW = 0%	NRW = 5%	NRW = 10%	NRW = 20%
$2.50 + \frac{2.08}{\sqrt{P}}$	$2.43 + \frac{1.97}{\sqrt{P}}$	$2.35 + \frac{1.86}{\sqrt{P}}$	$2.20 + \frac{1.97}{\sqrt{P}}$
2.55	2.47	2.39	2.24

P is the population in 1000s.

The equation shows that there is a relation between the DPF and the PI NRW (%). A critical note towards the equations is that the NRW (%) is not the most accurate PI and that only the real losses are included in the used NRW (%).Zhang (2005) has recommended that these equations should be validated to be able to use them in real life situations. The research of Zhang (2005) proves that the DPF is influenced by the performance of the real losses. Since the real losses

(l/con/day) are a component of NRW (l/con/day), it is assumed that NRW (l/con/day) also influences the DPF. The nature of the relationships is not specified.

The DPF is also influenced by the following variables:
- Number of consumers (Zhang, 2005);
- Consumption category (Trifunovic, 2006);
- Water price;
- Weather variables;
- Household composition (Arbués, et al., 2003);
- Cultural differences (Blokker, 2010).

7.2 Night peak factor (NPF)

The NPF is the ratio between the minimum night flow (MNF) and the average consumption, as is described in sub-chapter 6.1. No literature is found that describes the kind of influence of NRW (l/con/day), the real losses (l/con/day) and the apparent losses (l/con/day) on the NPF. The literature does show that the MNF is predominantly determined by leaks and therefore it can be assumed that the minimum night flow rate minus the consumption rate equals the maximum real losses. The consumption rate during the night is determined by the size of the area, thus influenced by the number of connections and the type of connections (Water Loss Task Force, 2007). It can be concluded that the NPF is influenced by the volume of the real losses and the size of the area (number of connections and type of connections).

It is unknown whether the NPF is also influenced by the performance of the real losses as it is influenced by the volume of the real losses, though the same relationship is assumed. Since the real losses (l/con/day) are a component of NRW (l/con/day), it is assumed that NRW (l/con/day) also influences the NPF. The nature of the relationship between the NPF and the performance of NRW and the performance of the real losses is not defined.

7.3 Pressure

Since the real losses is a component of NRW, the pressure should be taken into account as well, in fact Farley and Trow (2003) have stated that the pressure is the second most important factor determining the real loss volume. The pressure influences the flow rate in the system and thus the flow rate out of the leaks (Walski & Giustolisi, 2012). When the pressure is lowered, the flow rate will decrease, which leads to a drop in water leaking out of the system (Yeboah, 2008). The break frequencies are also influenced by the pressure. Figure 16 shows that there is a relationship between the reduction of the pressure and the reduction of the break frequency of the mains (Thornton, et al., 2008).

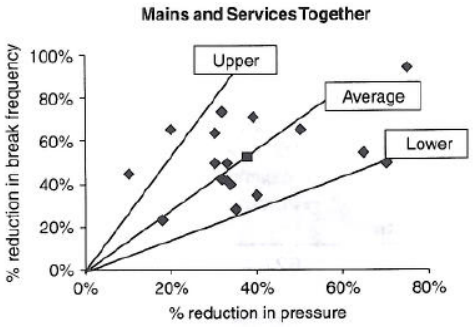


Figure 16: Plot of % reduction in pressure vs. % reduction in new break frequency (Thornton, et al., 2008)

Lambert & McKenzie (2002) have stated that the assumption of a linear relationship between the volume of the real losses and the pressure is most reliable for large systems with a mixed metal and non-metal pipe work which operate with a pressure in the range of 30-70 meters. Systems only using plastic pipes have a relationship of approximately the power of 1.5, while systems with metal pipes have a relationship of approximately the power of 0.5.

It is assumed that the relationship with the pressure will remain, even though the performance of the real losses will be used instead of the volume of the real losses. It is assumed that there is a relationship between NRW (l/con/day) and the pressure, since there is a relationship between the real losses (l/con/day) and the pressure. Since the performance is used instead of the volume it is unknown what the nature of the relationships is.

7.4 Connection density

The connection density can be obtained by dividing the number of connections by the length of the mains (km). Figure 17 shows that there is a non-linear relationship between the UARL (l/con/day) and the connection density, therefore it is assumed that there is a nonlinear relationship between the real losses (l/con/day) and the connection density. The same relationship is also expected between the NRW (l/con/day) and the connection density, since the real losses (l/con/day) are a component of NRW (l/con/day).

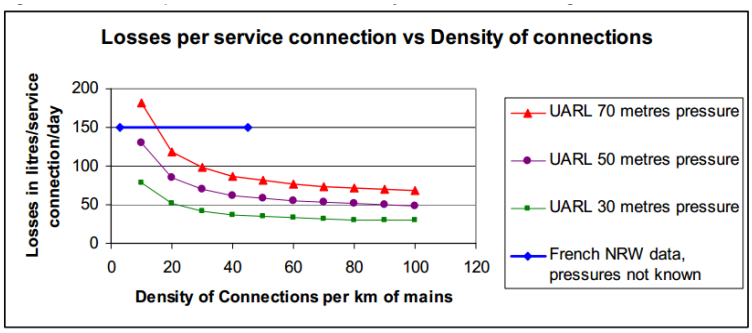


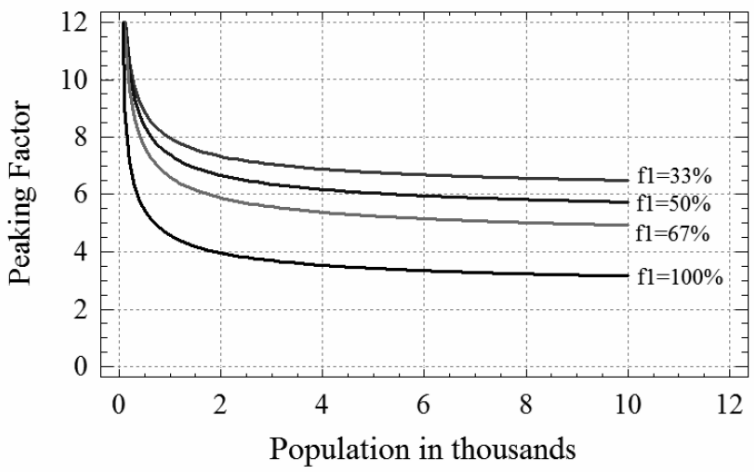
Figure 17: Losses per connection/day, UARL and Cemagref data (Fantozzi, et al., -)

7.5 Number of connections

Connection leaks are the most common type of leakage in the water distribution system. Hamilton, et al. (2006) stated that the real losses are influenced by the number of connections. Careful consideration of the number of connections could help reduce the effect and amount of real losses. Since the performance of NRW and its components is estimated by dividing the volume of the losses by the size of the area (specifically the number of connections), there is no relationship found between the performance of NRW and its components and the number of connections.

Sub-chapters 7.1 and 7.2 show that the DPF and NPF are also by other variables, one of these variables is the number of consumers. The nature of the relationship between the DPF and the number of consumers has been specified by Zhang (2005), Figure 18 shows this relationship. The nature of the relationship between the NPF and the number of consumers is not defined. Instead of using the number of consumers as a variable influencing the DPF and the NPF, the number of connections will be used. The number of connections can be monitored more closely than the number of consumers in an area, since new connections are registered by the water company. It is assumed that the same relationship will exists between the DPF and the number of connections as the one found for the DPF and the number of consumers. The nature of the relationship between the NPF and the number of connections is still unspecified.

Figure 18: Dimensionless peaking factor at the 99.9 percentile (Zhang, 2005)



7.6 Length of the mains

The longer the system, the more effort is needed to detect the small but running leaks that are unreported (Thornton, et al., 2008). Hamilton, et al. (2006) has stated that the length of mains influences the volume of the real losses and by considering the length of the mains the volume of NRW can be reduced. Since the performance of NRW and its components is determined by dividing the volume of the losses by the size of the area (specifically by the number of

connections), there is no relationship found between the performance of NRW and its components and the length of the mains.

The variable has no relationship with other variables and will not be used any further in the analysis.

7.7 Sector type

The sector type indicates what kind of development is present in the analyzed sector. The sector types can be divided in:

- Formal urban neat sectors: these sectors have a normal development, which can exist of housing, industry or a shopping area;
- Urban informal township style sectors: these are sectors that were illegally developed on the periphery of towns and cities, but are currently declared legal;
- Informal-untidy sectors: these sectors are the new illegal developments;
- Rural sectors: these sectors are located at the outskirts of the cities and towns.

No relationship is found between the sector type and the performance of NRW or its components.

Sub-chapter 7.1 shows that the DPF is not only influenced by NRW (l/con/day) and the real losses (l/con/day), but also by other factors, these factors include the consumption category, the income and the household composition. The sector type is used instead of these three variables, since it is difficult to determine these variables in areas where the administration is lacking unlike the sector type which can easily be determined by examining the sectors development and surroundings. All these variables differ for the different sector types, thus it is assumed that the sector type influences the DPF.

Sub-chapter 7.2 shows that the NPF is also influenced by the type of connections, since this is sometimes difficult to determine the sector type is used. Other types of connections are assumed in an untidy-informal sector, than in a formal urban neat sector. The former sector will probably have more standpipes than the latter sector.

7.8 Conclusion

It can be concluded that the real losses (l/con/day) are influenced by the connection density and the pressure. Since the real losses (l/con/day) are a component of the NRW (l/con/day) the same variables influence the NRW (l/con/day). A nonlinear relationship is specified between the connection density and the real losses (l/con/day) (and thus the NRW (l/con/day)). The nature of the relationship between the pressure and the performance of NRW and the real losses is not specified and should be examined.

The variables DPF and NPF are influenced by the number of connections, the sector type and the real losses (l/con/day) or NRW (l/con/day). The number of connections is used as influence on the DPF and the NPF, instead of the number of consumers. The sector type is used as influence on the DPF instead of the consumption category, the income and household composition. The sector type is used as influence on the NPF instead of the type of connections.

The DPF is also influenced by the cultural differences and the weather variables, these variables can be neglected since the data will be collected from one city. The nature of the relationship between the DPF and the number of connections is specified as nonlinear, all the other relationships should be examined to determine their nature.

There is little known about the variables that influence or are influenced by the apparent losses (l/con/day), therefore the same variables will be used for the apparent losses (l/con/day) as the ones found influencing or influenced by the real losses (l/con/day). Even though the relationship is not proven in the literature, the variables will be used to test their relationship with the apparent losses (l/con/day). The nature of the relationship between the performance of the apparent losses and the variables should be examined.

Table 6 shows the above described relationships. The top row displays the dependent variables, which values depend on that of other variables (Oxford dictionaries, 2013). The other rows show the independent variables; these are variables which value is not influenced by other variables (Oxford dictionaries, 2013).

Table 6: Variables

Non-revenue water (l/con/day) **Real losses (l/con/day)** **Apparent losses (l/con/day)**	**Day peak factor** **Night peak factor**
Pressure	Non-revenue water (l/con/day) Real losses (l/con/day) Apparent losses (l/con/day)
Connection density	Number of connections
	Sector type

8 METHOD

The relationships between the variables will be modeled in a structural model using a measured variable path analysis. The path analysis can be used to measure the direct and indirect effects one variable has on another variable. This chapter will discuss the method that will be used to determine the structural equations forthcoming from the structural model, as well as how these structural equations will be validated and which program will be used for the statistical analysis.

8.1 Choice of method

There are two general approaches that can be used to determine a structural equation; these two general approaches are the regression techniques and the machine learning techniques. Both techniques eventually determine the structural equation shown in equation 8-1.

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n \quad \text{8-1}$$

"Regression analysis is a statistical technique that attempts to predict the values of one variable using the values of one or more other variables" (Allen, 1997). The relationships that are determined using a regression analysis can be described in a mathematical form to enable one to predict the dependent variable. The β_n estimates in equation 9-1 are determined using the regression analysis. The regression analysis also assesses how accurate the independent variables can predict the dependent variable, by determining the proportion of the variation in the dependent variable that can be accounted for by the variation in the independent variables (R^2). And most importantly it analyses whether the relationship between variables is statistically significant and thus assesses whether there is a relationship between the variables. The different regression methods enable researchers to work with all kinds of data (Field, 2009).

The same structural equation can be determined using the machine learning techniques. Machine learning techniques are less concerned with the identification of the relationship between the variables, but still focus on producing an equation that generates useful predictions. The methods used in machine learning techniques are normally fitting very complex 'generic' models, which are not related to any theoretical understanding of the causal relationships (StatSoft, -). A research performed by Bougadis, et al. (2005) concludes that the machine learning technique outperformed the regression technique in modeling a short-term water demand forecast model. When the dataset has less than a hundred cases, there is not enough information to train the network (StatSoft, -). Since the data set dispose of a hundred cases a regression analysis will be used.

8.2 Application of regression analysis in water demand management

Regression analysis has been used for many disciplines; one of the disciplines is water demand management. The topic NRW is relatively new, the term NRW has only been introduced in the early 90's and the ILI method to analyze the real losses was introduced in 2000. No research has yet been performed towards predicting NRW or its performance, but the regression analysis has been used to forecast the water demand. Zhou, et al. (2000) has stated the following about forecasting the water demand with the help of regression analysis:

"The econometric approach is based on statistically estimating historical relationships between different factors (independent variables) and water consumption (the dependent variable) assuming that those relationships will continue into the future, getting forecasts of the factors related to water consumption and basing the water forecast on those."

Water demand forecasting is needed to efficiently operate and manage the water supply system. One of the used methods is the time series analysis (Zhou, et al., 2000, Zhou, et al., 2002, Bougadis, et al., 2005) in which water consumption is directly forecasted, without having to forecast other factors on which water consumption depends (Zhou, et al., 2000). The water demand forecast in OPIR is also based on the time series analysis (Bakker, 2012). Another method that is used to predict the water demand is the multiple regression method (Brekke, et al., 2002, Yasar, et al., 2012, Bougadis, et al., 2005, Babel, et al., 2006), this method is used to find the most suitable combination of independent variables to forecast water demand (Yasar, et al., 2012).

The regression analysis is never used to analyze models which include the performance of NRW, but it is used to analyze models predicting the water demand. Since NRW is a component of the water demand, it is assumed that the same method can be used to analyze the models which include the performance of NRW.

8.3 Regression analysis

The specifics of the variables and the relationship between the variables are examined to determine the right regression method. The regression model must meet the following specifications:

- Continuous dependent variable (gives a value for every observation and can take any value on the measurement scale that is used (Field, 2009));
- Continuous and nominal independent variables (nominal variables are variables with two or more categories, which do not have an intrinsic order (Field, 2009));
- More than one independent variable in the structural equations;
- Three dependent variables must be interpreted, of which one dependent variable is an independent variable used to predict the other dependent variables.

As discussed at the beginning of the chapter a path analysis will be used to develop a structural model of the relationships between the variables. The ordinary least squares (OLS) multiple linear regression method will be used to analyze the structural equations originating from the structural models. A backward elimination will be used to find the best combination of independent variables. A backward elimination is preferred as stepwise model since it is least likely to miss an independent variable that does predict the outcome. Variables are eliminated when the variable does not make a statistically significant contribution ($p>0.1$) to the prediction of the dependent variable. The final model must have a significance level of $p<0.05$, otherwise the model will be rejected. Dummy variables are used to analyze the nominal variable, sector type, in the multiple regression analysis (Field, 2009).

To determine whether the multiple regression analysis can be used the following assumptions must be met: (1) outliers must be examined, (2) there should be linearity in the dataset, (3) homoscedasticity must be present in the data, (4) as well as normality, (5) no independent errors should be present, (6) as well as no multicollinearity. When the assumptions are violated robust techniques are used (Field, 2009).

Appendix IV elaborates on the chosen regression method and the assumptions that must be met, it also discusses the robust techniques that will be used when the assumptions are violated.

8.4 Validation

A validation of the model should be performed to determine whether the structural equations that have been determined are correct; an internal and external validation will be performed.

An internal validation is needed to determine whether the structural equation is correct and accurate enough. First of all the adjusted R^2 is determined, this value describes the loss of predictable power or also called the shrinkage in the model. The lower the number, the less the variance of the independent variables influences the variance of the dependent variable.
The structural equations will also be cross-validated by splitting the data set in a training set and a validation set (Field, 2009). The training set will consists of 90% of the observations and will be used for the analysis. The other observations are selected for the validation set and will be used to predict the dependent variables using the structural equations resulting from the regression analysis. The cases for the validation and training set will be selected randomly. A plot will be created where the predicted values will be plotted against the observed values of the validation set; this will make it possible to visualize the preciseness of the model.

External validation will be executed to determine whether the structural equations can be used in other areas as well. Data is gathered from Malawi, Mozambique and Indonesia, but also from France and Belgium. Again the predicted and observed values will be plotted on the axis, even though there are too few data points to significantly determine whether the equations can be used for other areas, it might give a trend.

8.5 Program

There are a lot of statistical packages, which can perform a multiple regression analysis based on the OLS estimation. The statistical package that will be used is SPSS statistics. SPSS is one of the most widely used programs in statistical analysis. Not only statistical analysis can be performed in SPSS, but also data management and data documentation is featured in the software. A multiple regression analysis using a backward stepwise method can be performed in SPSS, as well as the robust techniques needed when an assumption is violated (Field, 2009).

9 DATA COLLECTION

This chapter will discuss the data collection. The first sub-chapter will discuss the data that is needed to perform the analysis. The second sub-chapter will describe the background of the collected data and the last sub-chapter will describe the requirements that must be met by the collected data.

9.1 Required data

Data must be collected to be able to perform a multiple regression analysis. Some of the data that is needed for the variables is pretty straight forward, while other variables need to be estimated. Table 7 shows for every variable the data that is needed to determine this variable.

Table 7: Required data for the estimation of the variables

Variable	Data required
NRW (l/con/day)	Non-revenue water (kl/day)
	Number of connections
Real losses (l/con/day)	Real losses (kl/day)
	Number of connections
Apparent losses (l/con/day)	Apparent losses (kl/day)
	Number of connections
Pressure	Pressure in mWc
Length of mains	Length of mains (km)
Connection density	Number of connections
	Length of mains (km)
Type of sector	Type of sector
Day peak factor	Maximum day flow (m^3/s)
	Average day flow (m^3/s)
Night peak factor	Minimum night flow (m^3/s)
	Average day flow ($m^3.s$)
Number of connection	Number of connections

Some data is used multiple times to determine several variables; Table 8 shows the list of data that is needed to perform the multiple regression analysis.

Table 8: Data required for multiple regression

Data required
Non-revenue water (kl/day)
Real losses (kl/day)
Apparent losses (kl/day)
Number of connections
Pressure in mWc
Length of mains (km)
Type of sector
Maximum day flow (m^3/s)
Average day flow (m^3/s)
Minimum night flow (m^3/s)

9.2 Collected data

The collected data is provided by the company JOAT. JOAT is a South-African company which specializes in all aspects of water management, one of their specializations is NRW. RHDHV and JOAT have a partnership concerning the topic of NRW outside South-Africa.

JOAT has provided two datasets which will be used; the datasets are from the areas of West and South Durban. Durban is the third largest city in South Africa and is the largest city in the province KwaZulu-Natal. Durban is a part of the eThekwini metropolitan municipality, which includes Durban and some surrounding towns. The total population in the city is nearly 3.5 million people. Durban and its surroundings are hilly, which could mean that more pressure is needed to distribute the water. Durban is located by the sea and has a warm subtropical climate (Wikipedia, 2013). Both West and South Durban have various types of sectors included in the dataset, though overall South Durban is more developed than West Durban.

Every dataset consists of different sectors for which the water balance is made and for which the characteristics are described. The areas (South and West Durban) each have their own engineer that analyzes the incoming data and determines the water balance. JOAT has strict companies' standards that are followed determining the volumes in the water balances. The water balances are made using measurements in the system, but also with the experience the engineers gained in the field. The values in the water balances are an approximation made by the engineers. Even though the engineers of JOAT have a lot of experience and there is a company standard it can happen that the water balances have incorrect values (Pena, 2013).

The dataset of West Durban provides almost all the data that is described in Table 8. The only data that is incomplete is the maximum day flow, which means that the DPF cannot be determined for all sectors.

The dataset of South Durban provides less information than the dataset of West Durban. The maximum day flow is not included for any sector, meaning that the DPF is not available. The type of area is also not described, though with the use of satellite images, Wikipedia and Google the type of area could be determined for the sectors in South Durban.

9.3 Cleaning the dataset

Data cleaning is needed to prevent highly misleading outcomes resulting from the regression analysis. First of all the sectors that did not have a water balance are deleted from the datasets.

The second cleaning step is performed by evaluating the water balances made by the engineers. This step is performed by analyzing the comments of the engineers, who evaluated the determined water balance. These comments are provided in the supplied data. The comments provided by the water balances of South Durban show that there are some questions regarding the input of the outlet meters, whenever there are questions regarding the outlet meters the sector is deleted. The comments provided by the water balances of West Durban are more specific, though the comments are not trustworthy. The dataset that is analyzed is from June 2012, but the comments in the dataset of August 2011 are the same. No sectors are removed from the West Durban dataset, but this does mean that the dataset should be examined thoroughly to exclude sectors with faulty water balances.

The third step is to prevent small sectors influencing the data too much. When a sector is small the consumers in the sector have a lot of influence on the drink water demand pattern. These values can differ a lot from sectors of normal proportions. To avoid the small sectors from influencing the dataset too much the sectors must meet the following rules:

- Number of connections > 100;
- Length of mains (km) > 5 km;
- System input volume (kl/day) > 250 kl/day.

When sectors do not meet these rules, the sectors are deleted from the dataset.

The fourth step is performed by examining whether there are sectors severely larger than all the other sectors. The water demand pattern in a large area is influenced by a lot of consumers all having their own demand pattern, this results in a pattern that has flattened peaks. The sectors must meet the following rule, if not the sector is deleted from the dataset:

- Length of mains (km) < 200 km.

The last step is performed by examining the dataset for out-of-range values. The first value that immediately catches the eye is the NPF values that are bigger than one. The minimum night flow must be exceeding the average day flow when this happens. The sectors where this happens do not have a normal water demand pattern and thus are excluded from the analysis. Another value that attracts the attention is the connection density that is double the value of the second highest connection density. This value is so out of range, that it is excluded from the analysis. Another sector that is excluded is the sector that only has negative NRW, real losses and apparent losses values.

When all the rules are applied the dataset exists of 99 sectors from which 98 sectors have a known NPF and 25 sectors have a known DPF. Since there are too few observations with a known DPF, the DPF will not be used as a variable in the analysis.

The ranges of the used data and the average value are described in Table 9.

Table 9: Ranges of the used variables

	Lowest value	Highest value	Average value
NRW (l/con/day)	10,85	9806,69	1402,05
Real losses (l/con/day)	15,43	4480,87	648,05
Apparent losses (l/con/day)	-374,52	4413,88	513,83
Pressure (mWc)	35	79	55,74
Connection density (# con/km)	3,83	121,09	73,01
Number of connections	227	10714	2132,19
Day peak factor	0,74	4,15	0,38
Night peak factor	0,022	0,897	1,96

10 RESULTS

The first sub-chapter will discuss the structural models that can be made using the relationships found in chapter 7. The second sub-chapter discusses the nature of the relationships that are modeled in the structural model. The third sub-chapter will analyze the equations forthcoming from the structural models. The last sub-chapter will validate the analyzed equations.

10.1 Structural model

Sub-chapter 7.8 concludes which variables are the dependent variables and which variables are the independent variables. The relationships between these variables are modeled in a structural model according to the rules of path analysis. The dependent variables are shown in ovals, while the independent variables are shown in rectangles. The straight lines show the direct effect of a variable on the other variable, for example km of mains influences NRW (Kline, 2011). Figure 19 shows the developed structural models all in one model, there are structural models developed for NRW (l/con/day), the real losses (l/con/day) and the apparent losses (l/con/day).

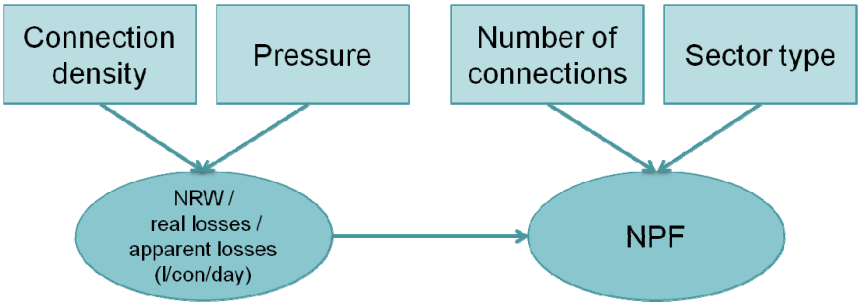


Figure 19: Model specification with dependent variables performance indicator (l/service connection/day) and the night peak factor and the day peak factor

The model specification leads to the structural equations shown in Table 10.

Table 10: Structural equations for model specification

Model	Structural equations
NRW (l/con/day)	$NRW\ (l/con/day) = \beta_0 + \beta_1 Pressure\ + \beta_2 connection\ density$
	$NPF = \pi_0 + \pi_1 NRW(l/con/day) + \pi_2 number\ of\ connections + \pi_3 Sector\ type$
Real losses (l/con/day)	$Real\ losses\ (l/con/day) = \beta_0 + \beta_1 Pressure\ + \beta_2 connection\ density$
	$NPF = \pi_0 + \pi_1 real\ losses(l/con/day) + \pi_2 number\ of\ connections + \pi_3 Sector\ type$
Apparent losses (l/con/day)	$Apparent\ losses\ (l/con/day) = \beta_0 + \beta_1 Pressure\ + \beta_2 connection\ density + \beta_2 Sector\ type$
	$NPF = \pi_0 + \pi_1 apparent\ losses(l/con/day) + \pi_2 number\ of\ connections + \pi_3 Sector\ type$

10.2 Relationships

Sub-chapter 7.8 concluded that the nature of some of the relationships could be defined by the reviewed literature. Scatter plots will be examined to define the nature of the relationships between the dependent and independent variables which were not given by the literature (Field, 2009). Appendix V provides the scatter plots used to examine the relationships between the dependent and independent variables. The dependent variable is plotted on the y-axis and the independent variable on the x-axis.

Table 11 shows the relationships between the dependent variables NRW (l/con/day), real losses (l/con/day) and apparent losses (l/con/day) and the independent variables connection density and the pressure. The literature defined the relationship between the connection density and the performance of the real losses and NRW as nonlinear. The relationship is further examined using the scatter plots, from the scatter plots can be concluded that the relationship can be described with a log function. The same relationship is found between the apparent losses (l/con/day) and the pressure. The scatter plots also show that a linear relationship exists between the performance of NRW or its components and the pressure, though it is less strong than the relationship between the volume of NRW or its components and the pressure.

Table 11: Relationships for dependent variables non-revenue water (l/con/day), real losses (l/con/day) and the apparent losses (l/con/day)

Dependent variables	Independent variables	
	Connection density	Pressure
NRW (l/con/day)	Nonlinear[a] log relationship	Linear
Real losses (l/con/day)	Nonlinear[a] log relationship	Linear
Apparent losses (l/con/day)	Nonlinear log relationship	Linear

[a] This relationship is defined in the literature

Table 12 displays the relationships between the dependent variable NPF and the independent variables. The natures of the relationships between the NPF and the independent variables have not been defined by literature. The independent variables NRW (l/con/day), real losses (l/con/day) and apparent losses (l/con/day) will not be used at the same time, since the variables belong to different structural models. The scatter plots show that the relationships between NRW (l/con/day) and the NPF, the real losses (l/con/day) and the NPF and the apparent losses (l/con/day) and the NPF differ, there is a linear relationship with the real losses (l/con/day) and the apparent losses (l/con/day) and a nonlinear relationship with NRW (l/con/day). The divergent relationship exists because the two linear relationships have an opposite slope. The scatter plots show that the NPF reduces when the apparent losses grow, while the NPF increases when the real losses is increasing (as is expected.) The nonlinear relationship between the NPF and NRW (l/con/day) can be expressed with a log function. The relationship between the NPF and the number of connections can be defined as nonlinear; the relationship is described with an exponential relationship. The box plot used to examine the relationship between the NPF and the sector type shows that there is a relationship. The mean values are close together for the urban neat sector, the informal-untidy sector and the rural

sector, while the township sectors' mean is a bit higher. The range of the numbers also differ, the range is the largest in the rural sector, while in the township sector it is the smallest. No outliers were found in the box plot.

Table 12: Relationships for dependent variables DPF and NPF

Dependent variables	Independent variables				
	NRW or components			Number of connections	Sector type
	NRW (l/con/day)	Real losses (l/con/day)	Apparent losses (l/con/day)		
Night peak factor	Nonlinear log relation	Linear	Linear	Nonlinear exp relation	Relationship

10.3 Analysis of the models

10.3.1 Assumptions

The assumptions are checked using the rules in appendix IV. The outliers are checked by examining the z-scores of the observations. The equations predicting NRW (l/con/day), the real losses (l/con/day) and the apparent losses (l/con/day) do not satisfy the rules, this can happen when the model is a poor representation of the actual data. Three outliers are found with a z-score bigger than 3.29. Two of the z-values are caused by the same sector (Folweni 2); the other outlier is caused by the sector Almond Road. The outliers are not influential outliers since the Cook's distance value is smaller than one and the centered leverage value is smaller than$(2\,(k+1)/n)$. Appendix VI displays the outliers per structural equation model and the Cook's distance value and the centered leverage value for the outliers > ±3.29.

The assumptions normality, homoscedasticity, no independent errors and no multicollinearity are tested using the methods described in Appendix IV. Appendix VI shows the plots and the outcome of the tests. From the plots and tests is concluded that some of the assumptions are violated. Table 13 shows which structural equation violates which assumption. When the normality assumption is violated a bootstrap technique will be used for the analysis and when the assumption of homoscedasticity is violated a weighted least squares regression will be used instead of the ordinary least squares regression.

Table 13: Assumption violated

	Normality	Homoscedasticity	No independent errors	Multicollinearity
NRW (l/con/day)	X	X		
Real losses (l/con/day)	X			
Apparent losses (l/con/day)	X	X		
NPF using NRW	X			
NPF using real losses	X			
NPF using apparent losses	X			

It should be noted that the assumptions have been checked for the structural equations that are specified in chapter 10.1. When independent variables are eliminated during the regression analysis, the assumptions should be checked again.

10.3.2 Results multiple regression analysis

Each paragraph will discuss one of the developed structural models. The paragraphs will first show the analyzed structural model plotted with the β-coefficients resulting from the regression analysis performed on the corresponding structural equations. The β-coefficients show how much influence an independent variable has on the dependent variable, the value refers to the number of standard deviations the value of the dependent variable will change, per standard deviation increase of the value of the independent variable. The β-coefficient can vary between the -1 and 1; a positive relationship exists when the value of the dependent variable increases due to an increase in the value of the independent variable. A negative relationship exists if the value of the dependent variable will decrease, when the value of the independent variable is increasing (Field, 2009). Then the results of the regression analysis per structural equation will be discussed. Ending the paragraph with a table showing the predicted value and the boundaries between which the predicted value lies using the 95% confidence intervals resulting from the regression analysis. The average values described in sub-chapter 9.3 are used to estimate the predicted value and its boundaries.

Model specification of non-revenue water (l/con/day)

The results for the structural model of NRW (l/con/day) are shown in Figure 20. The model shows that there is no relationship between the dependent variable NPF and the independent variables NRW (l/con/day), number of connections and the sector type. It does show that there is a relationship between the dependent variable NRW (l/con/day) and the independent variables connection density and pressure. The independent variables predicting the NRW (l/con/day) have an opposite slope; the connection density has the strongest relationship with NRW (l/con/day).

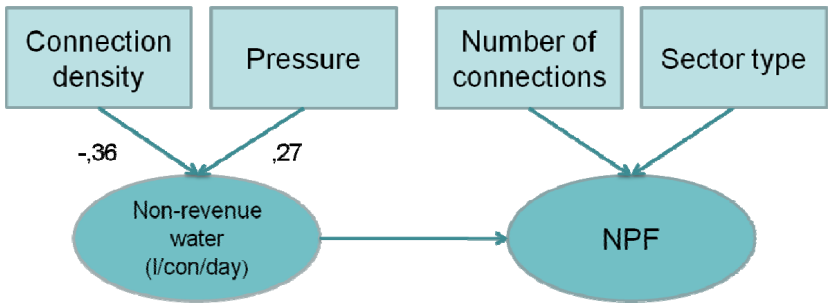


Figure 20: Results of the regression analysis on the model of non-revenue water (l/con/day)

The results of the regression analysis for the structural equation of NRW (l/con/day) show that the two independent variables explain 17,7 % of the variance (R^2=0.18, F=9.25, p<0.001). NRW (l/con/day) is predicted by both the connection density (β=-0.36, p<0.001) and the pressure

(β=0.27, p<0.05). Equation 10-1 shows the structural equation that is determined for the dependent variable NRW (l/con/day).

$$NRW\ (l/con/day) = 3390.10 - 2574.23\ \log(Connection\ density) + 37.31\ Pressure \quad \textbf{10-1}$$

The results of the regression analysis for the structural equation of the NPF show that the independent variables can be accounted for 23,4% of the variance in the NPF (R^2=0.23, F=4.70, p<0.01), but the independent variables number of connections (β=.13, p=.226) and NRW (l/con/day) (β=.12, p=.234) do not contribute statistically significant (p>.1) to the predicted NPF. The NRW (l/con/day) is the least significant independent variable; this variable is eliminated from the model.
The regression analysis is performed again using the independent variables number of connections and the sector type. The adapted structural equation does not violate the assumption of homoscedasticity or the assumption of no independent errors (Durbin-Watson = 1,595). The results show that the independent variables can be accounted for 22% of the variance in the NPF (R^2=0.22, F=5.49, p<0.01). The results also show that the number of connection still does not contribute statistically significantly to the model (β=.12, p=.13).
The structural equation left for the NPF only includes the independent variable sector type; this structural equation would just give an average NPF for each sector type. The regression analysis is not further used, since it would not provide a structural equation that is wanted.

Table 14 shows the boundaries between which the predicted value NRW (l/con/day) can be determined with a 95% confidence interval, using the average values of the provided data.

Table 14: Boundaries non-revenue water (l/con/day) model

	Predicted value	Lower bound	Upper bound
NRW (l/con/day)	672.98	-5921.27	11763.38

Model specification of real losses (l/con/day)
Figure 21 shows the structural model for the real losses (l/con/day) in which the relationships are defined using the β-coefficients. The model shows that there is no relationship between the pressure and the real losses (l/con/day). The relationships towards the NPF are all positive, when the independent variables increase the NPF will increase. The relationships that are found are average in strength; the relationship between the number of connections and the NPF is the least strong.

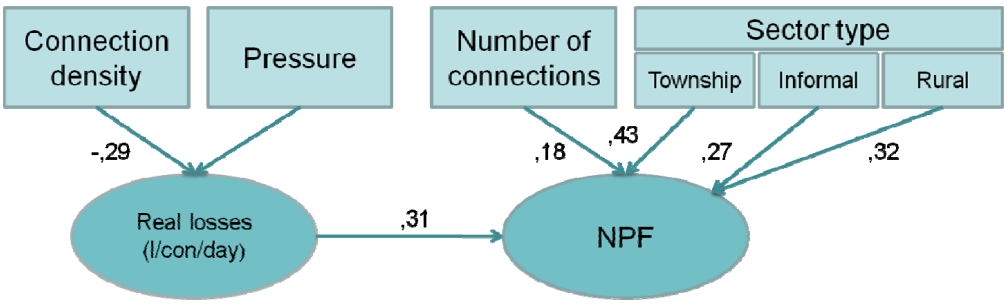


Figure 21: Results of the regression analysis on the model of the real losses (l/con/day)

The results of the regression analysis on the dependent variable real losses (l/con/day) show that the independent variables explain 8,8% of the variance (R^2=0.09, F=4.15, p<0.05), but the contribution of the independent variable pressure (β=.06, p=.52) is not statistically significant (p>.1).
The regression is performed again using only the connection density as an independent variable. The adapted structural equation does not violate the assumption of homoscedasticity (Koenker test p>.05) or the assumption of no independent errors (Durbin-Watson value=2.07). The connection density (β=-.29, p<.05) explains 8,4% of the variance in the real losses (l/con/day) (R^2=0.08, F=7.96, p<0.01). The corresponding structural equation for the real losses (l/con/day) is shown in equation 10-2.

$$Real\ losses\ (l/con/day) = 1802.38 - 739.86\ \log(Connection\ density) \qquad \textbf{10-2}$$

The independent variables predicting the NPF in this model explain 30,8% of the variance (R^2=0.31, F=6.84, p<0.001). The number of connections significantly predicts the NPF (β=.18, P<0.05), as does the real losses (l/con/day) (β=.31, p<0.01). Dummy variables are created to be able to use the nominal variable sector type in the regression analysis. The urban formal neat sector type is used as the baseline sector type. The results show that the urban formal neat vs. township significantly predicts the NPF (β=.43, p<.01), as does the urban formal neat vs. informal untidy (β=.27, p<.05) and the urban formal neat vs. rural (β=.32, p<.05). The structural equation shown in 10-3 results from the regression analysis; a one should be entered when the sector can be identified for the sector type variable in the equation, a zero when it cannot be identified as the specified sector type.

$$\begin{aligned} NPF = {} & -.242 + .237 * Township + .167 * Informal\ untidy + .177 * Rural + .417 \\ & * Number\ of\ connections^{0,00004483} + 9.847E^{-05} \\ & * Real\ losses\ (l/con/day) \end{aligned} \qquad \textbf{10-3}$$

Table 15 shows the boundaries between which the predicted variables values lie.

Table 15: Boundaries real losses (l/con/day) model

	Predicted value	Lower bound	Upper bound
Real losses (l/con/day)	423.74	-1680.29	2441.419
NPF	0.239	-0.575	0.974

Model specification for apparent losses (l/con/day)

The β-coefficients resulting from the regression analysis from the structural equations which define the structural model of the apparent losses (l/con/day) are shown in Figure 22. The structural model shows that there is no relationship between the apparent losses (l/con/day) and the NPF, the number of connections and the NPF and the sector type and the NPF. There is a relationship between the connection density and the apparent losses (l/con/day) and the pressure and the apparent losses (l/con/day). These relationships are of the same strength, but the variables have an opposite effect on the apparent losses (l/con/day).

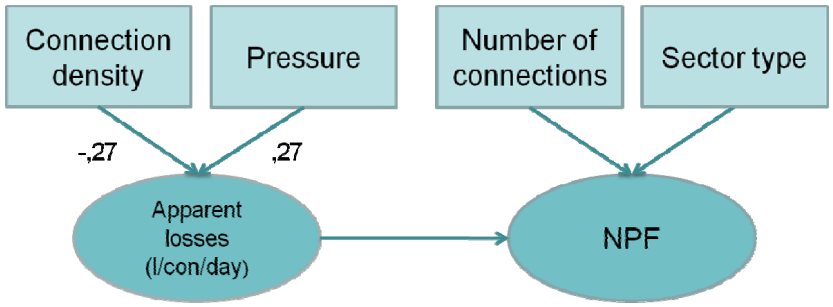


Figure 22: Results of the regression analysis on the model of the apparent losses (l/con/day)

The results of the regression analysis on the dependent variable apparent losses (l/con/day) indicates that the independent variables explain 13% of the variance (R^2=0.13, F=6.42, p<0.01). Both the connection density (β=-.27, p<0.05) and the pressure (β=.27, p<.05) significantly predict the apparent losses (l/con/day). The following structural equation is determined for the apparent losses (l/con/day).

$$Apparent\ losses\ (l/con/day) = 943.69 - 955.07\ \log(Connection\ density) + 19.24\ Pressure \quad \text{10-4}$$

The results show that the independent variables used to predict the NPF explain 22,1% of the variance (R^2=0.22, F=4.36, p<0.01), but the contributions of the variables number of connection (β=.11, p=.15) and the apparent losses (β=-.03, p=.73) are not significant (p>0.1). The apparent losses (l/con/day) are the first variable that is excluded as independent variable, leaving the same structural equation for the NPF in this model as for the NPF in the NRW (l/con/day) model. No structural equation is developed for the NPF in this model.

Table 16 shows the range between which the predicted value lies using a 95% confidence interval.

Table 16: Boundaries non-revenue water (l/con/day) model

	Predicted value	Lower bound	Upper bound
Apparent losses (l/con/day)	236.44	-2891.86	3741.25

10.2.3 Conclusion

Table 17 shows the structural equations and the corresponding R^2 value, the R^2 shows how much of the variability of the dependent variable is caused by the independent variables. The R^2 is the highest for the structural equation predicting the NPF and the lowest for the structural equation used to predict the real losses (l/con/day).

Table 17: Structural equations

Model	Structural equation	R^2
NRW (l/con/day)	$NRW\ (l/con/day) = 3390.10 - 2574.23\ \log(Connection\ density) + 37.31\ Pressure$	0.177
Real losses (l/con/day)	$Real\ losses\ (l/con/day) = 1802.38 - 739.86\ \log(Connection\ density)$	0.084
	$\text{NPF} = -.242 + .237 * \text{Township} + .167 * \text{Informal untidy} + .177 * \text{Rural} + .417 * \text{Number of connections}^{0,00004483} + 9.847\text{E}^{-05} * \text{Real losses (l/con/day)}$	0.308
Apparent losses (l/con/day)	$Apparent\ losses\ (l/con/day) = 943.69 - 955.07\ \log(Connection\ density) + 19.24\ Pressure$	0.130

10.3 Validation

10.3.1 Internal

The first step in the internal validation is assessing the adjusted R^2 values; this value describes the loss of predictable power or also called the shrinkage in the model. Wherry's adjusted R^2 shows how much of the variance in the dependent variable would be accounted for if the model was derived from the population from which the sample was taken. Stein's adjusted R^2 shows how much of the variance in the dependent variable would be accounted for if the model was derived from an entirely different data set (Field, 2009). The adjusted R^2 values and the R^2 values are shown in Table 18. The models are not responsible for a large part of the variance in the dependent variables (see R^2 value) and this is severely shrinking (see adjusted R^2 value). The variances for which the independent variables can be accounted for shrinks almost with 40% for the equations predicting the real and apparent losses (l/con/day) and with 30% for the equation predicting NRW (l/con/day). An exception is the shrinkage in the NPF; here the shrinkage is only shrinks 13%.

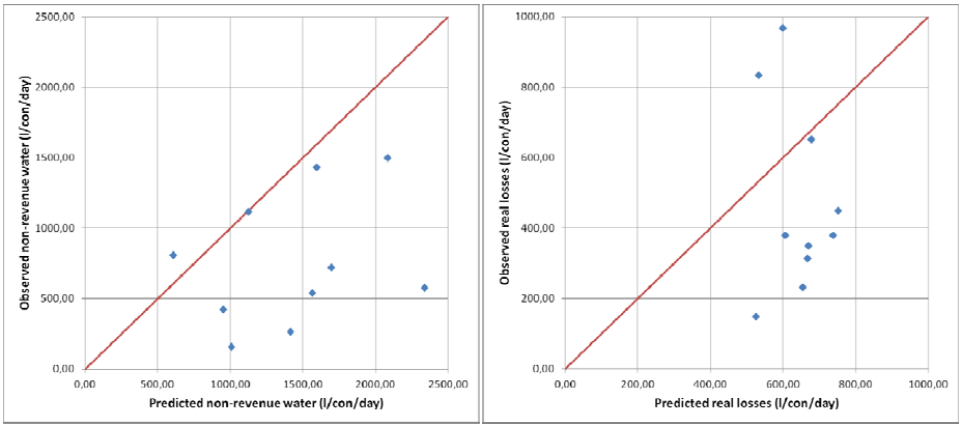
Table 18: Adjusted R^2 estimation

	R^2	Wherry adjusted R^2	Stein adjusted R^2
Non-revenue water (l/con/day)	0.177	0.158	0.128
Real losses (l/con/day)	0.084	0.073	0.052
Apparent losses (l/con/day	0.130	0.110	0.079
Night peak factor (using the model of the real losses (l/con/day))	0.308	0.213	0.266

The second part of the internal validation is carried out by examining the preciseness of the structural equations by plotting the predicted values on the X-axis and plotting the observed values on the Y-axis. The validation set is used to develop these plots; the data of the validation set is not used in the regression analysis.

Figure 23 shows the plots for the dependent values NRW (l/con/day), the real losses (l/con/day), the apparent losses (l/con/day) and the NPF. The blue dots are the prediction and observation of the relevant data points and the line shows where the observations should lie.

The plot of the dependent variable NRW (l/con/day) shows that the values are almost always overestimated. The variable real loss (l/con/day) is less often overestimated, but the data points do have a peculiar pattern. The predictions of the apparent losses (l/con/day) seem to follow the slope of the line, though not all the points are plotted on the same line and again most of the data is over predicted. The predictions of the NPF seem to fit the plotted line the best.



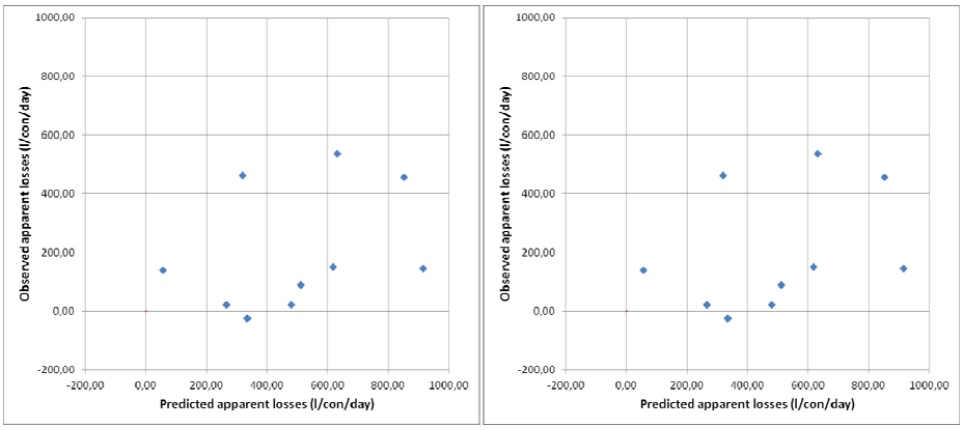


Figure 23: Observed by predicted value plots

10.3.2 External

The external validation is performed by predicting values for areas outside the project area. Table 21 shows the observed values of these areas and the predicted values estimated with the structural equations. The NPF could not be predicted since the real losses (l/con/day) are unknown for the areas where there is a known NPF.

Table 19: External validation

		NRW (l/con/day)		Real losses (l/con/day)		Apparent losses (l/con/day)	
		Observed	Predicted	Observed	Predicted	Observed	Predicted
Vietnam	Huta Galang	5133	-1152				
	Hoa Cuong	1265	-1074				
Belgium	Ophasselt	156	1990				
Mozam-bique	Maputo	15493	-3381	6851	-305	8623	-1487
France[a]	Sector 1-2			160	690		
	Sector 2			311	811		
	Sector 3			566	913		
Vietnam[b]		969	-1843	53	170	30	-933
Indonesia[b]		1031	-775	85	487	33	-542
Sri Lanka[b]		14505	-775	735	487	444	-542

[a] (Renaud, 2010)
[b] (Liemberger & McKenzie, 2005)

Table 21 shows that the predicted values of NRW (l/con/day) and the apparent losses (l/con/day) are almost always underestimated. This underestimation is caused by the pressure values of these areas, which are outside the boundaries of the pressure values used for the regression model. The real losses (l/con/day) are only predicted using the connection density;

the values show that the predicted values are closer to the observed ones, though almost all predicted values are overestimated. The under estimated predicted value is from Maputo, Mozambique; it deviates from the other values since there are around the 705 connections per km in this area, which is also outside the boundaries of the data used for the regression analysis.

11 DISCUSSION

This chapter will first discuss the limitations encountered during the research and the second sub-chapter discusses the results described in chapter 10.

11.1 Limitations of the research

In sub-chapter 1.2.3 is already discussed that the outcome of the analysis can be influenced by limitations experienced in the research. Even though the analysis is performed some limitations were noticed:

- It is difficult to collect NRW data that is verified, even collecting data which includes all the needed measurements is difficult. It occurs that water companies do not have the means to collect these measurements and when the companies have the means it occurs that NRW is none of their concerns, since there is a good performing system in place. JOAT determines the water balances every month for South and West Durban. The engineers determine these water balances based upon their experience and the company standards. Even though the data is complete, not all the data is measured as well as that it is not verified either. It is attempted to exclude wrongly captured measurements by cleaning the data. If there is still incorrect data included in the analyzed dataset, it could influence the reliability of the results and the results that are found;
- The second experienced limitation is the size of the dataset. The current dataset consists of 99 cases, from which 98 cases have a known NPF. The size of the dataset could influence the reliability of the results and the results that are found. The research already shows that the size of the dataset limits the method choice.

11.2 Results

The models show the following relationships:

- The relationship between the connection density and the real losses (l/con/day) and thus NRW (l/con/day) is described by Fantozzi, et al. (-) as a nonlinear relationship. This nonlinear relationship is described with a log function, the same relationship is found between the connection density and the apparent losses (l/con/day). The models show that the connection density has a negative relationship with NRW (l/con/day), the real losses (l/con/day) and the apparent losses (l/con/day). The relationships between NRW (l/con/day) and the connection density and the real losses (l/con/day) are expected since these relationships are described in the literature. However it should still be noted that the relationship between the apparent losses (l/con/day) and the connection density is not defined in the literature, but it seems logical that it would take more effort to check for example whether there are illegal connections (a component of the apparent losses) present in an area with only farms than in an city area;
- The models show that NRW (l/con/day) and the apparent losses (l/con/day) are influenced by the pressure, unlike the real losses (l/con/day);
    - The relationship between the pressure and the volume of the real losses is explained by Thornton, et al. (2008). It is assumed that the relationship would remain, even after transforming the real losses from volume to performance. However the results show that no relationship is found between the real losses (l/con/day) and the pressure. It is

peculiar that this relationship is not found, since it is proven by the literature that there is a strong relationship between the pressure and the real losses. The cause of this non-existing relationship might be found in the used dataset. As earlier discussed the dataset is small, as well as that the measurements in the dataset are not verified, both could have had an influence on the non-existing relationship;
  - The relationship between the apparent losses (l/con/day) and the pressure is not verified in the literature; still the relationship is analyzed to determine whether the apparent losses (l/con/day) might be influenced by the pressure. A logic explanation for this relationship is not found. Since the relationship is not verified, it is unknown whether the relationship resulting from the analysis is causal;
  - A relationship between NRW (l/con/day) and the pressure is expected, since it is assumed that there is a relationship between the real losses (l/con/day) and the pressure. However the results of the regression analysis show that there is no relationship between the real losses (l/con/day) and the pressure, meaning that the relationship between NRW (l/con/day) and the pressure exists based upon the relationship between the apparent losses (l/con/day) and the pressure. Since it is unknown whether this relationship is causal, it is also unknown whether the relationship between NRW (l/con/day) and the pressure is causal;
- The models show that the NPF can only be predicted using the real losses (l/con/day) model, the NPF cannot be predicted using the models of NRW (l/con/day) or the apparent losses (l/con/day):
  - It is expected that there is a relationship between the NPF and the real losses (l/con/day), because both variables are used to determine the minimum night flow. The results show that there is a relationship between the NPF and the real losses (l/con/day), which is expected;
  - The literature does not show whether the apparent losses (l/con/day) effect the NPF, this influence is examined with the analysis. As expected the analysis shows that the apparent losses (l/con/day) do not influence the NPF;
  - A relationship between the NPF and NRW (l/con/day) is expected, since it is expected that there is a relationship between the NPF and the real losses (l/con/day). The results show that there is no relationship between the NPF and NRW (l/con/day). The NRW (l/con/day) is not only determined by the real losses (l/con/day), but also by the apparent losses (l/con/day) which probably has a larger influence on NRW (l/con/day) than the real losses (l/con/day) in the used data set;
  - The literature shows that there is a relationship between the number of connections and the NPF. The scatter plot shows that the expected relationship can be defined as a nonlinear relationship using an exponent function to describe the relationship. Only in combination with the real losses (l/con/day) and the sector type, the number of connections can be used to predict the NPF. The influence of the number of connections on the DPF is not strong enough to only use this term to predict the NPF, this is unexpected since the literature does show that there is a relationship. This non-existing relationship could have been caused by the influence of the losses on the minimum night flow;

- The NPF is not only influenced by the real losses (l/con/day) and the number of connections, but also by the type of connections. Since it can be difficult to determine the type of connections in an area, the type of sector is used. The results show that the sector type can predict the NPF in combination with the real losses (l/con/day) and the number of connections.

The models are validated using an internal validation and an external validation.

- The boundaries between which the predicted variable lies, using a 95% confidence interval, is for all dependent variables really broad. An example is the NPF for which the lower boundary is negative and the upper boundary comes close to one. A negative value is not possible for the NPF and when the system is pressurized and has a continuous supply (the requirements for the analysis), it is unexpected that the NPF will come close to one;
- The R^2 values (which explains how much of the variability of a dependent variable can be explained using the independent variables) are for almost all structural equations below 0.25, except for the structural equation used to determine the NPF (R^2=0.31);
- The adjusted R^2 values (which describes the loss of predictable power or also called the shrinkage in the model) are for all equations below 0.25 and for two equations the values are even lower than 0.10;
- The plots made from the validation set show that the observation values and the predicted values mismatch. The structural equation that does predict the values the best is the equation used to predict the NPF, but even here there is a mismatch.

The described results show that the structural equations are unreliable, this unreliability could have been caused by the limitations in the research, the small dataset or the unverified dataset.

The results of the external validation show that when input values are used that are outside the boundaries of the used dataset, the predicted values can become negative. The other points show that the predicted values are way off from the observed values, which can be expected since the internal validation already shows that the structural equations are unreliable.

12 CONCLUSION AND RECOMMENDATIONS

12.1 Conclusion

This research is the first research to develop structural models including the performance of NRW or its components and to analyze these models using the regression analysis. The main question will be answered using the results and the discussion of chapter 10 and 0. The main question has been divided into six questions; the previous chapters answer these questions. The structural models has been developed using the literature study from the chapters 5 -9.

The main question of the research is:

> *What is the nature of the relationships between the performance of NRW or its components, the variables obtained from OPIR and the variables obtained at the beginning of the project and what is the predictive value of the models developed using these relationships?*

The first part of the question will be answered first. The performance of NRW and its components is expressed in liters/connection/day. The variables DPF, NPF and pressure are obtained from OPIR, while the variables that are obtained at the beginning of a project are the sector type, the number of connections and the length of mains (km), the connection density can be estimated using this data. The length of mains is not used as a variable in the model, since it has no relationship to the other variables. The DPF is excluded from the model, since the dataset does not have sufficient data points including the DPF. The nature of the relationships is examined using the performed literature study and by examining the scatter plots. The defined relationships are modeled in structural models and a multiple regression backward elimination analysis is applied on the equations resulting from the models. The following conclusions can be drawn from the performed steps, when the nature of the relationship is not defined in the conclusion a linear relationship is present:

- The model NRW (l/con/day) shows that the NRW (l/con/day) can be predicted using the pressure and the connection density, the relationship between NRW (l/con/day) and the connection density nonlinear and is described with a log function. No structural equation is developed, which can predict the NPF;
- The model real losses (l/con/day) shows that the real losses (l/con/day) can be predicted using the connection density, this relationship is nonlinear and is described with a log function. A structural equation is also made for the NPF, using the real losses (l/con/day), the sector type and the number of connections as independent variables. The relationship between the NPF and the number of connections is nonlinear and is expressed with an exponential function;
- The model of the apparent losses (l/con/day) shows that the apparent losses (l/con/day) can be predicted using the pressure and the connection density, the nonlinear relationship between the apparent losses (l/con/day) and the connection density can be expressed with a log function. No structural equation is developed, which can predict the NPF;

Some remarkable relationships result from the regression analysis:

- From the results of the regression analysis can be concluded that there is no relationship between the real losses (l/con/day) and the pressure, but this conclusion is questioned. It is peculiar that this relationship does not exists since Thornthon, et al. (2008) proves that there is a relationship between the real losses (kl/day) and the pressure, therefore it was assumed that there is a relationship between the real losses (l/con/day) and the pressure;
- The relationship between the apparent losses (l/con/day) and the pressure, resulting from the regression analysis, is not defined in the literature and thus it is unknown whether the relationship is causal;
- The causality of the relationship between NRW (l/con/day) and the pressure is based upon the relationship between the real losses (l/con/day) and the pressure. Since the results show that there is no relationship between the real losses (l/con/day) and the pressure, the relationship between NRW (l/con/day) is based upon the relationship between the apparent losses (l/con/day) and the pressure. This leads to questioning the causality of the relationship between NRW (l/con/day) and the pressure;
- The results show that there is a relationship between the NPF and the real losses (l/con/day). This relationship is expected, because both variables are used to determine the minimum night flow. This proven and now analyzed relationship gives opportunities for future research and to link OPIR with NRW;
- The sector type is used in the regression analysis to replace the type of connections. The results show that the sector type does have an influence on the NPF in combination with the real losses (l/con/day) and the number of connections.

The second part of the question asks whether the models can be used for predictive purposes. In the discussion is already stated that the outcome of the structural equations will be unreliable. This unreliability is proven by the broad 95% confident boundaries, the small proportion of the variance in the dependent variable that can be explained using the independent variables, the even smaller shrinkage levels and the validation plots, where the predicted values mismatch the observed values. The equations resulting from the structural models cannot be used for predictive purposes.

12.2 Recommendations

There are some recommendations for further research:

- There is a lot unknown about the variables that influence or are influenced by the apparent losses (l/con/day). Further research should invest in indentifying these variables, to be able to establish a structural model based on causal relationships.
- The analysis performed in this research shows that there is a relationship between the performance of the apparent losses and the pressure, but this relationship should be examined further to determine whether it is a causal relationship;
- It is recommended to analyze the not existing relationship between the real losses (l/con/day) and the pressure. It can be analyzed whether it happens more often that the pressure does not influence the real losses (l/con/day) and why this occurs;

- The variance in the dependent variables that could be explained by the independent variables is low for each structural equation that is determined. It is recommended that in future research more variables will be defined that influence the dependent variables and thus can be used in the regression analysis;
- The broad boundaries using the 95% confident interval might have been caused by the small data set; the outcome of the analysis might also be imprecise since the data set could not be verified. It is recommended for further research to use a larger data set and data that is verified;
- If a larger dataset is collected, the machine learning techniques can be used to determine the structural equations. The literature proves that the machine learning techniques outperformed the multiple regression analysis modeling a short-term water demand forecast model. However it should be noted that when using the machine learning technique the focus of the research should be to determine a prediction equation and not to examine the relationships between the variables in the model;
- For future research it is recommended to examine the relationship between the real losses (l/con/day) and the NPF over a time-period in an area. Most of the variables (like sector type, but also pressure) do not change a lot in the area, since the values of the variables are kind of stable. The changes in the real losses (l/con/day) value can be examined to determine whether these influence the changes in the NPF value.

LITERATURE REFERENCES

Allen, M. P., 1997. *Understanding regression analysis.* 1st ed. New York: Plenum Press.

Arbués, F., Garcia-Valinas, M. & Martinez-Espineira, R., 2003. Estimation of residential water demand: a state-of-the-art review. *Journal of Socio-Economics,* Volume 32, pp. 81-102.

Babel, M., Das Gupta, A. & Pradhan, P., 2007. A multivariate econometric approach for domestic water demand modeling: An application to Kathmandu, Nepal. *Water resource management,* Volume 21, pp. 573-589.

Bakker, M., 2012. *Introduction to OPIR* [Interview] (11 9 2012).

Blokker, E., 2010. *Stochastic water demand modelling for a better understanding of hydraulics in water distribution networks.* 1st ed. Delft, Netherlands: Water Management Academic Press.

Boomsma, J., 2013. *Non-revenue water in Albanie* [Interview] (11 1 2013).

Bougadis, J., Adamowski, K. & Diduch, R., 2005. Short-term municipal water demand forecasting. *Hydrological processes,* Volume 19, pp. 137-148.

Brekke, L., Larsen, M., Ausburn, M. & Takaichi, L., 2002. Suburban water demand modeling using stepwise regression. *American Water Works Association Journal,* 94(10), pp. 65-75.

Bristol water plc, 2009. *The 2009 Final Water Resource Plan.* [Online] Available at: http://www.bristolwater.co.uk/pdf/environment/wrp2010/APP6/Leakage%20control%20method%202007%20outturn.pdf [Accessed 27 10 2012].

Butler, D. & Memon, F., 2006. *Water Demand Management.* London, UK: IWA.

Charalambous, B., 2012. *The hidden costs of resorting to intermittent supplies.* Ferrara, Water Loss Europe 2012.

Cole, G. & Stewart, R. A., 2012. Smart meter enabled disaggregation of urban peak water demand: precursor to effective urban water planning. *Urban Water Journal,* pp. 1-21.

DAI, 2010. *The manager's non-revenue water handbook for Africa: A guide to understanding water losses,* -: DAI.

Department of Water Affairs, 2010. *National Non-Revenue Water Assesment,* -: -.

Dougherty, C., 2011. Nonlinear models and transformations of variables. In: *Introduction to econometrics.* Oxford: Oxford U.P., pp. 192-223.

ERM, 2007. *Climate change and water resources.* [Online] Available at: http://www.wateraid.org/documents/climate_change_and_water_resources_1.pdf [Accessed 22 1 2012].

Fantozzi, M., Lambert, A. & Liemberger, R., -. *Some examples of european water loss targets, and the law of unintended consequences.* [Online] Available at: http://www.miya-water.com/user_files/Data_and_Research/miyas_experts_articles/15jun2010/Some_Examples_of_European_Water_Loss_Targets_and_the_Law_of_Unintended_Consequences%20With%20Customer%20Water%20Conservation%20Programs.pdf [Accessed 08 02 2013].

Farley, M. & Liemberger, R., 2005. Developing a non-revenue water reduction strategy: planning and implementing the strategy. *Water Supply,* 5(1), pp. 41-50.

Farley, M. & Liemberger, R., -. Developing a non-revenue water reduction strategy part 1: Investigating and assessing water losses. -, -(-), pp. -.

Farley, M. & Trow, S., 2003. *Losses in Water Distribution Networks: A Practitioner's Guide to Assessment, Monitoring and Control.* 1st ed. London: IWA Publishing.

Farley, M. et al., 2008. *The manager's non-revenue water handbook: A guide to understanding water losses,* s.l.: USAID and Ranhill.

Feldman, M., 2009. *Aspects of energy efficiency in water supply systems.* Cape Town, South Africa, 5th IWA Water Loss Reduction Specialist Conference.

Field, A., 2009. *Discovering statistics using spss.* 3rd ed. London: SAGE publications Ltd.

Frauendorfer, R. & Liemberger, R., 2010. *The issues and challenges of reducing non-revenue water,* Mandaluyong City, Philippines: Asian Development Bank.

Freund, R. J., Wilson, W. J. & Sa, P., 2006. *Regression analysis: Statistical modeling of a respone variable.* 2nd ed. San Diego, California: Elsevier Inc..

Garson, G. D., 2012. *Weighted least squares.* - ed. -: Statistical associates publishing.

Gonzalez-Gomez, F., Garcia-Rubio, M. A. & Guardiola, J., 2011. Why is Non-revenue water so high in so many cities?. *International Journal of Water Resources Development,* Volume 27:02, pp. 345-360.

Graybill, F. A. & Hariharan, K. I., 1994. *Regression analysis.* 1st ed. Belmont, California: Duxbury Press.

Hamilton, S., McKenzie, R. & Seago, C., 2006. *A review of performance indicators for real losses from water supply systems.* [Online] Available at: http://www.miya-water.com/user_files/Data_and_Research/miyas_experts_articles/2_NRW/01_A_Review_of_Performance_Indicators_for_Real_Losses_from_Water_Supply_Systems.pdf [Accessed 08 02 2013].

Hesterberg, T. et al., 2010. Bootstrap methods and permutation tests. In: *Introduction to the practice of statistics.* New York: Freeman, pp. 14-2 - 14-70.

IDS Water, -. *Leakage Economics: Plugging the Knowledge Gap.* [Online] Available at: http://www.google.nl/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=0CC8QFjAA&url=http%3A%2F%2Fwww.idswater.com%2FCommon%2Fexhib_64%2FWWT%2520Leakage%2520Economics%2520Article%2520Feb%25202004.doc&ei=Y4i0UMjeIuXS0QXx14GIBg&usg=AFQjCNG3FF634O_jXEzsLq0FfB44G- [Accessed 27 11 2012].

Kingdom, B., Liemberger, R. & Marin, P., 2006. *The challenge of reducing non-revenue water (NRW) in developing countries; How the private sector can help: A look at preformance-based service contracting,* Washington, DC: The world bank.

Kline, R. B., 2011. *Principles and practice of structural equation modeling.* 3rd ed. New York: The Guilford Press.

Koenker, R. & Bassett, G. J., 1982. Robust tests for heteroscedasticity based on regression quantiles. *Econometrica,* 50(1), pp. 46-61.

Lambert, A., 2001. *Water Losses Management and Techniques.* Berlin, International Water Association.

Lambert, A. & McKenzie, R., 2002. *Practical experience in using the infrastructure leakage index.* Cyprus, IWA conference, leakage management - a practical approach.

Liemberger, R. & McKenzie, R., 2005. *Accuracy limitations of the ILI - Is it an appropriate indicator for developing countries.* Halifax, Nova Scotia, Canada, IWA Leakage 2005.

Mc Geown, J., 2012. *Unsustainable water use threatens agriculture, business and populations in China, India, Pakistan, South Africa and USA - global study.* [Online] Available at: http://maplecroft.com/about/news/water_stress_index_2012.html [Accessed 10 09 2012].

McKenzie, R., 1998. *SANFLOW User guide,* -: South African water reserach commission.

McKenzie, R., Buckle, H., Wegelin, W. & Meyer, N., -. *Water Demand Management Cookbook,* Pretoria, South Africa: Water Resource Planning and Conservation.

Mueller, R. O. & Hancock, G. R., 2008. 32 Best practices in structural equation modeling. In: 1st, ed. *Best practices in quantative methods.* California: Sage publications, inc., pp. 488-510.

Mutikanga, H. & Sharma, S., 2012. *Strategic planning for water loss reduction with imprecise data.* Manilla, Philippines, Water Loss 2012.

Niemczynowicz, J., 1999. Urban hydrology and water management - present and future challenges. *Urban Water,* pp. 1-14.

Osborne, J. W. & Waters, E., 2002. Four assumptions of multiple regression that researchers should always test. *Practical assessment, research & evaluation,* Volume 8, pp. 1-5.

Oxford dictionaries, 2013. *Dependent variable.* [Online] Available at: http://oxforddictionaries.com/definition/english/dependent%2Bvariable?q=dependent+variable [Accessed 06 02 2013].

Oxford dictionaries, 2013. *Independent variable.* [Online] Available at: http://oxforddictionaries.com/definition/english/independent%2Bvariable?q=independent+variable [Accessed 06 02 2013].

Page, B., 2005. Paying for water and the geography of commodoties. *Transactions of the Institute of British Geographers 30,* pp. 293-306.

Pearson, D. & Trow, S., 2012. *Comparing leakage performance using the frontier approach.* Manilla, Phillipines, Water Loss 2012.

Pena, I., 2013. *Collected data* [Interview] (10 01 2013).

Puust, R., Kapelan, Z., Savic, D. & Koppel, T., 2010. A review of methods for leakage management in pipe networks. *Urban Water Journal,* 7(1), pp. 25-45.

Renaud, E., 2010. *Towards a global performance indicator for losses from water supply systems.* Sao Paulo, Water Loss 2010.

Riemersma, M., 2013. *Meeting 28-1-2013* [Interview] (28 1 2013).

Roger, D. & Bettin, A., 2012. *EU LIFE PALM Project - Defining the optimum level of leakage.* Ferrara, Water Loss Europe 2012.

Royal HaskoningDHV, -. *Modules.* [Online] Available at: http://www.aquasuite.nl/en/opir/modules/ [Accessed 07 03 2013].

RoyalHaskoning DHV, -. *OPIR.* [Online] Available at: http://www.aquasuite.nl/en/opir/ [Accessed 14 1 2013].

RoyalHaskoningDHV, -. *Aquasuite leaflets.* [Online] Available at: http://www.aquasuite.nl/en/opir/leaflets/ [Accessed 19 1 2013].

Ruiters, G., 2005. *The age of commodity: Water privitization in Southern Africa.* 1 ed. Towerbridge: Cromwell Press ltd.

Seltman, H. J., 2009. *Chapter 9: Simple linear regression.* [Online] Available at: http://www.stat.cmu.edu/~hseltman/309/Book/chapter9.pdf [Accessed 14 02 2013].

Seralgeldin, I., -. *Water.* [Online] Available at: http://www.serageldin.com/Water.htm [Accessed 12 11 2012].

Shepherd, M., 2012. *Meeting Mark Shepherd* [Interview] (14 11 2012).

Shipley, B., 2000. *Cause and correlation in biology: A user's guide to path analysis, structural equations and causal inference.* 1st ed. Cambridge: Cambridge University Press.

StatSoft, -. *Model extremely complex functions, neural networks.* [Online] Available at: http://www.statsoft.com/textbook/neural-networks/ [Accessed 06 02 2013].

StatSoft, -. *What is data mining (predictive analytics, big data).* [Online] Available at: http://www.statsoft.com/textbook/data-mining-techniques/?button=1 [Accessed 06 02 2013].

Tabesh, M., Yekta, A. & Burrows, R., 2009. An integrated model to evaluate losses in water distribution systems. *Water resources management,* 23(3), pp. 477-492.

Thornton, J., Sturm, R. & Kunkel, G. P., 2008. *Water Loss Control.* 2nd ed. USA: McGraw-Hill.

Trifunovic, N., 2006. *Introduction to urban water distribution.* 1 ed. London, UK: Taylor & Francis Group.

Trifunovic, N. & Veenstra, S., 2012. *Bespreking afstudeerplan RHDHV, VEI en IHE* [Interview] (09 10 2012).

Trow, S., 2012. *Economic level of leakage modeling in the UK.* Ferrara, Italy, Water Loss Europe 2012.

Vouk, D., Halkijevic, I. & Vukovic, Z., 2012. *Water loss analysis - Croatian practice.* Ferrara, Water Loss Europe 2012.

Walski, T. & Giustolisi, O., 2012. *An overview of water demand: volume vs. pressure based demands.* Adelaide, Australia, Water Distribution Systems Analysis Conference.

Water Loss Task Force, 2007. *District Metered Areas: Guidance notes,* -: IWA Water Loss Task Force.

Water services planning and information, 2010. *National non-revenue water assessment,* -: Department of water affairs, South Africa.

Wikipedia, 2012. *Non-revenue water.* [Online] Available at: http://en.wikipedia.org/w/index.php?title=Non-revenue_water&action=history [Accessed 29 August 2012].

Wikipedia, 2013. *Durban.* [Online] Available at: http://en.wikipedia.org/wiki/Durban [Accessed 06 02 2013].

Wikipedia, 2013. *Non-revenue water.* [Online] Available at: http://en.wikipedia.org/wiki/Non-revenue_water#cite_note-7 [Accessed 18 1 2013].

Wikipedia, 2013. *Ordinary least squares.* [Online] Available at: http://en.wikipedia.org/wiki/Ordinary_least_squares [Accessed 30 01 2013].

World Water Assessment Programme, 2009. *The United Nations World Water Development Report 3: Water in a Changing World.,* Paris: UNESCO, and London: Earthscan: Unesco publishing.

Yasar, A., Bilgili, M. & Simsek, E., 2012. Water demand forecasting based on stepwise multiple regression analysis. *Arab J Sci Eng,* Volume 37, pp. 2333-2341.

Yeboah, P., 2008. *Management of non-revenue water: A case study of the water supply in Accra, Ghana,* Loughborough: Loughborough University.

Zhang, X., 2005. *Estimating peaking factors with poisson rectangular pulse model and extreme value theory,* Cincinnati, United States: University of Cincinnati.

Zhou, S., McMahon, T., Walton, A. & Lewis, J., 2000. Forecasting daily urban water demand: a case study of Melbourne. *Journal of Hydrology,* Volume 236, pp. 153-164.

Zhou, S., McMahon, T., Walton, A. & Lewis, J., 2002. Forecasting operational demand for an urban water supply zone. *Journal of Hydrology,* Volume 259, pp. 189-202.

APPENDIXES

Appendix I: Strategy to reduce non-revenue water

There is a gap between the future need for water in and the current availability of water, this gap can be solved with the help of three solutions, which are shown in Figure 24.

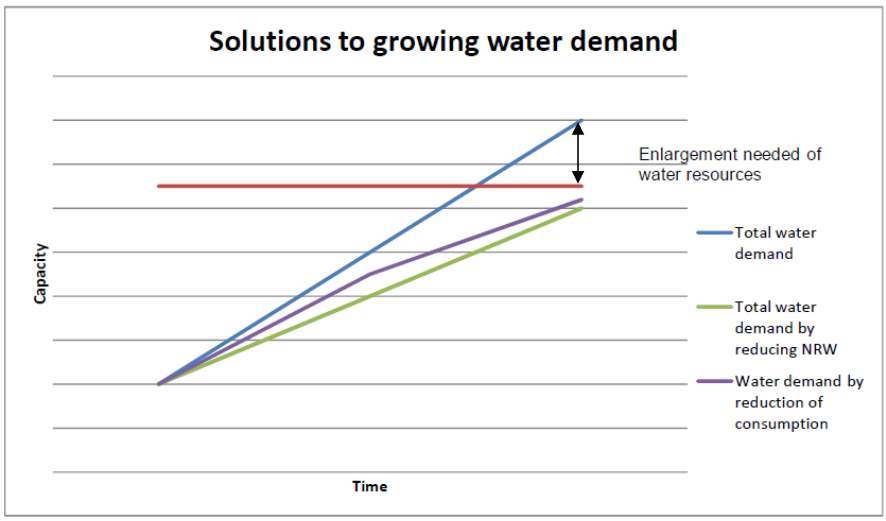


Figure 24: Solution to growing water demand

The first solution is to enlarge the water resources capacity and the transportation by adding reservoirs or pumping capacities or by increasing the treatment capacity and bringing water from another area into the area. The second solution is to reduce the need for water by reducing NRW. The last solution is to reduce the need for water by reducing the water demand of the consumers. The cost of the solutions needs to be evaluated for the specific area, to determine which solution is best suited. Faced with the construction of new treatment plants and even with desalination of water, NRW reduction might be chosen based on the evaluation of the cost (Farley & Trow, 2003).

When the decision is made to reduce NRW a project can be started, but reducing NRW is not a simple process. This chapter will discuss how to reduce NRW; the first sub-chapter will describe how the decision is made to invest in reducing NRW. Sub-chapter I.II discusses the importance of setting a leakage target. How to reduce apparent losses is described in sub-chapter I.III and sub-chapter I.IV describes how to reduce the real losses.

I.I. Economic Level of Leakage (ELL)

Economic Level of Leakage (ELL) is the point where the value of the water saved is less than the cost of making further reductions. The ELL needs to be determined at the beginning of a project

and should be monitored throughout the process of reducing NRW. The value of water changes over time as will the operating costs and the methods used to reduce NRW, thus the ELL can change over time. ELL has to be estimated from data the utility has available. At the start of a new NRW reduction project data might not be available, while during the process new data will be released and data will become more accurate (Farley & Trow, 2003).

There are two types of ELL, the short-term ELL and the long-term ELL. The short-term ELL is the steady state in which the marginal costs of active leakage control (ALC) effort is equal to the marginal cost of water saved by adopting ALC. In short-term ELL there are some key parameters fixed like the average pressure in the system, the condition of the mains and service pipes and the facilities available for collecting data. The level of leakage can be influenced by the number of personnel searching for leaks and repairing the leaks (Farley & Trow, 2003).

Long-term ELL is based on investment analysis. The investment in the facilities will have an impact on the short-term ELL. Long-term ELL takes in account the following questions:

- What is the current level of leakage?
- What is the short-term ELL?
- How will the short-term ELL change with the investment under consideration?
- What is the saving in water losses and the change in ALC resources from the proposed investment compared with the current policy?
- What is the cost of the proposed investment?
- What is the return on the investment?

The outcome of these questions will lead to an investment policy (Farley & Trow, 2003).

I.I.I. Determining ELL

The process to determine the ELL and set a target is shown in Figure 25.

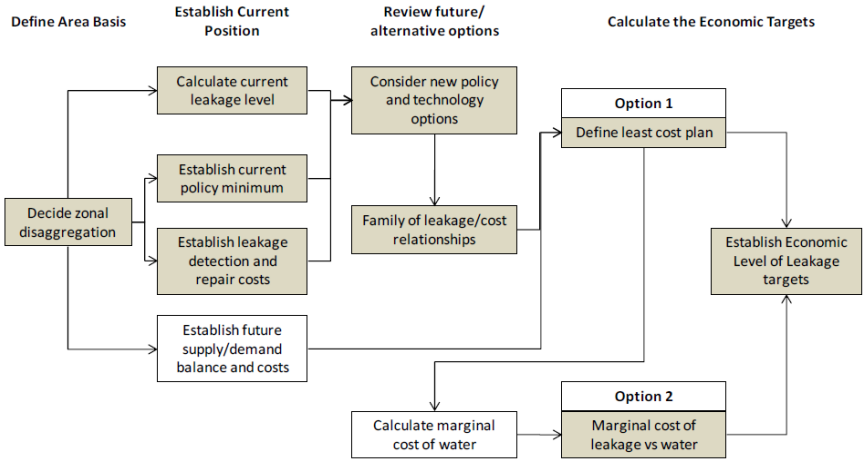


Figure 25: ELL target setting process map (Trow, 2012)

To determine the ELL two components must be determined (1) the cost of the water loss and (2) the costs of NRW management. Data is needed from the water balance as well as from other resources to determine the components. The cost for real losses is determined by multiplying the volume of real losses with the variable operational costs. The cost for apparent losses is found by multiplying the apparent losses with the average consumer tariffs. The total costs can be determined by adding the two cost components. Figure 25 shows that the minimum total cost is the economic level of NRW (Farley, et al., 2008).

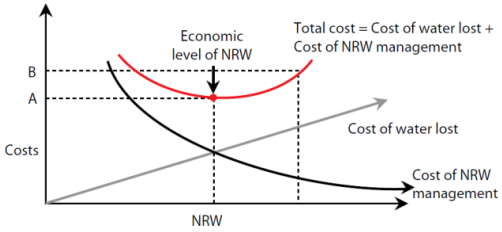


Figure 26: Identifying the economic level of non-revenue water (Farley, et al., 2008)

I.II. Setting leakage target

The ELL is the input for setting a leakage target. The target is set for a longer period of time, because large implementations can take time before the advantage of the implementation is noticed. The reduction of NRW can differ over time, these differences should be expected. The leakage target is area-specific and the target and the process to achieve the target is difficult to compare with other targets and processes set in other areas (Farley & Trow, 2003).

I.III. Reducing apparent losses

According to Farley, et al. (2008) water utilities should aim for a level of apparent losses around the 4 – 6 %. Utilities should focus on the apparent losses at the beginning of the NRW reduction program, since the activities take little effort and it immediately pays-back. The money saved can be used to implement other programs that reduce NRW (Farley, et al., 2008). Figure 27 shows that there are four pillars determining the apparent losses. The methods used to reduce the four pillars are discussed in the coming sub-chapters.

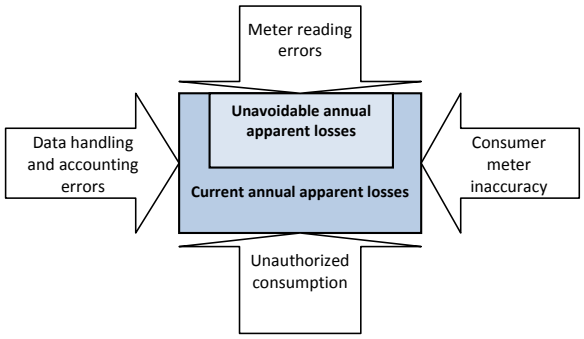


Figure 27: Four pillars of apparent losses (Farley, et al., 2008)

I.III.I. Unauthorized consumption

The unauthorized consumption will show up when a water balance is made. The utility needs to perform consumer surveys and leakage step test to determine the location of the flow that goes missing. To tackle the unauthorized consumption there are some more actions that can be taken. An action that can be taken is to implement a consumer awareness program. The consumer awareness program makes it possible for consumers to report illegal connections and illegal use of fire hydrants. Another factor that influences the volume of the unauthorized consumption is the meter readers, the meter redears can be a big help finding illegal connections, but can be corrupt as well. This can be reduced by rotating the routes of the meter readers on regular basis. Another action that can be taken is to introduce a policy with penalties for water theft (Farley, et al., 2008). Though at the start of a project there should be a policy to provide amnesties to people that are using illegal connections, this way the people can regularize their connections (Yeboah, 2008).

To ensure people are paying for the water that is use, prepaid meters can be installed. Prepaid meters are becoming quite popular by utilities dealing with large number of poor people. The advantages of prepaid meters are that there are no meter readings, no billing statements and errors that could occur and the moment of disconnection from the water is chosen by the consumer. The consumers experience that the meter is less flexible, the consumer needs to (re)charged the meter with hard cash or even with voucher bought in a store, which makes them dependent on the stores business hours (Ruiters, 2005).

I.III.II. Inaccurate meters

Meters that are inaccurate tend to under-register and sometimes over-register water consumption. Under-registration at meters placed by large consumers will cause a lot of water and money. Special attention should be given to these large consumers (Farley, et al., 2008).

Consumer meter inaccuracy has several causes which have their own solutions. The first point of attention is the proper installation of the meters. Meters do not register the flow, when the flow rate is lower than a specific minimum. The proper size of the meter is based on the water demand pattern. Before ordering meters, the water demand pattern of each type of consumer should be evaluated. Not only the size is important, but also the class of the meter (Farley, et al., 2008). There are different kinds of meter classes for different kind of meter situations, the class C and D are international standards (Farley & Liemberger, 2005). To be able to install a meter of high quality, utilities should purchase the meters on consumer's behalf. The meters can become inaccurate by a sediment flow through the meter. Utilities should monitor the water quality and clean the mechanical meters to minimize the sediment levels. When there is an intermittent supply, the meter can be damaged by the pressure changes in the mains. This can be avoided by transforming the intermittent supply to continuous supply. The meter condition should be monitored; this is also an aspect of the asset management. As well as having a plan to repair and replace the meters (Farley, et al., 2008).

I.III.III. Meter-reading errors

Meter readings are needed to be able to create a bill for the water that is used. Utilities should invest in training and motivating people to keep the staff recording and reporting the information effectively and efficiently. The utility can also implement auditors to control whether the numbers written down are correct (Farley, et al., 2008).

I.III.IV. Data handling and accounting errors

There are several errors that could occur in the process from reading the meter to delivering the bill. In many developed countries this is done digital, but in developing countries it is still done by hand. To preserve the utilities staff from making mistakes a robust billing database should be purchased. This database has built-in analysis functions that can identify potential data handling errors and reports them for verification (Farley, et al., 2008).

I.IV. Reducing real losses

There are four methods that could be used to reduce the real losses, the interaction between the methods and the real losses are shown in Figure 28. Each of these four methods has an economic level of investment. The economic level of investments provides the order and the extent in which the methods will be used. The extent in which the methods will be used, will determine the time that is needed to achieve the real loss reduction (Farley & Trow, 2003).

The sub-chapters will describe each of the methods, which are shown in Figure 29.

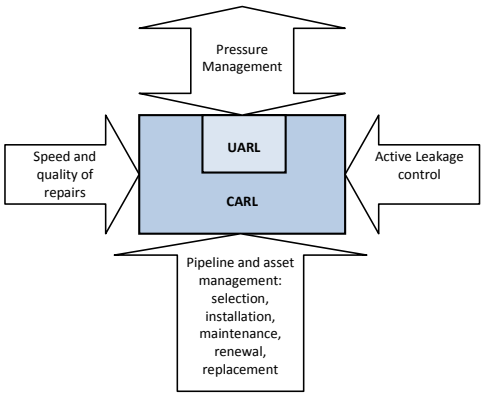


Figure 28: The four basic methods of managing real losses (Farley & Trow, 2003)

I.IV.I.Active leakage control

A utility has a passive control strategy, when it only responds on reported breaks. Effective ALC is applied when a utility has staff members which are hired to find leakages that are not reported by consumers or other means (Farley & Trow, 2003).

The intensity of the ALC is determined by the natural rate of rise (NRR). The NRR is the rate at which leakage increases within a system in the absence of any break repairs. The NRR exists of

two components, the new breaks in the system and the growth of existing leakage (increasing in volume). Some of the breaks will be visible and will be repaired, but a portion of the breaks will only be discovered using ALC (IDS Water, -).

Finding unreported breaks is performed by executing regular surveys and leakage monitoring. The regular survey is executed by starting at one end of the distribution system and proceeding to the other end. Survey techniques that are used are (1) listening for breaks on pipe work and fittings, (2) reading metered flows into temporarily zoned areas to identify high-volume night flows, (3) using clusters of noise loggers or (4) performing a step test (Yeboah, 2008).

Leakage monitoring is one of the most used and most cost-effective activities to reduce real losses. It monitors the flow that goes in to the areas, this makes it possible to quantify leakages and prioritize leak detection activities. In most developing countries leakage control is passive, these countries normally only mend visible breaks and conduct regular surveys of the system with acoustic or electronic apparatus (Farley & Liemberger, 2005).

I.IV.II. Pressure management

IWA has identified pressure management as:

"The practice of managing system pressures to an optimum level of service ensuring sufficient and efficient supply to legitimate uses and consumers, while eliminating or reducing pressure transients and variations, faulty level controls and reducing unnecessary or excess pressures, all of which cause the distribution system to leak and break unnecessarily" (Yeboah, 2008).

Farley and Trow (2003) have stated that pressure is the second most important factor in determining leakage levels, after infrastructure conditions. Pressure management is more cost effective than infrastructure management and it provides the fastest pay back.

Pressure management has several advantages. (1) Pressure influences the flow rate of leaks. When the pressure is lowered, the flow rate will decrease, which leads to a drop in water leaking out of the system. When the flow rate drops, the speed and quality of repairs influences the total leakage rate less (Yeboah, 2008). This makes pressure management an integral part of the strategy (Butler & Memon, 2006). (2) When breaks are mended the pressure will stay the same, if pressure management is implemented. Without pressure management the pressure will rise, when a break is mended. Figure 29 shows that with pressure management the water loss does significantly change. (3) Pressure management has an effect on the consumption level. Unwanted demand is present in almost every household; an example is taps that are leaking. Pressure management can be used to prevent or lower this unwanted demand. Demand management is a more cost-effective and environmentally sustainable method to meet the increased demand than system expansion. (4) Pressure management eliminates hydraulic impacts, because the pressure is kept at one lowest possible level, this will cause less damage to the system. Laboratory tests and tests on underground systems have proven that there is a relationship between the frequency of breaks and the pressure. When pressure management is

implemented the leakage flow rate as well as the breaks frequency will decrease (Yeboah, 2008).

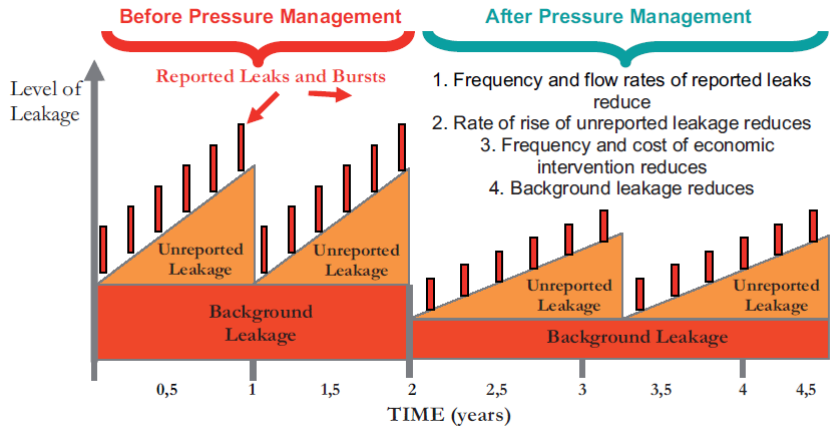


Figure 29: Components of real losses before pressure management (Yeboah, 2008)

Pressure management also has its disadvantages. Concerns rise about the relationship between pressure management and the fire flow. To fight fire with water a certain pressure is needed and lowering the pressure in the system can give problems. There should be no problems, when a good pressure management system is in place. Another concern is the filling of the reservoir. This concern can be solved by implementing pressure management on the smaller pipes and connections and allowing the normal pressures in the transmission mains (Yeboah, 2008).

The following tasks should be performed to implement pressure management on a particular system:

- Identify potential zones, installation points and consumer issues;
- Identify consumer types and control limitation through demand analysis;
- Field measurement of flow and pressure;
- Modeling of potential benefit using specialized models;
- Identify the correct control valves and control devices;
- Model the correct control regimes to provide a desired result;
- Perform a cost-benefit analysis (Farley & Liemberger, 2005).

Important to understand is that pressure management will only work if the system is maintained, thus a maintenance plan is needed when pressure management is implemented. If the system is not maintained the pressure will rise and fall outside the set limits (Farley & Trow, 2003).

I.IV.III. Speed and quality of repairs

Water losses do not only depend on the flow rate (thus the pressure), but also on the run time. The run time exists of three different components; Figure 30 shows the division of the three components of the run time. The first component of the run time is the awareness time; this is the time until the utility becomes aware that there is a leak. The following component is the location time, this is the time needed to precisely locate the leak. The third component is the repair time; this is the time between issuing for a repair job and the completion of the repair.
The runtime can be reduced, but a certain threshold of reduction in repair time will be reached. The costs of the reduction in repair time can become higher than the benefits from reducing the repair time (Frauendorfer & Liemberger, 2010).

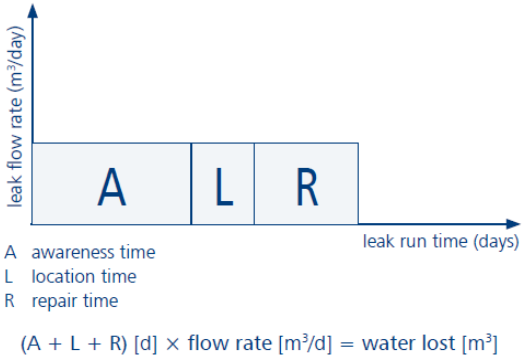


Figure 30: The impact of leak run time (Frauendorfer & Liemberger, 2010)

The quality of the repair is as important as the repair time for controlling the losses. The break can appear again, when the break is not properly fixed. A company in the Netherlands should meet the ISO 9001 quality standard to be permitted to mend breaks. ISO 9001 is a family standard which is related to quality management of systems and managements (Wikipedia, 2012). Contractors in other countries do not always have to meet any quality standards. The company that hires the contractor needs to assess his quality and should use contracts that oblige the contractor to deliver quality work. In Brazil leak detection and repair contracts are introduced. The contractors need to give a 24 month guarantee for the breaks that are detected and mended (Shepherd, 2012).

I.IV.IV. Asset management

Asset management is the most important part in reducing NRW. Asset management is a long term strategy in approaching the leakage management strategy and asset management does not immediately lead to payback (Farley & Trow, 2003).

The main drivers to replace a main on the ground of leakage reduction should be the consumer level of service and the operating costs (Farley & Trow, 2003). The following reasons justify the renewal or rehabilitation of the infrastructure:

- The internal condition of the main is affecting the water quality of the water delivered;
- The internal bore of the main has reduced due to corrosion or has become build of deposits;
- The pipe wall has been weakened and can no longer withstand the internal pressures of the water;
- The main cannot fulfill its current duty resulting from some external factor (Farley & Trow, 2003).

The replacement of the mains can have an adverse effect on the connection. The pressure will increase in the system due to the increased carrying capacity of the main, which can cause new leakages on connections (Farley & Trow, 2003).
The replacement of mains can be postponed till the system gets extended or rehabilitated for other purposes. This decision is mainly taken by the cost-effectiveness of the operation (Farley, et al., 2008). The following steps should be followed, carrying out a cost-effective main rehabilitation:

1. Identification of those mains that clearly needs replacement;
2. Identification of areas of high leakage record;
3. Cost-benefit analysis;
4. Consideration of other benefits;
5. Designing of the scheme;
6. Project management plan (Farley & Trow, 2003).

To decide whether the mains should be replaced or rehabilitated the following factors are in order:

- Compare the relative costs of continuing to make repairs with the cost of replacements;
- The impact on consumer service of continuing to accept the interruptions to supply whilst repair works is carried out;
- The requirement to meet the set target to reduce leakage, which will be assisted by the installation of a new main;
- The availability of finance and other sources (Farley, et al., 2008).

Appendix II: Performing a component analysis

Table 20 shows the basis of the component analysis. For every category the components are analyzed to determine the losses in the category. With the outcomes of the three categories the total real loss can be estimated.

Table 20: Parameters required for calculation of components of annual real losses (Thornton, et al., 2008)

Component according to water balance	Component of the system	Background (undetectable) losses	Reported breaks	Unreported breaks
Leakage on transmission and/or distribution mains	Mains	Length Pressure Min loss rate/km*	Number/year Pressure Average flow rate* Average duration	Number/year Pressure Average flow rate* Average duration
Leakage and overflows at utilities storage tanks	Service reservoirs	Leakage through structure	Reported overflows: Flow rates Duration	Unreported overflows: Flow rates Duration
Leakage on connections up to the consumers' meters	Connections main to edge of street	Number Pressure Min loss rate/conn.*	Number/year Pressure Average flow rate* Average duration	Number/year Pressure Average flow rate* Average duration
	Connections after edge of street	Length Pressure Min loss rate/km*	Number/year Pressure Average flow rate* Average duration	Number/year Pressure Average flow rate* Average duration

* At some standard pressure.

Most of the data is available in organized water utilities, though the average pressure is often difficult to estimate and thus is the data which depends on the average pressure. Because pressure is an important parameter in the component analysis, a good estimation of the average pressure is necessary (Farley & Liemberger, -).

The first category is the background losses. Background losses are difficult to define, because these leakages run slow and do not attract attention. Background losses are often unknown. The unavoidable background leakage rates per pressure can be determined using the values of the different components from Table 21 for an infrastructure correction factor (ICF) of 1 (Farley & Liemberger, -). The ICF is the ratio of an area's background losses to the nominal average values in England and Wales. Often the ICF factor is unknown, resulting from the unknown background losses. If the ICF is unknown a value of 1 should be used, this represents the average value found in England and Wales. A too high ICF leads underestimation of the true loss reduction potential (Bristol water plc, 2009).

Table 21: Unavoidable background leakage rates (Farley & Liemberger, -)

Infrastructure component	Background leakage at ICF = 1.0	Units
Mains	9.6	Liters per km of mains per day per meter of pressure
Connection – main to property boundary	0.6	Liters per km of mains per day per meter of pressure
Connection – property boundary to consumer meter	16.0	Liters per km of mains per day per meter of pressure

The reported and unreported breaks are determined in the same way. The reported and unreported breaks can be determined with equation II-1.

$$Number\ of\ breaks * average\ leak\ flow\ rate * average\ leak\ duration \quad \text{II-1}$$

The number of breaks is easily determined for the reported breaks. To determine the number of unreported breaks, the definition of unreported breaks must be known. Unreported breaks are the breaks that are not reported during the year, but found with other techniques (Thornton, et al., 2008). These techniques are described in appendix I.

The average leak flow rate should be based on the average pressure, when the flow rate is unknown the figures from Table 22 can be used (Farley & Liemberger, -).

Table 22: Flow rates for reported and unreported breaks (Farley & Liemberger, -)

Location of break	Flow rate for reported breaks (l/hour/m pressure)	Flow rate for unreported breaks (l/hour/m pressure)
Mains	240	120
Connection	32	32

The last factor of equation II-1 is the leak duration which can be split up in three elements:

- The awareness time;
- The location duration;
- The repair duration.

The awareness time for reported breaks is often short; when the breaks are unreported it can take a lot of time before the breaks are detected. The location time should be short, since the leak has been detected and only needs to be found. The repair duration for mains will be within 24 hours, but small leaks on connections can take up to 7 days.

Appendix III: Minimum night flow analysis

The real losses during the MNF period are determined using equation III-1

$$Real\ losses = Minimum\ night\ flow - consumption\ rate \quad \text{III-1}$$

When the value of the consumption rate is unknown, equation III-2 can be used to estimate the consumption rate.

$$Consumption\ rate = Total\ active\ population\ during\ the\ night * consumption\ during\ the\ night \quad \text{III-2}$$

The total active population is around 6% of the total population in the area. The consumption during the night is based on the standard toilet cistern of that area; normally it is around 10 l/head/h (McKenzie, 1998).

The real losses cannot be extrapolated since this would lead to overestimation of the real losses because the pressure during the night is higher than during the day. To calculate the real losses for a certain moment during the day equation III-3 developed by the Water Loss Task Force (2007) can be used:

$$\frac{L_1}{L_2} = \left(\frac{P_1}{P_2}\right)^{N_1} \quad \text{III-3}$$

In which L_1 is the leakage level at assessed time and L_2 is the maximum real loss value. P_1 is the pressure value at assessed time; P_2 is the pressure value during the maximum real loss value. The last factor that is used is N_1 which is an exponential value that indicates the sensitivity of the leak flow rates to the change of pressure. The average N_1 is in the order of 1.15. When N_1 cannot be determined the conservative value of 1.0 should be adopted (McKenzie, et al., -).

Equation III-3 only estimates the real loss value for a certain moment during the day. To determine the real loss value for the whole day, the night-day factor (NDF) should be estimated, this factor can vary between 18 to 24. Equation III-4 shows how to estimate the NDF; this equation is diminished from equation III-3 (Water Loss Task Force, 2007).

$$NDF = \left(P_1/P_0\right)^{N_1} + \left(P_2/P_0\right)^{N_1} + \left(P_3/P_0\right)^{N_1} + \ldots + \left(P_{24}/P_0\right)^{N_1} \quad \text{III-4}$$

When the relationship between the average leakage and the leakage rate is linear (N_1 = 1.0), than equation III-5 can be used to determine the NDF (Water Loss Task Force, 2007).

$$NDF = 24 * (\frac{Average\ daily\ pressure}{P_{min}}) \quad \text{III-5}$$

The total real loss value can be estimated with the help of equation 3-2.

$$Real\ losses = NDF * L_2 \quad \text{III-6}$$

Appendix IV: Method

This appendix will elaborate more on the methods that are used, as well as the assumptions that must be met.

IV.I Regression analysis

The regression analysis that will be chosen must meet the following specification:

- Continuous dependent variables;
- Continuous and nominal independent variables;
- More than one independent variable influencing a dependent variable;
- More than one dependent variable;
- A dependent variables influences another independent variable.

Allen (1997) states that whenever a dependent variable in one equation appears as an independent variable in another equation structural equation models (SEMs) should be used. The foundation of SEM is rooted in classical measured variable path analysis and confirmatory factor analysis. The models that are used in SEM can be either recursive or non-recursive models. A recursive model does not contain feedback loops and the relationships flow in a single direction (Allen, 1997). A non-recursive model has feedback loops and has correlated errors. The relationships described in sub-chapter 7.8 show that there are no feedback loops between variables. Since there are no feedback loops present and the relationships flow in a single direction, a recursive model is present in the research. When a recursive model is present a measured variable path analysis (MVPA) should be used to analyze the relationships. A MVPA is used when there are hypothesized structural/causal relations among directly measured variables (Mueller & Hancock, 2008).

Path analysis is superior to regression analysis, since it examines the causal processes underlying the observed relationship and estimates the importance of the alternative paths of influence. Path analysis can also be used to measure the direct and indirect effects that one variable has on another variable. Path analysis still makes use of the structural equation shown in equation; the only difference with regression analysis is that more equations are involved. Path analysis follows the following steps:

1. Specify the hypothesized causal structure of the relationships between the variables;
2. Translate this causal model in an observational model by writing down the linear equations and specifying which parameters must be estimated from the data and which are fixed;
3. Derive the predicted variance and covariance between each pair of variables in the model using covariance algebra;
4. Estimate the free parameters (Shipley, 2000).

Path analysis uses either the ordinary least-square (OLS) estimation or the maximum likelihood (ML) estimation to determine the estimates (β_n) in the structural equations. The method of OLS estimates the values of the estimators so that the least squares criterion is satisfied. This means that the sum of the squared residuals is as small as possible for the particular sample. OLS is a partial-information method or a limited-information method since it can only analyze the

equation for one criterion at a time (Wikipedia, 2013). The ML estimation is a full information method, since it can calculate all estimates (β_n) in all structural equations all at once. The outcome of the OLS and ML estimates (β_n) are often the same when used in multiple regression analysis (Kline, 2011). Since the model is a recursive model the unbiased estimates (β_n) of the structural equation model can be obtained using the OLS estimation procedures (Allen, 1997).

The OLS estimation is used in linear regression models, thus the multiple linear regression model can be used to determine the estimates (β_n) for the structural equations. It is important to use the best combination of the variables to predict the dependent variable. Stepwise regression is used to find the best combination of independent variables (Yasar, et al., 2012). A backward elimination will be used, since it is least likely to miss an independent variable that does predict the outcome. A backward elimination starts with placing all the independent variables in the model, when an independent variable is not making a statistically significant contribution (p>0.1) to how the model predicts the outcome variable it will be removed and the model will be re-estimated for the remaining independent variables. The independent variable with the highest p-value will be eliminated first etc. The final model must have a significance level of p<0.05, otherwise the model will be rejected (Field, 2009).

Most of the variables are continuous variables, but there is one variable (the sector type) which is a nominal variable. To be able to use this variable in the multiple regression analysis, dummy variables should be created. Dummy coding is performed by creating as many new variables as the number of groups minus one. One of the groups will be used as a baseline group against which the other groups are compared. The baseline group will always be coded zero. The value of 1 is assigned to the first group that is compared to the baseline group, this goes on till all groups are coded (Field, 2009).

IV.II Assumptions for using a regression model

There are several assumptions that must be met before an OLS multiple regression analysis can be used. These assumptions are (1) outliers must be examined, (2) there should be linearity in the dataset, (3) homoscedasticity must be present in the data, (4) as well as normality, (5) no independent errors should be present, (6) as well as no multicollinearity. When one of the assumptions is not met a fitting solution must be found. The assumptions are explained in the following paragraphs as well as how to test these assumptions.

First the dataset must be tested to determine whether there are outliers present in the dataset. An outlier is an observation that is different from the other observations in the dataset. These observations should be identified and examined. To identify the outliers the residual plot should be examined. The z-scores of the observations in the dataset should satisfy the following rules:

- 95% of the z-scores of should lie between -1.96 and 1.96;
- 99% of the z-scores of should lie between -2.58 and 2.58;
- 99.99% of the z-scores of should lie between -3.29 and 3.29.

When the first two rules are not satisfied, the model is a poor fit of the sample data. When a case has a standardized value greater than 3.29 the case should be examined, to determine whether the observations are influential observations. To determine the influence of the outlier

on the regression estimates, the Cook's distance and the average leverage needs to be examined. The outlier is an influential observation when the Cook's distance is greater than 1 or the average leverage is greater than (2(k+1)/n) (in which k is the number of independent variables and n is the number of participants). If the outlier is an influential observation, the observation must be examined to determine why the observation is divergent (Field, 2009). Outliers that influence the regression estimates to be substantially different from what it would be without the observations are influential observations (Freund, et al., 2006).

The second assumption is the assumption of linearity. Since the multiple linear regression method will be used, the relationships that are modeled must be linear ones. When a nonlinear relationship is present the variable can be transformed (Field, 2009), an example of an equation using nonlinear relations is shown equation.

$$Y = \beta_0 + \beta_1 X_1^2 + \beta_2 \sqrt{X_2} + \beta_3 \log X_3 \quad \text{9-1}$$

The nonlinear variables in the equation can be replaced for linear variables, so replacing X_1^2 for Z_1 etc. This replacements leads to a structural equation as is shown in equation

$$Y = \beta_0 + \beta_1 Z_a + \beta_2 Z_2 + \beta_3 Z_3 \quad \text{9-2}$$

The linearity assumption will still be met, but non-linearity is included in the model (Dougherty, 2011). The residual plots show whether the assumption of linearity is met. Figure 31 shows a plot where the assumption of linearity is met and a plot where the assumption is not met.

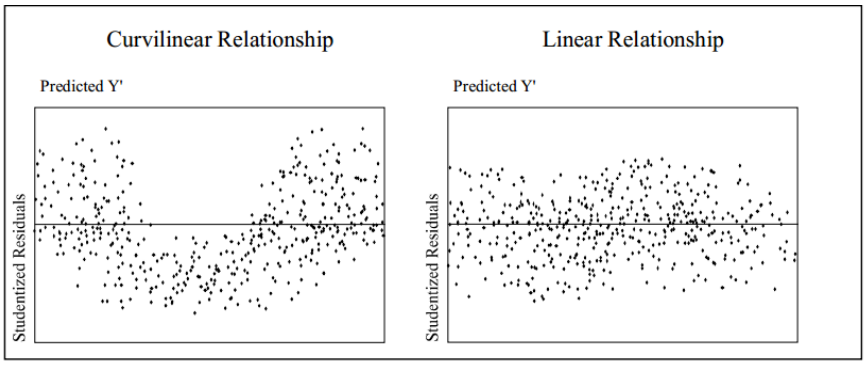


Figure 31: Example of curvilinear and linear relationships with standardized residuals by standardized predicted values (Osborne & Waters, 2002)

Normality is the third assumption that is examined, when the normality assumption is violated incorrect significance levels and confidence intervals might be estimated. It is assumed that the residuals are random and normally distributed variables with a mean of zero. The normality assumption must only be met by the dependent variable; the independent variables do not need to be normally distributed. The normal probability plot can be examined to determine whether the data meets the assumption of normality. Figure 32 shows a normal probability plot

where normality is assumed and one that violates the assumption of normality. Whenever it suspected that the normality assumption is not met, the Shapiro-Wilk test should be performed. The significance level (p<.05) indicates a deviation from normality (Field, 2009).
This problem can be avoided using the bootstrap technique; this is a robust technique which can be used when the normality assumption is violated. The bootstrap technique estimates the properties of the sampling distribution from the data set (Field, 2009). The data set is treated as a population from which small samples are taken. The data is put back, before a new sample is taken (sampling with replacement) (Hesterberg, et al., 2010). The mean is calculated for every sample and the sampling distribution can be estimated taking many of these samples. The confidence intervals and the significance tests can be computed using the standard error estimated from the standard deviation of the created sampling distribution (Field, 2009).

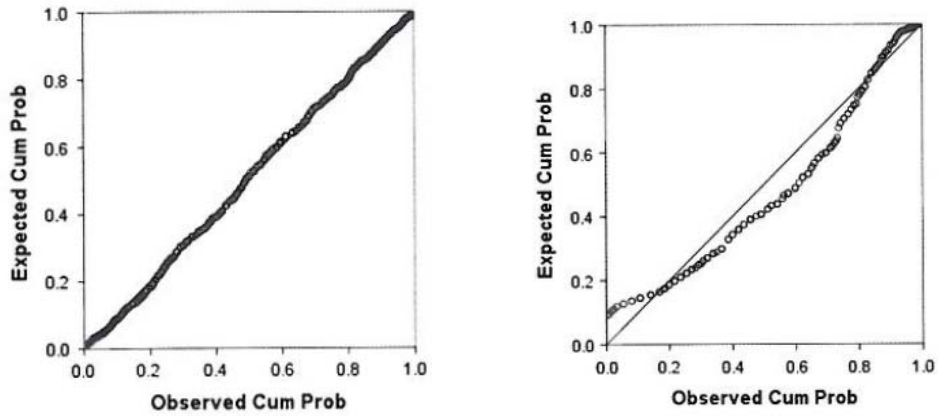


Figure 32: Normality assumption (Field, 2009)

The fourth assumption that must be met is homoscedasticity. *"Homoscedasticity means that the variance of errors is the same across all levels of the independent variable. When the variance of errors differs at different values of the independent variable, heteroscedasticity is indicated"* (Osborne & Waters, 2002). The findings are distorted and the analysis weakens when heteroscedasticity is present. This could lead to wrongly estimated standard errors of the coefficients (β_n), leading to a wrongly determined significant level. To determine whether homoscedasticity or heteroscedasticity is present the residual plot should be examined, which are also examined for the assumption of linearity. The residual plot shows homoscedasticity when the residuals (the points) are scattered randomly around the zero. Figure 33 shows examples of residual plots where homoscedasticity and heteroscedasticity is present. When the plots show that heteroscedasticity might be present, formal tests should be performed (Field, 2009). The test that will be performed is the Koenker test. Heteroscedasticity is present when the test has been found significant (p<.05). The Koenker test is still valid, whenever there is the dataset does not meet the assumption of normality (Koenker & Bassett, 1982).
To correct the presence of heteroscedasticity the weighted least squares regression will be used. *"Weighted least squares (WLS) regression compensates for violation of the homoscedasticity assumption by weighting cases differentially: cases whose value on the*

dependent variable corresponds to large variances on the independent variable(s) count less and those with small variances count more in estimating the regression coefficients. That is, cases with greater weights contribute more to the fit of the regression line. The result is that the estimated coefficients are usually very close to what they would be in OLS regression, but under WLS regression their standard errors are smaller" (Garson, 2012). The weighted least squares regression can be estimated by first calculating the weight estimation and then running the OLS regression using the weight estimation. The weight estimation is determined by searching for the power of the independent variable which maximizes the log-likelihood of the dependent variable, the reciprocal of the power of the independent variable is used to weight the cases. *"This procedure assumes that violation of homoscedasticity is characterized by the variance of the dependent increasing exponentially at some power function of the independent"* (Garson, 2012).

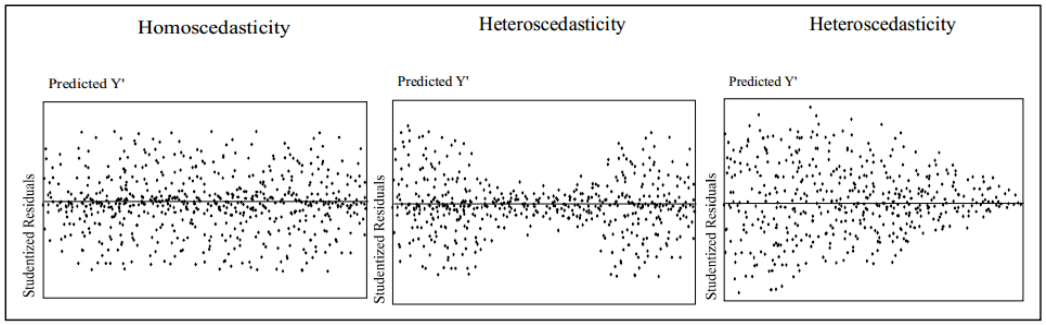


Figure 33: Examples of homoscedasticity and heteroscedasticity (Osborne & Waters, 2002)

The fifth assumption states that no independent errors should be present in the dataset. An independent error exists when the residual terms of two data points are correlated (Field, 2009). *"This assumption comes down to the idea that the error (deviation of the true outcome value from the population mean of the outcome from a given x value) for one observational unit (usually a subject) is not predictable from knowledge of the error for another observational unit"*(Seltman, 2009). There are two main causes for the existence of independent errors in the data. The first and most common cause is the one in which errors of adjacent observations are similar, also called serial correlation. The other less frequent cause occurs when observations are 'neighbors'. Seltman (2009) describes an example of this cause as follows: "*This assumption can be trivially violated if we happen to have a set of identical twins in the study, in which case it seems likely that if one twin has an outcome that is below the mean for their assigned dose, then the other twin will also have an outcome that is below the mean for their assigned dose (whether the doses are the same or different).*" The first cause can be solved by using time series models and the second cause can be solved by using mixed models (Seltman, 2009).

Multicollinearity is the sixth assumption which should be met. *"If there is perfect collinearity between predictors it becomes impossible to obtain unique estimates of the regression coefficients because there are an infinite number of combinations of coefficients that would*

work equally well." (Field, 2009). Multicollinearity present in the dataset produces untrustworthy estimates, limits the size of the measure of the multiple correlation between the independent variables and the outcome and the importance of the independent variables cannot be assessed. To determine whether there is multicollinearity present in the dataset the variance inflation factor (VIF) and its tolerance should be examined. The tolerance is the unique variance of the independent variable associated with the dependent variable. When the tolerance value is lower than 0.2 or the VIF is higher than four, multicollinearity could be present in the dataset (Field, 2009).

Appendix V: Scatter plots for determination of relationships

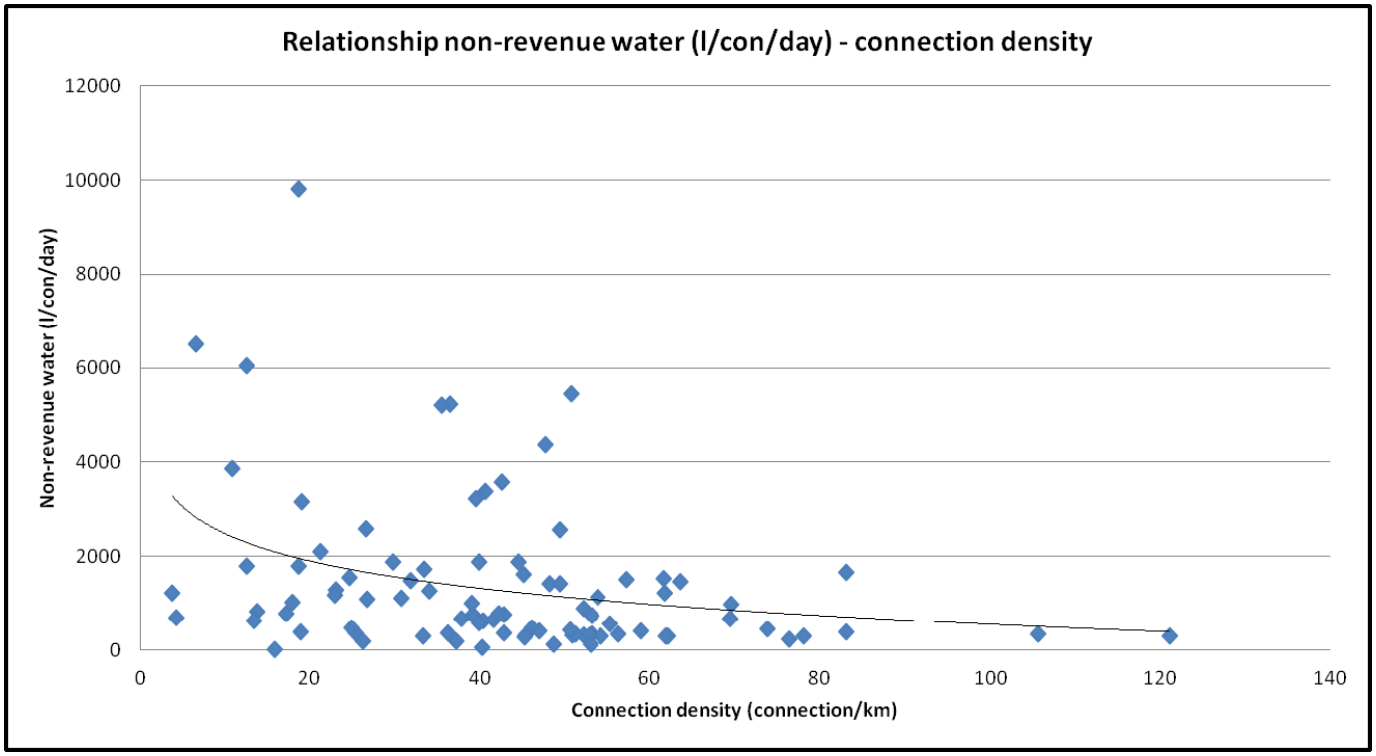


Figure 34: Relationships non-revenue water (l/con/day) - connection density

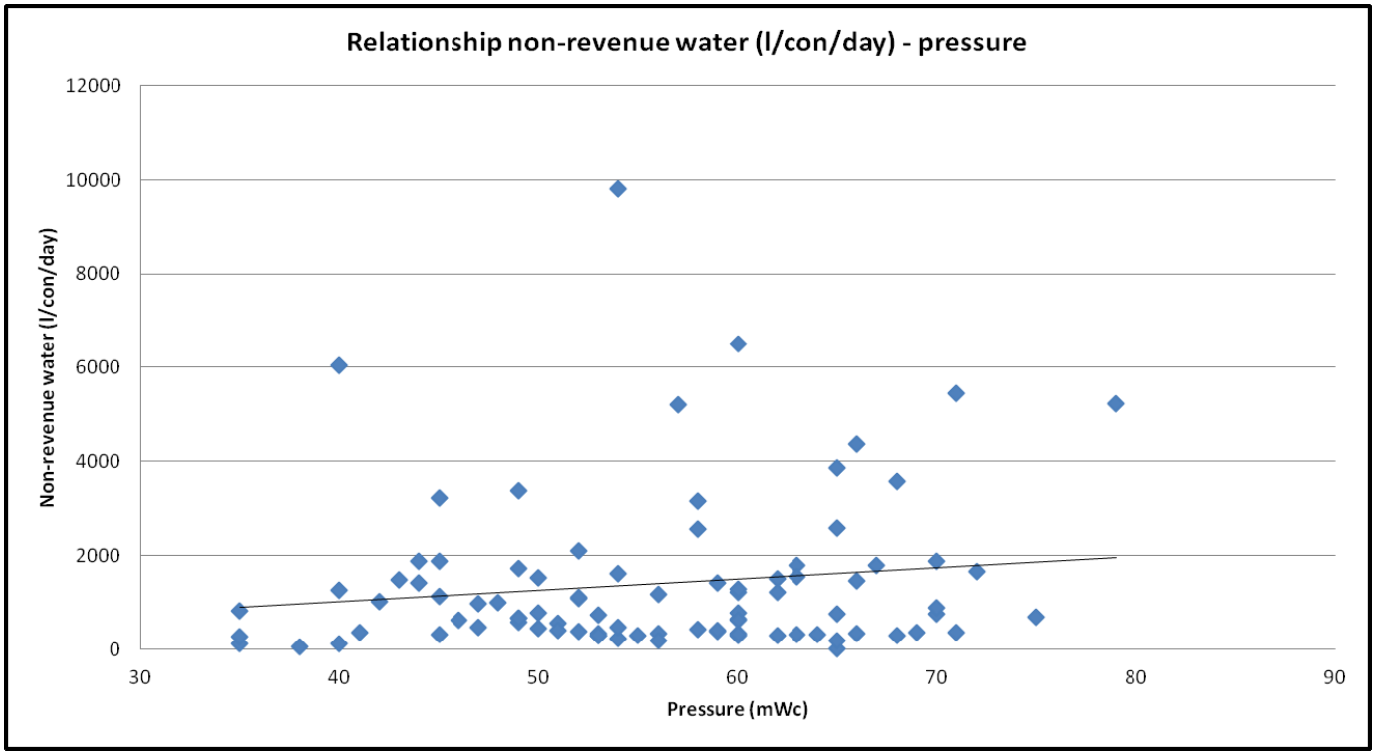


Figure 35: Relationships non-revenue water (l/con/day) - pressure

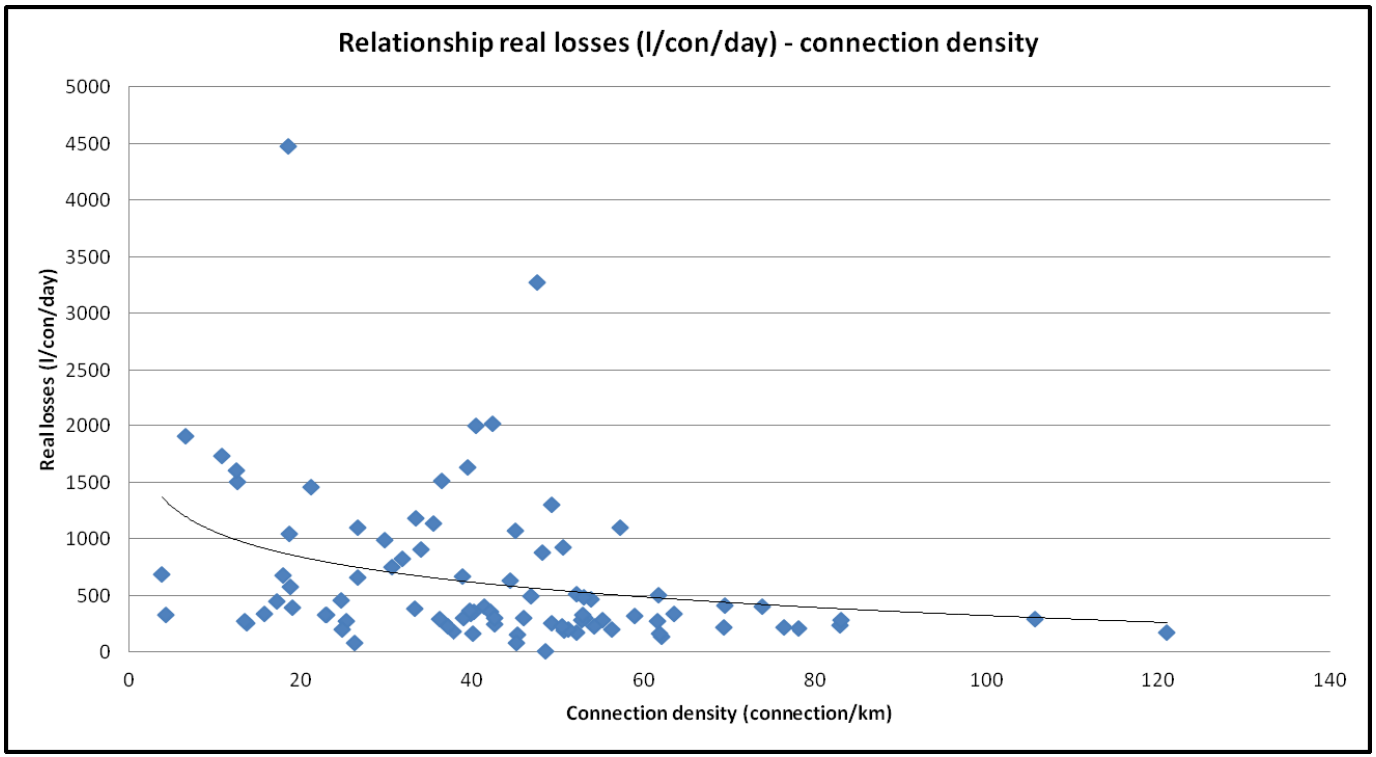


Figure 36: Relationships real losses (l/con/day) - connection density

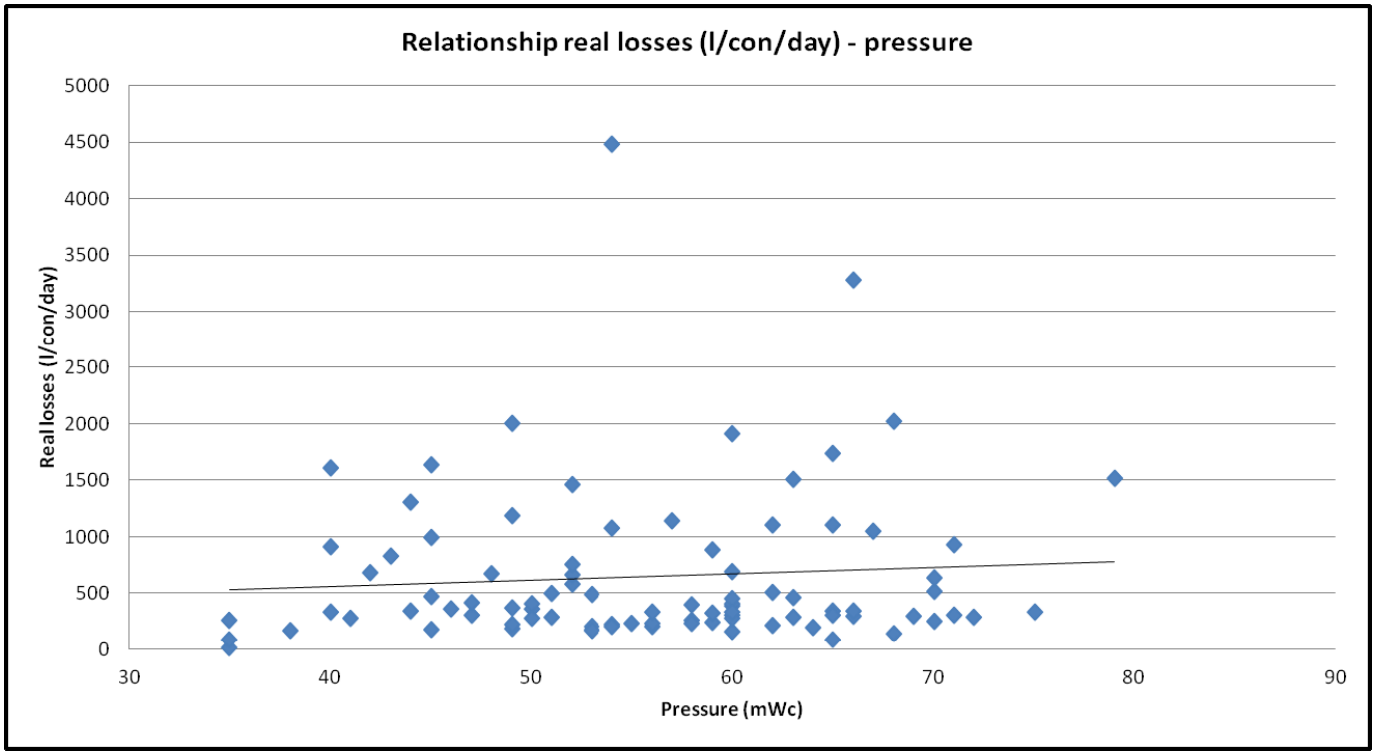


Figure 37: Relationship real losses (l/con/day) – pressure

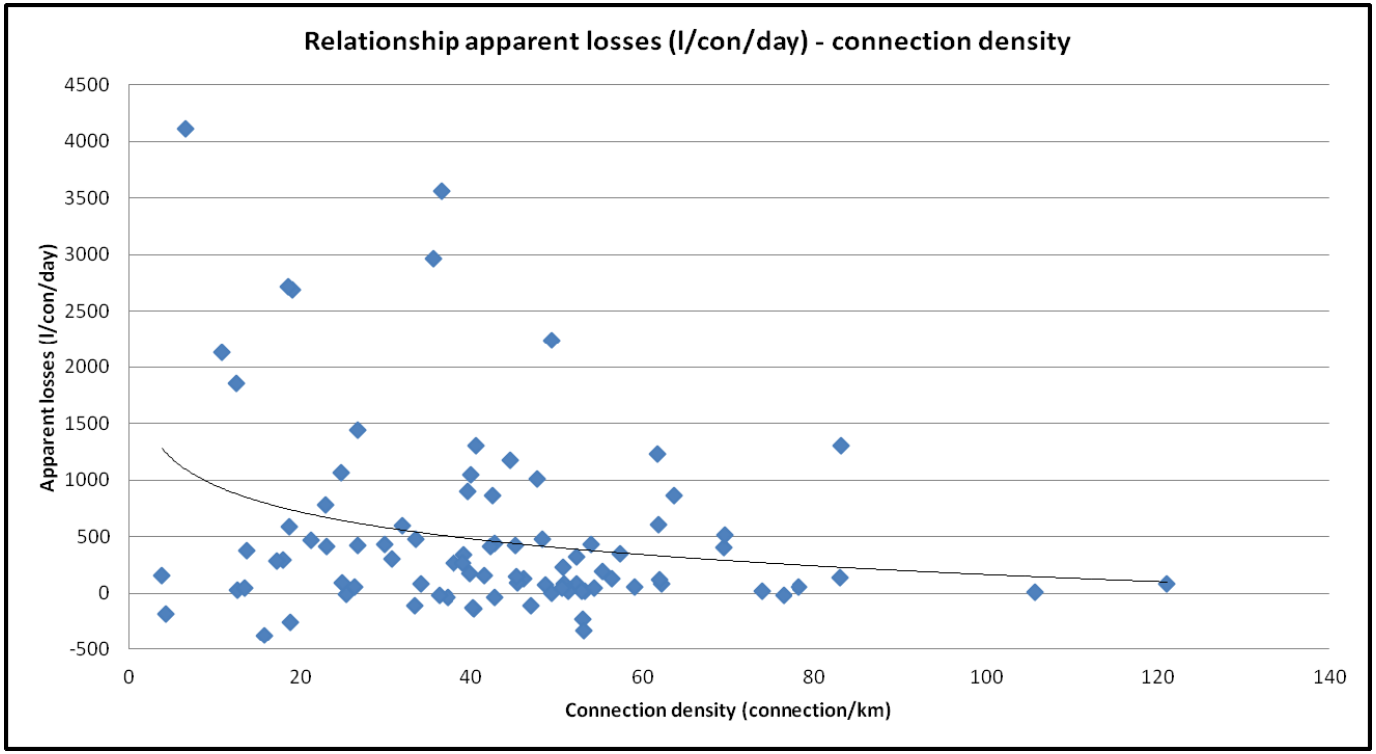


Figure 38: Relationship apparent losses (l/con/day) – connection density

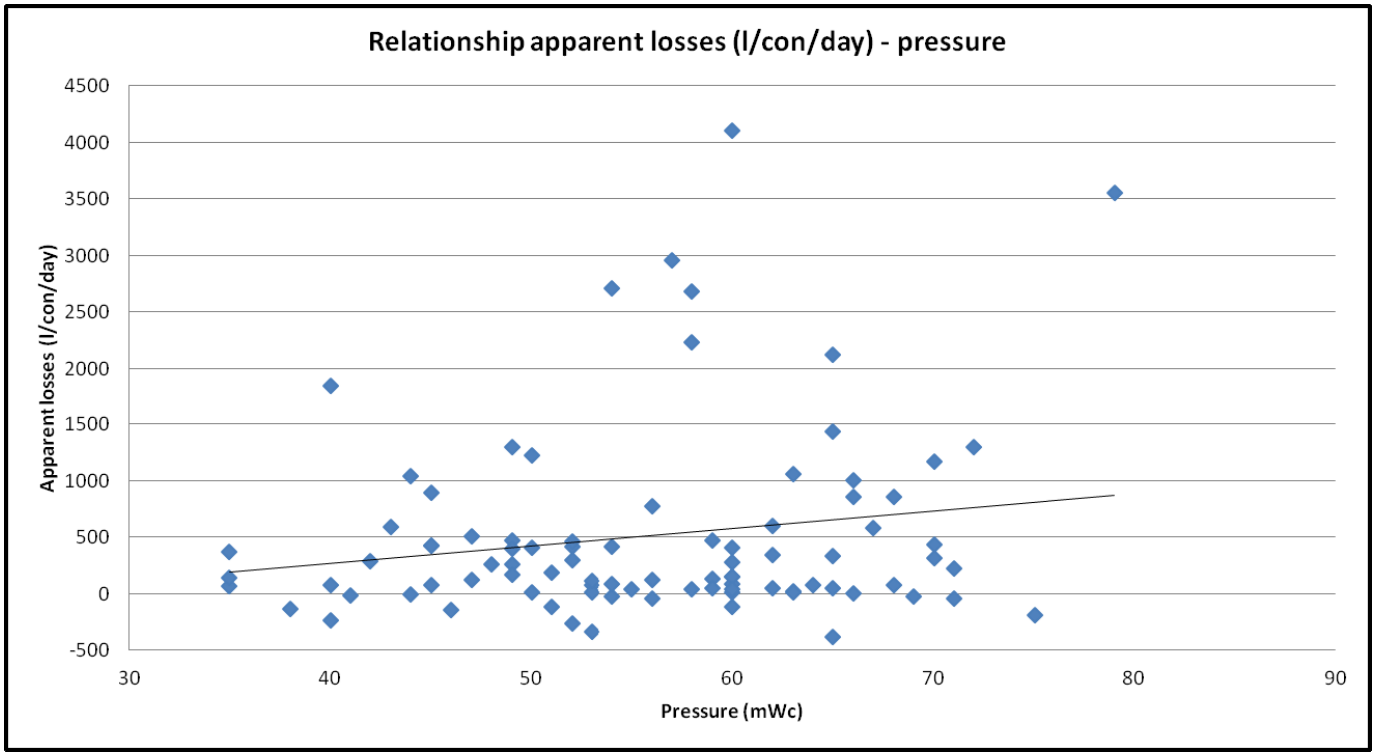


Figure 39: Relationship apparent losses (l/con/day) – pressure

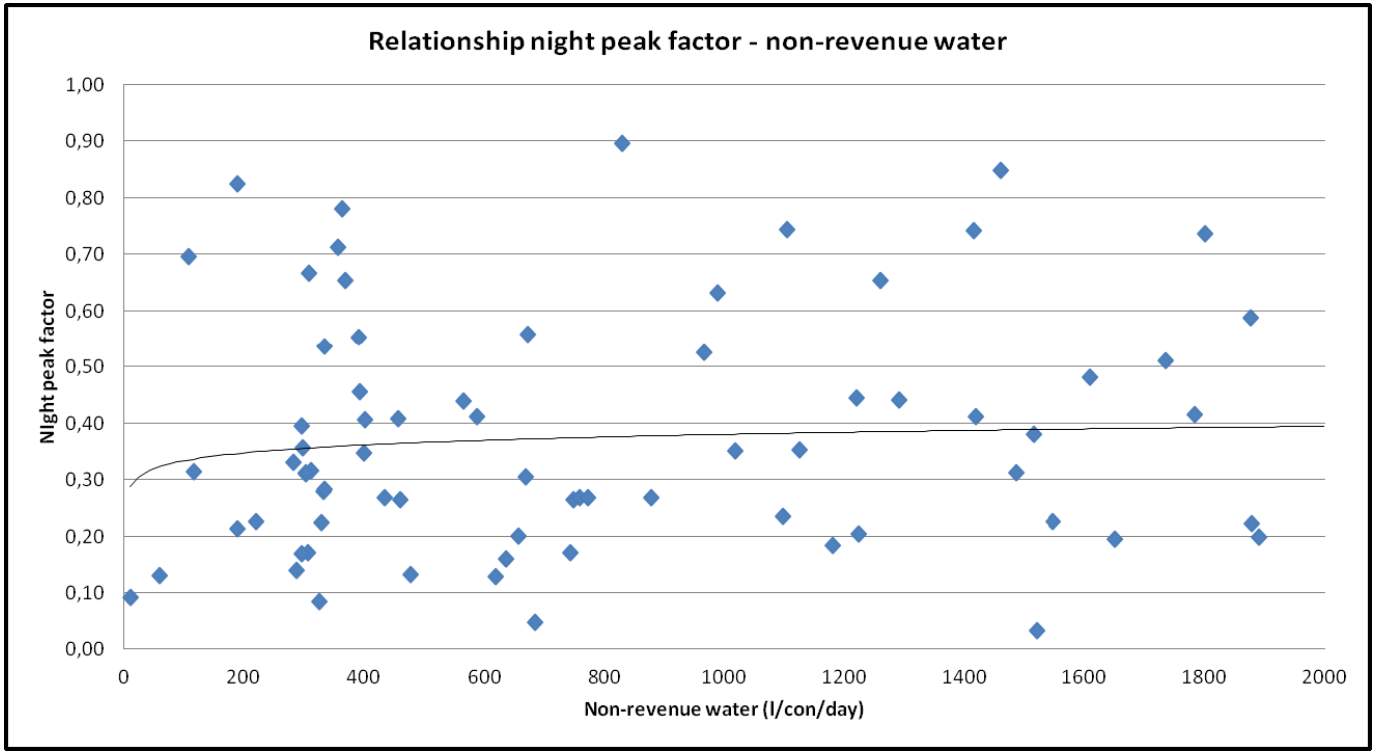


Figure 40: Relationship night peak factor – non-revenue water (l/con/day)

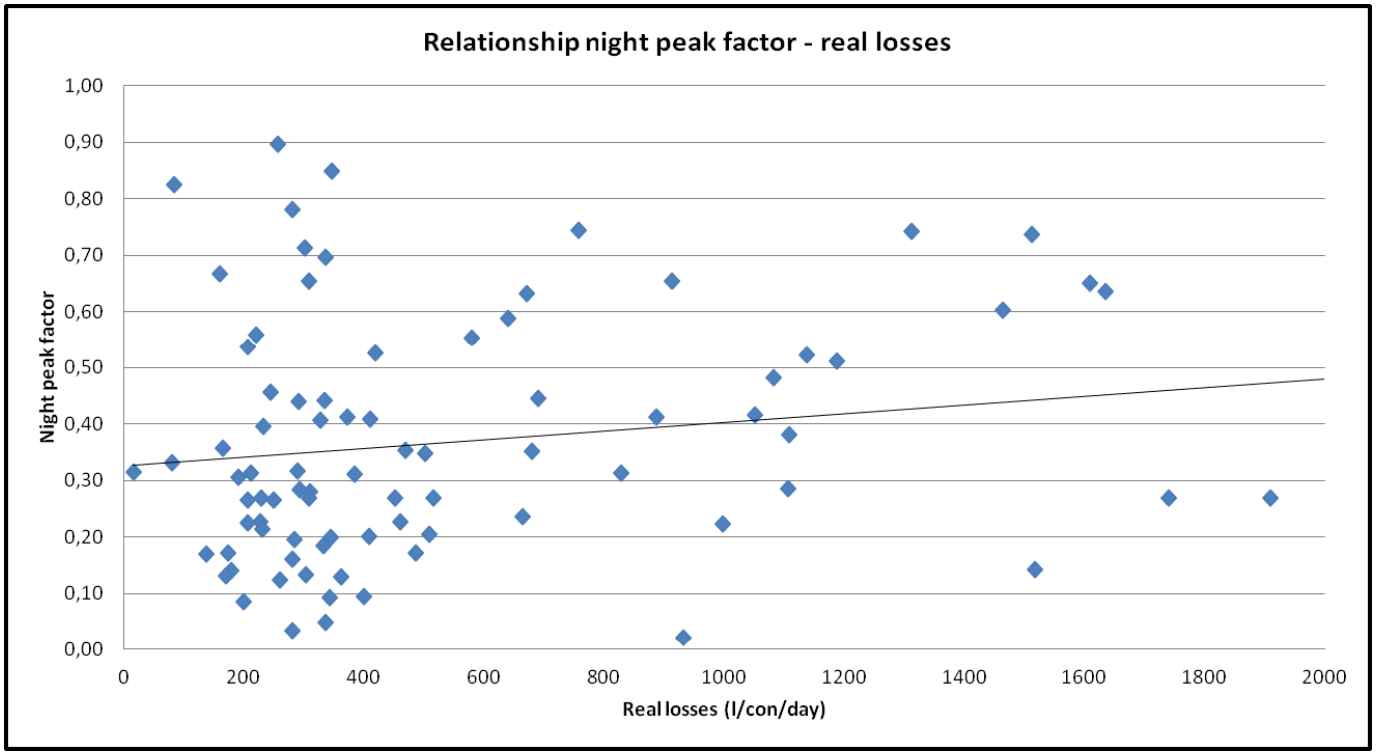


Figure 41: Relationship night peak factor - real losses (l/con/day)

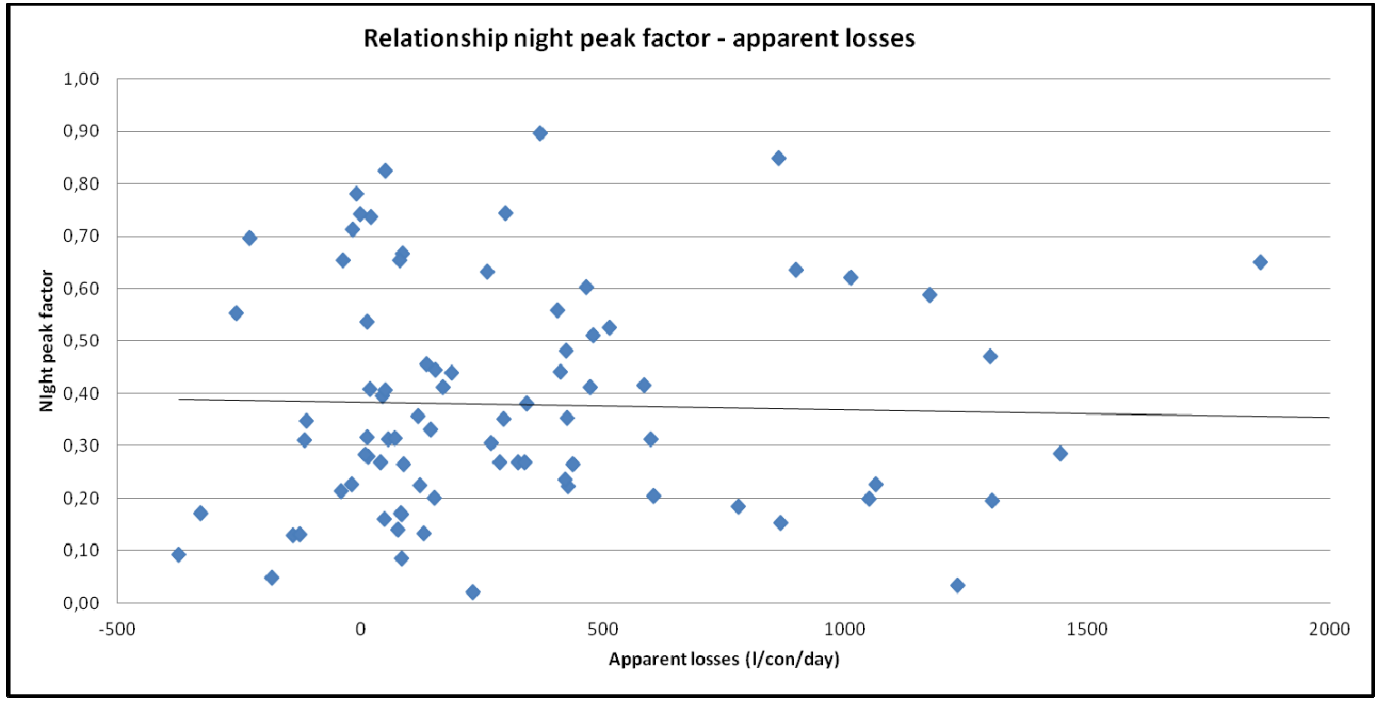


Figure 42: Relationship night peak factor - apparent losses (l/con/day)

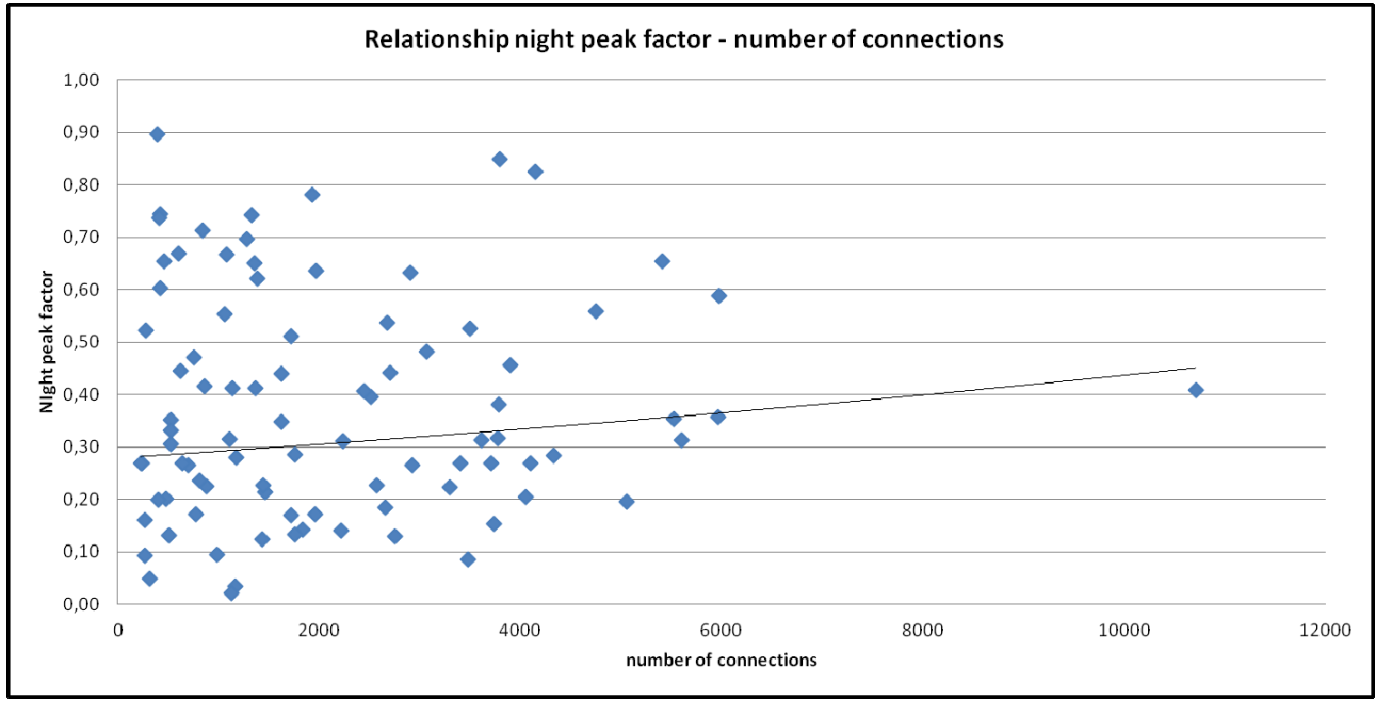


Figure 43: Relationship night peak factor - number of connections

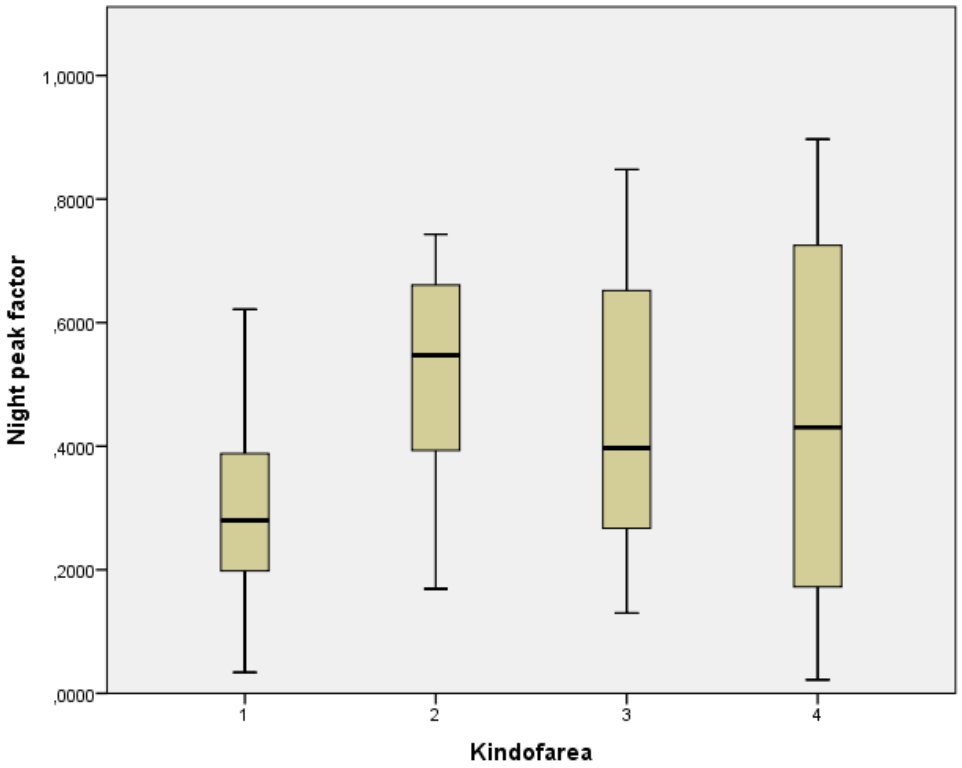


Figure 44: Relationship between night peak factor - sector type

Appendix VI: Assumptions

VI.I. Outliers

Table 23: Outlier identification for non-revenue water (l/con/day)

Case Number	Std. Residual	Non-revenue water (l/con/day)	Predicted Value	Residual
23	2,515	5458,3363	1450,682451	4007,6538318
27	4,950	9806,6556	1917,101986	7889,5535880
34	2,093	5227,5476	1891,670800	3335,8767798
49	2,263	6508,2819	2902,422663	3605,8592755
80	2,568	6038,8693	1945,886038	4092,9832716
84	2,361	5212,3913	1449,173660	3763,2176443

Table 24: Outlier identification for real losses (l/con/day)

Case Number	Std. Residual	Real losses (l/con/day)	Predicted Value	Residual
5	3,978	3277,5775	608,753298	2668,8242071
27	5,408	4480,8652	852,879348	3627,9858773
65	2,122	2006,5876	582,940804	1423,6468115
76	2,036	2020,0642	654,004235	1366,0599699

Table 25: Outlier identification for apparent losses (l/con/day)

Case Number	Std. Residual	Apparent losses (l/con/day)	Predicted Value	Residual
6	2,385	2687,9838	760,148833	1927,8349565
27	2,476	2711,8136	710,688983	2001,1246614
34	3,354	3558,1620	847,748148	2710,4138971
49	3,677	4113,8767	1142,267266	2971,6093864
70	2,216	2235,9024	445,004910	1790,8975290
84	2,993	2958,7681	540,194003	2418,5741135
89	-2,082	-182,6266	1500,346770	-1682,9733519

Table 26: Outlier identification for the night peak factor using non-revenue water (l/con/day)

Case Number	Std. Residual	Night peak factor	Predicted Value	Residual
2	2,432	,8968	,419703	,4771387
23	-2,303	,0217	,473576	-,4518593

Table 27: Outlier identification for the night peak factor using real losses (l/con/day)

Case Number	Std. Residual	Night peak factor	Predicted Value	Residual
2	2,748	,8968	,384220	,5126216
23	-2,364	,0217	,462559	-,4408416
38	2,118	,8479	,452847	,3950696
58	2,035	,8252	,445688	,3795551

Table 28: Outlier identification for the night peak factor using apparent losses (l/con/day)

Case Number	Std. Residual	Night peak factor	Predicted Value	Residual
2	2,367	,8968	,428435	,4684065
23	-2,105	,0217	,438216	-,4164990

Table 29: Cook's distance and Centered leverage value for the outlier cases

	Non-revenue water (l/con/day)		Real losses (l/con/day)	
	Cook's Distance	Centered Leverage Value	Cook's Distance	Centered Leverage Value
5			,14555	,01493
27	,21346	,01361	,25478	,01361

VI.II. Normality

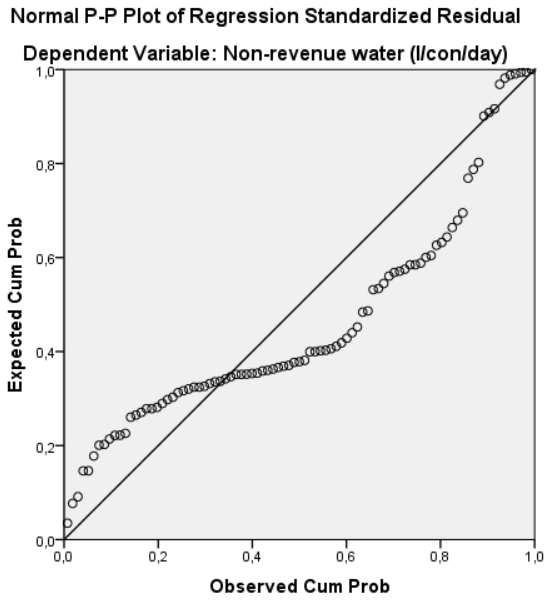


Figure 45: Normal probability plot for non-revenue water (l/con/day)

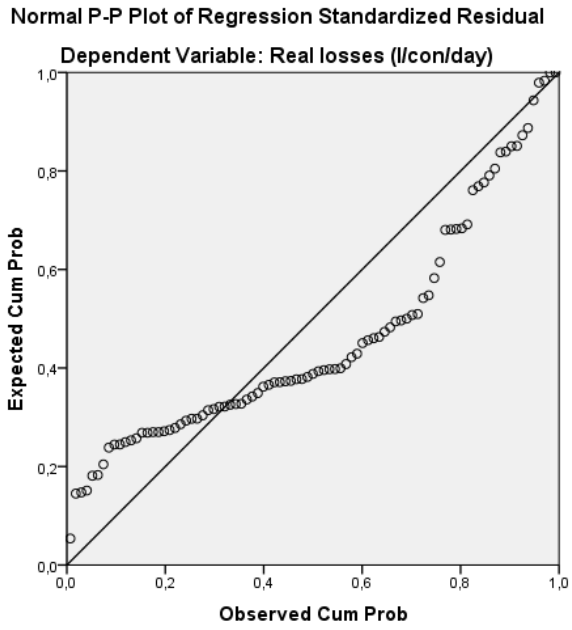


Figure 46: Normal probability plot for real losses (l/con/day)

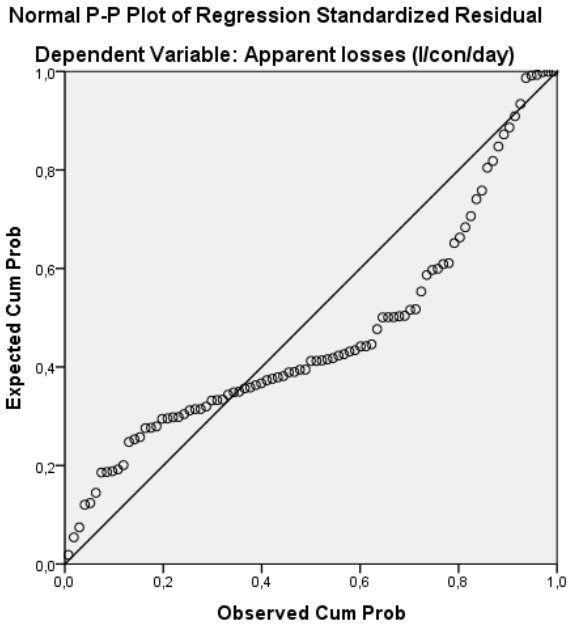


Figure 47: Normal probability plot for apparent losses (l/con/day)

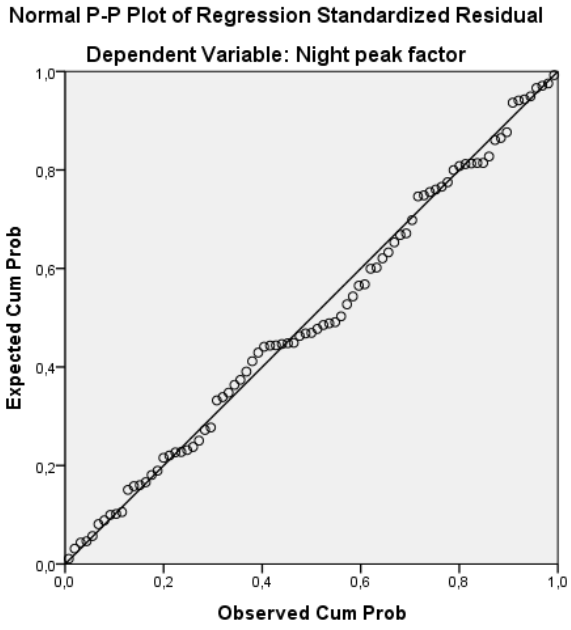


Figure 48: Normal probability plot for NPF using non-revenue water (l/con/day)

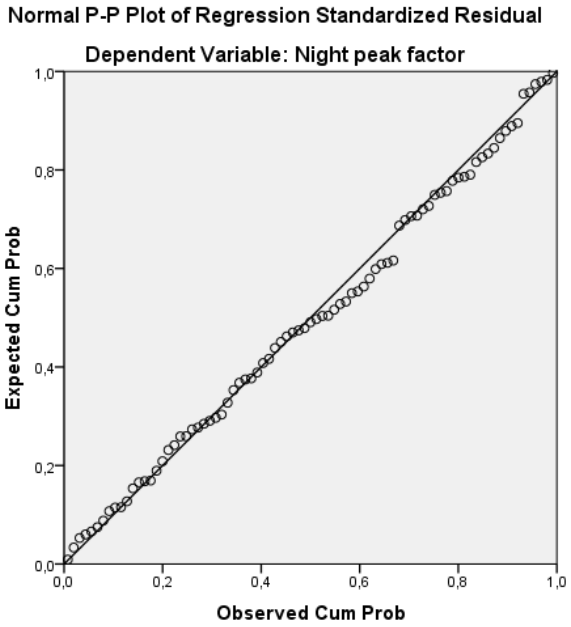


Figure 49: Normal probability plot for NPF using real losses (l/con/day)

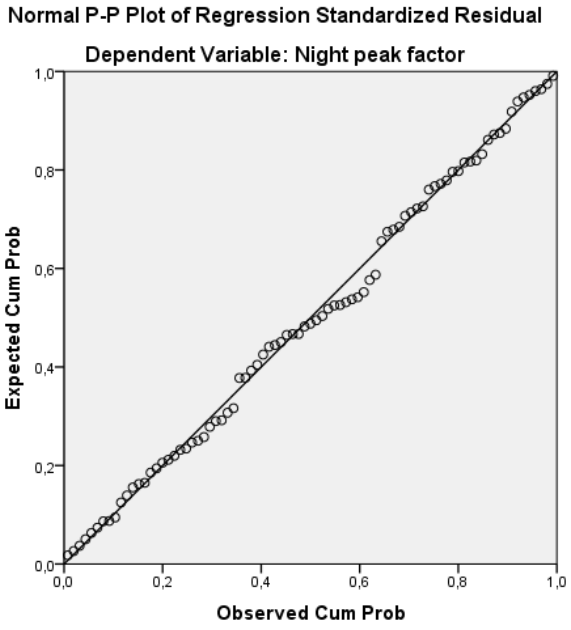


Figure 50: Normal probability plot for NPF using apparent losses (l/con/day)

Table 30: Normality test for dependent variables

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Non-revenue water (l/con/day)	,217	88	,000	,703	88	,000
Real losses (l/con/day)	,237	88	,000	,687	88	,000
Apparent losses (l/con/day)	,244	88	,000	,725	88	,000
Night peak factor	,119	88	,003	,953	88	,003

a. Lilliefors Significance Correction

VI.III. Homoscedasticity

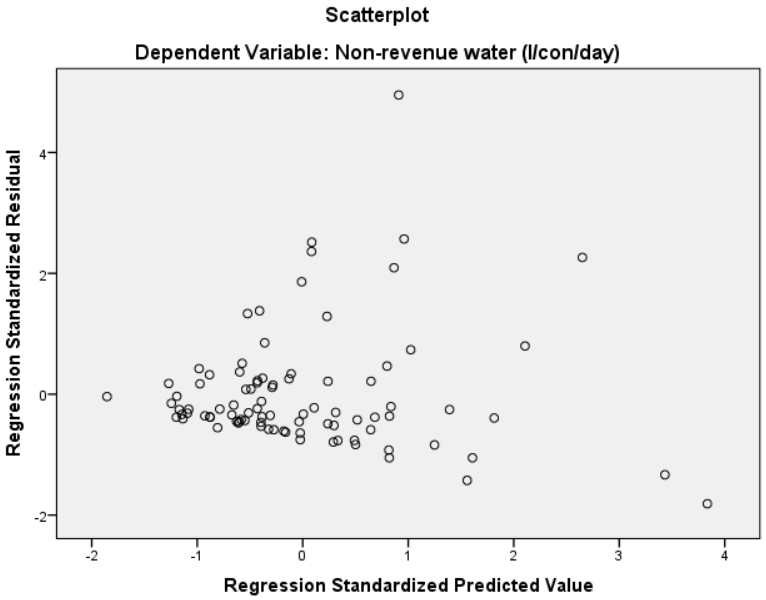


Figure 51: Residual plot for non-revenue water (l/con/day)

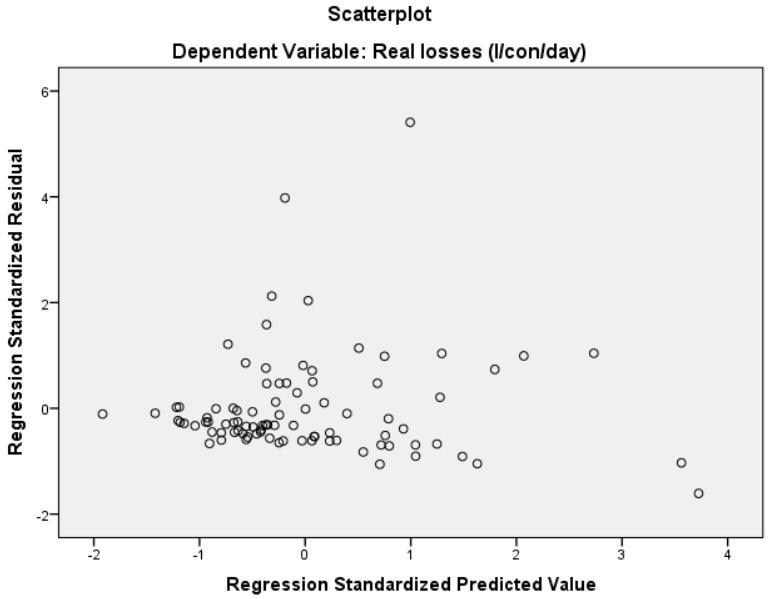


Figure 52: Residual plot for real losses (l/con/day)

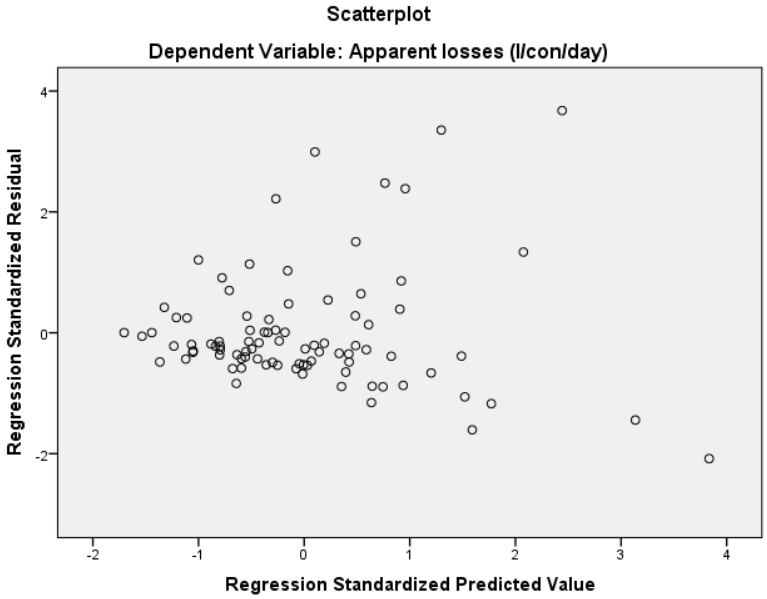


Figure 53: Residual plot for apparent losses (l/con/day)

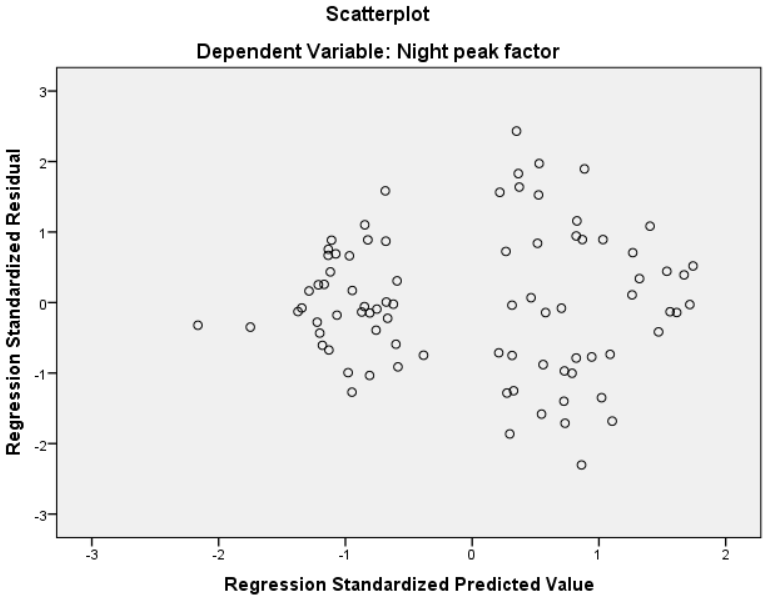


Figure 54: Residual plot for NPF using non-revenue water (l/con/day)

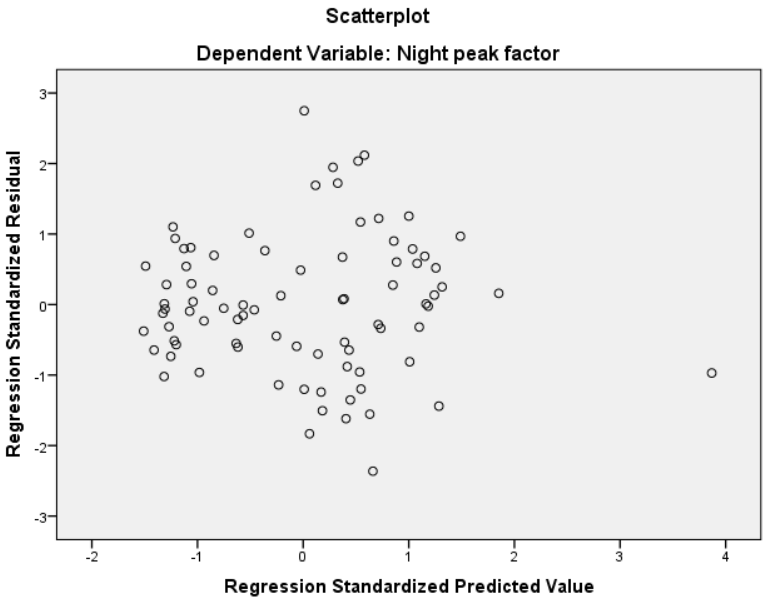


Figure 55: Residual plot for NPF using real losses (l/con/day)

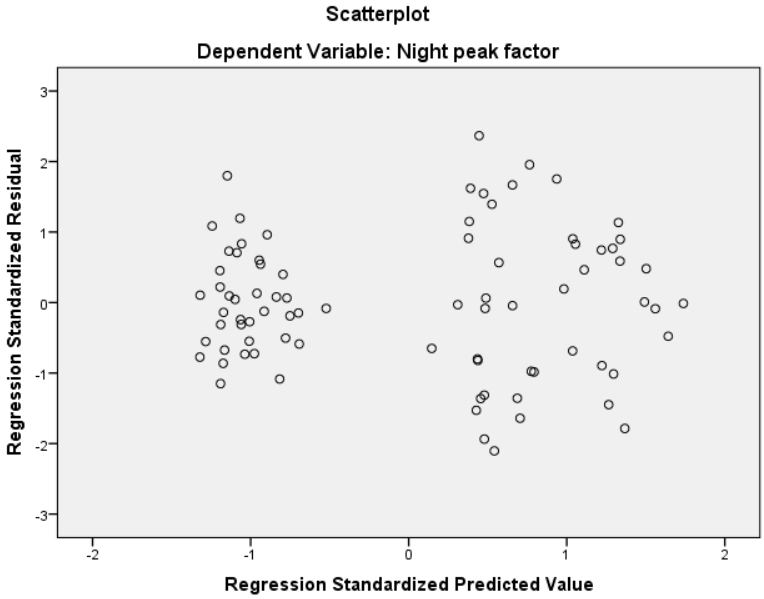


Figure 56: Residual plot for NPF using apparent losses (l/con/day)

Table 31: Heteroscedasticity test for dependent variables

	Koenker test		
	Statistic	df	Sig.
Non-revenue water (l/con/day)	7.351	89	.025
Real losses (l/con/day)	2.208	89	,331
Apparent losses (l/con/day)	19.110	89	,000

VI.IV. Independent errors

Table 32: Durbin-Watson value

	Durbin-Watson value
Non-revenue water (l/con/day)	1.866
Real losses (l/con/day)	2.029
Apparent losses (l/con/day)	2.034
Night peak factor using NRW (l/con/day)	1.594
Night peak factor using real losses (l/con/day)	1.584
Night peak factor using apparent losses (l/con/day)	1.593

VI.VI. Multicollinearity

Table 33: Collinearity statistics

	Tolerance	VIF
Non-revenue water (l/con/day)	.986	1.017
Real losses (l/con/day)	.997	1.003
Apparent losses (l/con/day)	.987	1.013
Night peak factor using NRW (l/con/day)	.968	1.033
Night peak factor using real losses (l/con/day)	.946	1.057
Night peak factor using apparent losses (l/con/day)	.979	1.021

Appendix VII: Multiple regression results

Table 34: Regression results dependent variable non-revenue water (l/con/day)

Model	Unstandardized Coefficients		Standardized Coefficients	Sig.	95,0% Confidence Interval for B		Correlations
	B	Std. Error	Beta		Lower Bound	Upper Bound	Zero-order
(Constant)	3390,10	1191,75		,006	1297,68	5781,62	
Pressure	37,31	15,55	,27	,025	9,96	64,05	,23
Connection density	-2574,23	627,05	-,36	,001	-4172,05	-1294,21	-,32

Table 35: Regression results dependent variable night peak factor using non-revenue water (l/con/day)

Model	Unstandardized Coefficients		Standardized Coefficients	Sig.	95,0% Confidence Interval for B		Correlations
	B	Std. Error	Beta		Lower Bound	Upper Bound	Zero-order
(Constant)	-,190	,301		,531	-,789	,410	
Urban neat vs. township	,247	,059	,450	,000	,130	,363	,314
Urban neat vs. Informal untidy	,156	,065	,254	,019	,026	,286	,127
Urban neat vs. Rural	,161	,060	,294	,009	,042	,280	,128
Number of connections	,286	,234	,125	,226	-,180	,752	,052
NRW (l/con/day)	,054	,045	,122	,234	-,036	,143	,140

Table 36: Regression results dependent variable night peak factor using number of connections adapted structural model

Model	Unstandardized Coefficients		Standardized Coefficients	Sig.	95,0% Confidence Interval for B		Correlations
	B	Std. Error	Beta		Lower Bound	Upper Bound	Zero-order
(Constant)	-,012	,219		,955	-,442	,404	
Urban neat vs. township	,246	,049	,449	,001	,139	,333	,312
Urban neat vs. Informal untidy	,164	,075	,267	,031	,025	,322	,127
Urban neat vs. Rural	,169	,075	,309	,031	,023	,310	,128
Number of connections	,264	,194	,116	,129	-,179	,724	,052

Table 37: Regression results dependent variable real losses (l/con/day)

Model	Unstandardized Coefficients		Standardized Coefficients	Sig.	95,0% Confidence Interval for B		Correlations
	B	Std. Error	Beta		Lower Bound	Upper Bound	Zero-order
(Constant)	1536,87	663,90		,021	293,22	2980,62	
Pressure	4,53	6,76	,06	,519	-8,81	18,86	,08
Connection density	-731,49	318,63	-,29	,023	-1451,53	-244,69	-,29

Table 38: Regression results dependent variable real losses (l/con/day) adapted structural model

Model	Unstandardized Coefficients		Standardized Coefficients	Sig.	95,0% Confidence Interval for B		Correlations
	B	Std. Error	Beta		Lower Bound	Upper Bound	Zero-order
(Constant)	1802,38	506,14		,001	1016,02	2878,96	
Connection density	-739,86	302,00	-,29	,014	-1447,00	-234,81	-,29

Table 39: Regression results dependent variable night peak factor using real losses (l/con/day)

Model	Unstandardized Coefficients		Standardized Coefficients	Sig.	95,0% Confidence Interval for B		Correlations
	B	Std. Error	Beta		Lower Bound	Upper Bound	Zero-order
(Constant)	-,242	-,022		,201	-,694	,126	
Urban neat vs. township	,237	,005	,433	,001	,148	,345	,312
Urban neat vs. Informal untidy	,167	-,001	,273	,013	,040	,295	,127
Urban neat vs. Rural	,177	,000	,324	,038	,028	,331	,128
Number of connections	,417	,017	,183	,018	,094	,848	,052
Real losses (l/con/day)	9,847E-05	6,005E-06	,305	,005	3,86E-05	,000	,285

Table 40: Regression results dependent variable apparent losses (l/con/day)

Model	Unstandardized Coefficients		Standardized Coefficients	Sig.	95,0% Confidence Interval for B		Correlations
	B	Std. Error	Beta		Lower Bound	Upper Bound	Zero-order
(Constant)	943,69	626,99		,154	-420,66	2313,44	
Pressure	19,24	8,80	,27	,034	4,19	35,13	,24
Connection density	-955,07	347,21	-,27	,012	-1451,53	-284,61	-,24

Table 41: Regression results dependent variable night peak factor using apparent losses (l/con/day)

Model	Unstandardized Coefficients		Standardized Coefficients	Sig.	95,0% Confidence Interval for B		Correlations
	B	Std. Error	Beta		Lower Bound	Upper Bound	Zero-order
(Constant)	,002	,219		,995	-,393	,403	
Urban neat vs. township	,245	,003	,448	,001	,138	,353	,312
Urban neat vs. Informal untidy	,163	,001	,265	,028	,032	,314	,127
Urban neat vs. Rural	,169	-,001	,310	,043	,000	,323	,128
Number of connections	,255	-,003	,112	,154	-,153	,671	,052
Apparent losses (l/con/day)	-7,66E-06	2,69E-06	-,030	,729	-5,23E-05	5,943E-05	-,051

MODELING THE RELATIONSHIPS EXISTING WITH THE PERFORMANCE OF NON-REVENUE WATER

Using structural models to link the performance of non-revenue water and its components with the variables that can be obtained

Author: Ing. C.B. (Bertine) Korevaar

Graduation program:
Construction Management and Urban Development 2012 – 2013

Graduation committee:
Prof. dr. ir. W.F. Schaefer
Dr. ir. B. Glumac
Ir. B. van Weenen
Ir. M. Riemersma (Royal HaskoningDHV)

Date of graduation:
14 March 2012

ABSTRACT

To manage drinking water more efficiently, non-revenue water should be reduced, which can only been done after analyzing non-revenue water. The current process of analyzing the performance of non-revenue water or its components should be more efficient in time. To take a step forward to a more time sufficient approach it is proposed, in cooperation with Royal HaskoningDHV, to examine the relationships between the performance of non-revenue water or its components and the variables that influence or are influenced by the performance. The performance of non-revenue water is expressed in liters/connection/day. The variables included in the model are the connection density, the pressure, the night-peak factor, the number of connections and the sector type. The relationships are shown in structural equation models created using the path analysis; the equations evolving from the models are analyzed using the multiple backward elimination analysis. An important relationship which is proven by the analysis is the influence of the performance of the real losses on the night peak factor. The model also shows that the pressure does not influence the performance of the real losses. The results from the 95% confident intervals, the variance and shrinkage in the model show that the structural equations are too inaccurate to be used for predictive purposes.

Keywords: Performance non-revenue water, multiple regression analysis, prediction models

INTRODUCTION

Currently the world population is growing, the increase in world population leads to an increase in water demand, putting a growing pressure on water as a resource. Water must be managed more efficiently since there are already regions where water is a scarce resource. Water can be managed more efficiently by reducing non-revenue water (NRW) (Gonzalez-Gomez, et al., 2011). "*Non-revenue water is the portion of water that a utility places in the distribution system*

that is not billed and, therefore, recovers no revenue for the utility" (Thornton, et al., 2008). NRW can effectively be reduced, when a reduction plan is made. This reduction plan can be made when the performance of non-revenue water and its components (apparent losses and real losses) is analyzed based upon the measurements performed in the system. The current process of gathering these measurements should be more efficient in time, so no time will get lost. To take a step forward to develop a more time sufficient analysis it is proposed, in cooperation with Royal HaskoningDHV, to analyze the relationships between the performance of NRW or its components and the variables that are influenced or are influencing this performance. Royal HaskoningDHV recognizes this as an opportunity to link one of their respected products with NRW to create a new business opportunity. The product that will be linked to the performance of NRW is OPIR (Optimal Production through Intelligent Control). Some of the variables will be obtained from OPIR, the other variables are data that can be obtained at the beginning of a project.

LITERATURE REVIEW

Non-revenue water and its components

NRW can be divided in three components (1) unbilled authorized consumption, (2) apparent losses and (3) real losses. Unbilled authorized consumption is the water that is used by utilities for operational purposes. The unbilled authorized consumption should be less than 1% of the system input volume, thus the volume of NRW consists mainly of the real losses and the apparent losses. The apparent losses are often the result of influences on the company, which are usually beyond the scope of the day-to-day operation practices. The real losses are caused by leakages in all parts of the systems and overflows at the utility reservoirs. The volumes of the components can be derived from either conducting measurements in the system or by verifying the numbers that are used by the administration, when both are not possible rules of thumb are used to determine the volumes of the components (Thornton, et al., 2008).

Analyzing performance of non-revenue water

The performance of NRW is analyzed using the performance indicators (PI). The PI helps utilities to understand the water loss better, to measure and compare the performance and to be able to define and set targets for improvement. Royal HaskoningDHV has determined that the PI that will be used in the analysis should be able to determine the performance of NRW, the real losses and the apparent losses and that it is not important whether the PI can be used to compare systems with each other. The PI that will be used in this research is the PI liters/connection/day, which satisfies the requirements posed by Royal HaskoningDHV. The PI is well-known by water utilities and can easily be transformed in a more detailed PI by dividing it by the average pressure in the system. The PI liters/connection/day is chosen, because it is assumed that the greater portion of the losses occurs on the service connections (Thornton, et al., 2008), since most of the projects are introduced in urban areas.

OPIR

Some of the variables from which the relationship with the performance will be analyzed can be obtained from OPIR. OPIR is a real-time software package that is used to optimize the operation of drink water systems, by predicting the future pattern. This intelligent control is established using the measured daily water demand patterns to predict the future patterns using statistical algorithms. The OPIR package can be expanded with a pressure model, this module controls the pressure for the weakest point in the system; the pressure in this point should never be lower than the minimum allowable pressure (RoyalHaskoning DHV, -).

The measured daily water demand patterns and the measured and controlled pressure values are the elements that are used to link OPIR with the performance of NRW. To variables can be created from the daily water demand pattern, the day peak factor (DPF) and the night peak factor (NPF). These two variables are estimated by dividing the maximum day consumption or the minimum night consumption by the average consumption of the day (Trifunovic, 2006). The pressure can directly be used as one of the variables.

Variables

The performance of NRW and its components will be expressed in liters/connection/day. The variables DPF, NPF and pressure will be obtained from OPIR, while the variables that can be obtained at the beginning of a project are the sector type, the number of connections and the length of mains (km), the connection density can be estimated using this data.

Thornton, et al. (2008) discusses the influence of the pressure, the number of connections and the length of mains on the real loss volume. Since the performance of the real losses is determined by dividing the volume by the size of the area (specifically the number of connections), it can be concluded that there is no relationship between the performance of the real losses and the number of connections or the length of the mains. It is assumed that there will be a relationship between the pressure and the performance of the real losses. By analyzing the relationships with the performance of the real losses another variable can be included as well, this is the connection density. Fantozzi, et al. (-) describes a nonlinear relationship between the real losses (l/con/day) and the connection density. The nature of the relationship between the real losses (l/con/day) and the pressure is undefined. Since the real losses (l/con/day) is a component of NRW (l/con/day), it is expected that the same variables have a relationship with NRW (l/con/day).The same variables will be used to test whether these variables have an influence on the apparent losses (l/con/day), since there is little known about the variables that influence the performance of the apparent losses.

The DPF and NPF are influenced by the real losses (l/con/day) and thus NRW (l/con/day). Zhang (2005) proves that the DPF has a relationship with the PI NRW (%) (in which only the real losses are used), therefore it is assumed that a relationship exists with the real losses (l/con/day). The relationship between the NPF and the real losses is proven since they are both needed to determine the minimum night flow (Thornton, et al., 2008). Since it is unknown whether the apparent losses (l/con/day) influences the DPF or NPF, these relationships will be analyzed as well.

The DPF is also influenced by the number of consumers (Zhang, 2005), the consumption category, the water price/income, the weather variables, the resident population per account (Arbués, et al., 2003) and the cultural differences (Blokker, 2010) and the NPF by the number of consumers and the type of connections (Thornton, et al., 2008). The relationship between the DPF and NPF and the number of connections will be analyzed, instead of the number of consumers. The number of connections is administrated by the water companies and thus can be accessed easily. Zhang (2005) shows that a nonlinear relationship exists between the number of consumers and the DPF, it is assumed that the same relationship exists with the number of connections. The nature of the relationship between the NPF and the number of connections is undefined. The influence of the sector type on the DPF and NPF will also be analyzed. The sector type replaces the variables consumption category, the income and the household composition for the DPF; it replaces the variable type of connections for the NPF. All these variables differ for the different sector types and thus it is expected that the sector type will influence the DPF and NPF. The sector types are (1) formal urban neat, (2) township, (3) informal-untidy and (4) rural. The remaining variables influencing the DPF are the weather variables and the cultural differences, these can be neglected since the data will be collected from one region.

METHOD

Choice of method

A path analysis will be used to create a structural equation model from the found relationships. Structural equations will evolve from this model. The regression analysis or the machine learning technique can be used to analyze the structural equations. The machine learning technique outperforms the regression analyses in modeling a short-term water demand forecast model (Bougadis, et al., 2005), but needs at least a hundred cases to train a network (StatSoft, -) Since the data set dispose of a hundred cases, the regression analysis is chosen.

With the help of a regression analysis the structural equation, shown in equation 1, can be determined. The relationships between the dependent and independent variables are expressed with the β_n estimates.

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n \qquad 1$$

The multiple linear regression method will be used to analyze the structural equations. A backward elimination will be used to find the best combination of independent variables for the structural equation. This method is least likely to miss an independent variable that does predict the outcome. Variables are eliminated when they are not contributing statistically significant ($p>0.1$) to the prediction of the dependent variable. The final model will be accepted when it has a significance level of $p<0.05$. The variable sector type is a nominal variable; dummy variables will be used to be able to use this variable in the multiple regression analysis (Field, 2009). To determine whether the multiple regression analysis can be used the following assumptions must be met: (1) outliers must be examined, (2) there should be linearity in the dataset, (3) homoscedasticity must be present in the data, (4) as well as normality, (5) no independent errors should be present, (6) as well as no multicollinearity.

Application of regression analysis in water demand management

The regression analysis is used for many disciplines; one of the disciplines is water demand management. The regression analysis is used to forecast the water demand in certain areas or in general. The methods that are used are time series analysis, but also multiple regression methods which determines the most suitable variables for forecasting the water demand (Bougadis, et al., 2005). The regression analysis is never used to analyze structural equations involving the performance of NRW, though the practice of the regression analysis in the water demand management shows that it is possible to use this technique to analyze the performance of NRW, since NRW is a component of the total water demand.

RESULTS

Data collection

The collected data is made available by the company JOAT. JOAT is a South-African company that specializes in all aspects of water management, one of their specializations is NRW. Two datasets are provided; the datasets are from the areas of West and South Durban, South Africa. Each dataset consists of several sectors for which a water balance is determined and for which the characteristics of the sector are described. Both dataset consists of various types of sectors, though overall South Durban is more developed than West Durban. The values in the water balances are an approximation determined by the engineers, each area has their own engineer that determines the water balances of the sectors. Even though the engineers of JOAT have a lot of field experience and there are company standards on how the water balance must be determined, it can happen that incorrect values are displayed in the water balances (Pena, 2013). The data is cleaned for sectors that are too small or too big, sectors that experienced out-of-range values in their data are also excluded. Since the data set does not provide all values of the maximum day flow, the DPF will not be analyzed. The type of sector is for some sectors not provided, but this variable can be determined using the internet.

Structural model

Figure 1 shows the structural model determined from the relationships. Three models will be analyzed the model of the performance of NRW, the second is the model for the real losses (l/con/day) and the last is the one for the apparent losses (l/con/day). The NPF and DPF are shown in the same structure, because they are both influenced by the same variables.

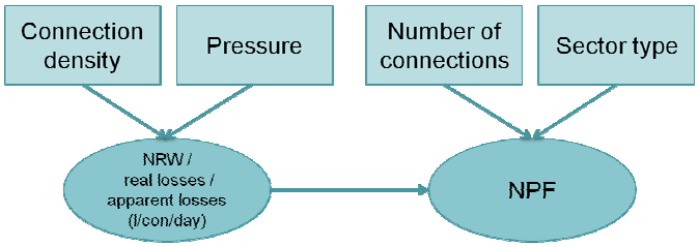


Figure 1: Model specification with dependent variables performance indicator (l/service connection/day) and the night peak factor and the day peak factor

Results regression analysis

The relationships are examined using the results from the literature and the scatter made for the relationships. The relationship between the connection density and the performance of NRW or any of its components is nonlinear (Fantozzi, et al., -), the log function is used to express this nonlinear relationship. The relationship between the NPF and NRW (l/con/day) is also nonlinear, but here an exponent function is used to express this relationship. The other relationships are all linear.

Whenever the assumptions are violated, robust techniques will be used. When the assumption of homoscedasticity is violated, the weighted least squares regression is used and when the assumption of normality is violated the bootstrap technique is used.

The first structural model that is analyzed is the NRW (l/con/day) model; the resulting model is shown in Figure 2. The values displayed in the model are the β-coefficients, these show how much influence an independent variable has on the dependent variable. The value refers the number of standard deviations the value of the dependent variable will change, per standard deviation increase of the value of the independent variable. The β-coefficient can vary between the -1 and 1 (Field, 2009).

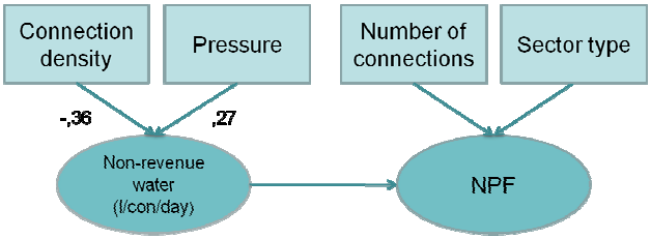


Figure 2: Results of the regression analysis on the model of non-revenue water (l/con/day)

The model shows that a structural equation is determined to predict NRW (l/con/day) using the connection density (β=-0.36, p<0.001) and the pressure (β=0.27, p<0.05). The two independent variables explain 17,7 % of the variance (R^2=0.18, F=9.25, p<0.001) of the variance in the NRW (l/con/day) value. The structural equation is shown in equation 2.

$$NRW\ (l/con/day) = 3390.10 - 2574.23\ \log(Connection\ density) + 37.31\ Pressure \qquad 2$$

The model also shows that the structural equation for the NPF is not determined, both the variables NRW (l/con/day) (β=.12, p=.23) and the number of connections (β=.12, p=.13) do not contribute significantly (p>.1) to the variance in the NPF. The regression analysis is not used any further, since the resulting structural equation would only be able to determine an average NPF for each sector type.

The second structural model analyzed using the regression analysis is the real losses (l/con/day) model for which the results are displayed in Figure 3.

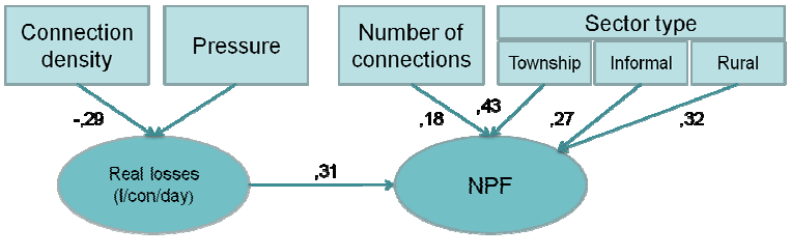


Figure 3: Results of the regression analysis on the model of the real losses (l/con/day)

The model shows that the real losses (l/con/day) can only be predicted using the connection density (β=-.29, p<.05), since the contribution of the pressure (β=.06, p=.52) is not statistically significant (p>.1). The connection density is accounted for 8,4% of the variance (R^2=0.08, F=7.96, p<0.01) in the dependent variable. The independent variables predicting the NPF in the real losses (l/con/day) model are accounted for 30,8% of the variance (R^2=0.31, F=6.84 and p<.001). It is found that the number of connections significantly predicts the NPF (β=.18, P<0.05), as does the real losses (l/con/day) (β=.31, p<0.01). Dummy variables are created to be able to analyze the nominal variable sector type. The urban formal neat sector type is used as the baseline sector type. The results show that the urban formal neat vs. township significantly predicts the NPF (β=.43, p<.01), as does the urban formal neat vs. informal untidy (β=.27, p<.05) and the urban formal neat vs. rural (β=.32, p<.05). Equation 3 and 4 show the structural equation belonging to the model in Figure 3.

$$Real\ losses\ (l/con/day) = 1802.38 - 739.86\ \log(Connection\ density) \quad (3)$$

$$NPF = -.242 + .237 * Township + .167 * Informal\ untidy + .177 * Rural + .417 * Number\ of\ connections^{0,00004483} + 9.847E^{-05} * Real\ losses\ (l/con/day) \quad (4)$$

The last analyzed structural model is that of the apparent losses (l/con/day), the results are shown in Figure 4.

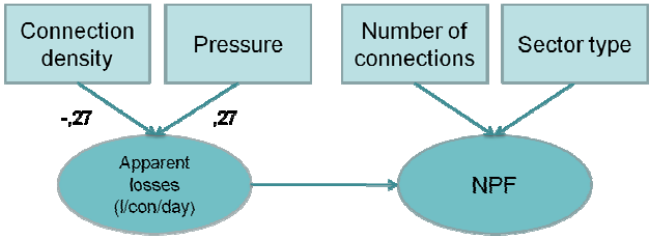


Figure 4: Results of the regression analysis on the model of the apparent losses (l/con/day)

Both the connection density (β=-.27, p<0.05) and the pressure (β=.27, p<.05) significantly predict the apparent losses (l/con/day), the variables explain 13% of the variance (R^2=0.13, F=6.42, p<0.01) of the apparent losses (l/con/day). Equation 5 shows the structural equation for the apparent losses (l/con/day).

$$Apparent\ losses\ (l/con/day) = 943.69 - 955.07\ \log(Connection\ density) + 19.24\ Pressure \quad (5)$$

The structural equation for the NPF is not determined using this model, since both the number of connections (β=.11, p=.15) and the apparent losses (l/con/day) (β=-.03, p=.73) do not contribute significantly (p>.1) to the NPF. The regression analysis is not used any further, since the resulting structural equation would only be able to determine an average NPF for each sector type.

Table 1 displays the range between which the predicted value lies using the 95% confidence interval resulting from the regression analysis and the average values from the collected data.

Table 1: Boundaries non-revenue water (l/con/day) model

	Predicted value	Lower bound	Upper bound
NRW (l/con/day)	672.98	-5921.27	11763.38
Real losses (l/con/day)	423.74	-1680.29	2441.42
Apparent losses (l/con/day)	236.44	-2891.86	3741.25
NPF using real losses (l/con/day)	0.239	-0.575	0.974

Validation

The internal validation is performed by examining the R^2 as well as the adjusted R^2. Two adjusted R^2 are estimated, Wherry's adjusted R^2 and Stein's adjusted R^2. Wherry's adjusted R^2 estimates the level of variance in the dependent variable that is accounted for if the model was derived from the population from which the sample was taken. Stein's adjusted R^2 estimates the level of variance in the dependent variable that is accounted for if the model was derived from an entirely different data set. The outcomes are shown in Table 2

Table 2: Adjusted R^2 estimation

	R^2	Wherry adjusted R^2	Stein adjusted R^2
Non-revenue water (l/con/day)	0.177	0.158	0.128
Real losses (l/con/day)	0.084	0.073	0.052
Apparent losses (l/con/day	0.130	0.110	0.079
NPF using real losses (l/con/day)	0.308	0.213	0.266

DISCUSSION, CONCLUSSION AND RECCOMENDATIONS

The following relationships are shown using the literature study and the regression analysis:

- The model NRW (l/con/day) shows that the NRW (l/con/day) can be predicted using the pressure and the log function of the connection density. No structural equation is developed to determine the NPF;
- The model real losses (l/con/day) shows that the real losses (l/con/day) can be predicted using the log function of the connection density. A structural equations is also made for the NPF, using the real losses (l/con/day), the sector type and an exponential function for the number of connections as independent variables;

- The model of the apparent losses (l/con/day) shows that the apparent losses (l/con/day) can be predicted using the pressure and the log function of the connection density. No structural equation is developed to determine the NPF;

Some remarkable facts are shown in the model results, these are:

- The non-existing relationship between the real losses (l/con/day) and the pressure, since the relationship between the real losses (kl/day) and the pressure is defined by the literature. The cause of this non-existing relationship could be the size of the data or the reliability of the data;
- The relationship between the apparent losses (l/con/day) and the pressure is not defined by literature and thus it is unknown whether the relationship is causal;
- The relationship between NRW (l/con/day) and the pressure is expected resulting from the relationship between the real losses (l/con/day) and the pressure. Since this relationship does not exist, the relationship between NRW (l/con/day) and the pressure is caused by the relationship between the apparent losses (l/con/day) and the pressure, thus it is unknown whether this relationship is causal;
- The results show that there is a relationship between the NPF and the real losses (l/con/day). This relationship is expected, because both variables are used to determine the minimum night flow. This proven and now analyzed relationship gives opportunities for future research and to link OPIR with NRW;

The regression analysis also shows whether the models can be used for predictive purposes:

- The proportion of the variance explaining the performance of NRW or its components using the independent variables is in all cases smaller than 25%, except for the equation predicting the NPF (30,8%);
- The shrinkage levels are all smaller than 25%; two of them are even below the 10%;
- The boundaries, determined using the 95% confident interval, are really broad.

The independent variables do not explain a large part of the variance of the dependent variable and the boundaries are really broad even becoming negative. When the equations resulting from the structural models are used to predict the dependent variables, they will be imprecise. The predicted values will not add any knowledge about the performance of NRW or its components in a new project area.

For future research it is recommended to:

- Identify more variables that influence the dependent variable, this may explain a larger proportion of the variance of the dependent variable;
- Use larger datasets and verify the data that will be used, this may lead to closer boundaries and thus a more precise structural equation. It should be noted that it is difficult to verify water balances;
- It is recommended that the machine learning techniques will also be used, when a large enough dataset is available;
- For future research it is recommended to examine the relationship between the real losses (l/con/day) and the NPF over a time-period in one area. The changes in the NPF value over a time-period can be examined to determine the influence of the real losses (l/con/day).

REFERENCES

Arbués, F., Garcia-Valinas, M. & Martinez-Espineira, R., 2003. Estimation of residential water demand: a state-of-the-art review. *Journal of Socio-Economics,* Volume 32, pp. 81-102.

Blokker, E., 2010. *Stochastic water demand modelling for a better understanding of hydraulics in water distribution networks.* 1st ed. Delft, Netherlands: Water Management Academic Press.

Bougadis, J., Adamowski, K. & Diduch, R., 2005. Short-term municipal water demand forecasting. *Hydrological processes,* Volume 19, pp. 137-148.

Fantozzi, M., Lambert, A. & Liemberger, R., -. *Some examples of european water loss targets, and the law of unintended consequences.* [Online] Available at: http://www.miya-water.com/user_files/Data_and_Research/miyas_experts_articles/15jun2010/Some_Examples_of_European_Water_Loss_Targets_and_the_Law_of_Unintended_Consequences%20With%20Customer%20Water%20Conservation%20Programs.pdf [Accessed 08 02 2013].

Field, A., 2009. *Discovering statistics using spss.* 3rd ed. London: SAGE publications Ltd.

Gonzalez-Gomez, F., Garcia-Rubio, M. A. & Guardiola, J., 2011. Why is Non-revenue water so high in so many cities?. *International Journal of Water Resources Development,* Volume 27:02, pp. 345-360.

Pena, I., 2013. *Collected data* [Interview] (10 01 2013).

RoyalHaskoning DHV, -. *OPIR.* [Online] Available at: http://www.aquasuite.nl/en/opir/ [Accessed 14 1 2013].

StatSoft, -. *Model extremely complex functions, neural networks.* [Online] Available at: http://www.statsoft.com/textbook/neural-networks/ [Accessed 06 02 2013].

Thornton, J., Sturm, R. & Kunkel, G. P., 2008. *Water Loss Control.* 2nd ed. USA: McGraw-Hill.

Trifunovic, N., 2006. *Introduction to urban water distribution.* 1 ed. London, UK: Taylor & Francis Group.

Zhang, X., 2005. *Estimating peaking factors with poisson rectangular pulse model and extreme value theory,* Cincinnati, United States: University of Cincinnati.

ING. C.B. (BERTINE) KOREVAAR

E: c.b.korevaar@student.tue.nl

E: Bertinekorevaar@gmail.com

This research is performed within the context of the KENWIB graduation program, and is the final part of the MSc. Construction Management and Urban Development. Royal HaskoningDHV gave me the opportunity to perform a research according to my interest.

2005-2009	Bachelor Civil Engineering, Hogeschool Utrecht
2007-2008	Internship DHV, Amersfoort
2008-2009	Graduation internship Witteveen & Bos, Deventer
2009-2010	Draftsmen Ballast Nedam Engineering, Nieuwegein
2010-2013	Master Construction Management and Engineering, TU Eindhoven
2011 – Recent	Draftsmen- Engineer Royal HaskoningDHV, Amersfoort
2012-2013	Graduation internship Royal HaskoningDHV, Amersfoort

HET MODELLEREN VAN DE RELATIES MET DE PRESTATIE VAN NON-REVENUE WATER

Met behulp van structurele modellen de prestatie van non-revenue water en zijn componenten linken aan de verkregen variabelen

Auteur: Ing. C.B. (Bertine) Korevaar

Afstudeerprogramma:
Construction Management and Urban Development 2012 – 2013

Afstudeercommissie:
Prof. dr. ir. W.F. Schaefer
Dr. ir. B. Glumac
Ir. B. van Weenen
Ir. M. Riemersma (Royal HaskoningDHV)

Afstudeerdatum:
14 maart 2013

SAMENVATTING

Één van de methodes om drinkwater efficiënter te beheren is het verminderen van non-revenue water in een drinkwatersysteem. Een actieplan voor het verlagen van non-revenue water kan worden gemaakt, wanneer de prestatie van non-revenue water is geanalyseerd. De huidige analysemethode neemt te veel tijd in beslag. Om een eerste stap te zetten naar een efficiëntere analyse is er in samenwerking met Royal HaskoningDHV voorgesteld om de relaties te analyseren die de prestaties van non-revenue water en zijn componenten heeft met andere variabelen. In het model wordt de prestatie van non-revenue water uitgedrukt in liters/aansluiting/dag. De variabelen die worden gebruikt in het model zijn: de aansluitingsdichtheid, de druk, de nacht piekfactor, het aantal aansluitingen en de sector type. De relaties worden weergegeven in structurele modellen, de vergelijkingen voortkomend uit de modellen worden geanalyseerd met behulp van de meervoudige backward elimination regressie analyse. Uit de resultaten komt naar voren dat de lekverliezen (een component van non-revenue water) de nacht piekfactor beïnvloed. Ook blijkt uit de 95% betrouwbaarheidsinterval, de variatie en de krimp in het model dat de structurele vergelijkingen te onnauwkeurig zijn om te worden gebruikt voor voorspellende doeleinden.

Trefwoorden: Prestatie non-revenue water, meervoudige regressie analyse, structurele modellen

De huidige groei in wereldbevolking leidt tot een toename van de vraag naar schoon water. Momenteel zijn er al gebieden waar schoon water schaars is, deze schaarste zal groeien door de huidige groei in de wereldbevolking. Om het tekort niet op te laten lopen zal water efficiënter beheerd moeten worden. Dit efficiëntere beheer kan worden bewerkstelligd door het verminderen van non-revenue water (Gonzalez-Gomez, et al., 2011). Non-revenue water, ook

wel niet-betaald water, is water dat wordt geproduceerd en uitgegeven door een waterbedrijf, maar niet wordt gefactureerd aan klanten. Non-revenue water kan worden opgesplitst in drie componenten, waarvan er twee verantwoordelijk zijn voor bijna het totale volume van het non-revenue water. Het eerste component omhelst de lekverliezen, deze worden veroorzaakt door lekkages in alle delen van de systemen en overstorten op de reservoirs van de waterbedrijven. Het tweede component zijn de commerciële verliezen, deze komen voort uit de invloeden op het bedrijf, die meestal ontstaan buiten het bereik van de dagelijkse processen. Een effectief non-revenue water reductieplan kan worden gemaakt wanneer de prestaties van het non-revenue water en zijn componenten zijn geanalyseerd op basis van metingen in het systeem (Thornton, et al., 2008). Er gaat tijd verloren aan het verzamelen van deze metingen, waardoor het reductieplan pas later dan gewenst kan worden geïmplementeerd in het desbetreffende project. Het huidige proces van het analyseren van de prestatie van non-revenue water en zijn componenten zou efficiënter moeten zijn in de tijd zodat het reductieplan eerder kan worden geïmplementeerd. Om een eerste stap te zetten naar een efficiënter proces is er in samenwerking met Royal HaskoningDHV voorgesteld om de relaties te analyseren tussen de prestatie van non-revenue water en zijn componenten en de variabele die daardoor worden beïnvloed of daar invloed op hebben. Enkele variabelen zullen worden verkregen met behulp van OPIR, dit is een systeem ontwikkeld door Royal HaskoningDHV. Door OPIR te koppelen aan non-revenue water kunnen nieuwe zakelijke kansen worden gecreëerd voor OPIR. De overige variabelen zijn variabelen die kunnen worden verzameld aan het begin van een project

Er zijn verschillende prestatie-indicatoren die kunnen worden gebruikt om de prestaties van non-revenue water uit te drukken. In dit onderzoek zal de prestaties van non-revenue water en zijn componenten worden uitgedrukt in liters/aansluiting/dag. Deze indicator wordt door vele waterbedrijven gebruikt om de prestatie van non-revenue water en zijn componenten te beschrijven. Daarnaast kan deze prestatie-indicator gemakkelijk worden omgebouwd in een gedetailleerdere prestatie-indicator door het te delen door de gemiddelde druk (m wk) in het gebied. De prestatie-indicator liters/aansluiting/dag kan worden gebruikt om de verschillende maand waarden met elkaar te vergelijken, maar enkel als de druk niet teveel fluctueert over de te vergelijken maanden (Thornton, et al., 2008).

De variabele dag piekfactor, nacht piekfactor en druk worden verkregen uit OPIR. De piekfactoren worden berekend door het maximale dagverbruik of minimale nachtverbruik te delen door het gemiddelde dagverbruik. De variabelen type sector, de lengte van het systeem en het aantal aansluitingen kunnen worden verzameld aan het begin van het project, de aansluitingsdichtheid kan worden berekend met behulp van deze variabelen. De literatuur bewijst dat de lekverliezen (l/aansl./dag) worden beïnvloed door de druk en de aansluitingsdichtheid (Thornton, et al., 2008) (Fantozzi, et al., -), aangezien de lekverliezen onderdeel zijn van het non-revenue water wordt er aangenomen dat ook deze wordt beïnvloed door deze variabelen. De DPF en NPF worden beide beïnvloed door de lekverliezen (Thornton, et al., 2008) (Zhang, 2005) , er wordt wederom aangenomen dat zij ook worden beïnvloed door het non-revenue water. De DPF en NPF wordt ook nog beïnvloed door enkele andere variabelen, deze worden in deze analyse tot uitdrukking gebracht door de variabelen: (1) het aantal aansluitingen en (2) het type sector. Er is weinig bekend van de variabelen die worden

beïnvloed of invloed hebben op de commerciële verliezen, daarom is er besloten om dezelfde relaties aan te houden als voor de lekverliezen, er moet echter niet uit het oog verloren worden dat deze relaties niet zijn gedefinieerd in de literatuur.

De relaties worden gemodelleerd in structurele modellen, voortkomend uit deze structurele modellen zijn de structurele formules. Deze structurele formules worden geanalyseerd met behulp van een meervoudige lineaire regressie. De regressie analyse is verkozen boven het machinaal leren (machine learning techniques), omdat de data set kleiner is dan 100 data punten en het machinaal leren dan niet mogelijk is (StatSoft, -). Een backward elimination techniek wordt gebruikt, hierbij worden eerst alle onafhankelijke variabelen geplaatst in de analyse, wanneer een variabele niet statistisch significant is (p>.1) wordt deze verwijderd uit de analyse, dit gebeurt op volgorde van significantie. De regressie analyse kan alleen worden gebruikt wanneer de aannames niet worden geschonden, de aannames zijn (1) geen invloedrijke uitschieters, (2) lineariteit in de dataset, (3) homoscedasticiteit in de dataset, (4) de dataset moet een normaal verdeling hebben, (5) er zijn geen onafhankelijke fouten aanwezig of (6) multicollineariteit. Wanneer dit wel het geval is zal er gebruik worden gemaakt van robuuste technieken (Field, 2009).

De data die gebruikt is in het onderzoek is beschikbaar gesteld door JOAT, een Zuid-Afrikaans bedrijf. De data bestaat uit twee datasets van twee verschillende gebieden, Zuid Durban en West Durban. De DPF data is incompleet voor beide datasets, deze variabele zal niet worden gebruikt in het model. Het sector type is niet gegeven in de dataset van Zuid Durban, maar kan worden verkregen door het gebruik van internet.

De resultaten van de regressie analyse laten het volgende zien:

- Het non-revenue water (l/aansl./dag) model laat zien dat het non-revenue water (l/aansl./dag) kan worden voorspeld met behulp van de druk en de aansluitingsdichtheid. Er is geen structurele formule vastgesteld waarmee de NPF kan worden voorspeld;
- Uit het structurele model van de lekverliezen (l/aansl./dag) kan worden geconcludeerd dat de lekverliezen (l/aansl./dag) kunnen worden voorspeld met behulp van de aansluitingsdichtheid. De NPF kan worden voorspeld met behulp van de lekverliezen (l/aansl./dat), het sector type en het aantal aansluitingen;
- Het model van de commerciële verliezen (l/aansl./dag) laat zien dat de commerciële verliezen (l/aansl./dag) kunnen worden voorspeld met behulp van de druk en de aansluitingsdichtheid. De structurele formule voor de NPF is niet gegeven in dit model.

Er komen enkele opmerkelijke relaties voort uit de regressie analyse, deze zijn kort beschreven:

- Het meest opmerkelijke feit is dat er geen relatie is tussen de druk en de lekverliezen (l/aansl./dag), terwijl er wel een relatie is waargenomen tussen de lekverliezen (kl/dag) en de druk in zowel de literatuur als in de dataset. Gezien deze relatie is er aangenomen dat er een relatie is tussen de lekverliezen (l/aansl./dag) en de druk. Er wordt aanbevolen om te onderzoeken of deze relatie structureel niet aanwezig is en zo ja waardoor er geen relatie is. Een oorzaak van deze niet-bestaande relatie kan de grootte van de dataset zijn of de betrouwbaarheid van de gebruikte data;

- Er is een relatie waargenomen tussen de druk en de commerciële verliezen (l/aansl./dag), deze is niet gefundeerd in de literatuur en er wordt geadviseerd om de causaliteit van de relatie te onderzoeken;
- Er is aangenomen dat er een relatie is tussen het non-revenue water (l/aanls./dag) en de druk, omdat er volgens de literatuur ook een relatie is tussen de lekverliezen (l/aansl./dag). Deze relatie is echter niet aanwezig, waardoor de relatie tussen het non-revenue water (l/aansl./dag) en de druk wordt bepaald door de relatie tussen de commerciële verliezen (l/aansl./dag) en de druk. Aangezien het onbekend is of deze relatie causaal is, is het ook onbekend of de relatie tussen het non-revenue water (l/aansl./dag) en de druk causaal is;
- De relatie tussen de lekverliezen (l/aansl./dag) met de NPF is nog niet eerder bewezen met behulp van een regressie analyse, hij is wel al eerder beschreven in de literatuur. Doordat deze relatie bewezen is, kan er met deze relatie worden gewerkt in toekomstige onderzoeken en wordt het mogelijk OPIR te koppelen aan non-revenue water.

De waarde van de R^2 voortkomend uit de regressie analyse geeft aan hoeveel procent van de variatie in de afhankelijke variabele wordt veroorzaakt door de gebruikte onafhankelijke variabelen. Deze waarde is voor bijna alle formules lager dan 25%, alleen de formule die de NPF voorspelt met behulp van het model van de lekverliezen (l/aansl./dag) heeft een percentage van 31%. De aangepaste R^2 is ook berekend om te bepalen hoe de formules standhouden buiten de gebruikte dataset. Deze waarde laten zien dat het percentage voor alle formules daalt tot onder de 25%, voor twee formules daalt het zelfs onder de 10%. Daarnaast zijn de 95% betrouwbaarheidsintervallen gebruikt om de grenzen te berekenen waar tussen de voorspelde waarde 95% van de tijd ligt. Deze grenzen liggen ver bij elkaar vandaan en de onderste grens krijgt in alle gevallen ook een negatieve waarde. Uit de bovenstaande feiten kan worden opgemaakt dat de formules die voortkomen uit de regressie analyse, niet kunnen worden gebruikt voor voorspellende doeleinden.

Voor toekomstig onderzoek wordt aanbevolen om de volgende zaken te onderzoeken:

- Identificeer meer onafhankelijke variabelen die de afhankelijke variabele beïnvloeden, door de juiste variabelen te identificeren zal de R^2 groter worden;
- Maak gebruik van grotere datasets en als het mogelijk is van geverifieerde datasets. Beide zijn moeilijk om te bemachtigen, maar dit maakt het onderzoek wel betrouwbaarder;
- Wanneer een grotere dataset beschikbaar is, is het aan te raden om de machinale leertechniek te gebruiken;
- Voor toekomstig onderzoek wordt aangeraden om de relatie tussen de lekverliezen (l/aansl./dag) en de NPF over een tijdsperiode te analyseren. Waarbij er wordt geanalyseerd wat voor invloed de variatie in lekverliezen (l/aansl./dag) heeft op de variatie in de NPF.

LITERATUUR

Fantozzi, M., Lambert, A. & Liemberger, R., -. *Some examples of european water loss targets, and the law of unintended consequences.* [Online] Available at: http://www.miya-water.com/user_files/Data_and_Research/miyas_experts_articles/15jun2010/Some_Examples_of_European_Water_Loss_Targets_and_the_Law_of_Unintended_Consequences%20With%20Customer%20Water%20Conservation%20Programs.pdf [Accessed 08 02 2013].

Field, A., 2009. *Discovering statistics using spss.* 3rd ed. London: SAGE publications Ltd.

Gonzalez-Gomez, F., Garcia-Rubio, M. A. & Guardiola, J., 2011. Why is Non-revenue water so high in so many cities?. *International Journal of Water Resources Development,* Volume 27:02, pp. 345-360.

StatSoft, -. *Model extremely complex functions, neural networks.* [Online] Available at: http://www.statsoft.com/textbook/neural-networks/ [Accessed 06 02 2013].

Thornton, J., Sturm, R. & Kunkel, G. P., 2008. *Water Loss Control.* 2nd ed. USA: McGraw-Hill.

Zhang, X., 2005. *Estimating peaking factors with poisson rectangular pulse model and extreme value theory,* Cincinnati, United States: University of Cincinnati.