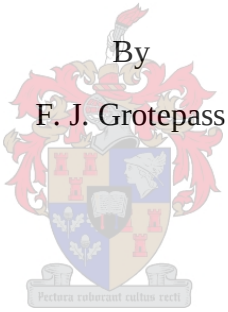
A study to determine the relationship between the maximum capacity of an urban water reticulation network and its physical characteristics

By

F. J. Grotepass



Supervisor: C. Loubser

Faculty of Engineering

Department of Civil Engineering

Division of Water and Environmental Engineering

December 2019

Stellenbosch University https://scholar.sun.ac.za

"Concern for man and his fate must always form the chief interest of all technical endeavours. Never forget this in the midst of your diagrams and equations."

Albert Einstein
(1879-1955)

Stellenbosch University https://scholar.sun.ac.za

Plagiarism declaration

Surname and initials: ……..

Student number: ……..

Module: …….

1. I hereby declare that I know what plagiarism entails, namely to use another's work and to present it as my own without attributing the information to the correct source.
2. I know that plagiarism is a punishable offence because it implies theft.
3. I know that plagiarism is harmful for the academic environment and that it has a negative impact on any profession.
4. I am aware that a lecturer can give me a zero for any of my assignments for my module if I should plagiarise. If I do declare a dispute about it, the normal disciplinary procedures will apply.
5. I declare that all the information that I present in my assignments will be my own. In cases where I use somebody else's work, I will refer to the source in an acceptable manner.

---------- ----------

Signature Date

(Used with permission from the Department of Afrikaans and Dutch)

Copyright © 2019 Stellenbosch University
All rights reserved

Stellenbosch University https://scholar.sun.ac.za

Abstract

Water distribution systems around the world typically comprise a large number of pipelines designed to distribute pressurised water. These pipelines stretch over long distances and vary in diameter, material, wall thickness, internal roughness and age. By nature, these systems are extremely complex and require expert knowledge and specialised tools to be modelled and designed correctly. Furthermore, these networks encompass multiple parameters such as available supply pressures, internal roughness coefficients, frictional losses, annual average daily demands and instantaneous peak hour factors. In order to efficiently manage all of these variables, engineers require a substantial volume of data and sophisticated computer modelling software.

Considering this, the following research question was identified: “Can one develop an urban network capacity model by considering only the network’s physical characteristics?” Or more simply, if sufficient knowledge about a reticulation network’s physical parameters are known, can these parameters be used to model certain other parameters associated with a water network?

Therefore, the aim of this project was to develop a “network capacity model” by analysing the relevant physical parameters of many existing water reticulation network models. The parameters that were identified and could potentially impact the overall supply zone capacity include: total pipeline length, total pipeline volume, average pressure, supply zone topology, supply zone shape, supply zone area, land use and distance from supply position to the centroid of the supply zone. Three linear regression approaches, namely Multi Linear Regression, Principal Component Analysis and Partial Least Squares were used to test this relationship and determine the most accurate model.

Stellenbosch University https://scholar.sun.ac.za

The model that was generated following application of these analyses, presents significant advantages to engineers, and enable options that were never before possible. If the future water demand of an area is known or can be estimated, the model could be used to reverse engineer a list of the required pipe diameters and associated pipe lengths that could meet this demand. For the first time, it now becomes possible to provide a fairly accurate water network cost estimate for future development areas, without the availability of a street layout.

This model also holds the potential to be implemented in developing countries where the necessary skills or resources are not always available to compile computerised models of water distribution networks. In these developing areas, a manual model with simple input parameters can be a reliable and useful tool to manage and plan for expanding water networks.

Furthermore, it has the potential for application in the field of asset management to provide a breakdown of the various pipe diameters and their respective pipe lengths (for purposes of establishing a technical asset register). In this sense it could be used in areas where water networks exist, but where the water network drawings or detailed water network models may not be available. In these instances, the model may be used to provide an estimate of the pipelines and overall replacement cost of the water reticulation network.

Afrikaans translation

Waterverspreidingsnetwerke wêreldwyd bestaan tipies uit ‘n groot aantal pyplyne wat ontwerp word om water onder druk te vervoer. Hierdie pyplyne strek oor lang afstande, en varieer in diameter, materiaal, wanddikte, interne ruheid en ouderdom. Gegewe die inherente komplekse aard van hierdie netwerke, word spesialiskennis en –sagteware benodig vir die modellering en ontwerp daarvan. Parameters soos beskikbare verspreidingsdruk, interne ruheidskoëffisiënte,

Stellenbosch University https://scholar.sun.ac.za

wrywingsverliese, gemiddelde jaarlikse waterverbruik en piekfaktore moet tipies in ag geneem word. Om die impak van al hierdie veranderlikes korrek te bestuur, benodig die ingenieur toegang tot ‘n groot hoeveelheid data en gevorderde rekenaarpakkette.

Gegewe hierdie uitdagings, is die volgende navorsingsvraag geïdentifiseer: “Kan ‘n netwerkkapasiteitsmodel ontwikkel word deur bloot die fisiese eienskappe van die netwerk in ag te neem?” Of meer simplisties, indien genoegsame inligting bestaan rakende die fisiese eienskappe van ‘n gegewe waternetwerk, kan hierdie eienskappe gebruik word om sekere ander, onbekende eienskappe te bepaal?

Die mikpunt van hierdie studie is dus om ‘n stedelike netwerkkapasiteitsmodel te ontwikkel deur ontleding van die toepaslike veranderlikes van ‘n groot aantal bestaande waternetwerke. Die parameters wat geïdentifiseer is wat moontlik die kapasiteit van ‘n waternetwerk kan beïnvloed sluit in totale pyplynlengte, totale pyplynvolume, gemiddelde druk, topologie van die verspreidingsarea, vorm van die verspreidingsarea, oppervlakte van die verspreidingsarea, grondgebruik en die afstand van die punt van lewering tot by die sentroïde van die verspreidingsarea. Drie verskillende analises is toegepas, naamlik multi-liniêre regressie, hoofkomponent regressie en gedeeltelike minste vierkante regressie. Die oogmerk was om moontlike verwantskappe te identifiseer en die mees akkurate model te ontwikkel.

Die model wat op sodanige wyse ontwikkel was, hou groot voordele vir ingenieurs in, en ontsluit moontlikhede wat tot op hede nie beskikbaar was nie. Indien die toekomstige waterverbruik van ‘n area beskikbaar is of beraam kan word, kan die gebruiker die totale benodigde pyplengte en verwante diameters met ‘n groot mate van sekerheid voorspel. Dit word sodoende vir die eerste keer moontlik om ‘n redelike akkurate watermeesterplan vir toekomstige uitbreidings te ontwikkel,

Stellenbosch University https://scholar.sun.ac.za

selfs in omstandighede waar die toekomstige straatuitleg nie beskikbaar is nie. Dit bemoontlik ook kosteberamings en verwagte konstruksietyd vir die ontwikkeling van waternetwerke vir hierdie areas.

Die model het ook moontlike toepassing in onwikkelde lande, waar die nodige kundigheid en hulpbronne wat benodig word om waternetwerk rekenaarmodelle te ontwikkel dikwels ontbreek. In hierdie omstandighede vergemaklik die eenvoud van die toepassing van die model die bestuur en beplanning van die uitbreiding van waternetwerke.

Die model kan ook gebruik word vir batebestuur, veral in areas waar waternetwerke bestaan, maar die inligting of netwerkmodelle ontbreek. In hierdie omstandighede kan die model gebruik word om met redelike sekerheid te bepaal hoe die netwerk waarskynlik daaruit sien en wat die totale batewaarde van sodanige waternetwerk behoort te wees.

Stellenbosch University https://scholar.sun.ac.za

Acknowledgements

A special thank you to my supervisor and mentor Mr Carlo Loubser for guiding me and providing support throughout my thesis. Also, it must be acknowledged that the principal ideas and concepts surrounding the thesis were his idea.

Dr Erik Loubser and Mr Andre-Hugo Van Zyl alongside the entire GLS team for providing the water distribution systems data and continuous support when working in Wadiso. Data from four sources were used, including George Municipality, City of Tshwane Metropolitan Municipality, Ekurhuleni Metropolitan Municipality and Mbombela Municipality. The data was used anonymously protecting the intellectual property of each source.

My parents for the financial support to enrol for a Masters at Stellenbosch University. And most importantly for the emotional aid when work efforts resulted in meaningless results or simply incorrect conclusions.

Stellenbosch University https://scholar.sun.ac.za

Table of contents

Stellenbosch University https://scholar.sun.ac.za

Stellenbosch University https://scholar.sun.ac.za

Stellenbosch University https://scholar.sun.ac.za

Stellenbosch University https://scholar.sun.ac.za

List of Figures

Stellenbosch University https://scholar.sun.ac.za

Stellenbosch University https://scholar.sun.ac.za

List of Tables

Stellenbosch University https://scholar.sun.ac.za

Stellenbosch University https://scholar.sun.ac.za

List of Abbreviations

AADD	Annual Average Daily Demand
AM	Area Meters
DCM	Domestic Consumption Monitors
DDA	Demand-Driven Analysis
DFS	Demand Forecasting System
EPS	Extended Period Simulation
HDA	Head-Driven Analysis
IWS	Intermittent Water Supply
LCH	Low-Cost Housing
MLR	Multi Linear Regression
MPH	Mean Pressure Head
PCA	Principal Component Analysis
PLS	Partial Least Squares
TFA	Transient Flow Analysis
UAW	Unaccounted For Water
VIF	Variance Inflation Factor
WDS	Water Distribution System

Stellenbosch University https://scholar.sun.ac.za

SECTION 1

INTRODUCTION

1.1. Water Distribution Systems

Savic describes a Water Distribution System (WDS) as a crucial part of any urban area's infrastructure (Savic and Banyard, 2011). These systems consist of various connections, valves, pipes, and storage facilities transporting potable water to consumers. This potable water also has to meet the necessary pressure ratings and water quality standards to be safe and convenient for users (Savic and Banyard, 2011).

To ensure smooth operation, these systems need to be analysed using simulation programmes such as EPANET (Loubser, 2018). The software is used to determine the different pressures and flow rates at predefined points in the system, namely nodes and links respectively. As the network is expanded with more nodes and links, the simulations become larger and more complex. These simulation programmes also enable the user to determine the total capacity of the network.

Loubser stated that one of the limitations of these simulations is that they are often restricted to larger cities where parameters such as pipe diameters are easily available from municipal records and as-built drawings (Loubser, 2018). Often, in smaller urban areas, available data pertaining to the total capacity and exact pipe diameters of the network may be limited.

Furthermore, when new developments are planned or undertaken, a broad outline of the WDS is often based on the nature of the development and the surrounding area. This is challenging in many future developments, as often the planning has not yet progressed to a level where other infrastructure components, such as potential street layouts, are available.

Stellenbosch University https://scholar.sun.ac.za

1.2. Problem Statement

Water reticulation network models are used in many of the major cities in South Africa. These models provide a virtual representation of the existing WDSs. With knowledge of the existing infrastructure, modellers can add various pipelines, reservoirs, pumps and valves to the simulation. Furthermore, with user demands known, pipe pressures and flow rates can be predicted at various points in a WDS. As an urban area expands via new developments, the network simulation model can be expanded with new pipelines and appropriate infrastructure components. With the model setup completed, an analysis can be performed and the total network capacity, among other parameters, can be determined. It can however be, that for a certain urban development, there exists no water reticulation information. Or in other instances, there may be information, but the information is either incomplete, or a physical network model has not been compiled or is not available.

Furthermore, in areas with no existing infrastructure, like street layouts, planning of the total pipeline length and pipeline layout becomes challenging. Engineers then typically rely on experience to design a broad outline of the network. This often involves a form of educated guesswork, that in most cases do not yield reliable outcomes.

A hypothesis was thus developed, which states that if the physical parameters of a large number of water reticulation networks could be analysed, it may be possible to create a link (or several links) between the physical network characteristics and the maximum capacity of these networks. If these links could be established, it may become possible to generate these physical network parameters if the required maximum capacity of a future network could be established.

Stellenbosch University https://scholar.sun.ac.za

1.3. Objectives

This study aims to analyse various reticulation network models of individual supply zones within cities across South Africa. This would aid in deriving a model which can accurately estimate the capacity of a reticulation network based on the physical characteristics of the network. The hydraulic models analysed were provided by GLS (GLS Consulting, 2018), an engineering company based in Stellenbosch. GLS has a large database of the WDSs for various types of supply zones within South Africa. The WDSs were analysed using the Wadiso software. With this large pool of water network data, it was envisaged that comparisons of multiple networks and their associated capacities could yield a viable outcome.

In addition, it was envisaged that this model could be used in reverse, to provide a reasonable outline of the physical WDS parameters, based on the demand of the area or alternatively, the number of capita. The objectives of the study are:

- Conduct a thorough literature review on how WDSs function and are modelled;
- Investigate various appropriate statistical methods that can be used to develop a model;
- Analyse multiple WDS network models from South Africa, as provided by GLS;
- Identify physical network characteristics that could potentially influence the capacity of a WDS;
- Statistically analyse the different WDSs and the network characteristics, in order to find a model which can predict a reasonably accurate capacity for each WDS;
- Test the newly derived mathematical model, in order to determine whether it can accurately predict the capacity of other WDS models that have been deliberately kept aside for testing purposes.

Stellenbosch University https://scholar.sun.ac.za

1.4. Scope and Limitations

The study was conducted using data from urban models that were analysed using demand-driven analyses in Wadiso. Demand-driven analyses are when network demands influence the pressures within the network. Reticulation network models that were analysed using demand-driven analyses are common in South Africa, and data was made available for the study by GLS. WDS capacities computed from other software packages or models developed by other consultants for the same study areas were not available, which if available, could lead to slightly different outcomes.

1.5. Structure of Report

The report is structured as follows:

- Chapter 2: Literature review describing water distribution systems;
- Chapter 3: Literature review analysing and investigating different linear regression approaches;
- Chapter 4: Methodology of how the regression models are applied to the WDS data;
- Chapter 5: Comparing the linear regression models derived from the transformed WDS models;
- Chapter 6: Comparing the linear regression models derived from the untransformed WDS models;
- Chapter 7: Selecting the most accurate and precise model that best represents real-world water reticulation networks;
- Chapter 8: Concluding comments about the regression model and its performance;
- Chapter 9: Recommendations to improve the usability of the model and potential future studies.

Stellenbosch University https://scholar.sun.ac.za

SECTION 2

WATER DISTRIBUTION SYSTEMS

2.1. Water Distribution Systems: History

When one consider WDSs, these seem to be very simple networks. The general field of water reticulation design has even been criticised to have lacked ingenuity and development in recent years as network design tends to stay relatively similar from year to year. This is however not strictly true. Savic explains that modern WDSs are not much older than 100 years and that the first dedicated use of purification plants was only introduced at the turn of the 20 century (Savic and Banyard, 2011). To put that into perspective, WDSs as we know these today, were introduced at approximately the same time as when Titanic was built. The world has only had safe WDSs as long as there have been automobiles, cruise liners and skyscrapers, which are some of the marvels that represent the modern advancements of mankind.

The first known form of a piped water system was during the Roman Empire in 100 AD. These systems were the brainchild of Frontinus (Savic and Banyard, 2011). Frontinus made use of both surface and underground aqueducts to transport water around Rome. This water was transported over long distances to supply water features and fountains around the city. After the fall of the Roman Empire, these networks fell into poor shape and were no longer functional. As time passed, no other European countries adopted these water networks and relied rather on water sourced from rivers and other sources.

Without WDSs that could transport water around cities, city inhabitants adopted a lifestyle of uncleanliness and a lack of hygiene. Events such as the black plague in London in the 1930s

Stellenbosch University https://scholar.sun.ac.za

emphasises this, as people lived in cramped quarters with no potable water to clean themselves and their environment (Savic and Banyard, 2011). This trend continued for multiple years before Dr Snow made the link between cholera and dirty water in 1855. Once the public became aware of the importance of sanitation and clean water, WDSs started developing rapidly. By 1895, Fuller had developed filtration with coagulation to filter water and by 1900, chlorine was introduced to kill off the remaining bacteria in the water (Savic and Banyard, 2011). Since these methods were introduced, water networks have evolved and been optimised to provide users with safer cleaner drinking water.

2.2. Outline of system

2.2.1 Overview

A WDS is the complete system responsible for transporting water from a source to the user. There are three main stages or phases to acquire, clean and distribute the water. The first phase is acquiring a source for the water. Surface water sources include rivers, dams and lakes. Aquifers and the use thereof via boreholes and other means are typical examples of subsurface sources (Lee, 2017).

After the water has been extracted from a source, it is transported to a treatment plant, typically via a pipeline making use of a pumping station. At the water treatment plant, the water is filtered and purified to meet quality measures (Lee, 2017).

Once the water has been filtered, it is ready to be distributed to users. This final phase has multiple factors which need consideration, including pipe diameters, storage volume, storage elevation and placement of fittings such as valves. As this section of the WDS falls within the scope of this study, a more detailed discussion of this portion of a WDS will be included.

Stellenbosch University https://scholar.sun.ac.za

2.2.2 Distribution network pressures

Pressure drives water distribution, and therefore a clear understanding of how the water is distributed to users is necessary. When engineers design WDSs, storage reservoirs located at higher elevations than the area being supplied are constructed. This, in turn, creates a static head or pressure within the network. Furthermore, system pressures in the pipelines need to be considered to ensure that the pressures are between the maximum and minimum range to meet pipe design specifications and acceptable consumer as well as fire flow design requirements respectively. If the pressures within the pipelines are above the maximum specified pressure, problems such as pipe bursts and leaks can occur.

Conversely, negative pressures could cause pipeline collapse or contaminants being drawn from outside into a pipe. These conditions often occur in Intermittent Water Supply (IWS) systems, when supply to an area is turned off, while consumers are still drawing water from the system (Ghorbanian et al., 2015).

Points or nodes located at higher elevations especially when far from the supply points tend to have lower pressures. The pressure at these critical points are often the governing minimum pressures of a WDS (Jacobs and Strijdom, 2009). The Minimum Pressure Head (MPH) is calculated at these points for their respective maximum demands. From this, the node with the lowest pressure can be identified and used as the basis for the minimum pressure the system can handle (Jacobs and Strijdom, 2009). In South Africa, a minimum peak pressure head of 20 m and a recommended maximum static pressure head of 60 m is prescribed by *The Neighbourhood Planning and Design Guide* (Department of Human Settlements, 2019). The maximum pressure occurs when the demand is at its lowest when few users are abstracting water from the network. This generally occurs after midnight and before dawn, when most consumers are asleep.

Stellenbosch University https://scholar.sun.ac.za

2.2.3 Negative pressures in network

Negative pressures typically occur under conditions of intermittent water supply, but can also occur in a normal continuous WDS. This occurs when pumps are turned off, valves are closed and if a point's demand is greater than the system's capacity to meet that demand (Zeng et al., 2016).

When the pressure within the network becomes negative or falls below atmospheric pressure, it induces additional stresses on the pipes and network components. When pipes crack, the cracks provide a means for intrusion of contaminants when negative pressures occur (Erickson et al., 2017).

If the pressure inside the network is below the vapour pressure of the water, water hammer will occur (Wang et al., 2014). More simply, the negative pressure in the system results in water vapour and air in the pipes. This newly induced pressure within the pipeline then starts to oscillate within the pipes, thus causing water hammer. The addition of air valves can reduce this risk, as the air that is entrapped in the pipe can be released (Wang et al., 2014).

Air within the WDS can also rise to higher locations within the network. If an air valve is not present, this can result in negative pressures at this location. This additional air at these locations can also cause water meters to return inaccurate readings, or even lead to accelerated wear.

When new WDSs are designed and modelled, negative pressures are often found at various points around the network. These negative pressures could be a sign that the network is simply not able to meet the necessary demand users place on the system. Potential solutions to solve this problem include:

- Increase the reservoir elevation to raise the pressure within the WDS network;
- Select alternative pipes with smoother wall linings with a lower friction coefficient;

Stellenbosch University https://scholar.sun.ac.za

- Increase pipe diameters to decrease head losses;
- Install additional pumps to increase pressure for water being supplied to high lying areas.

2.2.4 Water Distribution System hydraulics

WDSs are typically gravity-fed and require positive pressure within the pipelines to ensure user demands can be supplied. Positive pressures are generated by locating reservoirs that supply the network on a hill above the points of demand. This difference in elevation, principally ensures that there is a positive pressure in the network (Rathore, 2015).

The maximum pressure as described in Section 2.2.2 is the pressure in the pipe network when the demand is at its lowest, typically during the night. If the user demand is 0 and there are no leaks, static pressure is present in the system. Static pressure then simply becomes the difference in elevation between the water level of the reservoir and the water level within a specific point in the pipeline (Savic and Banyard, 2011).

As soon as there is a leak or someone opens a tap, there is a reduction in pressure in the pipe network. This happens as pressure is required to ‘push’ or force the water from the pipe to the point of extraction. Pressure also drops in the network due to frictional forces experienced by water flowing in the pipes. Therefore, pressure decreases at nodes further from the source, because the water travels further along the pipe network and thus experiences greater friction (Chadwick et al., 2013).

In summary, as the demand increases during the morning and evening hours when users shower, cook and wash clothing, the pressure in the system decreases. This decrease in pressure is generally caused by frictional losses and to a lesser extent by transitional losses.

Stellenbosch University https://scholar.sun.ac.za

2.2.5 Optimisation of Water Distribution Systems

Forecasts of future water demands can be predicted based on past records. With this information known, management authorities can plan for the long-term management and short-term optimisation of WDSs (Romano and Kapelan, 2014).

This technology is currently limited worldwide, but has been tested and implemented in major cities such as London. The Demand Forecasting System (DFS) uses nonlinear models of past records to predict WDSs demand in the future. Three main time frames are considered, namely long term, medium term and short term. The long-term forecasting helps with the design of the WDS and possible expansions which might be necessary for the future. The short-term forecasting is important as it deals with hours and days as opposed to years, which enables management to optimise the system and ensure smooth daily operation (Romano and Kapelan, 2014). This includes monitoring the pumps to ensure optimal pressures and flow rates within the network.

2.2.6 Pipe materials used

In 1850, cast iron pipes were commercially used in water distribution networks for the first time (Savic and Banyard, 2011). These pipes were not common and no standard sizes existed. Asbestos cement pipes were also used as an alternative to cast iron. These two materials were the only pipe materials available until the introduction of ductile iron pipes in the 1970s (Mora-Rodriguez et al., 2014). Asbestos cement pipes were discontinued due to their safety hazards during manufacturing. Currently, the most widely used pipe materials are PVC, HDPE, ductile iron and concrete (Mora-Rodriguez et al., 2014).

PVC is the most widely used pipe material today. PVC can be easily manipulated as it is a thermoplastic and its mechanical properties make it ideal to withstand water pressures.

Stellenbosch University https://scholar.sun.ac.za

Furthermore, it is nontoxic, odourless and chemically inert (Mora-Rodriguez et al., 2014). Also, as PVC pipes have a smooth lining, frictional losses within the pipe can be limited.

HDPE has gained ground in recent years and has become a fairly widely used pipe material in WDS networks. It is extremely good at withstanding corrosion due to chemicals both inside and outside of the pipe. Furthermore, it has a low friction coefficient, which helps reduce head losses in the WDS. Lastly, due to its low modulus of elasticity, it can easily be manipulated when on site to bend around turns and fit in smaller spaces (Mora-Rodriguez et al., 2014).

Ductile iron pipes are regarded as the most robust pipes. This makes them ideal when handling on site, as pipe damage during installation is limited (Robor, 2018). Ductile iron pipes can however corrode. They are coated with zinc or bitumen to reduce the rate of oxidation (Savic and Banyard, 2011). If well protected against rust, ductile iron pipes have a long life span and can be economical (Robor, 2018).

There are two types of concrete pipes often used in WDSs, pre-stressed steel reinforced concrete pipes and post-tensioned concrete pipes. The post-tensioned concrete pipes are generally stronger and can handle higher pressures than the pre-stressed pipes. This is also partly due to the deterioration and corrosion rate of steel in the reinforced pipes (Mora-Rodriguez et al., 2014). The high pH levels of soils also play a significant role in the corrosion of concrete pipes, as the pH of the pipes and the mortar lining within the pipes starts to decline till a point of corrosion (Mora-Rodriguez et al., 2014).

Pipe bursts are caused by two main factors, namely hydraulic and non-hydraulic factors (Wang et al., 2014). In these two factors, there are four main types of stresses, namely: longitudinal and circumferential tensile stresses as well as circumferential bending and socket cracking stresses

Stellenbosch University https://scholar.sun.ac.za

(Wang et al., 2014). The different causes for each tensile stress were researched by Mora-Rodriguez in the *Urban Water Journal* (Mora-Rodriguez et al., 2014).

Other hydraulic factors include effects such as water hammer which could result in severe cracks and pipe bursts due to the large forces placed on sections of the network. Corrosion is a major non-hydraulic factor to consider. Groundwater quality, as well as the voltage of power lines above the pipes, increase the corrosion rate of metal pipes. As the pipes corrode, their capacity to resist stresses also decreases, resulting in cracks and pipe bursts (Wang et al., 2014).

2.3. Water Distribution System Design

When designing a WDS, there are several requirements that need to be considered from the planning phase. This is essential to ensure that the system will function optimally and have sufficient capacity to cater for periods of peak water demand.

Savic proposes an outline of design requirements (Savic and Banyard, 2011). These include the adequacy, serviceability and efficiency of the WDS. Adequacy is associated with the quality, pressure and flow of the WDS. Serviceability describes how well each component of the WDS is managed and how users are affected by the management of these assets. Lastly, efficiency monitors the performance of the system and how optimally each component is functioning.

Furthermore, allowance must be made during design for fire protection and emergency supply situations. In the early stages of development, information regarding user consumption, population growth trends and changes in the topology are all considered.

The level to which these requirements are enforced and managed, also influence the level of service of the designed system. Therefore, these requirements are often used as a benchmark to ensure the WDS is of sufficient standard (Department of Human Settlements, 2019).

Stellenbosch University https://scholar.sun.ac.za

2.3.1 Designing an optimal Water Distribution System

When designing a WDS, no definite method exists to attain the perfect system with exact pipe dimensions and pump types (Savic and Banyard, 2011). Traditionally a trial and error approach is used, whereby a network solver considers multiple scenarios until an optimum situation is found.

When designing for a specific scenario, a problem is set up that needs to be solved. These problems are essentially mathematical functions with various constraints, independent variables and objective functions that all need to be satisfied. The constraints are the system conditions that need to be met to meet the required design criteria. These include, for example, the minimum nodal pressures and maximum flow velocities. The independent or design variables are the variables that the designer can change and manipulate until the constraints are satisfied. These typically include, for example, the diameter of the pipes used. The objective function is the variable that must be maximised or minimised. This can include the cost of the project or water quality requirements.

In optimising WDSs in the past, network modellers have typically adopted a single objective optimisation approach. In this approach, the objective function was primarily associated with the reduction of capital costs, with the design variables being the pipe diameters and pipe lengths. This approach is however insufficient for the design of WDSs when there is a change in the population and number of users. To solve this, a multi-objective approach has recently been adopted, which considers multiple objectives with different weightings (Savic and Banyard, 2011). Therefore, not only cost is considered, but objectives such as water quality and sustainability as well, with each of these variables been giving a certain weighting of importance.

When designing a WDS, it is no longer sufficient to purely adopt a lowest cost approach. A WDS needs to be designed to cope and adapt to the changes in the needs and conditions imposed on the

Stellenbosch University https://scholar.sun.ac.za

network. These changes can include population growth, urbanisation, age and functioning of infrastructure and variable climate conditions. These parameters should be included in the design stages of a WDS, to avoid the network being insufficient or strained in the future.

This design approach can also be seen as designing for uncertainty. When designers are not sure to what extent the population will grow, or whether urbanisation will happen at the predicted rate, then this design approach should cater for that. It is thus essential that designers give sufficient attention to the predictive models and statistics surrounding the prediction of the changes in these variables (Savic and Banyard, 2011).

2.3.2 Flexible design approach

A flexibly designed WDS can be seen as a system where the requirements can be changed without drastically increasing the complexity of the system (Savic and Banyard, 2011). Thus, a flexible WDS can still meet the necessary user demands by making simple alterations to the system's infrastructure. This is ideal for cities or water distribution zones where changes are constantly occurring. A well-designed system can continue to meet the necessary requirements imposed on the network with small pipe, pump and valve changes. This helps to prolong the useful and functional life of the network.

A four-step approach to flexible design of water networks is proposed by Savic (Savic and Banyard, 2011).

- Describe the uncertainty of the network and how to limit it. This is usually done by analysing past records and formulating predictive algorithms to present more accurate estimations of future values;

Stellenbosch University https://scholar.sun.ac.za

- Identify and decide which components in the network should be altered to give the safest or optimal change in the network;
- Assign values to each option, making each option a quantitative choice with regard to benefits and difficulties;
- Lastly, the best network option should be picked and thoroughly tested before implementation.

2.4. Estimating and forecasting water demand

Water is crucial for all living things to survive. People need water not only to drink, but for washing, working and personal health. Furthermore, migration of people towards cities has increased drastically in recent years (Savic and Banyard, 2011). As groups of people become more densely populated, sufficient water supply is necessary to meet the increased demand. Local sources are often insufficient and other sources further away need to be found and developed. Thus, accurate estimations of demands are necessary to ensure that adequate supply can be sourced to meet the consumers' needs.

When estimating the demand for water, one can assume that the demand will vary constantly (Savic and Banyard, 2011). Trends can however be found by considering historic data to find how and when the demands change. Also, these trends do not only vary hourly, but seasonally. This continuously varying demand is critical in ensuring that sufficient supply is available at normal times as well as peak periods (Savic and Banyard, 2011).

2.4.1 Estimating demand

Household water meters are typically used to measure the water demands of end-users. When estimating the demand for unmeasured households, two approaches are typically used. These

Stellenbosch University https://scholar.sun.ac.za

include Domestic Consumption Monitors (DCM) and Area Meters (AM) (Savic and Banyard, 2011).

The DCM method makes use of a sample of households to estimate the water demand for an area. The DCM method is not the preferred method to use, as it has many requirements. This includes the need to have a sample group from each income group to accurately determine the variability in results from different areas. Another requirement is a sufficient number of households for forecasting.

The AM method is used for estimating demands for large areas. The method analyses the night time demand which occurs in the early hours of the morning when demands are low and fewer unforeseen events occur. Legitimate night time industrial water use and leaks are then considered to calculate the residential water demand (Savic and Banyard, 2011). The AM method then considers census counts for the area analysed to determine the demand per capita. The method considers land use to accurately divide the water demand between different income and social groups. For example, if a university is present, the AM method will assume most residents surrounding the university are students (Savic and Banyard, 2011).

2.4.2 Forecasting demand

When forecasting demand, commercial and domestic demands are calculated differently. Commercial demands are forecasted in two ways, in an empirical statistic and process deterministic way. The empirical statistical method uses historical data to forecast demand. Common trends and correlations are found from the historical data to predict future demands. The process deterministic method is more complicated as it considers either inputs or outputs to determine the relations between them. The method relies on disaggregation, as it breaks each

Stellenbosch University https://scholar.sun.ac.za

component into smaller parts. For example, it considers the number of employees and water demand to find the water demand per person for commercial areas (Savic and Banyard, 2011).

When forecasting domestic demands, the demand is linked to the population. With a known household consumption from water meters, the average number of residents per household is necessary to determine the demand per capita. The number of residents per household is attained from population forecasts considering migration of people and other analytical studies to determine the demand for different households. These studies consider individual households and the number of residents in each household to find a per capita demand value. The houses are typically sample houses in a specific group which give insight into the area as a whole. Factors such as income and lifestyle habits are recoded to further divide the variances in demand for different household types (Savic and Banyard, 2011).

2.5. Project appraisal for Water Distribution Systems

Engineering projects need to be appraised to assess whether they are worth pursuing. Or more simply, a project needs to be assessed to check if its outcome is worth the resources necessary to produce it (Savic and Banyard, 2011).

Projects need to be evaluated following accepted methodologies to ensure that its outcomes can be validated and are sound. For engineering projects, both economic and technical appraisals are often performed. These point out any risks or unknowns which a project might have. Once an appraisal is completed for each project alternative, the most favourable option can be selected.

When conducting an economic appraisal of a project, the project's economic impacts are assessed. These impacts are not just the projects own income and expenses, but what effect the project has on the local and national economy. These impacts also branch further to fields such as society and

Stellenbosch University https://scholar.sun.ac.za

the environment. A civil water project often affects an entire community; this appraisal type is necessary to see the greater effects of a project on persons not directly involved in its construction.

There are several appraisal types used to assess civil projects. An outline of each follows.

2.5.1 Least-cost analysis

The least-cost analysis considers all the different project alternatives to find the one with the smallest capital cost. When using this analysis approach, a predefined output is set and each alternative is required to meet this standard. From this, each option is normalised.

2.5.2 Cost-benefit analysis

The cost-benefit analysis establishes a quantitative value to benefits. From this, it is possible to assess whether the extra financial outlay for different alternative projects can be recovered in their benefits. These benefits can incorporate both social and environmental benefits which ordinarily don't hold any monetary advantages.

2.5.3 Cost-effectiveness analysis

The cost-effectiveness analysis considers each alternative and how they achieve the specified end result. From this, the alternative that is the most cost-effective for its specific outcome is selected.

2.5.4 Life cycle analysis

The life cycle analysis considers the full life of a project, from its beginning stages to its eventual end. It considers all the goods and services used to produce the product or project and the resources used to manage and operate it.

Stellenbosch University https://scholar.sun.ac.za

2.5.5 Whole life costing

Whole life costing is different from the life cycle analysis in that it does not see a project as a one-time event, but rather sees a project as an ongoing event spanning over several years. The reason for this is that often projects are expanded at later stages or that sections of the necessary infrastructure already exist. Because of this, the appraisal tool is often used in water engineering when networks are constantly upgraded and expanded.

2.6. Network modelling

WDSs are highly complex schemes. Its complexity is partly due to the size of the network and all the different variables which are dependent on one another. For example, the different operating conditions of pumps and valves could have major effects on the pressures and flows within the network. Furthermore, as these pressures and flows change, the hydraulic forces within the pipe network change (Rathore, 2015).

It is because of this complexity, that WDS models become crucial tools in understanding how the system behaves. To fully understand the system's behaviour, the different pressures and flows at different locations in the network can be simulated. With this knowledge, informed decisions can be taken to both improve and optimise the network's performance (Nyende-Byakika et al., 2012).

Before modelling the network, a full understanding of the network and all the various components is necessary. These components typically include the layout of the pipes along with each pipe's characteristics, the elevations of specific points along the network and the position of the various valves, pumps and reservoirs (Nyende-Byakika et al., 2012).

When modelling a WDS, there are key inputs to describe the network itself and the pipes that form the network. To model a network, it must be schematized, built and calibrated to represent the

Stellenbosch University https://scholar.sun.ac.za

prototype. For the pipes that form the basis of the network, various hydraulic inputs are necessary. These include the start and end nodes, the pipe diameters, pipe lengths and roughness coefficients. The model outputs include the flow velocities, flow rates and head losses within each section of a pipe (Nyende-Byakika et al., 2012).

Models are a vital tool to plan future expansions of a network, as well as which existing components must be upgraded to meet future demand requirements. Furthermore, once the basic network has been constructed, it can be expanded as more areas are developed and network alterations occur.

Results attained from these simulations should be compared to data attained directly from the network. The reason for this is that the simulations rely on mathematical formulas with constant variables. To satisfy these constant variables, the models often display unrealistic answers (Nyende-Byakika et al., 2012). These concepts will however be further explored for each of the different models.

There are three main types of analyses which can be performed, namely the Hardy Cross method, Demand-Driven Analyses (DDA) and pressure/Head-Driven Analyses (HDA) (Nyende-Byakika et al., 2012). These three approaches are herewith discussed regarding their workings, input parameters and implementations in network modelling in South Africa.

2.6.1 Model types

2.6.1.1 Steady-state simulation

A steady-state simulation has been described as a snapshot of a WDS (Savic and Banyard, 2011). Variables such as demands and operational constraints are kept constant. In a real WDS, these variables are constantly changing, for example, the demand at a house changes depending on

Stellenbosch University https://scholar.sun.ac.za

whether a toilet is being flushed or a washing machine is operating. In a steady-state analysis, a moment in time is considered, as to give a fixed value to the demand at a house at that moment. This simulation approach, during the peak demand instance, is generally used to size the pipes within the WDS. A more thorough explanation of steady-state analysis follows.

2.6.1.2 Extended period simulations

In an Extended Period Simulation (EPS), there are changes in the demands of a network (Savic and Banyard, 2011). These demand changes cause the operational conditions of the network to fluctuate. From this, knowledge about how a network reacts to these changes can be found. By changing the demands, the pressures and flow rates in the system change, thus altering the reservoir levels. Along with the emergency storage and bulk supply flowrate, this gives details of how large a reservoir should be and if it is adequate in meeting the varying demands of the network.

To perform an EPS, the following additional information about the network is required:

- Pump curve and valve settings;
- Reservoir starting levels, Reservoir depth /volume curves;
- Duration of simulation;
- Time-varying demands.

2.6.1.3 Water quality simulation

As water distribution software packages have advanced, the ability to simulate water quality has been added. Modern programmes such as Wadiso consider source water mixing, chlorine decay, water age and contamination spread (Loubser, 2018). These programmes can now be used to see the effects of increased chlorine concentrations on water quality and how the water quality decreases further from the source.

Stellenbosch University https://scholar.sun.ac.za

2.6.2 Equations used in models

WDSs make use of two equations to compute the various hydraulic conditions in the system. These equations are the Continuity equation and the Energy equation (Nyende-Byakika et al., 2012). Although both equations are relatively simple, as a network grows in size and complexity, with more connections and fittings, these equations become more difficult to solve. It is for this reason that network models are typically used for larger systems, with hand computations being impractical to use. Furthermore, pipe networks are solved by considering loops within the network. This results in there being more unknowns than equations. To resolve this, an iterative solution is necessary to solve the various unknowns and satisfy the equations (Savic and Banyard, 2011).

An outline of each of the various equations follows. If a deeper understanding of the different terms and methods are needed, consult *Hydraulics in Civil and Environmental Engineering* (Chadwick et al., 2013).

2.6.2.1 Continuity equation

The Continuity equation is based on the fact that no matter how complex a system is, or in which direction fluids flow, the fluid mass in the system is conserved (Fishxing, 2006). If the theory of conservation of mass is applied to a system under steady flow conditions, continuity of flow occurs. What this simply means is that if for example the cross-sections of the flow in a pipe are taken at various points, the product of each cross-section's area and flow velocity is equal. This relationship is expressed in Equation 1.

$$Q = Velocity1 * Area1 = Velocity2 * Area2 \quad (1)$$

Furthermore, inflows are treated as positive values, whereas outflows are treated as negative values (Chadwick et al., 2013).

Stellenbosch University https://scholar.sun.ac.za

2.6.2.2 Energy equation

The Energy or Bernoulli equation calculates the energy of a system by considering the pressure, elevation and velocity of the fluid within the system. Care must be taken when using the formula, as it is considering an environment with only inviscid fluids.

When considering a WDS, energy can either be added to the network or withdrawn from the network. Energy is typically added to the network with elevated service reservoirs and pumps. These methods indirectly increase the pressure in the system, thus adding energy. Energy is lost or extracted from the network due to friction from pipes, valves and elevated points of demand. The Energy formula is presented as Equation 2.

$$\frac{V1^2}{2g} + \frac{P1}{\gamma} + h1 = \frac{V2^2}{2g} + \frac{P2}{\gamma} + h2 + \Sigma hL \qquad (2)$$

The ΣhL term accounts for the head losses in a system. There are two types of head losses, namely frictional losses and local losses (Savic and Banyard, 2011). Frictional losses are caused by the rough inner surface of the pipes. Local losses ae caused by fittings such as valves. Local losses are however small for large networks and typically neglected in calculations.

2.6.2.3 Friction Losses

Two equations are typically used to calculate frictional losses. These include the Darcy-Weisbach equation and the Hazen-Williams equations.

The Darcy-Weisbach equation is presented as Equation 3.

$$hf = \lambda \frac{L}{D} \frac{V^2}{2g} \qquad (3)$$

Stellenbosch University https://scholar.sun.ac.za

The λ term is a frictional factor and is dependent on the pipe material and Reynold's number. If the Reynold's number is below 2000 and laminar flow conditions exist, λ is solved as in Equation 4.

$$\lambda = \frac{64}{Re} \tag{4}$$

If the Reynold's number is above 4000, then turbulent flow conditions are present. This requires the use of the Colebrook-White equation to solve for λ. Furthermore, when turbulent flow conditions exist, the pipe roughness has a far greater effect on the λ term. The Colebrook-White equation is presented as Equation 5.

$$\frac{1}{\sqrt{\lambda}} = -0.8686 \ln\left(\frac{6}{3.7D} + \frac{2.51}{Re\sqrt{\lambda}}\right) \tag{5}$$

The Hazen-Williams equation was originally introduced in 1902 to solve frictional losses (Savic and Banyard, 2011). Due to its empirical nature, the equation is not dimensionally homogenous and should be used with care (Savic and Banyard, 2011). Equation 6 represents the Hazen-Williams equation.

$$hf = 10.67 \frac{Q^{1.85}}{C^{1.85} D^{4.87}} L \tag{6}$$

2.6.2.4 Local Losses

Local losses in a WDS typically occur at bends, reducers and valves. By making reference to Figure 1, these losses are explained. Savic provides an overview of how local losses are caused (Savic and Banyard, 2011). When a valve is present in a pipe, the area of the pipe effectively decreases at the valve's location. Thus, when water flows past this point, its velocity has to increase to maintain constant flow according to the Mass equation.

Stellenbosch University https://scholar.sun.ac.za

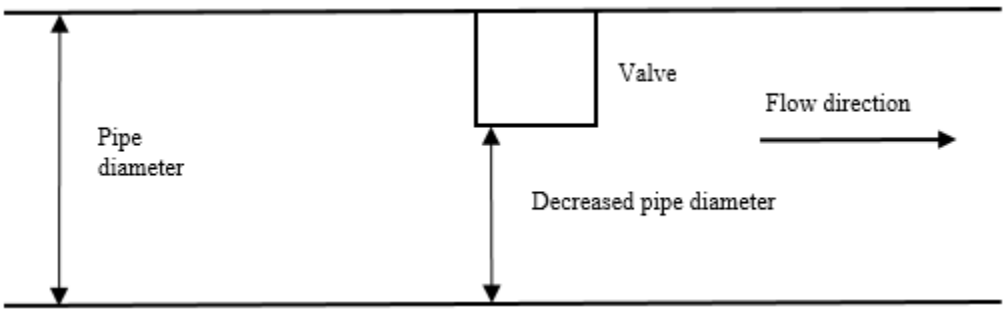


Figure 1: Valve cross-section. The cross-sectional pipe area decreases at a valve.

When this occurs, pressure energy is converted to kinetic energy. Once the water has passed the valve, it steadily slows down and forms eddies. When this occurs, the kinetic energy is not fully converted back to potential energy, thus resulting in an energy loss.

As mentioned earlier, these losses are typically very small values when considering a large network and are omitted to simplify computations.

2.6.3 Steady flow analysis of networks.

A steady flow analysis of a network is a widely used approach when designing WDSs. This method makes use of the continuity and energy equations discussed earlier, to solve the hydraulic variables at each node in a system (Savic and Banyard, 2011). Three analysis approaches are described, namely the Hardy Cross method, the DDA method and the HDA method.

In each of these three methods, the network makes use of three types of nodes, namely fixed-head, variable-head and ordinary nodes or junctions (Savic and Banyard, 2011). Fixed-head nodes are used to represent features such as lakes, which have a very marginal change in water elevation. Variable-head nodes are used to represent reservoirs, which have a greater change in water elevation more regularly. Lastly, the ordinary nodes are used for all other features in a network like pipe connections, points of demand and valves (Savic and Banyard, 2011).

Stellenbosch University https://scholar.sun.ac.za

2.6.3.1 Hardy Cross method

The Hardy Cross method to solve for nodal demands and pressures in a WDS was introduced in 1936 (Savic and Banyard, 2011). The method relies on applying the Continuity equation at each node in a network and applying the Energy equation for each independent loop in the same system (Savic and Banyard, 2011).

The method relies on making iterative guesses of the flows in a pipe, until the sum of the head losses in a loop is 0. Once this condition is met, the next loop can be solved by applying the same principle. After all the loops are completed, the different flows within the network are known. From this, the pressure heads at each node in the network can be solved (Savic and Banyard, 2011).

This analysis method later evolved to become the Newton-Raphson method, which adjusted both flows and heads in each loop simultaneously to reach hydraulic equilibrium within the network (Savic and Banyard, 2011).

The Hardy Cross method was successful in analysing smaller networks, but became limited in larger networks with multiple unknowns. As computers became available, the Hardy Cross method was used in multiple software programmes, which made it possible to apply the method to larger networks. Some of these programmes include GINAS and EPANET (Savic and Banyard, 2011).

2.6.3.2 Demand-driven analysis

DDA is the most widely used method in WDS modelling. The analysis relies on the assumption that pressures within the network are dependent on network demands, thus demands are the independent variables and pressures are the dependent variables (Nyende-Byakika et al., 2012).

When using this method, demands are fixed at the various nodes that make up the network. The simulation then calculates the nodal pressures that satisfy the nodal demands (Nyende-Byakika et

Stellenbosch University https://scholar.sun.ac.za

al., 2012). The model relies on the Continuity and Energy equations to determine the pipe flows and hydraulic heads at each respective node, assuming demands are fixed and pressures can change. This model approach yields accurate and realistic answers when pressures are high and the network can supply the required demands. The model however yields unrealistic answers when demands are higher to such an extent that unrealistic negative pressures are calculated. The reason for this is that the model keeps the demands fixed, regardless of whether the network can meet the demands. To satisfy the Continuity and Energy equations for this fixed demand, pressures are computed as negative values (Nyende-Byakika et al., 2012).

In many WDSs, nodal pressures are not always dependent on demands. For networks with excessive demands or inadequate pressure such as that found in intermittent water supply networks, the pressures within the network are generally lower. When these situations arise, the network is physically not capable of meeting the demands placed on the system. Thus, for situations like these, the pressure is no longer demand dependent, but demand becomes dependent on network pressures (Nyende-Byakika et al., 2012).

For a DDA, the pressure at a node needs to be greater or equal to the pressure threshold in order to meet the specified demand. The pressure threshold is the pressure of a node if it were to provide adequate pressure without extra pumping to the highest building in the vicinity (Tanyimboh and Templeman, 2000). When using a DDA, if the required pressure is below this threshold pressure, the demand will not be met. However, because the demand is constant, the model solution computes an unrealistic pressure rather than a smaller demand that can be supplied (Nyende-Byakika et al., 2012).

Stellenbosch University https://scholar.sun.ac.za

When using the DDA, care must be taken to analyse several hours. In periods of low demand when pressures are high, the solutions will be accurate and realistic. It is only during peak periods of high demand that pressure results are sometimes unrealistic. Thus, periods of low and high demand should be compared to establish which nodes are really experiencing negative pressures and what solutions are necessary (Nyende-Byakika et al., 2012). DDA can also be manipulated to yield more accurate results. For example, if demands are reduced, nodes with negative pressures will systematically become positive till a point is reached when the network can meet the demands. This can then be viewed as the maximum demand the system can practically meet (Nyende-Byakika et al., 2012). Also, Peak Factors can be adjusted until the number of nodes with negative pressures converge. At this point, more realistic demand values are attained (Hirst, 2017).

2.6.3.3 Pressure/head driven analysis

The primary difference between a DDA and a HDA is that in a DDA, the higher the outflow, the lower the pressure (Nyende-Byakika et al., 2012). This changes for a HDA, as the higher the pressure, the higher the outflow (Nyende-Byakika et al., 2012). Furthermore, in a DDA the demand is the independent variable whereas the pressure is the dependent variable. This changes again for the HDA as the pressure is the independent variable and demand is the dependent variable (Nyende-Byakika et al., 2012).

The HDA considers a link between pressure and demand to make computations like a DDA. The HDA does however show a decrease in nodal demands considering the available pressures. Or more simply, the HDA makes use of iterations to constantly adjust both pressures and demands till both variables are satisfied and the network is in a state of equilibrium (Nyende-Byakika et al., 2012).

Stellenbosch University https://scholar.sun.ac.za

This iterative processing makes HDA ideal in situations where the water reticulation network is operating at lower pressures. Like mentioned before, a DDA will compute the nodal pressures to mathematically satisfy the nodal demands. This method however fails if the system has lower pressures, as the full demand can physically not be supplied by the network. A HDA considers this fact and will therefore decrease the nodal demand to a value that the network can provide. From this, more accurate and realistic nodal pressures and demands can be computed. Fundamentally, with a HDA, the resultant pressures cannot be negative. The supply will just be less than the demand (Nyende-Byakika et al., 2012).

This more accurate process of a HDA is however much more difficult to implement in network models. The main reason for this is that a large pool of data will need to be collected on site to accurately depict the network. Furthermore, the model will have to implement and solve the complex pressure-flow link at each and every node, which is highly computational intensive to solve and find a solution for (Nyende-Byakika et al., 2012).

2.6.4 Unsteady flow analysis of networks

The Continuity equation changes for unsteady flow analysis. The equation states that the difference between the inflow and outflow of a hydraulic stowing structure is equal to its change in storage (Savic and Banyard, 2011).

This approach is used for EPS and transient flow analysis (TFA). EPS can be seen as the summation of multiple steady-state analyses at fixed time intervals. From this, the analysis of reservoirs and other structures can be considered when there is a small change in flow over a long period of time (Savic and Banyard, 2011). Or more simply, the simulation shows how variables change over time and not what their status is at one specific time.

Stellenbosch University https://scholar.sun.ac.za

2.6.5 Network modelling data

When building a model, information of existing infrastructure can be found using as-built drawings, GIS, surveys and drawing archives. This information is often limited and if records are found, they must ideally be verified using calibration tests, to ensure the model is accurate.

To find information about user demands is also problematic, as not every household, business and water user are metered.

2.6.6 Model building

The purpose of a WDS model is to assist in the design and analysis of a WDS. Furthermore, the model can be used to monitor operations which can be used to optimise the performance of the network and to schedule routine maintenance. For example, new installations and infrastructure can be tested and simulated to see what effects they have on each network component before any major system alterations take place.

When building a model, the level of detail that is required changes. If a master plan is compiled, it generally proposes an outline of the system. Smaller pipes and finer details are often omitted to simplify the construction and analysis of the model. As more detail is acquired, the model can be expanded to give a more realistic representation of the real WDS.

2.6.7 Calibration

The purpose of calibration is to improve a model at depicting reality and to reduce any uncertainties in the system. This can only be done by collecting data at relevant points on the real system. This data includes flows and pressures at different nodes and at pumping stations. Data is typically collected for at least a duration of a week to obtain accurate results.

Stellenbosch University https://scholar.sun.ac.za

Once the field data has been collected, the calibration process is initiated. Two approaches are typically used. The first is an iterative approach, whereby flows and pressures are iteratively updated until the model and reality match. The second is an implicit approach, whereby the input parameters are automatically updated in a trial sequence until the modelled and real system match (Savic and Banyard, 2011).

2.7. Summary

A thorough explanation of how WDSs function and are designed has been provided. The typical components that a WDS consists of have been discussed, together with an overview of how these components are modelled. It has been indicated, that due to the complexities involved, modelling is essential to ensure that these systems are designed and built in a cost effective manner and, where possible, in a way to account for future demands on the network.

The process of building models to simulate these networks are tedious, as each network consists of many components. Models can be simplified by excluding smaller features or features that have a negligible effect on the system as a whole.

Furthermore, the formulas and simulation techniques used to model these networks are computationally extensive and require adequate computer power to solve. It was also highlighted that some of the techniques used may yield unrealistic results. These results are often attributed to incorrectly assumed variables, such as pipe roughness, average water demands, leaks and instantaneous peak factors. It is thus the responsibility of the modeller to understand the workings of the software, and have sufficient knowledge of potential parameter impacts to understand where network problems may arise.

Stellenbosch University https://scholar.sun.ac.za

Software packages such as Infoworks, EPANET and Wadiso that are specifically designed to simulate WDSs have evolved in recent years. These packages do not only predict pressures and water velocities in the network, but are now able to perform computations regarding water quality and future design parameters for undeveloped areas.

Stellenbosch University https://scholar.sun.ac.za

SECTION 3

STATISTICAL REGRESSION METHODS

A regression model is a model that finds the relationship between variables (Montgomery et al., 2014). These models are evident in everyday life. Simple examples include the relationship between vehicle mass and fuel consumption or the size of a house and the number of residents living in the house.

Furthermore, a model has independent or predictor variables which cannot change and dependent or response variables which can. For example, the engine capacity can be considered to be the independent variable and the vehicle's fuel consumption the dependent variable. As the vehicle's engine capacity increases, normally so does the fuel consumption. Consequently, the fuel consumption is dependent on the engine capacity and changes if the engine capacity changes.

Regression models are also useful, as the relationship found between variables can be used to make predictions on new variables (Montgomery et al., 2014). For example, if a relationship was found between the vehicle's engine capacity and its fuel consumption, then this relationship can be used to predict how much fuel another vehicle could use, by considering its engine displacement.

By considering a simple relationship between one independent variable and one dependent variable, the two variables can be plotted on a 2D graph. A line with regression coefficients consisting of an intercept and slope can be used to represent the relationship between the variables (Montgomery et al., 2014).

Auret, an engineering statistics lecturer from Stellenbosch University (Auret, 2018), recommended to research three accepted statistical regression approaches. These included Multi

Stellenbosch University https://scholar.sun.ac.za

Linear Regression (MLR), Principal Component Analysis (PCA) and Partial Least Squares (PLS). These methods were recommended because all three consider multiple independent variables and perform the regression analysis using different techniques.

3.1. Multi Linear Regression

3.1.1 Introduction

A MLR model is a regression model that has more than one independent or x-variable. The typical model is represented by Equation 7. The β_0 variable is the intercept where the rest of the β variables are the regression coefficients. These variables represent the change in Y for a unit change in x, while the other x-variables are kept constant (Montgomery et al., 2014).

$$Y = \beta 0 + \beta 1x1 + \beta 2x2 \ldots + \epsilon \quad (7)$$

If some variables seem to have a greater influence on Y, then interaction effects can be incorporated into a linear regression model. Interaction effects include the cross-product of more than one variable and are represented by a new term. By incorporating these new terms, the magnitude of influence that the more important variables have on Y, is increased. Thus, the more important variables have a greater effect on changing Y than the other variables (Montgomery et al., 2014).

$$Y = \beta 0 + \beta 1x1 + \beta 2x2 + \beta 12x1x2 \ldots + \epsilon \quad (8)$$

3.1.2 Least-squares regression

The least-squares regression model is the model most often used to estimate the regression coefficients or β terms. The model differs from ordinary least-squares in that more than one x-variable can be considered.

Stellenbosch University https://scholar.sun.ac.za

Statistical regression model

The least-squares model estimates the regression coefficients by minimising the residual, ϵ values. If the residuals are small, the estimated and observed Y values are similar and thus the model is fairly accurate. Equation 9 is used to minimise the residuals. The n term is the number of samples.

$$L = \sum_{i=1}^{n} \epsilon i^2 = \sum_{i=1}^{2} (yi - \beta 0 - \sum_{j=1}^{k} \beta j x i j) \tag{9}$$

By minimising the L term, while considering each of the individual β terms, each of the equations represented by Equation 10 are found. The exact and extensive mathematics used to find these terms are not fully explained. If a more in-depth explanation is required, the reader is referred to *Applied Statistics and Probability for Engineers* (Montgomery et al., 2014).

$$\begin{aligned} 2\widehat{\beta 0} + \widehat{\beta 1} \sum_{i=1}^{n} xi1 + \widehat{\beta 2} \sum_{i=1}^{n} xi2 \ldots \widehat{\beta k} \sum_{i=1}^{n} xik &= \sum_{i=1}^{n} yi \\ \widehat{\beta 0} \sum_{i=1}^{n} xi1 + \widehat{\beta 1} \sum_{i=1}^{n} xi1^2 & \\ + \widehat{\beta 2} \sum_{i=1}^{n} xi1xi2 \ldots \widehat{\beta k} \sum_{i=1}^{n} xilxik &= \sum_{i=1}^{n} xi1yi \\ \widehat{\beta 0} \sum_{i=1}^{2} xik + \widehat{\beta 1} \sum_{i=1}^{n} xikxi1 & \\ + \widehat{\beta 2} \sum_{i=1}^{n} xikxi2 \ldots \widehat{\beta k} \sum_{i=1}^{n} xik^2 &= \sum_{i=1}^{n} xikyi \end{aligned} \tag{10}$$

This method will produce as many equations as there are unknown β terms. Once these equations are ready, simultaneous equations are used to solve each of the unknown β terms and thus the linear regression model can be constructed.

Stellenbosch University https://scholar.sun.ac.za

3.1.3 Matrix representation of multi linear regression model

This model is the same as the least-squares model discussed previously, but expresses the mathematical relationships in a matrix form rather than a scalar form (Montgomery et al., 2014). The Y, β and ϵ terms are vectors while X represents a matrix of all the x terms.

$$Y = X\beta + \epsilon \tag{11}$$

For a full decomposition of the matrix MLR model, refer to *Applied Statistics and Probability for Engineers* (Montgomery et al., 2014).

3.1.4 Hypothesis testing

After the regression model has been developed, hypothesis tests can be done to check the adequacy of the regression coefficients and the model itself.

3.1.4.1 Regression

Regression tests are performed to assess if there is a linear relationship between the independent variables and the dependent variable. The magnitude of this linearity can be measured using the coefficient of multiple determination or R^2 statistic. Simply stated, the R^2 measures the quality of the fit of the model. This indicator is however problematic as its magnitude increases with increasing number of variables (Montgomery et al., 2014).

To solve this, a R^2adj indicator is used. The magnitude of this coefficient only increases if the variable added decreases the error mean square. This coefficient is thus ideal for checking that the model is not over-fitted or that multicollinearity does not exist (Montgomery et al., 2014). Multicollinearity is when different variables are showing the same thing.

Stellenbosch University https://scholar.sun.ac.za

3.1.4.2 Tests on individual regression coefficients

To test whether a regression coefficient is adequate or should be removed to improve the model's accuracy, one can consider the p-value of each β variable. A significance value of 0.05 is the general statically accepted value to represent the important variables (Montgomery et al., 2014). Thus, if a regression coefficient has a p-value greater than 0.05, it can be regarded as having less significance in making the model accurate and can be removed from the model if needed.

3.1.5 Confidence intervals

Confidence intervals can be constructed for the individual β terms. These intervals thus provide an interval in which the regression coefficients can safely occur.

$$\hat{\beta}j - t\ a/2, n-p\sqrt{\hat{\sigma}^2 Cjj} \leq \beta j \ \leq \ \hat{\beta}j + t\ a/2, n-p\sqrt{\hat{\sigma}^2 Cjj} \tag{12}$$

Confidence intervals can also be constructed for the y-variables for fixed x values.

$$\begin{aligned} &\hat{\mu}Y|Xo - t\ a/2, n-p\sqrt{\hat{\sigma}^2 Xo'(X'X)^{-1}Xo} \leq \hat{\mu}Y|Xo \\ &\leq\ \hat{\mu}Y|Xo + t\ a/2, n-p\sqrt{\hat{\sigma}^2 Xo'(X'X)^{-1}Xo} \end{aligned} \tag{13}$$

3.1.6 New observation predictions

Similar to the confidence intervals for y or the mean response, one can also find prediction intervals. These intervals are wider than the confidence intervals, as these do not only consider the error in estimating a value, but also the error in predicting a new future value at fixed x values (Montgomery et al., 2014).

$$\begin{aligned} &\hat{y}0 - t\ a/2, n-p\sqrt{\hat{\sigma}^2(1 + Xo'(X'X)^{-1}Xo)} \leq \hat{y}0 \\ &\leq\ \hat{\mu}Y|Xo + t\ a/2, n-p\sqrt{\hat{\sigma}^2(1 + Xo'(X'X)^{-1}Xo)} \end{aligned} \tag{14}$$

Stellenbosch University https://scholar.sun.ac.za

These intervals are only for one y point, which could be a limiting factor if one desires a confidence interval for the full range of results. For the full derivation of the confidence and prediction intervals refer to *Applied Statistics and Probability for Engineers* (Montgomery et al., 2014).

3.1.7 Checking model adequacy

3.1.7.1 Analysis of Residuals

To analyse the residuals, one can use residual plots. This is an easy method for visually determining whether the model under or over predicts values. Furthermore, it can show whether the residuals are normally distributed or whether these follow another form of distribution. If another distribution profile is present, it is an indication that the model is perhaps not linear and that additional or new terms are necessary (Montgomery et al., 2014).

A normal residual plot and a non-normal residual plot are represented in Figure 2. The left graph shows a residual plot that follows a normal distribution. The points stay relatively close to the prediction line and do not vary in an unordered or strange manner, like the right graph.

Before analysing the residuals and constructing the residual plots, it is common practice to standardise residual values. This is done using Equation 15.

$$di = \frac{ei}{\sqrt{\hat{\sigma}^2}} \tag{15}$$

Another method to scale the residuals is to studentise them. This follows a more complicated procedure that underestimates the magnitude of the residuals. This method is thus ideal for detecting outliers, as the method has specific boundaries in which values can occur (Montgomery et al., 2014). This method is outlined in Equation 16.

Stellenbosch University https://scholar.sun.ac.za

Statistical regression model

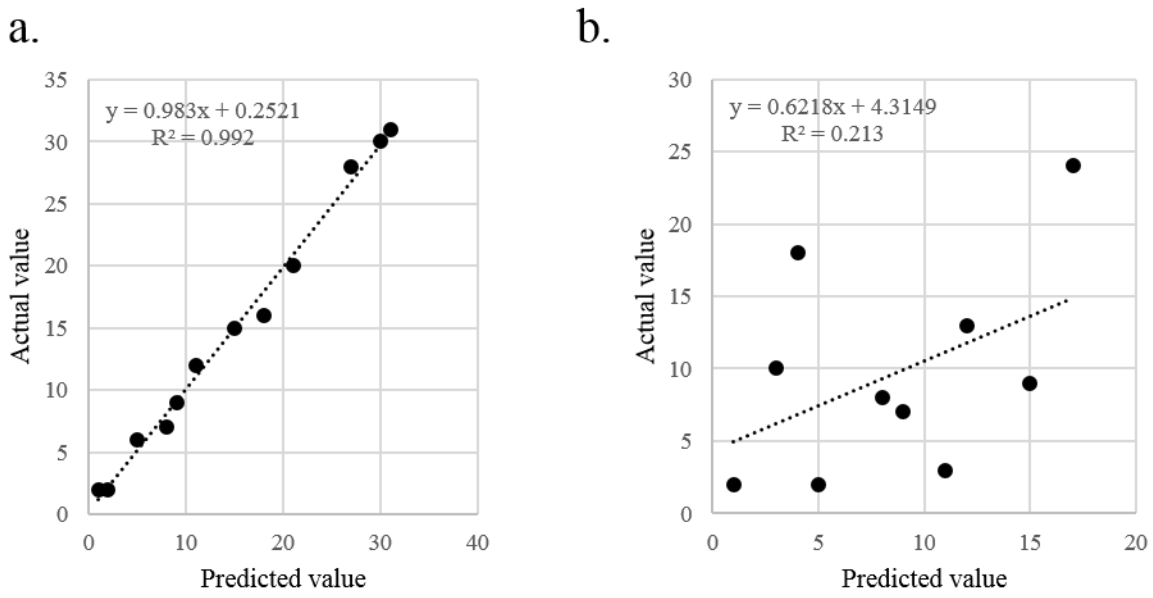


Figure 2: Types of residual plots. The plot (a) follows a normal distribution, while the plot (b) follows a non-normal distribution.

$$hii = Xi'(X'X)^{-1}Xi$$

$$ri = \frac{ei}{\sqrt{\hat{\sigma}^2(1-hii)}} \quad where\ i = 1,2,3 \ldots n \tag{16}$$

$$0 < hii < 1$$

3.1.7.2 Measuring the influence of observations

When plotting the linear regression model, it often occurs that certain observed points lie in more remote regions of the graph. At first glance, one would regard these points as outliers. But this is not always the case as these points could be valid observations that are crucial in determining the accuracy of the model. One method to check these points is Cook's method, recommended by Montgomery (Montgomery et al., 2014). This method considers the square distance between the β term for all the observations and the $\hat{\beta}$ term when the potential outlying point is removed. If the D value exceeds one, then the point is influential and requires additional inspection.

Stellenbosch University https://scholar.sun.ac.za

$$Di = \frac{(\beta i - \hat{\beta})X'X(\beta i - \hat{\beta})}{p\hat{\sigma}^2} \quad (17)$$

3.1.8 Characteristics to consider with multi linear regression modelling

3.1.8.1 Nonlinear models

If a nonlinear relationship exists between x and y, then additional or new terms should be added to the model. This concept was mentioned briefly with the inclusion of interaction variables. When using interaction variables, the model is however still linear. One can also construct polynomial regression models which can account for nonlinear effects. An example of the outline of a nonlinear equation is represented by Equation 18.

$$Y = \beta 0 + \beta 1x + \beta 2x + \beta 11x^2 \ldots + \epsilon \quad (18)$$

3.1.8.2 Categorical variables

Often one can encounter variables which are not quantitative. This could for example include a yes/no answer. One method to account for qualitative answers is to use indicator variables (Montgomery et al., 2014). This method assigns a quantitate value to qualitative values. For example, all yes responses could be 1 while all no response could be 0.

3.1.8.3 Selection of variables

When picking the number of variables to use for a model, one does not always want to use the maximum number of variables. The reason for this is that if the number of variables increases, there could be more noisy data and the maintenance costs to obtain the values of the variables could be costly (Montgomery et al., 2014). Therefore, several methods exist that tests the model to determine the ideal and most effective number of variables to use. The model should still contain enough variables to be as accurate and precise as desired.

Stellenbosch University https://scholar.sun.ac.za

All regression parameters

This method determines a regression model for all possible combinations of x-variables. From this, the model with the highest R^2 value is picked. This method is computationally intensive as 2^k models are created (Montgomery et al., 2014). For example, if the model contains 7 x-variables, 128 models are created. For this reason, the method is typically executed via a computer and is rarely used for hand calculations.

PRESS

Another method for determining the number of components needed, is the cross-validation PRESS method (Geladi and Kowalski, 1985). The PRESS or Prediction Residual Sum of Squares method computes the sum of the residuals squared, R^2 for each number of components. This is then again plotted against the number of components, in order to find a point where a minimum value exists. The number of components correlating to this minimum PRESS value is then chosen. This is illustrated in Figure 3.

Stepwise

The stepwise method is the most commonly used method for variable selection (Montgomery et al., 2014). The method iteratively adds variables to the model and tests the p-values of all the variables. The variable with the highest p-value is then removed and the process is repeated. P-values are commonly regarded as significant if they fall below the chosen significance level, rendering the variables important. Only variables with p-values greater than the chosen significance level are removed, regardless of whether they have the highest p-value. For example,

Stellenbosch University https://scholar.sun.ac.za

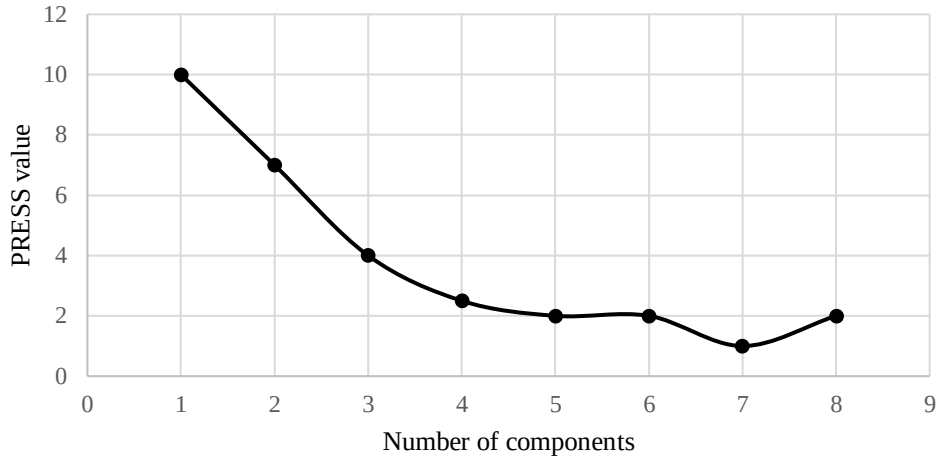


Figure 3: Typical PRESS test output graph. The necessary number of components can be found at the minimum PRESS value.

if a significance level of 0.05 is chosen, even if the highest p-value is 0.045, the variable is not removed.

Forward stepwise

The forward stepwise method is similar to the standard stepwise method. The method adds additional variables to the model until a point is reached where the R^2 term does not change greatly (Montgomery et al., 2014). This method is however somewhat flawed as the model does not remove or re-assess variables added earlier in the process. From this, one can see that the order in which you add variables has a great impact on the model. If less important variables are added early in the process, they could be considered, whereas they would not be considered if they were added later in the process. This was found from personal experience using the forward stepwise method.

Stellenbosch University https://scholar.sun.ac.za

Backward stepwise

The backward stepwise method starts with all the variables and iteratively removes the variable with the highest p-value that is greater than the chosen significance level. This is done until a point is reached where all the variables can be considered as significant.

Multicollinearity

If two x-variables have a strong linear relationship, multicollinearity exists. This in effect means that both variables are showing the same thing, therefore one of the variables can potentially be eliminated. One method to test for this multicollinearity is the VIF test. A VIF test simply tests the correlation between two predictor or x-variables (Bock, 2018). If this VIF value between two predictor variables is high, they are essentially showing the same information and multicollinearity exists. If a VIF value of greater than 4 is found, multicollinearity is present and the variable with the weaker linear relationship to y can be eliminated (Montgomery et al., 2014).

$$VIF(\beta j) = \frac{1}{1 - Rj^2} \tag{19}$$

3.1.9 Summary

MLR is the most commonly used method to find a regression model relating variables. Furthermore, it can consider multiple x and multiple y variables at the same time.

There are also various ways of detecting outliers and assessing which variables are important. The method is also made easy as commonly used data software such as Excel have built-in functionality to perform these analyses. These software packages also provide a host of additional information to assess the quality of the model. These include residual plots, p-values and f-values.

Stellenbosch University https://scholar.sun.ac.za

As this method of regression is well established, it is highly recommended as a quick and efficient way of attaining regression models. The method can also account for nonlinear relationships by considering interaction effects. Prediction interval equations also exist, which aids the user in using models for safe and accurate forecasts.

3.2. Principal Component Analysis

3.2.1 Introduction

PCA is a statistical method of simplifying confusing datasets and extracting the crucial or most vital and influential information (Shlens, 2014). It does this by reducing the number of variables that expresses the data. Or more simply, it reduces the problem to a lower dimension which makes it easier to visualise and analyse the data, with the added benefit of reducing meaningless or noisy data. By identifying the most important data points, PCA can find new variables or components to express the data. PCA is only performed on the x-variables.

3.2.2 Vital mathematical proofs necessary for PCA

In order to understand the derivation of the PCA formulas, certain key concepts need to be explained. These concepts consist of various ideas and proofs which manipulate and highlight key characteristics of matrices. These proofs can be found in Appendix 10.1.

3.2.3 Mathematical framework and foundation of PCA

3.2.3.1 Example

Before elaborating on the derivation of PCA, an example will be explained in part. Many parts and reasons illustrated in the example will be defined and explained in full in the subsequent sections. The example should however present a basic idea of when and why PCA is used.

Stellenbosch University https://scholar.sun.ac.za

Statistical regression model

Imagine a dataset with 40 samples and 3 variables per sample describing the appearance of a car. The three variables are the length and height of the car, and whether the car has a sunroof. Considering this amounts to 120 data points, it can be difficult to comprehend how such a modal can be analysed. If we consider the variables, it becomes apparent that some variables are more important than others. For example, the length and height variables are more important in describing how the car looks, as opposed to whether it has a sunroof or not. One can imagine that there is a greater variance in the length and height, as these variables change significantly for different cars. For example, if we know the car has a short length and short height it could be a hatchback and if it has a long length and short height it could be a sedan.

In PCA, new variables are created by taking weightings of the x-variables. Thus, to describe a car, the length and height will have a greater weighting than whether it has a sunroof or not. So, each new variable is made up of a summation of each x-variable multiplied by its respective weighting.

Furthermore, when determining these new variables, one wants it to describe as much of the data as possible. This is where PCA becomes imperative. PCA enables one to solve the weightings that will amount to the new variables which describe the dataset the best. Or more simply, accounts for the greatest variation within the dataset and thus the most information. After calculating the first new variable, subsequent variables can be found to describe the remaining data. Thus, the first variable or component describes the greatest variation in the dataset, the second variable the second most etc. If the first variable describes a sufficient proportion of the data, for example, 80 %, then the subsequent variables can be disregarded.

Stellenbosch University https://scholar.sun.ac.za

3.2.3.2 Change of basis

As stated before, the primary goal of PCA is to find the best basis to re-express the data and account for the greatest variance in the dataset, thus describing most of the data points (Auret, 2018). In a more technical way, PCA finds a new basis, which is made up of a linear combination of the original basis.

This transformation from the original basis to the new basis can be expressed as:

$$T = XP \tag{20}$$

X is a matrix summarising the original dataset. Each column of X is one of the original variables with each row values representing one of the samples. T is the transformation of matrix X and expresses the original dataset in a new way. Matrix P transforms matrix X to T. It does this by rotating and stretching the dataset. The rows of X are multiplied by the columns of P to give the new values in the T matrix. P represents the principal components. Each column of P is a principal component with each row value representing the weighting factor for the x-variables.

This idea of multiplying the rows of X with the columns of P is the same concept of weighted averages defined earlier. These linear combinations are a way of combing the original variables in a linear way to attain the new variables. Thus, the P matrix summarises the weighted averages. With these weighted averages known, we can form a new basis that best represents the data. The T matrix is also known as the scores matrix and the P matrix is known as the loading matrix.

This concept may seem obscure at first but becomes clearer when the problem is expressed in a visual way via Figure 4.

Stellenbosch University https://scholar.sun.ac.za

Statistical regression model

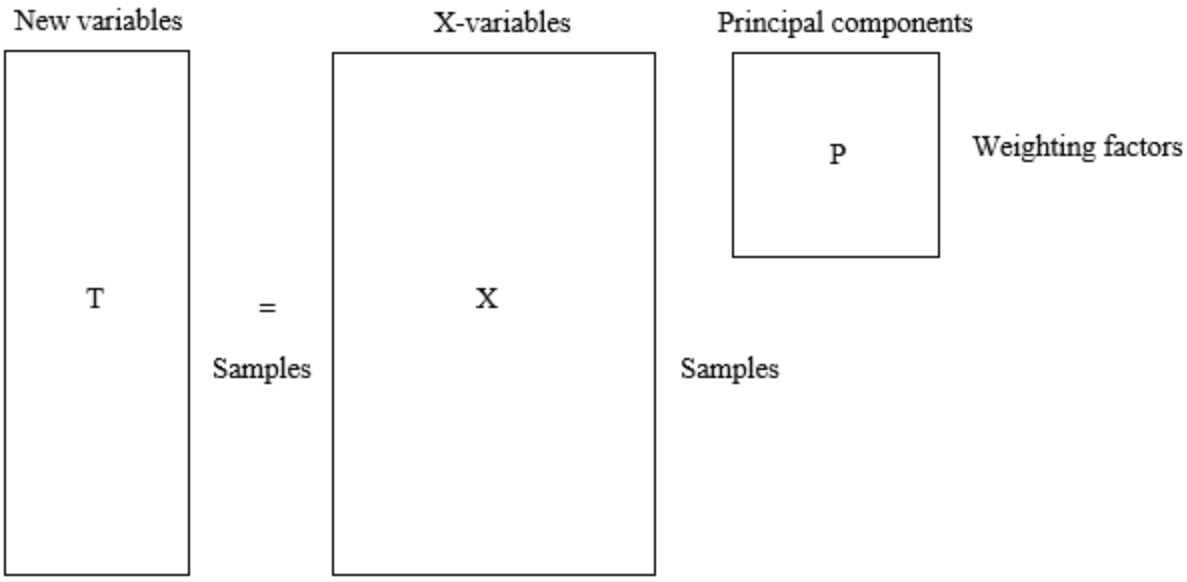


Figure 4: PCA basis transformation. The transformation of the x-variables or X matrix to the T matrix using the principal components outlined in the P matrix.

3.2.3.3 Variance

Knowing that we want to find a new basis to express the data, the question arises: What aspects of the dataset should the new basis represent to best depict the data? This is quite simply the greatest variance of the data (Shlens, 2014). Therefore, one wants to choose the best weighted averages to depict the greatest variance in the data. The various components of the variance that outline its importance are described hereafter.

Noise and rotation

When analysing the data, we consider the direction with the greatest variance as important (Shlens, 2014). This direction is often referred to as a signal. The reason for this is that this direction accounts for the largest range of values and shows how the values from a sample can vary.

Noise can simply be thought of as meaningless data, or data with a small variance (Rouse, 2010). This is represented graphically in Figure 5.

Stellenbosch University https://scholar.sun.ac.za

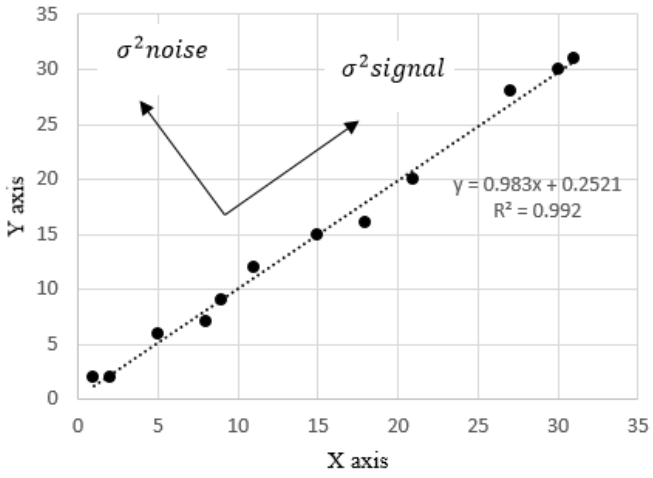


Figure 5: Noise and variance. The signal direction represents the direction of greatest variance in the data while the noise direction is simply perpendicular to the signal direction.

Although there is no direct measure of noise, its variance can be compared to the signal strength variance of the data to establish a relative ratio. This ratio is represented in the Signal to Noise Ratio equation.

$$SNR = \frac{\sigma^2 signal}{\sigma^2 noise} \tag{21}$$

From this, the quality of the data can be assessed. If the ratio is greater than 1, the data can be considered as precise whereas a ratio smaller than 1 indicates noisy data. By maximising the SNR ratio, and hence the variance, the appropriate rotation of the basis is found. This rotation therefore accounts for the greatest variance in the data.

Redundancy

Data or information can be considered redundant if it is not necessary or not adding any value to the problem (Shlens, 2014). For example, if two variables give the same information, it would be meaningless to consider both, as the same outcome could be attained by simply considering one.

Stellenbosch University https://scholar.sun.ac.za

Statistical regression model

If two variables are directly proportional, either one could be used to represent the outcome. Therefore, these show the same information and one could simply be eliminated. For example, if we consider the volume of a reservoir or the mass of water inside, the two variables will be directly proportional and give the same information.

Covariance matrix

Covariance simply measures the linear relationship between two variables and how strong this relationship is (Shlens, 2014). A large positive or negative value indicates a high degree of positively or negatively correlated data. The covariance matrix is represented in Equation 22 with n representing the number of samples.

$$Cx = \frac{1}{n} XX^T \tag{22}$$

Some of the properties of Cx include:

- Cx is a square matrix;
- The diagonal terms represent the variance of a specific variable as the column and row positions are equal;
- The off-diagonal terms represent the covariance between different variables.

Thus, the covariance matrix represents the covariance between all measurement pairs. The size of these covariance values shows the level of noise and redundancy. For example, if a value of 0 is found, the measurements are not redundant and each contributes to accurately depicting the data. If a value of 1 is found, both measurements show the same aspect of the data and essentially show the same thing.

Stellenbosch University https://scholar.sun.ac.za

Diagonalising the Covariance matrix

The covariance matrix should ideally have large diagonal values to best depict the variance of data and small off-diagonal values to eliminate redundancy. The optimal Ct matrix should thus be a diagonal matrix with 0 values in all the off-diagonal positions. This would represent the data in a way where redundancy is eliminated, and variance is optimised. The matrix should also have its diagonal values arranged in descending degree of variance. This process of making Ct a diagonal matrix is also called decorrelating T. To diagonalise Ct to form this ideal matrix, PCA has to make certain assumptions. The first being that the rows of matrix P or the basis vectors are orthonormal. This assumption is made as P rotates matrix X so that a new axis is made in the direction of maximum variance. The steps to achieving this include:

1. Finding a vector p1 which represents the normalised direction in an m-dimensional space which includes the greatest variance of X's values;
2. Find vectors p2 etc. which are orthogonal to p1 and account for second and third etc. greatest variance of X's values.

These p-vectors represent the principal components which make up matrix P. The p-vectors are ranked in descending order of how principal they are in accounting for the variance of the data.

3.2.3.4 Using eigenvector decomposition to solve PCA

Before approaching the next part of the problem, a quick summary of what has been discussed thus far and what needs to be achieved, is presented. An orthogonal matrix P to solve T = XP to best represent the data needs to be found. This is achieved by ensuring that matrix Ct = $\frac{1}{n}TT^T$ is a diagonal matrix. The P matrix that satisfies this condition will represent the principal components.

Stellenbosch University https://scholar.sun.ac.za

Statistical regression model

To ensure that Ct is a diagonal matrix, we need to break it up into its unknown components.

$$
\begin{aligned}
Ct &= \frac{1}{n} TT^T \\
&= \frac{1}{n} (XP)(XP)^T \\
&= \frac{1}{n} PXX^T P^T \\
&= \frac{1}{n} (\frac{1}{n} XX^T) P^T \\
&= \frac{1}{n} CxP^T
\end{aligned}
$$

From the proofs outlined in Appendix 10.1, we know that a symmetric matrix can be diagonalised by an orthogonal matrix consisting of its eigenvectors. Knowing that Cx is a symmetric matrix, we can present P as a matrix consisting of its eigenvectors.

Ct	= PCxP^T	Cx is a symmetric matrix
	= P(EDE^T)P^T	E is an orthogonal matrix with the eigenvectors of Cx
	= P($P^T DP$)P^T	P is selected to have rows equal to the columns of E
	= (PP^T)D(PP^T)	
	= (PP^{-1})D(PP^{-1})	P is orthogonal because E was orthogonal, thus $P^T = P^{-1}$
	=IDI	
	=D	

Therefore, in order to ensure that Ct is a diagonal matrix, the principal components of X or the columns of P are simply the eigenvectors of Cx. In other words, the diagonal values of Ct are the variance of X along each principal component p1, p2 etc.

Stellenbosch University https://scholar.sun.ac.za

3.2.4 Practical aspects of PCA

3.2.4.1 Processing raw data

Before a PCA can be performed, the input data needs to be processed (Bro and Smilde, 2014). This is essential to make the variables comparable to each other. Autoscaling is the accepted tool used to achieve this. Autoscaling standardises the data and essentially ensures that each variable has an equal opportunity to be analysed and modelled.

Firstly, the mean is computed for each set of variables. This is then subtracted from the specific variable to attain a new value. This process is performed to attain the mean centre for the data. With this process completed, each new term is divided by the standard deviation of its corresponding set of variables. This scales the variables to ensure that they are all of similar magnitude.

3.2.4.2 Determining the required number of components

When using PCA one wants to determine a set of components that are smaller than the original number of variables, which still depict the data accurately. The various methods proposed by Bro (Bro and Smilde, 2014) include:

Exploratory studies

If one wants to identify the main variation within the data, then considering the first few components is sufficient. The components are nonetheless ranked in descending order of variation per direction, so by picking the first few components, most of the variation per direction will be accounted for. If one is to analyse the outliers, then it is important to quantify the number of outliers used.

Stellenbosch University https://scholar.sun.ac.za

Eigenvalues

Since the data is autoscaled, all variables are comparable to each other. From this, an eigenvalue of 1 means that its component describes 1 variable. Furthermore, if a component has an eigenvalue of greater than 1, it describes more than 1 variable. Thus, one could pick all the components with eigenvalues greater than 1, knowing that they simplify the original data and describe the variation of more than one variable.

Scree test

The Scree test assumes that relevant data is greater than noisy data. Furthermore, the more noise that exists in a dataset, the faster the eigenvalues of the components converge. Thus, once the eigenvalues of the components converge to relatively small values, it can be assumed that only noise is left and a sufficient number of components have been chosen.

Broken stick

This method is an extension of the Scree test and adds a broken line to the plot. The broken line shows the eigenvalues that would be expected for random data. Equation 23 outlines how the line is plotted, were J represents the number of pieces the line is broken up into.

$$br = \sum_{j=r}^{J} \frac{1}{j} \tag{23}$$

Fraction of variation

The number of components that are picked are linked to the amount of noise in the data and the variation described by the components. If the model has 1 % noise only, and the first few components describe 50 % of the data only, then more components are needed. Similarly, if a

Stellenbosch University https://scholar.sun.ac.za

model has 50 % noise and the first few components account for 90 % of the variation in the model, then the model is over fitted. In such a case it would be more beneficial to have fewer components, as the extra components are accounting for meaningless data.

Cross validation

Cross validation leaves out sections of the data that appears to contain outliers. This left-out data is then approximated for the rest of the PCA modal. If the estimated left-out section is independent of the actual left-out section, then one knows that the data was not linked. From this, the model is more reliable, in that over fitting becomes less probable.

3.2.4.3 Outlier detection

Outliers can simply be thought of as data that is unusual and does not match the same trend as the rest of the data (Auret, 2018). This can mean that the data is simply ‘bad’ or ‘wrong’ samples; however, this is not always the case. If the data is not ‘wrong’ samples and is a true and accurate reflection of what the data is representing, then more of these samples are necessary. If more of these ‘fake’ outliers are detected and incorporated into the input data, the accuracy of the output data is expected to increase.

The various methods well suited to PCA proposed by Bro (Bro and Smilde, 2014) for detecting outliers are included hereafter. If a more in-depth explanation is required, the reader is referred to *Principal component analysis* (Bro and Smilde, 2014).

Stellenbosch University https://scholar.sun.ac.za

Statistical regression model

Inspection of data

Once the input data has been auto-scaled, it can be analysed to detect outliers before the full PCA analyses commences. Generally, scatter plots and histograms are used to visualise the data and detect outliers.

Score plots

It can happen that outliers are missed in the inspection phase. The PCA method does luckily provide information later that also helps in the identification of outliers. This is typically represented by the score values, or the points plotted on the new principal axis. When points are identified that appear to be outliers, it is recommended to remove them and repeat the analysis. If the output data changes significantly compared to the original analysis, then the outlier data should be removed.

Hotelling's T^2

When analysing the scores as described in the score plots, it can become more problematic if the model has many components. The reason for this is that the data is difficult to visualise and only a handful of components are thus analysed at one moment. To solve this, the Hotelling's T^2 equation can be used. This equation uses confidence limits to show the range of values which are reliable. The score values plotted on the new component axes are represented by the T matrix. It is furthermore recommended to only use the equation for samples from the same population, as this leads to the most useable results.

$$Ti^2 = \frac{R(I-1)}{I-R} FR, I-R, \alpha \tag{24}$$

Stellenbosch University https://scholar.sun.ac.za

Score contribution plots

It can be interesting to investigate outliers once they have been identified, in order to find out what caused these strange values. These outliers can be broken down into their variables to determine what fraction of each variable they are composed of. To do this, a score contribution plot depicted in Equation 25 is used.

$$Cj^{D} = \sum_{r=1}^{R} \frac{tr^{new} xj^{new} pjr}{tr^{T} tr/(I-1)} \tag{25}$$

From Equation 25, the variables accounting for the extreme values is given. From this, the variable can be investigated for other samples, and a decision to remove the data point or entire variable can be made.

Lonely wolfs

Distance measurements can be done between the centre of points and their nearest neighbours. From this, the set of distances can be plotted to determine which are larger and potentially account for outliers.

Residuals

For this method of outlier detection, the sum squared residuals are plotted alongside the T^2 values to form the graph known as the influence plot. From this plot, outliers are easy to identify. Care must however be taken, as the number of components chosen influence both axes of the graph.

3.2.5 Summary

The PCA is an ideal way of simplifying a problem if there are many variables. The PCA method analyses all of these variables simultaneously to find what effect each one has. Furthermore, by

Stellenbosch University https://scholar.sun.ac.za

making use of several statistical proofs and concepts, the PCA manages to compare these variables while minimising the covariance and redundancy between them.

The PCA is designed to locate a new set of axes that display the greatest variance, and thus most information of a set of data.

3.3. Partial Least Squares

3.3.1 History

Regression by PLS was developed primarily by Wold in the 1960s (Geladi and Kowalski, 1985). Wold developed the regression method specifically for problems in the field of econometrics, but later refined the method to be applied in other fields of science (Geladi and Kowalski, 1985).

During development, it was found that PLS proved more robust than the then established PCA and MLR approaches (Wold et al., 1984). This robustness emanates from the model parameters showing little variation when having to adapt to new calibrations of the same model (Wold et al., 1984). Or more simply, if the method is applied to calibrated samples of the same original population, the original components determined by PCA will change very little. This also highlights that the model estimates a fairly accurate regression model after very few iterations.

3.3.2 Description

PLS is related directly to PCA. The primary difference is that PLS is performed on both the X and Y variables. From this, the score matrix for both the X and Y variables become the primary data matrices. These two matrices are then compared to each other to find a regression model relating them (Wold et al., 1984).

Stellenbosch University https://scholar.sun.ac.za

This analysis technique can be thought of as consisting of both an outer and inner component (Geladi and Kowalski, 1985). The outer relationship is performing the PCA on each of the X and Y datasets, while the inner relationship is linking these two datasets by finding a simple regression coefficient for each component in the corresponding score datasets (Geladi and Kowalski, 1985). The score dataset is the T matrix as described in Section 3.2.3.

This regression approach differs from others, in that the decomposition of both the X and Y datasets considers information from each other (Siong, 2013). Methods such as PCA only analyse the variance in the X dataset and then relates it to the Y dataset. PLS breaks X and Y up into its score and loading matrices, while considering information from the corresponding X or Y dataset.

These concepts will become more apparent in the succeeding sections when this method is explained visually.

3.3.3 Procedure

PLS regression consists of multiple steps. The first is standardising and organising the dataset. This is done by mean centering and scaling the data. Following this, the outer and inner relationships between the X and Y datasets are performed. With the outer and inner relationships available, the mixed relationship relating both the outer and inner parts is performed. This results in the regression model, which allows one to estimate y given multiple x-variables.

3.3.3.1 Mean centering and scaling

Mean centering and scaling is the same as the autoscaling procedure outlined earlier. From mean centering and scaling, the different variables become easier to compare with one another. As the different variables have different units, some may have very large values and others may have small values. Mean centering and scaling aids in modifying all the variables to have comparable

Stellenbosch University https://scholar.sun.ac.za

magnitudes. For example, if one variable is in meters and another is in kilometres, the values of the latter may be smaller due to the unit of measurement. Mean centering and scaling allows these two variables to be adjusted and scaled to have similar magnitudes, making it easier to compare (Geladi and Kowalski, 1985).

3.3.3.2 Outer relationship

The outer relationship of PLS is obtained by performing a PCA on both of the X and Y datasets. This also means that one can analyse more than one Y variable at a time. PCA breaks the X matrix up into the score and loading matrices, which allows one to perform dimension reduction. The loading matrix (P) expresses each of the principal components as loading factors of each of the x-variables. The score matrix is then simply the matrix multiplication of the X matrix and the loading matrix. This is the score or T matrix.

When the same procedure is performed on the y or dependent variables, the same score (U) and loading (Q) matrices are generated. This transformation is visually expressed in Figure 6.

This method is however imperfect, as the PCA analysis is performed separately on the X and Y datasets. Ideally one wants to provide information about each dataset to the other, while the analysis is being done (Geladi and Kowalski, 1985). Or more simply, when performing PCA on the X dataset, the scores from the Y dataset are considered and when performing PCA on the Y dataset, the scores from the X dataset are considered (Geladi and Kowalski, 1985). In order to achieve this, the t and u variables are swopped. This means that when PCA is performed on the X matrix t is swopped for u and vice versa for the Y matrix computation.

Stellenbosch University https://scholar.sun.ac.za

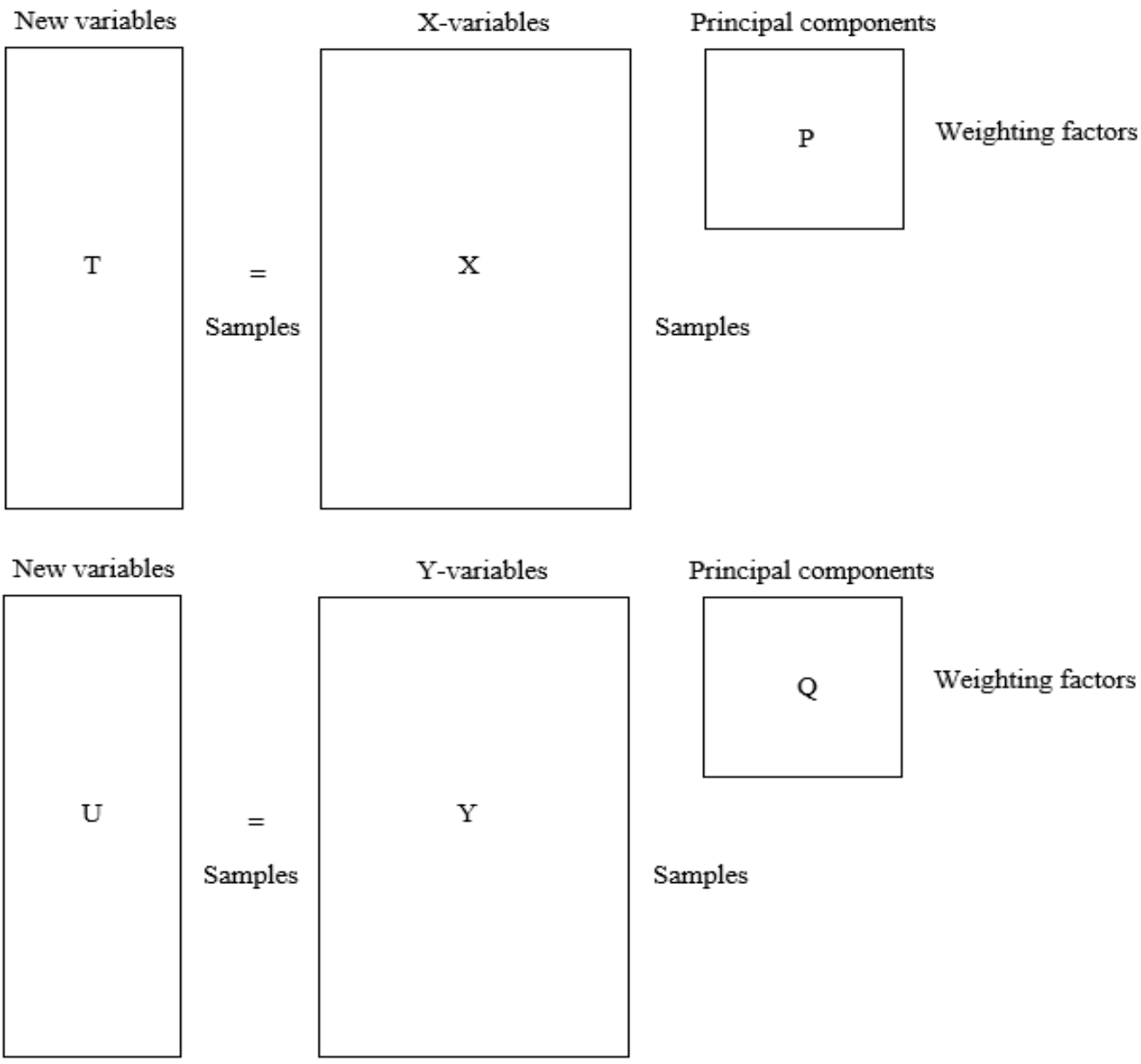


Figure 6: Outer relationship of PLS. The transformation of the X matrix to the T matrix using the P matrix and the transformation of the Y matrix to the U matrix using the Q matrix.

By incorporating the scores of the other PCA analysis, the new principal components are slightly rotated as information from the other dataset is considered. These new rotated matrices result in more accurate regression models (Geladi and Kowalski, 1985).

Stellenbosch University https://scholar.sun.ac.za

This process of swapping the t and u variables are summarised as follows:

For the X PCA analysis:

1. $t\ start = xj$
2. $p' = \frac{t\prime X}{t\prime t} = \frac{u\prime X}{u\prime u}$
3. $p'new = \frac{p\prime old}{\|p\prime old\|}$
4. $t = \frac{Xp}{p\prime p}$
5. Compare the t values in step 2 and 4. If they are equal, the iteration stops, otherwise continue iteration from step 2.

After the same procedure is performed on the Y matrix, the rotated components for both the X and Y datasets are found.

With these new rotated components, another problem arises, namely that the t values are no longer orthogonal. This means that the new principal components are not orthogonal to each other. This is due to the PCA's order of computations having changed (Geladi and Kowalski, 1985). Fortunately, by considering $p'new = \frac{p\prime old}{\|p\prime old\|}$ a new t can be found from $t = \frac{Xp}{p\prime p}$. This new t is however a scalar multiplication of the p found in step 2. Knowing this, it becomes clear that the t does not strictly have to be orthogonal. One can simply rescale the t and w terms using equation $t\ new = \frac{told}{\|p\prime old\|}$ and $'\ new = \frac{w\prime old}{\|p\prime old\|}$.

3.3.3.3 Inner relationship

Having found the score matrices for both the X and Y datasets, a regression analysis between the two matrices can be done. This is simply done by finding the regression coefficient that relates

Stellenbosch University https://scholar.sun.ac.za

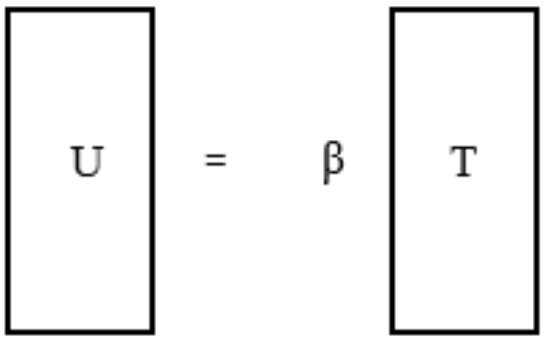


Figure 7: Inner relationship of PLS. Regression coefficient β relating matrix U to matrix T.

each variable in the T and U matrices (Figure 7). The regression coefficient β can be considered to be similar to the m variable in the familiar y = mx + c equation (Geladi and Kowalski, 1985).

$$U = \beta T \tag{26}$$

3.3.3.4 Mixed relationship

Having computed both the inner and outer relationships, these can be linked using the mixed relationship. This is done by combining the individual equations and forming one new model combining all the individual matrices and coefficients found.

$$\begin{aligned} Y &= UQ' \\ Y &= T\beta Q' \\ Y &= XP\beta Q' \end{aligned} \tag{27}$$

From this equation, an accurate prediction of Y can be made with the known x terms, β correlation coefficient and loading matrices P and Q.

Stellenbosch University https://scholar.sun.ac.za

3.3.4 Determining the required number of components

When formulating the final model, one can pick how to relate X and Y. For example, X and Y can be related linearly, exponentially etc. If a simple linear relationship exists, the number of components needed typically equals the number of dimensions (Geladi and Kowalski, 1985). That means that if one has five x-variables describing a y-variable, you will require five components in the equation. This is because all five components are necessary to describe the given y-variable.

If one wants to eliminate certain components to avoid noisy data or multicollinearity problems, certain tests and procedures can be followed. One such test is the F-test. The F-test simply plots the residuals' F value against the number of components. As the number of components increases, the residuals' F value decreases as the model is accounting for greater variance. One simply picks a threshold value and checks where the F value drops below that point. At that same point, the number of components can be read off (Figure 8).

Another method for determining the number of components needed, is the cross-validation PRESS method (Geladi and Kowalski, 1985). This method was explained earlier when picking the number of variables to consider in MLR.

3.3.5 Summary

PLS is a relatively new method of regression analysis. The method is closely related to PCA, but performs the same analysis on both the x and y datasets. It is this combined analysis that enables the method to consider the variance of one dataset while incorporating effects of the other dataset at the same time.

Stellenbosch University https://scholar.sun.ac.za

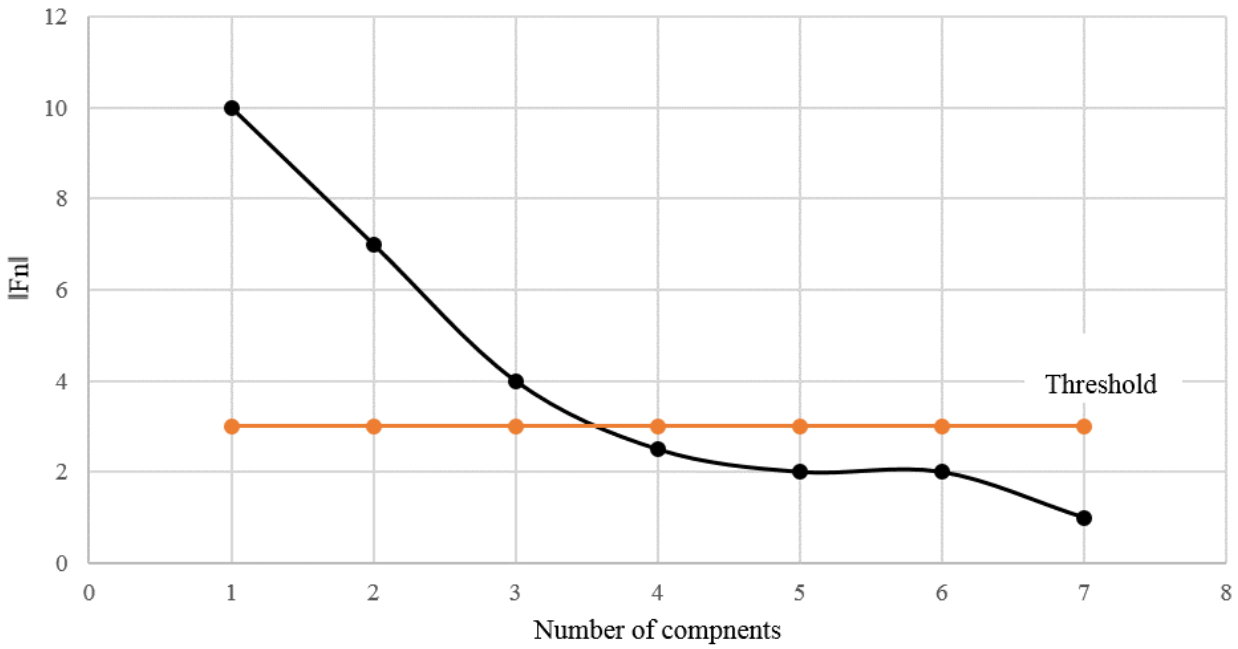


Figure 8: Required number of components from F-test graph. The necessary number of components can be found when the F value falls below the threshold.

This method can be computationally extensive and requires adequate software to perform the analyses. Furthermore, the user is expected to understand the inner workings of the method, as user input errors can easily occur.

PLS is an exciting method to use, as it incorporates multiple statistical theories and methods to present results that consider multicollinearity and redundancy. It is because of this that the method is ideal in regression analysis.

Stellenbosch University https://scholar.sun.ac.za

SECTION 4

METHODOLOGY

4.1. Overview

From the literature review, a clear overview of WDSs was given. Furthermore, the literature review investigated the various facets of WDSs, including how they are designed and the operating conditions that engineers need to consider before completing final designs.

One aspect that became apparent while investigating these WDSs was that initial master plans are typically done by experienced modellers relying on past experience and completed plans. No guide exists that gives engineers and city planners a rough outline of how a WDS should look, as there are multiple variables such as topography and cadastral layout. When interviewing companies that specialise in modelling WDSs such as GLS in Stellenbosch, South Africa, it became apparent that this really is the case. From this and thorough research into how other companies design these systems, it was decided to find a solution to this problem. That, in turn, led to the concept of a WDS capacity model that was originally introduced at the start of this thesis.

Having gained the necessary knowledge of WDSs, the next step would be deciding how to construct this urban capacity model and what steps would be required to achieve this.

It was recommended by Loubser that several existing hydraulic models would have to be analysed to create a sufficient sample space to be able to construct a meaningful regression model to represent a typical WDS. Furthermore, the WDSs analysed would have to be comparable to each other. Therefore, initially the hydraulic models had to be transformed in order to make them comparable to one another.

Stellenbosch University https://scholar.sun.ac.za

Once this had been completed, the variables that determine the capacity of a WDS had to be identified. With all these variables for each sample available, an accepted statistical approach had to be found that can compare the variables. After an interview with Auret (Auret, 2018), it was decided to use MLR, PCA and PLS as the three methods for attaining a regression model between the WDS components and its output. From this the capacity of the network could be computed by considering the network parameters, and vice versa.

With different regression models attained, each one would need to be tested against a set of testing data, which was reserved for testing from the onset. From this, the most accurate and precise regression model could be selected. This model would require multiple tests to ensure that it can be used safely and with an acceptable margin of safety in the industry, when designing urban WDSs.

4.2. Data acquisition

The WDS data was obtained from a consulting engineering company, GLS (GLS Consulting, 2018). GLS has specialised in analysing and planning of WDSs since 1989. It is because of their extensive lists of projects, mostly for water services providers in South Africa, that a large pool of WDS data was available.

The chosen networks were all located in South Africa, as most of these networks were probably designed according to the *Guidelines for human settlement, planning and design* (National Water Act, 1998) design requirements. The so-called ‘Red Book’ provides a set of guidelines for WDSs and details the requirements that civil engineers in South Africa must abide by.

The network models that were selected for analysis included several zones from the following water services providers:

Stellenbosch University https://scholar.sun.ac.za

- George Municipality
- City of Tshwane Metropolitan Municipality
- Ekurhuleni Metropolitan Municipality
- Mbombela Municipality

In the end, the model dataset included 165 individual water supply zones.

4.3. Standardising networks

Loubser suggested a way of standardising or transforming the WDSs. He recommended that the fixed constraints of the network should be the pressure head at the critical node and the maximum flow velocity in the pipe network. The critical node can be considered to be the node in the model that experiences the lowest supply pressure. The pressure head at the critical node represents the minimum pressure available to users within the network and thus shows when the network is at peak demand capacity.

From a previous master's degree study by Strijdom (Strijdom, 2016), it was found that a minimum operating water pressure for residential areas in South Africa can be considered to be 18 m, which is similar to the 20 m specified by *The Neighbourhood Planning and Design Guide* (Department of Human Settlements, 2019). The work done by Strijdom showed that an 18 m pressure head is sufficient for general house appliances such as washing machines and dishwashers to function effectively.

With these two fixed variables ensuring that the networks are functioning at full capacity, other variables needed to be chosen that can be manipulated to reach this equilibrium point. It was recommended that user demands be adjusted until the pressure head at the critical node was 18 m. As a theoretical capacity was being investigated, user water demands were manipulated until the network was functioning at full supply capacity. Once the user water demands corresponding to

Stellenbosch University https://scholar.sun.ac.za

this critical pressure head was found, the pipes where the flow velocity exceeded 2 m/s were adjusted to larger pipe diameters. If the water flow velocity in a single pipe exceeds 2 m/s, that pipe was replaced with a larger pipe which led to reduced flow velocity in that pipe. From this it was found that each of the 165 networks analysed had two fixed and two adjustable parameters that could ensure that they are comparable. The fixed parameters were the pressure head of 18 m at the critical node and the maximum flow velocity of 2 m/s. The two adjustable parameters were the user demands and the pipe diameters. This process of standardising models resulted in several months of constant iteration and adjustments to the networks. Two network examples will be explained to demonstrate the process.

Step 1

When opening a model, the user demands needed to be set to a standard. This was done by selecting the Annual Average Daily Demand (AADD) from the water meters and adding the unaccounted for water (UAW). The unaccounted for water includes the water lost due to leaks etc. This combined value was then multiplied by a peak factor. An outline of the different peak factors used by GLS for different land uses is presented in Table 1.

After considering the peak factor, a user-defined multiplication factor was added that could adjust all the demands in the WDS by the same proportion. The final water demands were calculated following Equation 28.

$$Total\ peak\ hour\ demand = (AADD + UAW) * Peak\ factor * User - defined\ factor \quad (28)$$

Stellenbosch University https://scholar.sun.ac.za

Table 1: Peak factors (GLS Consulting, 2018)

Predominant land use	AADD (kl/d)	PWF	PDF	PHF
Low cost housing (LCH)	<1000	1.50	1.90	3.60
	1000 - 5000	1.40	1.80	3.40
	5000 - 10000	1.35	1.70	3.30
	10000 - 15000	1.30	1.50	3.20
	15000 - 20000	1.25	1.40	3.10
	>20000	1.25	1.40	3.00
Residential (Csir)	<1000	1.80	2.20	4.60
	1000 - 5000	1.65	2.00	4.00
	5000 - 10000	1.50	1.80	3.60
	10000 - 15000	1.40	1.60	3.50
	15000 - 20000	1.35	1.50	3.30
	>20000	1.30	1.50	3.00
Business/Commercial/Industrial (BCI)	<5000	1.45	1.70	3.30
	5000 - 10000	1.30	1.60	3.15
	>10000	1.25	1.50	3.00
Large single consumers (LRG)	>500	1.45	1.70	2.50
Inner City CBD (CBD)	<5000	1.30	1.60	2.00

By applying a user-defined factor, all the user demands could be adjusted by the same proportion. What this means is that even if one WDS had multiple land uses with different peak factors, their final demands could be increased or decreased by the same proportion rather than by a fixed consumption. At the onset of the analysis, the user-defined factor was set to 1.0.

Step 2

Once the user demands had been adjusted considering a user-defined factor of 1.0, the simulation was performed. This process often took multiple minutes if there were many individual WDSs in one model. For example, Tshwane Central had over 50 individual WDSs, each with its own dedicated reservoir or reservoirs. It was because of these individual reservoirs that each network could be analysed individually. If adjustments were made to one WDS, it had no effect on the other WDSs, as each system had its own reservoir.

Stellenbosch University https://scholar.sun.ac.za

After the simulation was performed, a list of nodal pressure heads and pipe flow velocities could be attained. The nodal pressure heads were ranked in ascending order and the water flow velocities in each pipe were ranked in descending order. Take note that the current, and not future water demands were considered.

Step 3a (nodal pressure below 18 m)

If the lowest nodal pressure head was below 18 m, then the user-defined factor for all nodes in the WDS was decreased to a value of less than 1.0. As the user demands decrease, the pressure heads increase. This was done in increments of 0.05 until the pressure head at the critical node was 18 m.

After this, the flow velocities in the pipes were analysed. If the flow velocity in a pipe exceeded 2 m/s, then a new pipe diameter was selected by considering the continuity equation, represented by Equation 29. Velocity1 represented the water flow velocity over 2 m/s, with the current inner pipe radius set as r1. Velocity2 was set as 2 m/s and the equation was balanced to find a new r2 value. The pipe with a flow velocity exceeding 2 m/s was then replaced with a new pipe with inner radius r2.

$$Velocity1 * Area1 = Velocity2 * Area2$$
$$Velocity1 * \pi r1^2 = Velocity2 * \pi r2^2 \quad (29)$$

Step 3b (nodal pressure above 18 m)

If the pressure head at the critical node was above 18 m, then the user-defined factor for all nodes in the WDS was increased to a value above 1.0. This was again done in increments of 0.05 until the critical node had a pressure head of 18 m.

Stellenbosch University https://scholar.sun.ac.za

Once a value of 18 m was found, the pipes with a flow velocity exceeding 2 m/s were analysed. If the user demands were increased, it often resulted in water flow velocities increasing. From this, pipes were almost always adjusted to larger pipe diameters if the user-defined factor was increased.

Step 4

The simulation was then performed again to find the critical nodal pressure head and flow velocity. When pipe diameters were increased, it resulted in the nodal pressure heads increasing, as the pipe friction decreased. From this, nodal pressure heads that were originally adjusted to be 18 m, then increased to over 18 m. Step 3 would then be repeated by increasing the user-defined factor until the critical node's pressure head was 18 m and the pipe diameters adjusted until the maximum water flow velocity was below 2 m/s.

The user demands and pipe diameters were constantly adjusted in new iterations until the pressure head of 18 m at the critical node and the maximum flow velocity of 2 m/s, was achieved.

4.4. Modelling terrain

The individual nodes in each of the WDSs were used to model the terrain. This was done by using the x, y and z coordinates of the nodes and constructing a surface plot in Matlab. An example is illustrated in Figure 9. The reservoir location is specified by the red X.

A terrain model was constructed for all 165 WDSs. From these surface plots it became apparent that simply using the range between the highest and lowest nodal elevations would not be sufficient to describe the terrain. The reason for this is that the range would not account for the number of smaller hills and terrain fluctuations. For example, if an area has a small range, but has many small hills, then the range would not be sufficient in describing the terrain topography.

Stellenbosch University https://scholar.sun.ac.za

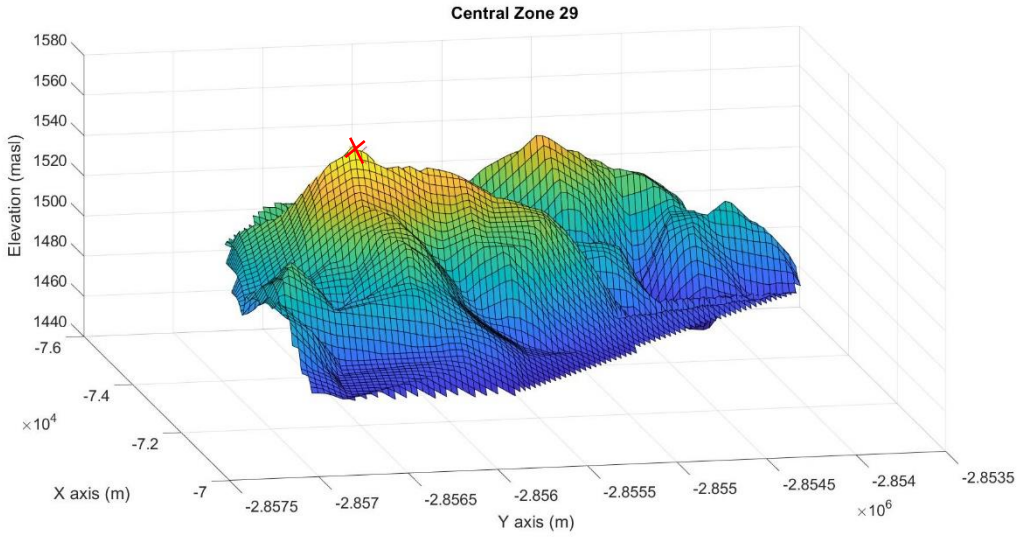


Figure 9: Terrain model in Matlab. The physical terrain profile of a WDS with the lower elevations illustrated in blue and the higher elevations illustrated in yellow.

From this it was decided to consider both the range and standard deviation of the node's elevations to rank the terrain types. By considering the standard deviation, the fluctuations of the nodal elevations were accounted for.

The process of construction of the terrain index follows:

1. Compute the range for each WDS;
2. Compute the standard deviation for each WDS;
3. Find the highest and lowest values for the range and construct 5 equal bins to span this range;
4. Find the highest and lowest values for the standard deviation and construct 5 equal bins to span this range.

By using the index tables represented in Table 2 and 3, each WDS could be classified. It was generally found that the range and standard deviation index value was the same for each individual WDS. For zones where this was not the case, an average value was taken to classify the terrain. For example, if a WDS had a range index value of 4 and a standard deviation index value of 3, the terrain was classified with an index value of 3.5.

Stellenbosch University https://scholar.sun.ac.za

Table 2: Terrain range index

Range bins (m)	Index	Colour
10 – 40	1	
41 – 70	2	
71 – 100	3	
101 – 130	4	
131 - 160	5	

Table 3: Terrain standard deviation index

Std dev bins (m)	Index	Colour
0 – 6	1	
6.1 – 12	2	
12.1 – 18	3	
18.1 – 24	4	
24.1 - 30	5	

4.5. Area of Water Distribution Systems

Shapefiles from Wadiso were used to represent each individual WDS zone's outline and total area. A problem which became apparent is that many zones contained large areas without any pipelines. This varied from zone to zone. Some zones contained pipelines for the whole area of the zone, while others only contained pipelines for a fraction of the total area. It was for this reason that the shapefiles could not be used to represent the area of a WDS, and another method had to be found.

To solve this, each area was considered as an ellipse. The reason for this was that none of the areas was perfectly circular or square in shape.

Firstly, only the area containing pipelines was considered. From this, a line joining the furthest most two points was drawn and measured. Another line was drawn perpendicular to this line at its midpoint. This second line again joined the corresponding two points furthest from each other.

Stellenbosch University https://scholar.sun.ac.za

With these two known lengths, Equation 30 was used to compute the area, assuming the area was an ellipse.

$$Area = \pi r1 r2 \quad (30)$$

This method was not ideal in representing the area of each water distribution zone, but proved to be consistent and provided a fast and user-friendly way to attain a rough area.

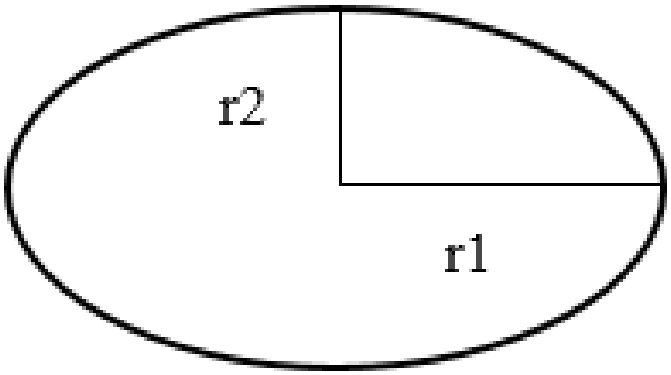


Figure 10: Ellipse area. The area of an ellipse can be computed by: πr1r2.

4.6. Final model parameters

Once all the WDS models had been transformed, the various model parameters could be extracted from each model. These included:

- Total peak hour demand (l/s)
- Total pipeline length (m)
- Total pipeline volume (m^3)
- Area (km^2)
- Reservoir distance from the centre of area (m)
- Reservoir elevation above mean terrain elevation (m)
- Land use

Stellenbosch University https://scholar.sun.ac.za

- Shape factor ratio
- Terrain index

A description of how each parameter was obtained follows:

Total peak hour demand: The total peak hour demand was taken as the sum of the demands at each of the individual nodes. This denotes the (AADD + UAW), multiplied by the peak factor and user-defined factor, and is essentially the peak water demand.

Total pipeline length: The total pipeline length was computed by summing the lengths of each of the individual pipes.

Total pipeline volume: To compute the total pipeline volume, Microsoft Excel (Excel) pivot tables were used. From this, the total pipe length of each pipe diameter could be found. With this known, the inside area of each pipe diameter could be computed and multiplied with the associated pipe length to attain a volume.

Reservoir distance from the centre of area: The centre of the area was considered to be the point where the r1 and r2 lengths used to compute the area, crossed. The distance of this point from the reservoir or supply point was then measured.

Reservoir elevation above mean terrain elevation: The mean elevation for each area was attained by considering the nodal elevations. Furthermore, the Wadiso models provided reservoir elevations with water levels. From this the difference in height between the reservoir water level and average nodal elevation was found. The reservoir elevation above the highest node was not considered as occasionally the models included nodes that were higher than the reservoir. To avoid unrealistic values, the average nodal elevation was used as an alternative.

Stellenbosch University https://scholar.sun.ac.za

Land use: This variable was given in the model output files. Some land uses included residential and Low-Cost Housing (LCH) areas.

Shape factor ratio: This was attained by dividing the r1 and r2 values used to attain the area. If the answer was 1, the WDS was more or less round in shape. Any value greater than 1 showed that the area was elongated to one side.

4.7. Removing outliers

Outliers are values which appear abnormal or extreme in magnitude compared to the rest of the observations (Santoyo, 2017). The removal of these values is a subjective process, as many values which appear to be outliers are often impartial values which are equally as important as the rest of the observations. Thus, the identification and analysis of these extreme values are imperative to assess whether they should be removed or not.

Outliers can result from multiple causes. These typically include human errors during collecting, sampling and measuring of values. True outliers not caused by human errors do exist and are the points we wish to identify. These are referred to as novelties (Santoyo, 2017).

Multiple outlier detection techniques were discussed in the literature review. Many of these techniques were specific to the type of regression analysis being performed. For example, when score plots are used in the PCA analysis, outliers can be identified when they are plotted onto the new principal components. Some other outlier detection techniques, which can be used on sample data include (Santoyo, 2017):

- Z-scores
- Linear regression model
- Visual inspection

Stellenbosch University https://scholar.sun.ac.za

- Scatter plots and histograms

The Z-score method as well as scatter plots were used to identify outliers for each of the model parameters. These methods were picked as these are commonly used and are statistically accepted ways of identifying outliers from datasets (Santoyo, 2017). The outliers were identified and removed before the three regression analyses were performed. This was done to ensure that each regression model considered the same data and that no additional outliers were removed at a later stage.

4.7.1 Z-scores

The Z-score method of accessing outliers checks how far points deviate from the mean. This method can only be used for data which follows a Gaussian distribution. Outliers can be identified as being three or more standard deviations from the mean (Montgomery et al., 2014).

The software programme language Python was used to check if the sample data for each model parameter followed a Gaussian distribution. It was found that this was not always the case. It was recommended to solve this problem by taking the log values of each of the sample values (Auret, 2018). By taking the log values, the distribution for each model parameter more closely resembled a Normal or Gaussian distribution. From this, the mean and standard deviation for the model parameter log values could be computed. With these values known, the model parameter log values could be scaled using Equation 31.

$$Z = \frac{x - \mu}{\sigma} \tag{31}$$

By considering the Z-scores of 3 and -3, it was assumed that 99.7% of the data would fall into this interval (Montgomery et al., 2014).

Stellenbosch University https://scholar.sun.ac.za

The Z-scores proved to be an easy and accurate method for identifying values which appeared extreme in magnitude. After analysing each of the model parameters, the total number of WDSs containing outliers, amounted to 15. The various outliers that had a Z-score of more than 3 or less than -3, are illustrated in yellow in Table 18 in Appendix 10.2. Table 18 illustrates the raw data of the analysis. As each dataset was assumed to follow a parametric distribution, a secondary method for outlier detection was also performed, namely visual inspection via scatter plots.

4.7.2 Scatter plots

Scatter plots were used as a visual aid in providing a secondary check to identify outliers. The scatter plots were ideal in representing values which were far higher or lower than the rest of the sample space. As the capacity model under development was based on real WDSs, a minimum number of outliers were removed. The reason for this was that the model had to simulate reality and that outliers do exist in real WDSs. Furthermore, the model would have to be applicable to any WDS, and thus would have to be able to make accurate summations and predictions even if the system is not necessarily 'ideal.' A further 5 WDSs were removed, as their transformation procedure was compromised and a maximum flow velocity of 2 m/s was not achieved. Thus, a total of 20 WDSs were removed.

4.8. Regression analysis

Once all the outliers had been removed, the dataset was ready for analysis. The original dataset shrank from 165 WDSs to 145. Another 30 WDSs were removed and reserved for testing, thus the remaining 115 models were used to develop each regression model. The WDSs in Table 18 in Appendix 10.2 from number 105 to 134, were reserved for testing and the data outlined in Table 19 in Appendix 10.3.1 was used to construct the regression models.

Stellenbosch University https://scholar.sun.ac.za

Three different regression techniques were considered. These included: MLR, PCA and PLS. Each of these methods was discussed in the literature review, outlining how they function and how the techniques are performed.

Regression models were developed for the general case where all 115 WDSs were analysed, as well as models specific to residential and LCH areas. This was done in order to make the models more specific to the land use of each WDS. Furthermore, the land use was the only qualitative model parameter and by separating the models into different land uses, every model parameter would be accounted for.

It was recommended to develop models where the only known model parameters include total user demand, total pipeline length, total pipeline volume, land use and area. The reason for this is that the total pipeline length and total pipeline volume are the fundamental parameters that the model needs to solve to determine its capacity. Also, when considering a development, the area is typically known. The other model parameters like reservoir elevation above mean terrain elevation or the terrain index value would typically require additional financial outlay, as these variables would need to be estimated using contour and boundary information.

For example, if a new housing development was to be designed, the designer would know the area of the development, the number of proposed housing units and the expected water consumption of each unit. Thus the total peak hour demand and area would be known. The total pipeline length and total pipeline volume would be the only unknown variables, and could be solved to determine the distribution of pipeline diameters which could indicate the construction costs.

In the end, models were developed using all of the chosen model parameters as well as by using the minimum number of parameters. The models with all the chosen parameters had seven x-variables and the models with the minimum number of parameters had three x-variables.

Stellenbosch University https://scholar.sun.ac.za

4.8.1 Multi Linear Regression

4.8.1.1 Considering all model parameters

The regression analysis was performed using Excel regression analysis tool. This tool simply asks the user to highlight the y-variables and x-variables and from this computes a regression model with an intercept and coefficient value for each x-variable. Excel performs the regression using the least-squares regression method. From this, a regression model was attained considering all of the model parameters. The total peak hour demands were considered to be the y-variable and the remaining model parameters were considered to be the x-variables.

4.8.1.2 Removing multicollinearity

The analysis was also performed after variables contributing to multicollinearity were removed. This was done using a Variance Inflation Factor or VIF test (Auret, 2018).

Excel correlation tool was used to compute the correlation between all the model parameters. From this, if a correlation value between two variables exceeded 0.8, then it was highlighted as having a strong correlation. This value of 0.8 was user-defined and recommended by Auret to represent a strong correlation (Auret, 2018).

Thereafter, a regression analysis was performed for each of these highlighted variables. In each case, one of the highlighted variables was chosen as the y-variable and the remaining predictor variables were considered as the x-variables. Excel regression tool was then again used to find the regression model. The output also includes the coefficient of multiple determination or R^2 statistic. The VIF was computed using Equation 32.

$$VIF(\beta j) = \frac{1}{1 - Rj^2} \tag{32}$$

Stellenbosch University https://scholar.sun.ac.za

This process was repeated for each of the variables with high correlation values. As discussed in the literature review, if a VIF value exceeding 4 is found then multicollinearity is present. In these instances, the variable correlating to the highest VIF value was removed from the analysis, on the basis of its VIF value exceeding 4.

Thereafter, the Excel regression tool was applied again, by considering the total peak hour demand as the y-variable and the remaining model parameters, excluding the variable with the high VIF value, as the x-variables.

This process of removing variables with multicollinearity was not performed for the analysis with only three x-variables, as it had been established by then that the three dominant variables were all necessary for computations.

4.8.1.3 Removing less significant variables

As discussed in the literature review, several processes exist which can be used to remove variables of lower significance. Some of these processes include the PRESS method and backward stepwise method. By removing these variables, noise within the dataset is removed (Auret, 2018). The remaining variables should, however, still accurately represent the model. The backward stepwise method was chosen to remove less significant variables, as it is commonly used in data analyses and can be performed in a quick and measurable way (Auret, 2018). A significance level of 0.05 was selected, as it is commonly used in data analysis (Auret, 2018).

The analysis started by performing a regression analysis on all the x-variables. After each analysis, the p-value of each variable was analysed. If this value exceeded 0.05, then the variable was considered as less significant. The variable with the highest p-value, considering that it exceeded 0.05, was removed. The regression analysis was then repeated using only the remaining variables. This process was repeated until no variables with p-values exceeding 0.05 was found. A summary

Stellenbosch University https://scholar.sun.ac.za

of the R^2 statistic for the different MLR models is tabulated in Table 4. An outline of each model with its intercept and coefficient values is available in Table 20 in Appendix 10.3.2.

Table 4: Multi Linear Regression summary

7 x-variables					
General land use		**Residential land use**		**Low-Cost Housing land use**	
Method	**Adj R^2**	**Method**	**Adj R^2**	**Method**	**Adj R^2**
Standard regression	0.744	Standard regression	0.768	Standard regression	0.837
Removing multicollinearity	0.608	Removing multicollinearity	0.770	Removing multicollinearity	0.817
Low p-values	0.593	Low p-values	0.779	Low p-values	0.823
3 x-variables					
General land use		**Residential land use**		**Low-Cost Housing land use**	
Method	**Adj R^2**	**Method**	**Adj R^2**	**Method**	**Adj R^2**
Standard regression	0.723	Standard regression	0.721	Standard regression	0.851
Low p-values	0.718	Low p-values	0.723	Low p-values	no value

4.8.2 Principal Component Analysis

The PCA considers all the x-variables and determines new axes that account for the greatest variance in the sample space. The new components are constructed by taking weighting factors of the x-variables, and thus represents a linear combination of all the x-variables.

A thorough explanation of PCA was provided in Section 3.2 and will not be repeated here. After the PCA analysis was performed, the x-variables were substituted into each principal component equation to transform the data. A MLR analysis was then performed on the transformed data.

The PCA was performed using an Excel add-in product named Analyse-it (Analyse-it, 2018). This product allows users to not only attain the principal component formulas, but also provides biplots and monoplots to aid in the interpretation of the relationship between variables. The program also indicates to users what percentage of the variance each principal component accounts for.

Stellenbosch University https://scholar.sun.ac.za

As there are various graphs and charts, for the purpose of this thesis, only one regression model will be presented. A detailed presentation for the model with seven x-variables and a general land use is included.

4.8.2.1 Correlation between variables

Analyse-it required the user to select the x-variables. The x-variables included total pipeline length, total pipeline volume, reservoir distance from the centre of area, reservoir elevation above mean terrain elevation, area, shape factor ratio and terrain index. From this, the programme could execute a PCA analysis. The first output was the correlation coefficient between the variables. This was a quick and easy way to see the strength of the relationship between the variables and patterns which existed. The correlation value ranged between -1 and 1, describing a negative linear relationship and a positive linear relationship (Analyse-it, 2018). A correlation value of 0 indicates that no correlation or linear relationship exists. Table 5 shows the relationship between the different x-variables. A positive relationship is represented by a blue cell and a negative relationship with a red cell. Furthermore, the magnitude of the value is represented by the intensity of the colour, as well as, obviously, by the value of the cell.

From Table 5, a few strong linear relationships become apparent. A strong correlation exists between total pipeline length and total pipeline volume, with a correlation coefficient of 0.884. Furthermore, both of these variables are also strongly correlated with area. The reservoir elevation above mean terrain elevation and shape factor ratio, appears to have a very weak correlation to all the other variables.

Stellenbosch University https://scholar.sun.ac.za

Table 5: Correlation between variables

Pearson's r	Reservoir elevation above mean terrain elevation	Total pipeline length	Total pipeline volume	Reservoir distance from the centre of area	Shape factor ratio	Area	Terrain index
Reservoir elevation above mean terrain elevation	-	0.251	0.331	0.345	-0.081	0.325	0.637
Total pipeline length	0.251	-	0.884	0.429	-0.041	0.850	0.215
Total pipeline volume	0.331	0.884	-	0.512	-0.114	0.851	0.311
Reservoir distance from the centre of area	0.345	0.429	0.512	-	-0.056	0.378	0.204
Shape factor ratio	-0.081	-0.041	-0.114	-0.056	-	-0.159	-0.045
Area	0.325	0.850	0.851	0.378	-0.159	-	0.366
Terrain index	0.637	0.215	0.311	0.204	-0.045	0.366	-

Table 6: Principal components

	Principal components						
	1	**2**	**3**	**4**	**5**	**6**	**7**
Reservoir elevation above mean terrain elevation	-0.309	-0.623	0.024	-0.078	0.713	-0.037	-0.014
Total pipeline length	-0.470	0.338	0.091	0.161	0.136	0.291	0.727
Total pipeline volume	-0.496	0.242	0.024	0.068	0.016	0.508	-0.657
Reservoir distance from the centre of area	-0.335	0.001	0.086	-0.894	-0.250	-0.121	0.065
Shape factor ratio	0.087	0.020	0.989	0.060	0.024	-0.068	-0.068
Area	-0.483	0.203	-0.042	0.285	-0.043	-0.787	-0.142
Terrain index	-0.290	-0.631	0.053	0.281	-0.639	0.131	0.104

Stellenbosch University https://scholar.sun.ac.za

4.8.2.1 Principal components

The principal components are presented in Table 6. Each column represents a principal component with the linear combination of x-variables that forms it. The magnitude of the weighting factors shows how much each x-variable contributes to the principal component. It is evident that the first principal component has relatively large weighting factors for all the x-variables, except for that of the shape factor ratio. Thus, the shape factor ratio does not contribute as much as the other variables to the first principal component.

The first principal component can be expressed as:

$$\begin{aligned} PC1 = {} & -0.309 \\ & * Reservoir\ elevation\ above\ mean\ terrain\ elevation \\ & - 0.470 * Total\ pipeline\ length - 0.496 \\ & * Total\ pipeline\ volume - 0.335 \\ & * Reservoir\ distance\ from\ the\ centre\ of\ area \\ & + 0.087 * Shape\ factor\ ratio - 0.483 * Area - 0.290 \\ & * Terrain\ index \end{aligned} \tag{33}$$

4.8.2.2 Percentage of variance explained

The cumulative percentage variance that each principal component accounts for is presented in Table 7. The data was originally standardised to have a variance of one. Therefore, from the variance column it is apparent that the first principal component explains the variance in over three of the original x-variables. The second principal component explains the variance for more than one variable. From the third to the seventh principal component, variance in less than one variable is explained. Notably, the variance column should add up to the number of original x-variables.

Stellenbosch University https://scholar.sun.ac.za

Table 7: Percentage variance explained by principal component

Component	Variance	Proportion	Cumulative proportion
1	3.419	0.488	0.488
2	1.277	0.182	0.671
3	0.991	0.142	0.812
4	0.733	0.105	0.917
5	0.342	0.049	0.966
6	0.136	0.019	0.985
7	0.102	0.015	1.000

The first principal component accounts for 48.8 % of the variance in the dataset. The first two principal components account for 67.1 % variance and so forth. If seven principal components are used, 100 % of the variance in the seven x-variables are accounted for.

Correlation monoplot

The correlation monoplot expresses similar information as the correlation matrix. The monoplot enables the visual presentation of the relationship between variables. Each vector represents one of the x-variables. The angle between these vectors illustrates the correlation between them. A small angle represents a strong positive correlation, an angle of 90 degrees represents a correlation of 0 and an angle of 180 degrees represents strong negative correlation or a correlation of -1. For example, the angles between the total pipeline length, total pipeline volume and area vectors are small, showing that the three variables are strongly positively correlated. Lastly, the length of the vector shows how well it is described, or explained. For example, the shape factor ratio variable is short, and is therefore not well represented. The shape factor ratio should thus not solely be used to make assumptions about the model.

Stellenbosch University https://scholar.sun.ac.za

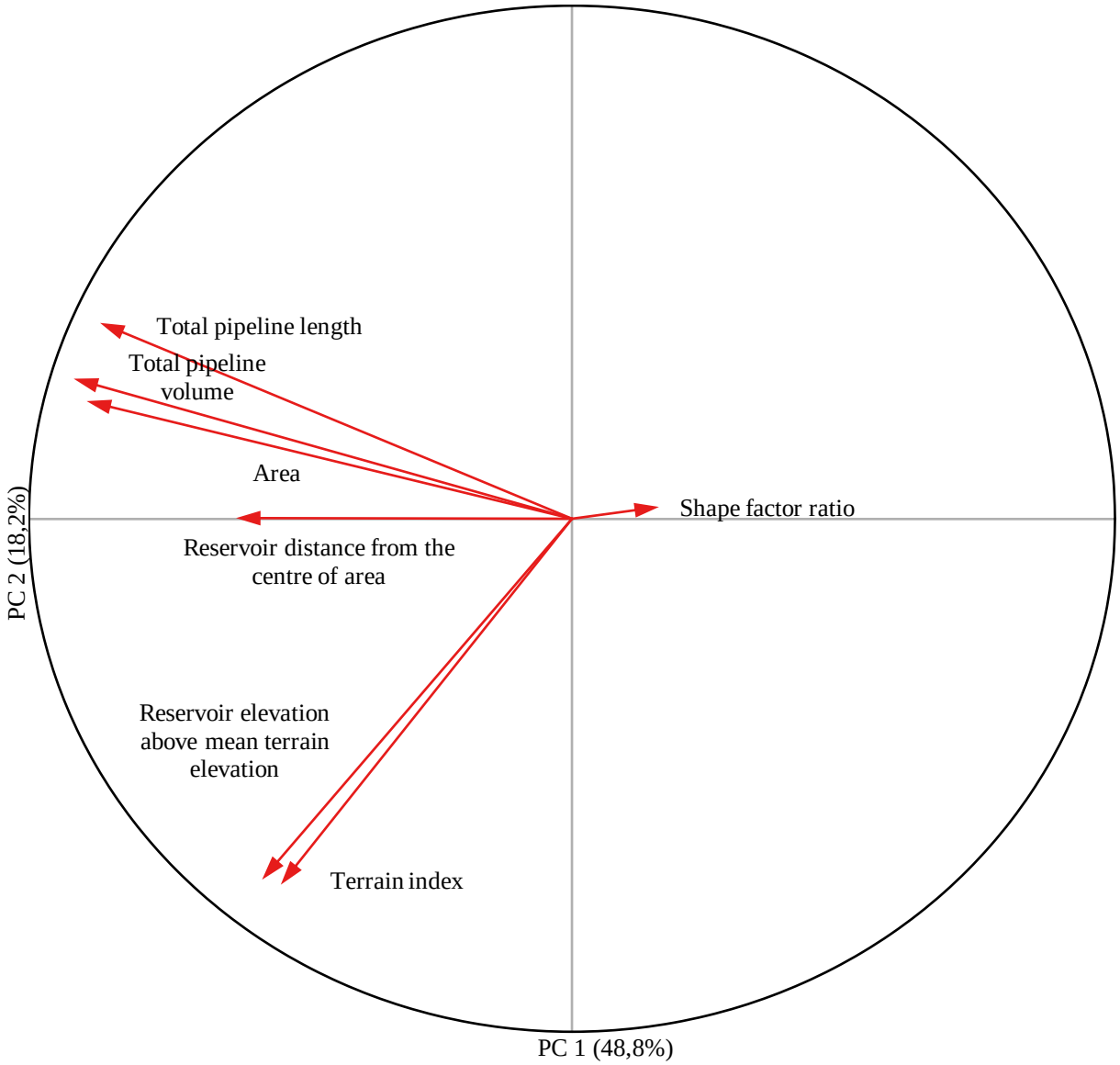


Figure 11: PCA Monoplot. Each x-variable is represented by a vector in red. The smaller the angle between vectors, the stronger their correlation.

Biplot

The monoplot in Figure 11 describes the relationship between the various x-variables. To describe the similarities between the x-variables, a biplot is typically used. In this method of visualisation, each x-variable is presented as an axis and shows the various observed data points in space. A biplot of the data is presented in Figure 12. The biplot in Figure 12 is a two-dimensional representation of the sample space as if it were represented in a seven-dimensional space.

Stellenbosch University https://scholar.sun.ac.za

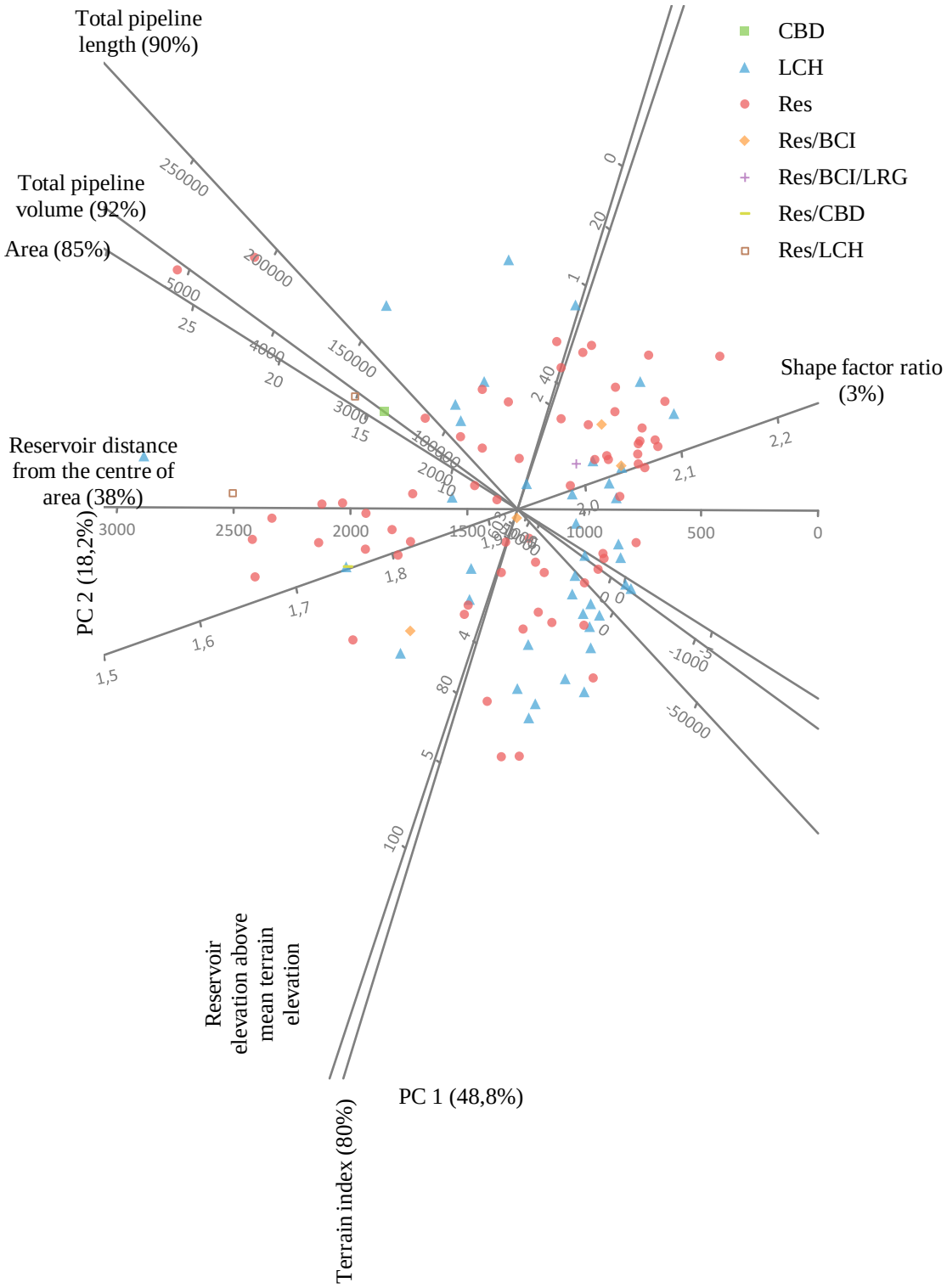


Figure 12: PCA Biplot. Each x-variable is represented by an axis, with the data points in space. As there are 7 x-variables, the data points are in a 7-dimensional space.

Stellenbosch University https://scholar.sun.ac.za

Like with the monoplot, similar observation points are close to each other. Also, the greater the similarity between variables, the smaller the distance between the lines.

Rotated biplot

Biplots can also be rotated so that the x-variable which is best described, is presented horizontally (Figure 13). From the rotated biplot it becomes easier to see each of the individual observations.

4.8.2.3 Results

The R^2 statistic results from the PCA analysis are presented in Table 8. The complete formula associated with each principal component and land use type is presented in Table 21 and 22 in Appendix 10.3.3.

Table 8: Principal Component Analysis results

7 x-variables					
General land use		**Residential land use**		**Low-Cost Housing land use**	
Method	**Adj R^2**	**Method**	**Adj R^2**	**Method**	**Adj R^2**
Standard regression	0.744	Standard regression	0.768	Standard regression	0.837
Low p-values	0.739	Low p-values	0.775	Low p-values	0.847
3 x-variables					
General land use		**Residential land use**		**Low-Cost Housing land use**	
Method	**Adj R^2**	**Method**	**Adj R^2**	**Method**	**Adj R^2**
Standard regression	0.723	Standard regression	0.721	Standard regression	0.837
Low p-values	no value	Low p-values	0.724	Low p-values	0.847

4.8.3 Partial Least Squares

PLS regression is based on PCA. PLS performs a PCA analysis on both the x dataset and y dataset. From this, a scoring matrix for both the x dataset and y dataset is attained. These two matrices are then compared to find a coefficient that relates them. A broader outline of PLS was presented in Section 3.3. This method is ideal for dealing with datasets with more than one y-variable.

Stellenbosch University https://scholar.sun.ac.za

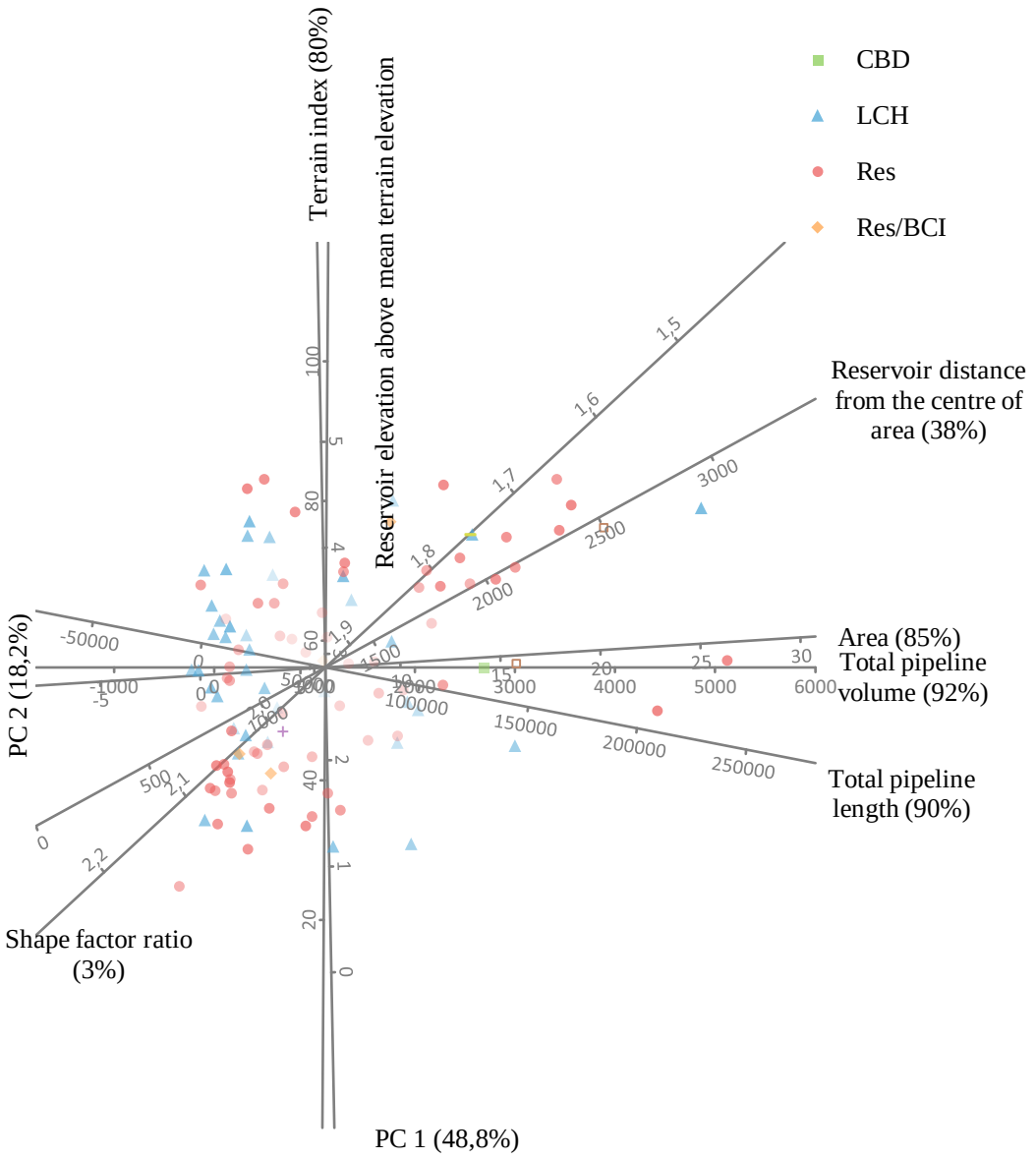


Figure 13: PCA Rotated biplot. Each x-variable is represented by an axis, with the most descriptive x-variable represented by the horizontal axis.

PLS regression was computed using an Excel add-in named Xlstat (Xlstat, 2018). This programme only required the user to select the respective x and y datasets. The output included various graphs and tables describing the quality and accuracy of the model.

Only the model considering all seven x-variables for the general land use will be discussed here. The remaining model outputs can be found in Table 23 in Appendix 10.3.4.

Stellenbosch University https://scholar.sun.ac.za

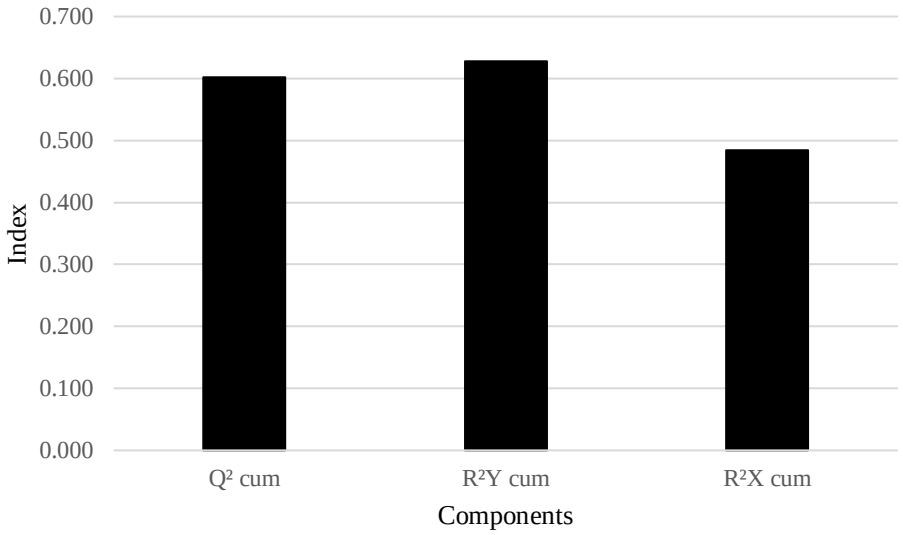


Figure 14: Model quality by number of components. The R^2 value for the y-variable is represented by the middle column, while the R^2 value for the x-variables is represented by the right column.

4.8.3.1 Quality of model

As the model only had one y-variable, the Xlstat output only included one component. Figure 14 presents the quality of this model, with a R^2 value of 0.628 for the y-variable and a R^2 value of 0.484 for the x-variables. Neither had a particularly strong R^2 value.

4.8.3.2 Correlation between variables

Figure 15 presents the correlation between the x and y-variables. The x-variables are represented by red points and the y-variable by a blue point. It is clear that the shape factor ratio has a weak correlation with the total peak hour demand, as it is the furthest x-variable from y.

4.8.3.1 VIP plot

The programme Xlstat also provides users with a VIP plot. The VIP plot is presented in Figure 16. If the variable has a VIP value exceeding one, then it is considered important. It is clear that total pipeline volume, total pipeline length and area have VIP values exceeding one, and are thus the most important variables in finding a correlation with y.

Stellenbosch University https://scholar.sun.ac.za

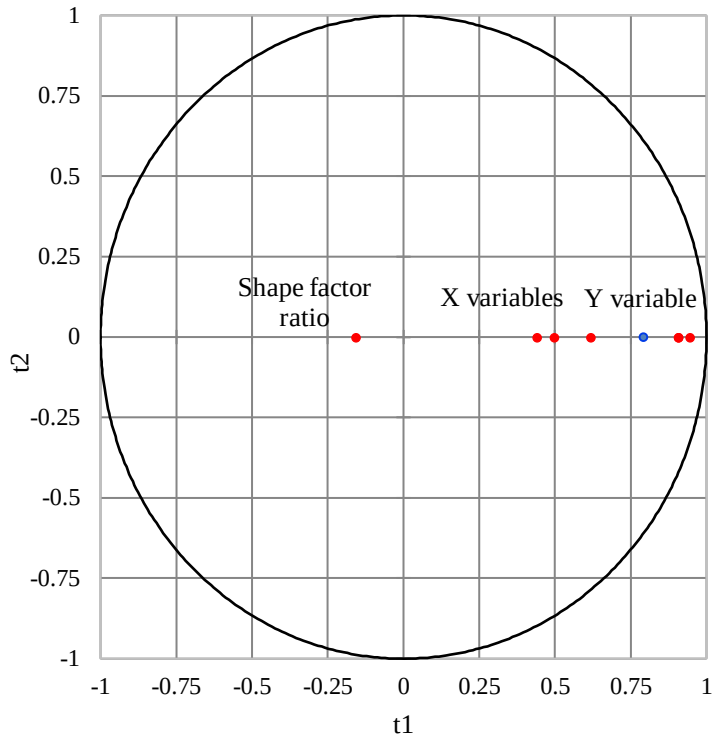


Figure 15: Correlation between the x and y variables. The greater the horizontal distance between points, the weaker their correlation.

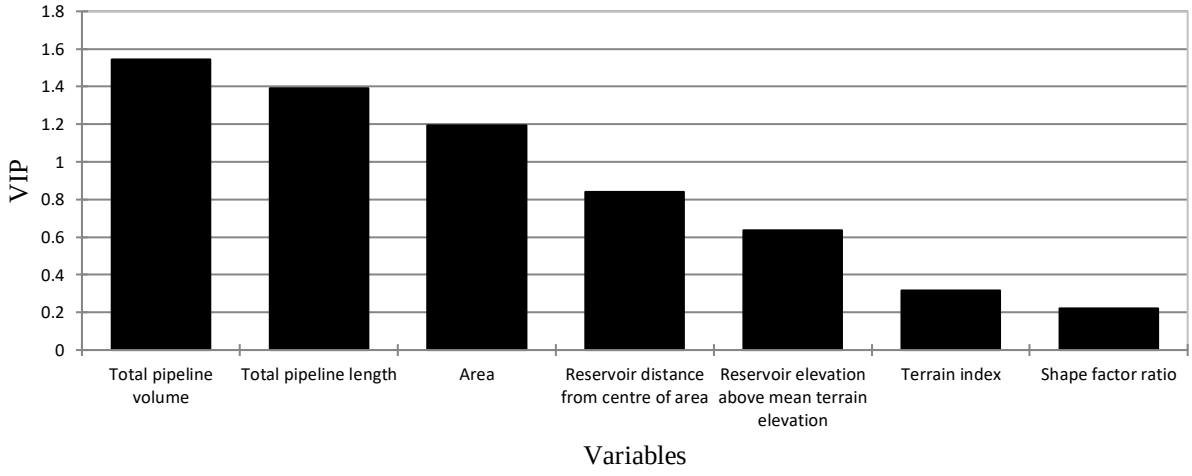


Figure 16: VIP scores. The VIP value for each x-variable is illustrated by the columns. A VIP value exceeding 1 shows that an x-variable is important.

Stellenbosch University https://scholar.sun.ac.za

4.8.3.2 Model parameters

The output showing the model parameters that describe the regression model are presented in Table 9.

Table 9: Model parameters for PLS regression

Variable	Output
Intercept	-69.804
Reservoir elevation above mean terrain elevation	1.319
Terrain index	9.961
Total pipeline length	0.001
Total pipeline volume	0.041
Reservoir distance from the centre of area	0.034
Shape factor ratio	-10.059
Area	6.089

4.8.3.3 Model relatability

Xlstat also provides a plot of the predicted values versus the observed values. Ideally, a linear relationship is desired as it shows that the prediction model closely mirrors the observed values. It is apparent from Figure 17 that a perfect linear relationship does not exist. Furthermore, it is evident that the residuals, or difference between the observed and estimated values, are quite large in many cases. This shows that the model had a low degree of accuracy. The coefficients of multiple determination for the analysis is presented in Table 10. It is evident that the R^2 values yielded by PLS are notably lower when compared to the other regression techniques, such as MLR. The reason for this is that PLS is typically used for finding regression models with multiple y-variables. This is however alright, as it was the expected outcome and both the MLR and PCA methods yielded good results.

Stellenbosch University https://scholar.sun.ac.za

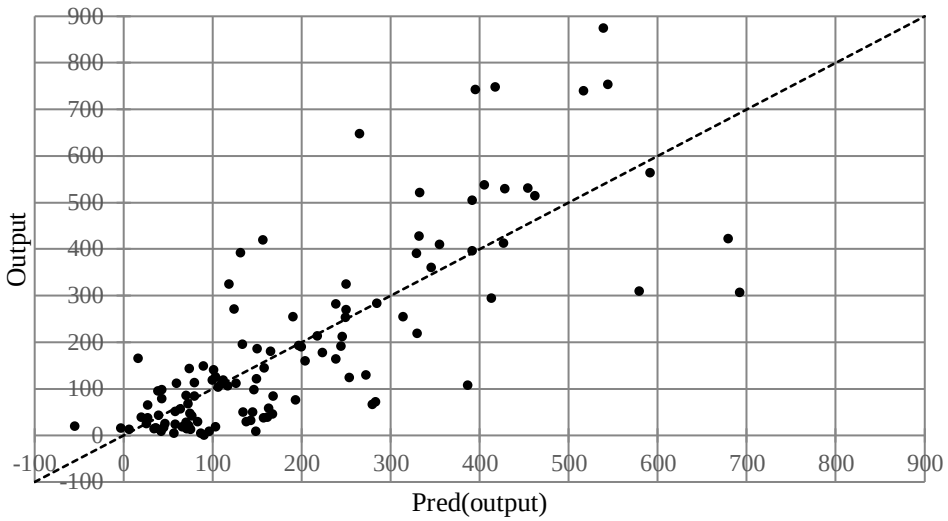


Figure 17: Estimated versus observed values. The estimated y-values from the PLS model is compared to the observed y-values.

Table 10: Partial Least Squares Regression

7 x-variables					
General land use		**Residential land use**		**Low-Cost Housing land use**	
Method	**Adj R^2**	**Method**	**Adj R^2**	**Method**	**Adj R^2**
PLS	0.628	PLS	0.696	PLS	0.670
3 x-variables					
General land use		**Residential land use**		**Low-Cost Housing land use**	
Method	**Adj R^2**	**Method**	**Adj R^2**	**Method**	**Adj R^2**
PLS	0.631	PLS	0.700	PLS	0.670

4.9. Summary

A structured and logical way to transform the WDSs was found. These transformed hydraulic models could potentially be used in the development of a regression model to find a relationship between the total peak hour demand of a network and the network characteristics. Once all the hydraulic models had been transformed and manipulated to be functioning at the same peak demand level while yielding similar outputs, network characteristics that appeared abnormal had to be removed. These outliers were identified using Z-scores, were Z-scores of 3 and -3 indicated outliers.

Stellenbosch University https://scholar.sun.ac.za

Furthermore, 20 % of the data was removed specifically for testing. This was recommended by Auret and was necessary to ensure that any models developed could be tested and validated for accuracy.

Three different regression methods were used to find the regression model that most accurately predicted the linear relationship between total peak hour demand and the network characteristics. Microsoft Excel add-ins in the form of Analyse-it and Xlstat aided in the computation of the PCA and PLS regression approaches.

By using three different regression approaches, the most precise and accurate regression model could be selected to ensure that the model can be used with confidence and safety in the engineering workplace.

Stellenbosch University https://scholar.sun.ac.za

SECTION 5

TRANSFORMED DATA

5.1. Summary of regression models

Table 11 presents a summary of the results for all the regression models. The adjusted coefficients of multiple determination or R^2 values were used to test and quantify the quality of each model and measure the fraction of variability that each model accounted for. The adjusted R^2 values were used, so that all the models were analysed in an objective manner, regardless of the number of variables used in the model. The regression models that resulted in the highest R^2 statistic is highlighted in bold.

From Table 11 it is evident that the MLR standard regression model that considered all seven x-variables, had the highest R^2 value for a general land use type. The MLR model that only considered three x-variables had a similarly high R^2 value, differing by only 2.1 %. Both of these methods yielded models with a fairly good fit, as the adjusted R^2 values were close to unity.

The adjusted R^2 values did increase when the residential and LCH areas were considered separately. The most notable increase was for the LCH, as the adjusted R^2 values increased by more than 10 % compared to the general land use type. It must be noted however, that only 39 models were considered in the LCH regression analysis. This smaller sample space could account for the increased R^2 values as there was less variability within the smaller dataset.

Before making any final recommendation on which model to use, each model was tested on a dataset of 30 models reserved for testing.

Stellenbosch University https://scholar.sun.ac.za

Transformed data

Table 11: Results for all regression models using transformed data

7 x-variables				
Land use		**General**	**Residential**	**Low-Cost Housing**
Method		**Adj R^2**	**Adj R^2**	**Adj R^2**
MLR	Standard regression	**0.744**	0.768	0.837
	Removing multicollinearity	0.608	0.770	0.817
	Low p-values	0.593	**0.779**	0.823
PCA	Standard regression	0.744	0.768	0.837
	Low p-values	0.739	0.775	**0.847**
PLS	PLS	0.628	0.696	0.670
3 x-variables				
Land use		**General**	**Residential**	**Low-Cost Housing**
Method		**Adj R^2**	**Adj R^2**	**Adj R^2**
MLR	Standard regression	**0.723**	0.721	**0.851**
	Low p-values	0.718	0.723	no value
PCA	Standard regression	0.723	0.721	0.837
	Low p-values	no value	**0.724**	0.847
PLS	PLS	0.631	0.700	0.670

5.2. Testing model

Each of the regression models was tested using 30 test datasets. These hydraulic models were reserved for testing from the onset and are completely independent from the datasets from which the regression models were derived. The results are tabulated in Table 12.

The various regression models were tested for general and residential land use types. The LCH land use type was removed, as only five of the 30 test datasets were LCH areas. The adjusted R^2 values were relatively low when the regression models were applied to the test data. On closer inspection, it was learnt that specific models that performed the best in the development phase did not necessarily perform the best on the test data. For example, the standard MLR model that considered all seven x-variables for the general land use case did not yield the highest adjusted R^2 value during testing. The model that removed multicollinearity with VIFs and the model that removed variables with large p-values, yielded far higher adjusted R^2 results. This could indicate that there was multicollinearity in the test data and that some of the variables were redundant.

Stellenbosch University https://scholar.sun.ac.za

Table 12: Test results using transformed data

7 x-variables			
Land use		**General**	**Residential**
Method		**Adj R^2**	**Adj R^2**
MLR	Standard regression	0.529	0.495
	Removing multicollinearity	0.567	0.486
	Low p-values	0.620	0.468
PCA	Standard regression	0.529	0.495
	Low p-values	0.503	0.508
PLS	PLS	0.502	0.530
3 x-variables			
Land use		**General**	**Residential**
Method		**Adj R^2**	**Adj R^2**
MLR	Standard regression	0.479	0.511
	Low p-values	0.448	0.500
PCA	Standard regression	0.479	0.510
	Low p-values	no value	0.506
PLS	PLS	**0.589**	0.472

The PLS regression model performed the best with an adjusted R^2 value of 58.9 %. This shows that the model accounted for almost 60 % of the variability in the dataset. This value is however not particularly high and does not give a conclusion of which is the most reliable model to use.

Stellenbosch University https://scholar.sun.ac.za

UNTRANSFORMED DATA

Due to the inconclusive results obtained during the testing phase of the regression models, the overall concept of the WDS capacity model was re-evaluated. When transforming the hydraulic models, the total peak hour demand and pipe diameters were manipulated to satisfy the condition of a pressure head of 18 m at the critical node and a maximum flow velocity of 2 m/s. When the model is represented in a mathematical way, the flaws of this original concept become apparent.

$$\begin{aligned} Total\ peak\ hour\ demand = {} & \beta 0 + \\ & \beta 1 * Total\ pipeline\ length + \\ & \beta 2 * Total\ pipeline\ volume + \\ & \beta 3 * Reservoir\ elevation\ above\ mean\ terrain\ elevation + \\ & \beta 4 * Reservoir\ distance\ from\ the\ centre\ of\ area + \\ & \beta 5 * Area + \\ & \beta 6 * Shape\ factor\ ratio + \\ & \beta 7 * Terrain\ index \end{aligned} \tag{34}$$

During the transformation phase, the total peak hour demands were adjusted by a factor until the pressure head at the critical node was 18 m. Furthermore, the pipes which had a flow velocity exceeding 2 m/s were increased in diameter until an equilibrium point was reached. This equilibrium point represented the state of the WDS at full capacity. With reference to Equation 34 and from experience while performing the standardisation exercise on all the hydraulic models, it has to be noted that the total peak hour demand term was adjusted substantially to reach the state of full capacity. The only other variable which changed during the transformation exercise was the

Stellenbosch University https://scholar.sun.ac.za

total pipeline volume. The pipes where the water flow velocity exceeded 2 m/s were changed by increasing diameters; thus only a handful of pipes were increased in size in each case. It was realised that the network characteristics or x-variables did not change significantly when the network was transformed. The pipes that were adjusted had a small impact on the total pipeline volume as only a handful of pipe diameters were changed. In contrast, when adjusting the total peak hour demands, all of the demands were adjusted. In summary, during the standardisation or transformation procedure, the independent variables did not change significantly, but the resultant change to the dependent variable, namely the total peak hour demand, was great.

For this reason, and at this point, it was decided that the entire analysis had to be repeated on the untransformed hydraulic models, as they were before the standardisation exercise. The untransformed hydraulic models also represent real-world WDSs, and not WDSs that are forced to supply their theoretical maximum capacity. In essence, these models would in many instances have critical nodes where the lowest pressure head exceeds 18 m.

The differences between the standardised data and original data were discussed with Auret (Auret, 2018). It was recommended to use the original untransformed dataset. In order to develop a prediction model, real-world recordings would provide a more accurate and reliable dataset. Considering this, the analysis was repeated, using the hydraulic models before standardisation.

6.1. Summary of regression models

The MLR, PCA and PLS regression analyses were subsequently repeated on the original untransformed data. This dataset can be found in Table 24 in Appendix 10.4.1. These methods have already been discussed, and therefore only a summary of the adjusted R^2 values for each is presented in Table 13. The full computation for each regression model is attained in Table 25 to 28 in Appendix 10.4.

Stellenbosch University https://scholar.sun.ac.za

Untransformed data

Table 13: Results for all regression models using untransformed data

7 x-variables				
Land use		**General**	**Residential**	**Low-Cost Housing**
Method		**Adj R^2**	**Adj R^2**	**Adj R^2**
MLR	Standard regression	0.889	0.912	0.918
	Removing multicollinearity	0.827	0.884	0.838
	Low p-values	0.829	0.881	0.840
PCA	Standard regression	0.889	0.912	0.918
	Low p-values	**0.891**	**0.914**	**0.923**
PLS	PLS	0.763	0.809	0.799
3 x-variables				
Land use		**General**	**Residential**	**Low-Cost Housing**
Method		**Adj R^2**	**Adj R^2**	**Adj R^2**
MLR	Standard regression	**0.890**	**0.909**	**0.921**
	Low p-values	No value	0.905	0.918
PCA	Standard regression	0.890	0.909	0.921
	Low p-values	No value	0.906	No value
PLS	PLS	0.814	0.871	0.786

From Table 13 it is clear that the adjusted coefficients of multiple determination improved significantly when analysing the untransformed model data. The regression models that yielded the highest R^2 value for each of the different scenarios are highlighted in bold. These models all had an adjusted R^2 value of approximately 90 %. Furthermore, it is important to note that the regression model with the highest adjusted R^2 value was the same when considering all seven x-variables or only the essential three x-variables for the general land use. This shows that it was unnecessary to consider all the variables when deriving a model. The regression models attained when only the total pipeline length, total pipeline volume and area were considered had a fit which was equally good to the regression models that considered all seven of the physical network characteristics.

When analysing the standardised models, the regression models had far higher adjusted R^2 values for the LCH land use compared to the general land use. This was not the case when analysing the

Stellenbosch University https://scholar.sun.ac.za

untransformed model data. The quality of the fit stayed relatively constant for all three land uses, ranging by only 3 %.

6.2. Testing models

The various regression models developed were tested on the original test dataset of 30 samples. The resultant adjusted R^2 values are presented in Table 14. The LCH analysis was excluded as only five samples had this land use, resulting in inconclusive results.

The adjusted R^2 values were far higher when applying the regression models on the original untransformed model data. Furthermore, for the general land use, the specific models which yielded the best fit when developing the regression models were the same models with the highest R^2 values when testing. This is ideal, as these regression models consistently yielded reliable results. This was the PCA model using seven x-variables with low p-values and the standard MLR model for the case with three x-variables.

There was however a significant drop in the adjusted R^2 values between the analysis data and the test data for the general land use. This drop is comparable to that found on the analysis of the standardised models. The adjusted R^2 dropped by approximately 16 % for the general land use case.

The adjusted R^2 for the analysis data and test data did not change significantly for the residential land use. On closer inspection, it was however found that this regression model over predicted the y-variable by 50 percent. Therefore, although the model had a high R^2 value, the model predicted y values that were far higher than the observed y values.

Stellenbosch University https://scholar.sun.ac.za

Untransformed data

Table 14: Test results using untransformed data

7 x-variables			
Land use		**General**	**Residential**
Method		**Adj R^2**	**Adj R^2**
MLR	Standard regression	0.729	0.885
	Removing multicollinearity	0.618	0.848
	Low p-values	0.617	0.874
PCA	Standard regression	0.729	0.885
	Low p-values	**0.728**	0.882
PLS	PLS	0.618	**0.923**
3 x-variables			
Land use		**General**	**Residential**
Method		**Adj R^2**	**Adj R^2**
MLR	Standard regression	**0.724**	0.913
	Low p-values	No value	0.922
PCA	Standard regression	0.724	0.913
	Low p-values	No value	0.921
PLS	PLS	0.695	**0.937**

By considering all the results, the MLR model for the general land use that only considered three x-variables, proved to be the most accurate and consistent model. To ensure that this regression model is indeed the preferred model, several additional analyses were performed on the model.

Stellenbosch University https://scholar.sun.ac.za

FINAL REGRESSION MODEL SELECTION AND ANALYSIS

The regression model attained via the MLR analysis on the three essential x-variables for a general land use, was chosen as the final model. This model proved to be fairly reliable in describing the variability in the dataset with an adjusted R^2 value of 0.890. Furthermore, although the regression model was simple in that it only considered three variables, it provided equally accurate results when compared to the models using more variables. Surprisingly, despite the inherent simplicity of the MLR method, compared to the more complex PCA and PLS methods, it yielded very good results.

Equation 35 describes the model numerically. The total peak hour demand is in l/s, total pipeline length in m, total pipeline volume in m^3 and area in km^2.

$$\begin{aligned} &Total\ peak\ hour\ demand \\ &\quad = 9.855 + 0.00189 * Total\ pipeline\ length + 0.0845 \\ &\quad * Total\ pipeline\ volume - 7.253 * Area \end{aligned} \tag{35}$$

7.1. Testing model adequacy

7.1.1 Estimated y vs observed y

Figure 18 presents the relationship between the estimated y attained via Equation 35 and the observed y values. It is clear that a strong linear relationship exists. Certain residual values are however large. This is clear from how far many of the points are from the fitted mean line. As there are consistently many points relatively far below and above the line, the R^2 value is reasonably high. This demonstrates visually why the R^2 value must be analysed with caution.

Stellenbosch University https://scholar.sun.ac.za

Final model selection and analysis

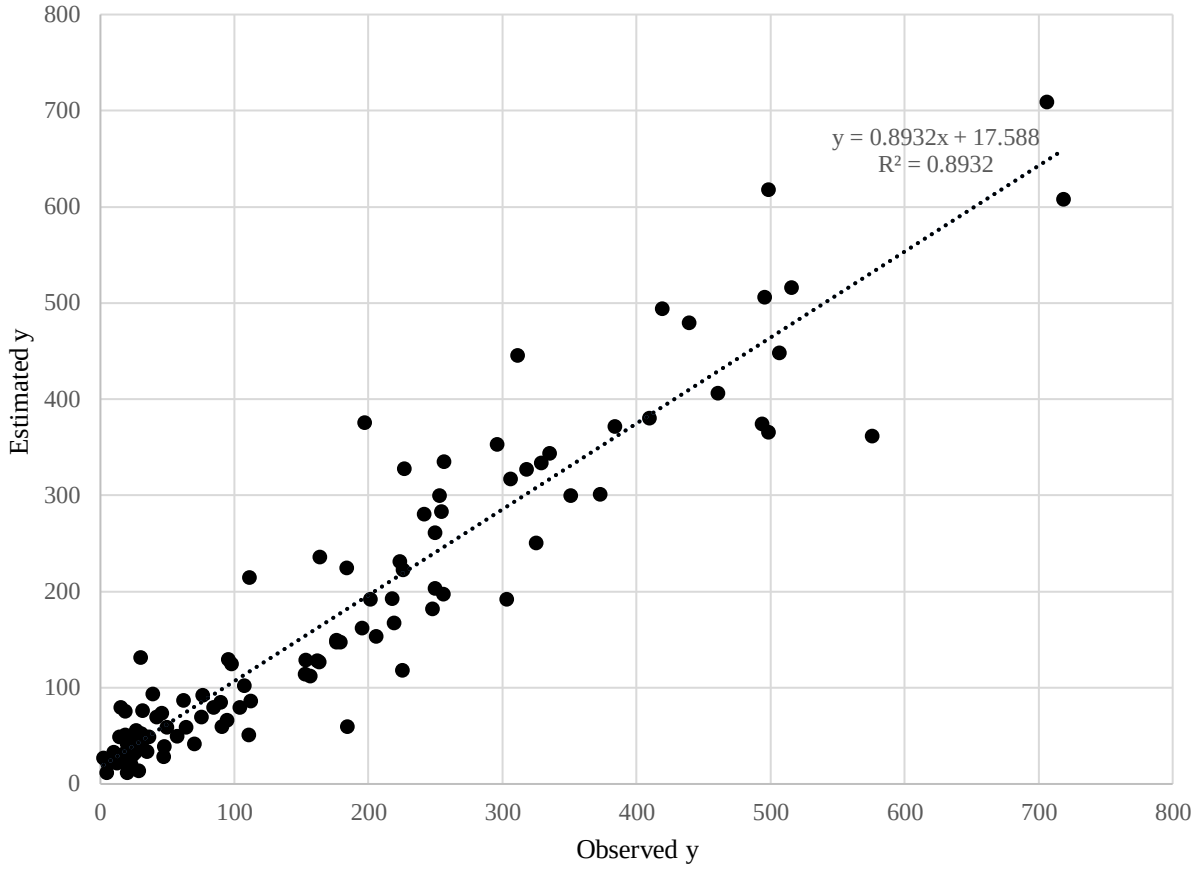


Figure 18: Estimated y vs observed y. The estimated y-values using the MLR model are compared to the observed y-values.

7.1.2 Test for normality

The normality test, tests if the residuals are normally distributed. This was achieved by plotting a quantile-quantile plot (De Souza and Junqueira, 2005).

Before the quantile-quantile plot could be plotted, the input data needed to be prepared and ordered. This involved ordering the residuals in ascending order and finding the Z-scores of the corresponding percentage points from the standard normal distribution tables. Once these two parameters were known, a graph could be plotted. If the points form a straight line, a normal distribution can be assumed. If the points form a curve, the data is not normally distributed and needs to be transformed to log format to improve the distribution (De Souza and Junqueira, 2005).

Stellenbosch University https://scholar.sun.ac.za

Equation 36 governs the calculation of the percentage point values.

$$ci = \{\frac{\left(i - \frac{3}{8}\right)}{\left(n + \frac{1}{4}\right)}\} \tag{36}$$

After plotting the quantile-quantile plot, the corresponding correlation coefficient was calculated. This value should be greater than the Rcrit value for the chosen α value. The Rcrit value for a significance level of 5 % was found to be 0.989 where n represents the number of samples.

$$Rcrit(n) = 1.0063 - \frac{0.1288}{\sqrt{n}} - \frac{0.6118}{n} - \frac{1.3505}{n^2} \; for \; \alpha = 0.05 \tag{37}$$

The normality plot in Figure 19 illustrated that the residuals were not normally distributed, because the points did not plot on a straight line. Because the residuals followed a nonlinear distribution, they formed a curve rather than a straight line (De Souza and Junqueira, 2005). It is because of this curvature that the R^2 value for the normality plot was 0.937 and not 1. Also, this R^2 value of 0.937 was below the critical value of 0.989. From this one can confidently say that the residuals do not follow a normal distribution. This was not necessarily the desired outcome, but as the R^2 value was close to the critical value, no alterations to the model were made.

7.1.3 Test for homoscedasticity

Homoscedasticity is a crucial element in accessing the quality of a linear regression model. A model that is homoscedastic is one where the noise in the data, or the error terms are similar for all the independent variables (Statistic Solutions, 2018). Ideally, one wants the error terms to be consistent in the model. Thus, if the magnitude of these terms fluctuates across the independent variables' ranges, then the errors are not constant and the model is not homoscedastic.

Stellenbosch University https://scholar.sun.ac.za

Final model selection and analysis

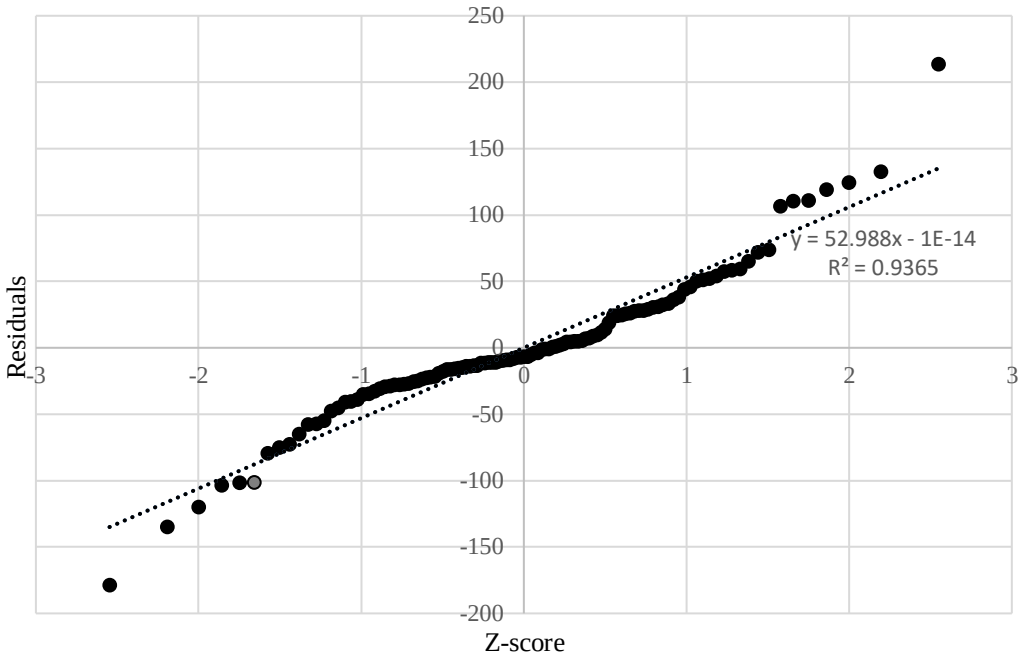


Figure 19: Test for normality

To perform the test for homoscedasticity, a hypothesis test was necessary. The null hypothesis was that the model is homoscedastic and the alternative hypothesis was that the model is not homoscedastic (De Souza and Junqueira, 2005).

The absolute value for the difference between each observed y value and the mean of all the observed y values was computed. The same was done on the set of estimated y values. A single factor Anova test was then performed on the two new datasets. The F-statistic could then be used to assess which hypothesis can be accepted and which must be rejected. The calculated F-value of 0.246 was found to be smaller than the critical F-value of 3.88, indicating that the null hypothesis could be accepted. Thus it was established that the variances within the model were homogenous.

7.2. Safety Factor

To ensure that the regression model can be used with confidence, a safety factor was developed. This safety factor was designed to be multiplied with the y term, thus increasing the y term and indirectly the x terms as well.

Stellenbosch University https://scholar.sun.ac.za

From Figure 19, it is evident that the residuals are not normally distributed, thus a constant safety factor would not be adequate in compensating for the spread and variation of the y values. To solve this, a percentage error was proposed.

The percentage error represents the residuals as fractions of the estimated y terms (University of Iowa, 2017). If the percentage error term is close to 0, then the residual is small and the computed y value is close to the target y value.

Once the percentage errors had been calculated, they were converted to a safety factor. This was done by converting the percentage errors to fractions and adding one. A value of one was added, as the safety factor is multiplied into the y-variable and needs to compensate for the original value and the safety margin. The safety factor chart is presented in Figure 20.

To use the chart, one considers the y value and finds the corresponding safety factor using the green line. It is evident that the variance in the data decreases with increasing y values, thus the safety factor decreases as y increases. The red line can also be used as an ultimate safety factor. For example, if the user knows the total peak hour demand and area, then the y value can be upscaled using the red line. This will lead to an increased confidence level that the pipe lengths and diameters in the network will be adequate. It is however recommended that the operator generally uses the lower safety factor line, in order to avoid unnecessary costs and overdesign of the pipe network.

Stellenbosch University https://scholar.sun.ac.za

Final model selection and analysis

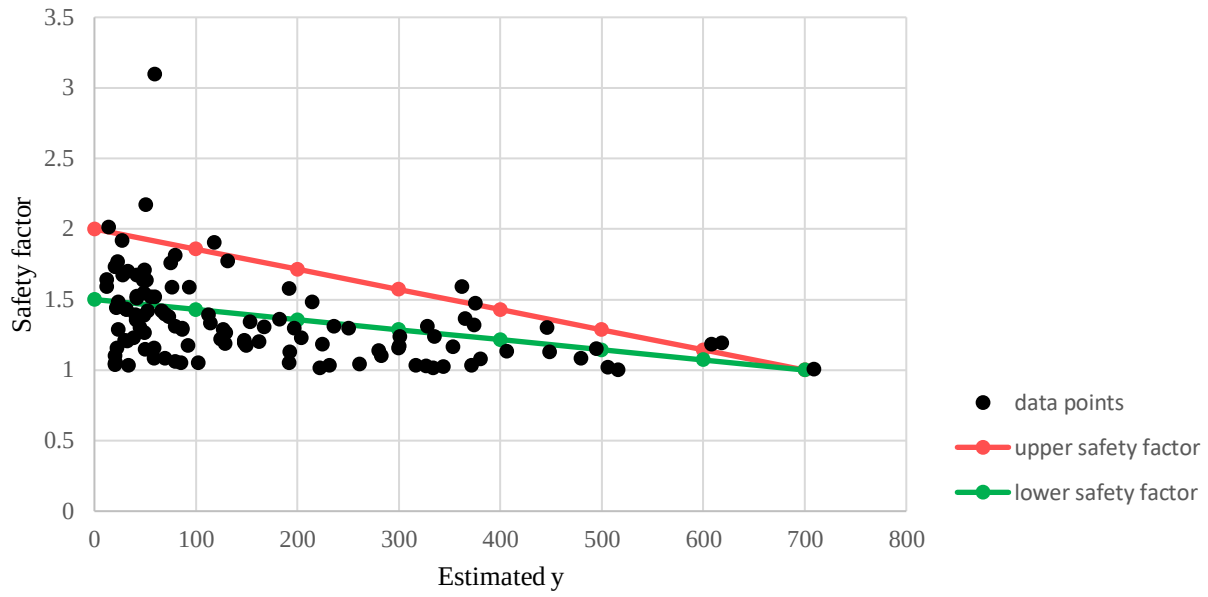


Figure 20: Safety factor. The green and red lines represent the lower and upper safety factor lines.

7.3. Pipe distribution

When constructing the regression model, the total pipeline length and total pipeline volume were considered. There was however no explanation of the different pipe diameters necessary to attain the total pipeline volume, given the total pipeline length.

To find this relationship, the length of each nominal pipe diameter for each WDS was required. These values were attained for all the different WDSs to find what proportion of the total pipeline length could be attributed to each pipe diameter present in the network.

Graphs were developed for various area sizes, terrain profiles and land uses. These include small areas, moderately sized areas, large areas, flat areas, partially hilly areas, hilly areas, general areas, residential areas and LCH areas. These results can be found in Figure 21 to 29.

Stellenbosch University https://scholar.sun.ac.za

Small areas (< 10 km^2)

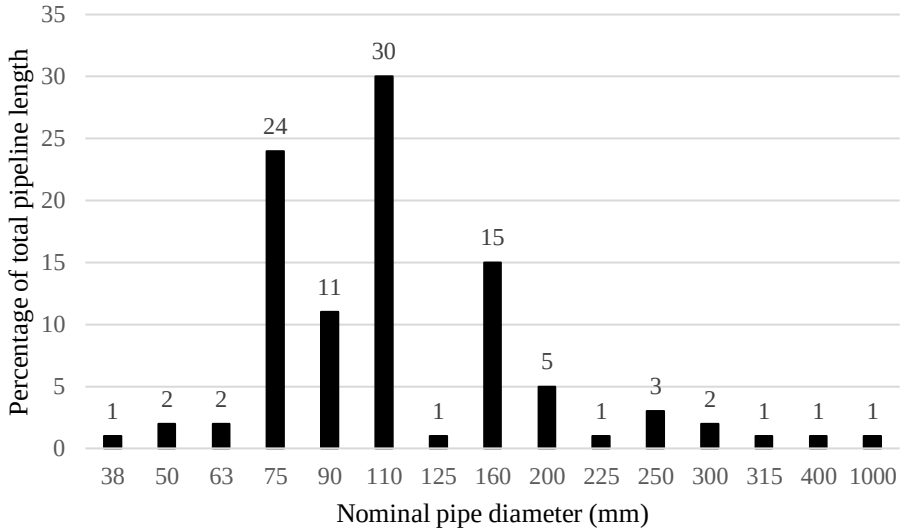


Figure 21: Small areas pipe diameter distribution

Moderately sized areas (between 10 and 20 km^2)

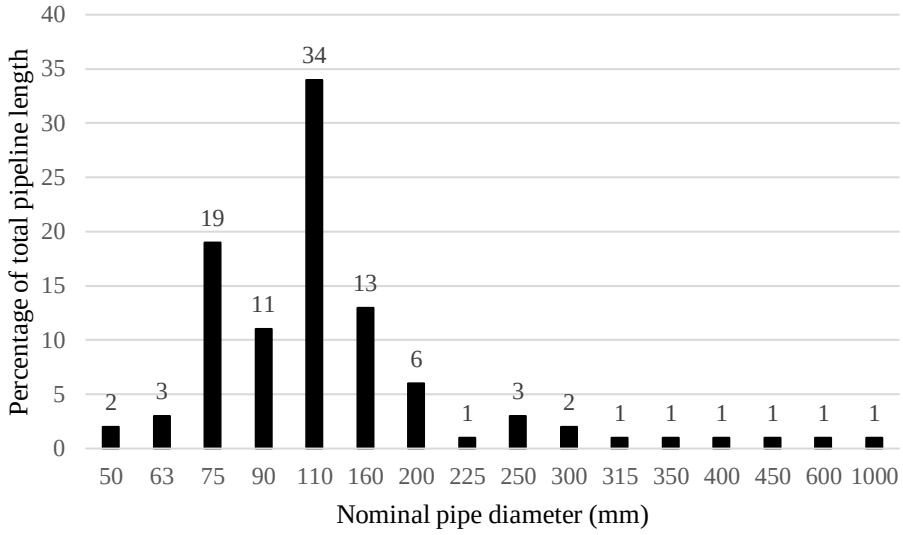


Figure 22: Moderately sized areas pipe diameter distribution

Stellenbosch University https://scholar.sun.ac.za

Final model selection and analysis

Large areas (> 20 km^2)

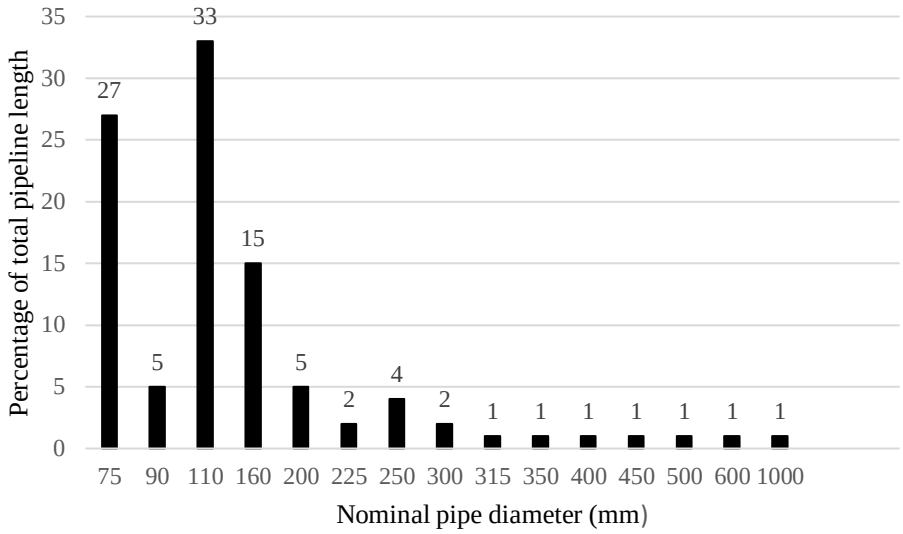


Figure 23: Large areas pipe diameter distribution

Flat areas

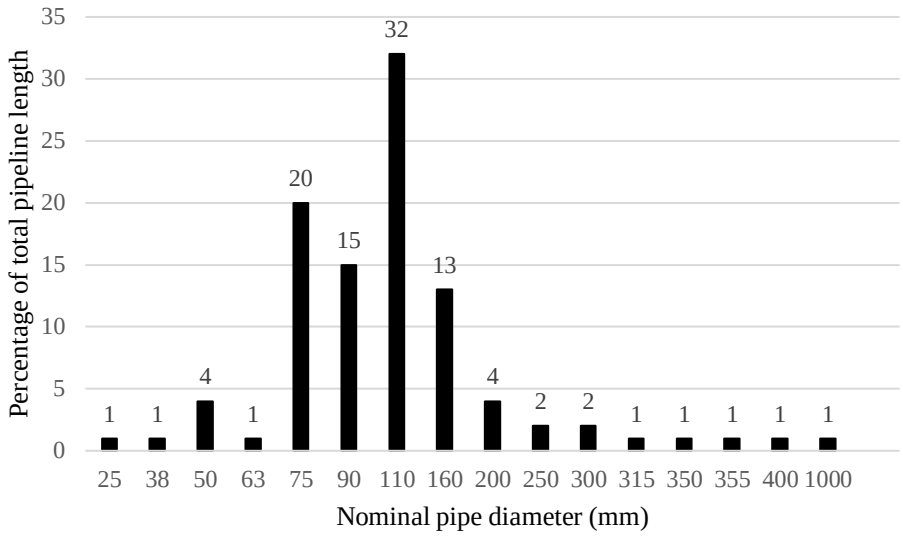


Figure 24: Flat areas pipe diameter distribution

Stellenbosch University https://scholar.sun.ac.za

Partially hilly areas

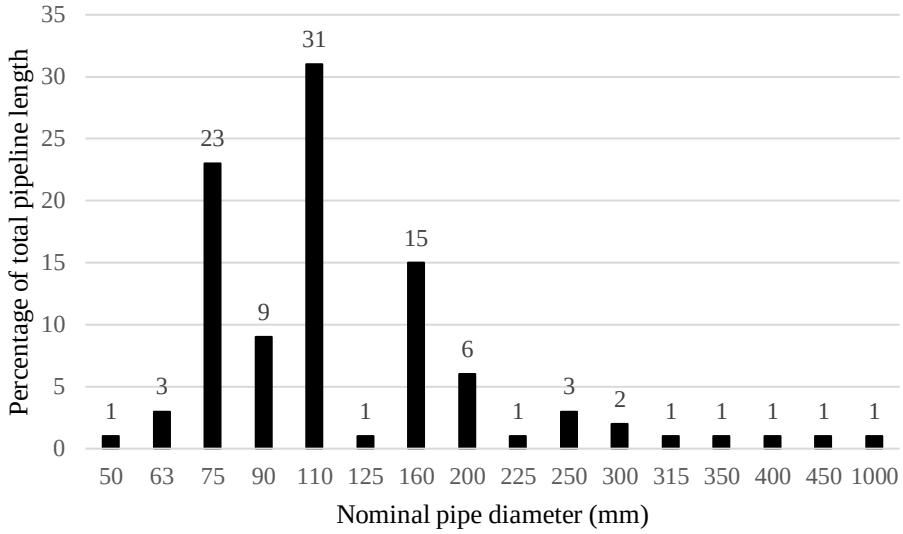


Figure 25: Partially hilly areas pipe diameter distribution

Hilly areas

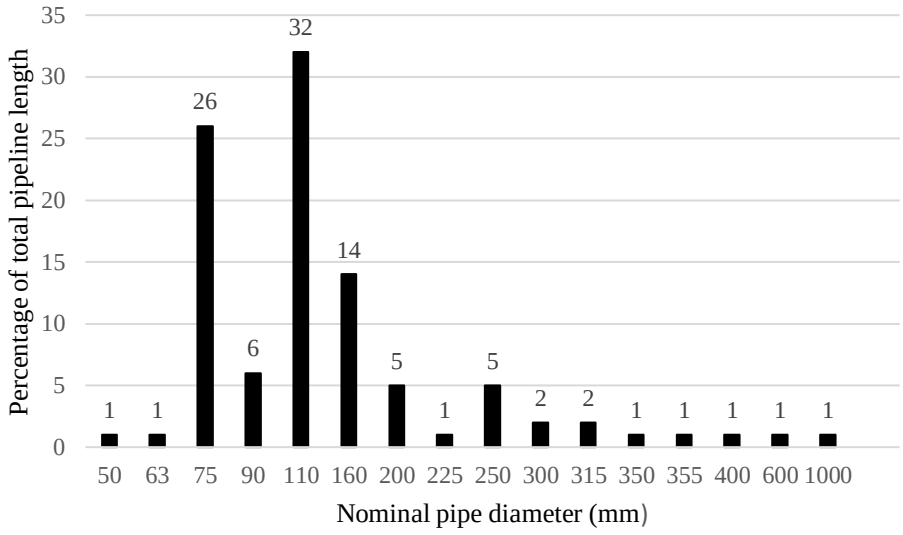


Figure 26: Hilly areas pipe diameter distribution

Stellenbosch University https://scholar.sun.ac.za

Final model selection and analysis

General areas

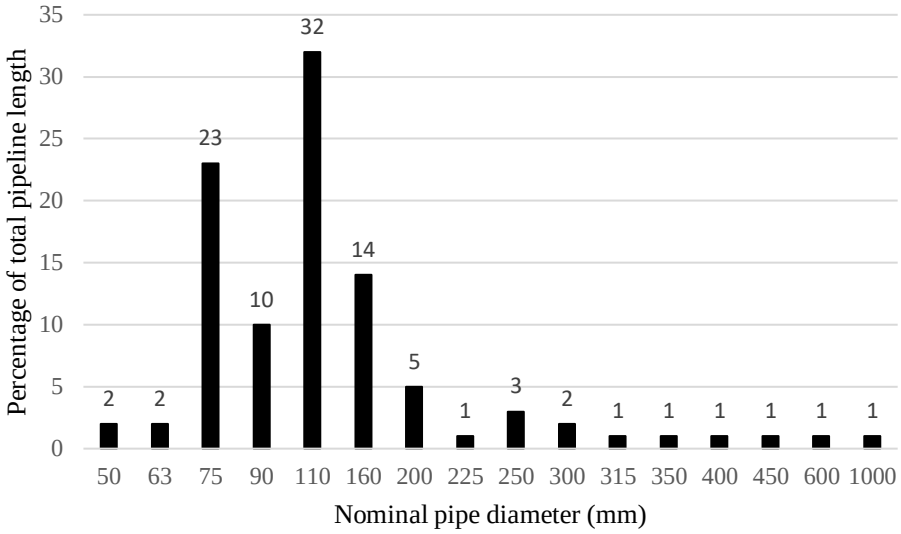


Figure 27: General areas pipe diameter distribution

Residential areas

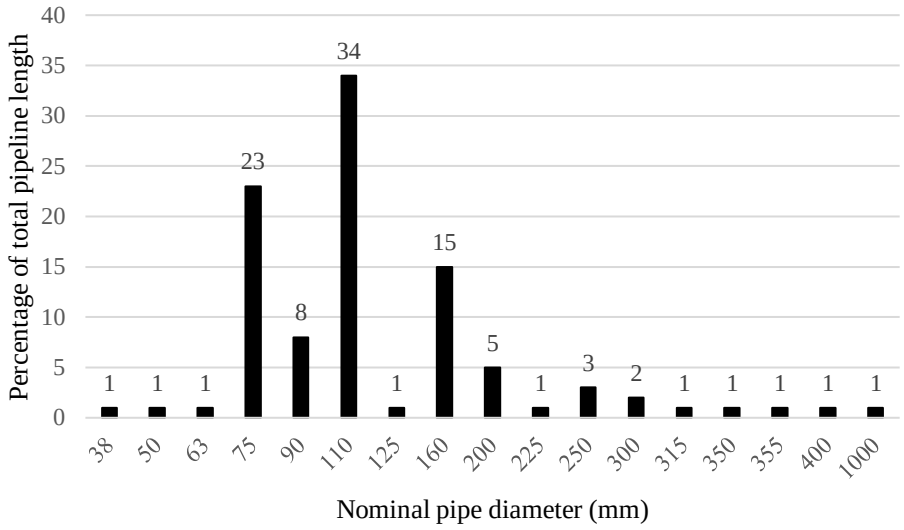


Figure 28: Residential areas pipe diameter distribution

Stellenbosch University https://scholar.sun.ac.za

Low-Cost Housing areas

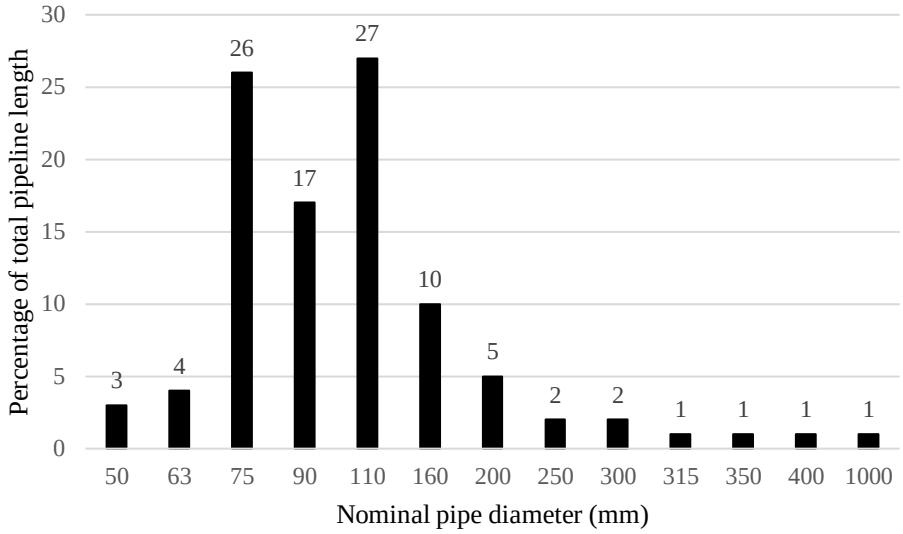


Figure 29: Low-Cost housing areas pipe diameter distribution

To classify the topology, the terrain index outlined in Section 4.4. was used. Models that had an index value of 1, 1.5 and 2 were classified as flat. Areas with an index value of 2.5, 3 and 3.5 were regarded as partially hilly and areas with index values between 4 and 5 were classified as hilly. When classifying the area, the range of different land areas was considered. Small areas were classified as smaller than $10km^2$. Medium sized areas were classified as being between 10 and 20 km^2 with large areas measuring over 20 km^2.

The City of Tshwane's *Standard Specifications for Municipal Civil Engineering Works* (Department of Water & Sanitation, 2005) stipulates that only certain pipe diameters approved by the Director are allowed in definite areas. Even so, the minimum pipe diameter for PVC and Polyethylene pipes is governed to 110 mm. By considering this minimum pipe diameter and only considering commercially available PVC pipe diameters in South Africa, the various distribution tables were adjusted. This was done by summing the proportion of all pipe diameters smaller than 110 mm and adjusting each uncommon pipe diameter to the closest larger commercially available

Stellenbosch University https://scholar.sun.ac.za

pipe diameter. Summaries of these different pipe distributions are found in Figure 30 to 38 in Appendix 10.5. It must be noted that these adjusted graphs are only applicable to urban networks in South Africa and that the graphs outlined in Figure 21 to 29 be used if other design criteria are requisite. From the pipeline diameter distribution graphs, it was also possible to find the ratio between total pipeline volume and total pipeline length for each case. Equation 38 demonstrates this relationship for the general land use case. A summary of these ratio factors illustrating the relationship for each pipe diameter distribution graph is available in Table 15. It is clear that there is a non-constant relationship between total pipeline length and total pipeline volume.

Table 15: Ratio factor relating total pipeline volume and total pipeline length

Pipe size distribution type	**Ratio of total pipeline volume to total pipeline length (RF)**
Small areas	0.0226
Moderately sized areas	0.0279
Large areas	0.0290
Flat areas	0.0231
Partially hilly areas	0.0254
Hilly areas	0.0287
General areas	0.0277
Residential areas	0.0247
LCH areas	0.0218

$$Total\ pipeline\ volume = 0.0277 * Total\ pipeline\ length \quad (38)$$

Stellenbosch University https://scholar.sun.ac.za

7.4. Final model implementation and recommendations for use

An outline of how the final WDS capacity model is implemented follows.

If a new housing development is proposed and the number of housing units, as well as the estimated AADD of each unit is known, then the y-variable or total peak hour demand can be estimated. This can be found using Equation 39. This is similar to Equation 28, but as the untransformed models were used, the user-defined factor is equal to one. The peak factor can be found using Table 1 for the specific land use and size.

$$Total\ peak\ hour\ demand = (AADD + UAW) * Peak\ Factor \quad (39)$$

With the total peak hour demand known, Figure 20 can be used to find the safety factor. This factor is not requisite, but can be multiplied with the total peak hour demand to improve safety. Furthermore, the area of the new development is typically also known.

With the area and total peak hour demand known, Equation 40 can be implemented to solve for the unknown variables.

$$\begin{aligned} &Total\ peak\ hour\ demand \\ &= 9.855 + 0.00189 * Total\ pipeline\ length + 0.0845 \\ &* Total\ pipeline\ volume - 7.253 * Area \end{aligned} \quad (40)$$

Currently, both the total pipeline length and total pipeline volume are unknown. The ratio factor best describing the site, can be used to relate the total pipeline volume and total pipeline length. This is illustrated in Equation 41.

$$\begin{aligned} &Total\ peak\ hour\ demand \\ &= 9.855 + (0.0845 * RF + 0.00189) \\ &* Total\ pipeline\ length - 7.253 * Area \end{aligned} \quad (41)$$

Stellenbosch University https://scholar.sun.ac.za

Final model selection and analysis

Equation 41 can be further simplified by considering a final factor that accounts for the weightings of the total pipeline length and pipeline volume, as well as the ratio factor relating them. This final interpretation of the user demand model is illustrated in Equation 42 and the final factors are illustrated in Table 16.

$$Total\ peak\ hour\ demand = 9.855 + FF * Total\ pipeline\ length - 7.253 * Area \quad (42)$$

Table 16: Final factors for final user demand model

Pipe size distribution type	**Final ratio factors (FF)**
Small areas	0.00381
Moderately sized areas	0.00425
Large areas	0.00435
Flat areas	0.00385
Partially hilly areas	0.00404
Hilly areas	0.00432
General areas	0.00424
Residential areas	0.00398
LCH areas	0.00374

Thus, by considering a general area, the model can be illustrated as:

$$Total\ peak\ hour\ demand = 9.855 + 0.00424 * Total\ pipeline\ length - 7.253 * Area \quad (43)$$

From this, the equation only has one unknown variable, namely the total pipeline length. This variable can then be solved directly. With the total pipeline length known, the pipe diameter

Stellenbosch University https://scholar.sun.ac.za

distribution graphs, can be used to calculate what length of each nominal pipe diameter is required to construct the pipe network.

If the location is a large hilly LCH area, then multiple final factors and pipe diameter distribution graphs can be used to construct different pipeline models. From this, the pipe network with the most adequate combination of pipe diameters to meet the system requirements, can be chosen.

With the length of each pipe diameter known, the construction costs can be estimated. Table 17 illustrates the construction costs for different pipe diameters used by GLS (GLS Consulting, 2018). This table was used during 2019 with rates susceptible to annual changes. This is a tool which can be highly effective in developing rough estimate costs if minimal knowledge of a proposed WDS is known. An example calculation is present in Appendix 10.6.

Table 17: Cost per pipe diameter. A combination of PVC, HDPE, Steel and Concrete pipes are considered (GLS Consulting, 2018)

Nominal Pipe Diameter (mm)	Public open space costs (R/m)	Nominal Pipe Diameter (mm)	Public open space costs (R/m)
50	452	600	4456
63	483	650	5175
75	518	700	5628
90	569	750	6988
110	610	800	7415
125	667	850	7846
140	737	900	8300
160	830	950	8694
200	1069	1000	10493
250	1425	1100	11413
315	1937	1200	12569
355	2398	1300	15054
400	3009	1400	15965
450	3383	1500	21416
500	3752	1600	23185
550	4082	1800	25624

Stellenbosch University https://scholar.sun.ac.za

SECTION 8

CONCLUSION

8.1. WDS regression model development and use

On initial inspection, the likelihood of a linear relationship between the physical characteristics of a WDS and the capacity of the system seemed possible. With access to many of the WDS models in South Africa via GLS, the possibility to evaluate whether this relationship exists was made possible.

These models had already been calibrated to simulate peak demands, depending on the land use and land area of each WDS. In order to further standardise these models, a state of maximum supply capacity where the pressure head at the critical node is 18 m, and the maximum flow velocity in the system is limited to 2 m/s, was enforced. This was done by modifying the user demands until a point was reached where the pressure requirements were met and the pipes with a flow velocity exceeding 2 m/s were adjusted to larger pipe diameters.

With 165 hydraulic models standardised, an analysis of the WDSs could be performed to assess whether this linear relationship between the physical network characteristics and network capacity, does indeed exist. It was also necessary to check whether the relationship is consistent and can be used in industry with confidence.

Three statistical approaches were performed to develop different linear regression models. This included the MLR analysis directly on the network characteristics. Linear regression models were also constructed using the PCA and PLS methods combined with MLR. The various regression models varied significantly with regards to their adjusted R^2 statistic. It was interesting to note that models which considered seven physical network characteristics or only three dominant

Stellenbosch University https://scholar.sun.ac.za

characteristics necessary to construct an initial pipeline distribution, achieved equally accurate results. These models were tested and results were verified, as the same models did not yield consistently accurate results when using the analysis and test data. This raised the concern of the usability and accuracy of the models.

To assess the initial standardisation process and the prospective change in network characteristics, the original models which had not been transformed were tested again. These hydraulic models were the untransformed Wadiso files, with peak demands simply correlated to peak factors.

After performing the same three linear regression approaches, the outcomes changed significantly. The regression models had far higher adjusted R^2 values, showing that a greater amount of variance within the data was explained. This was a major issue in the methodology approach, as the initial judgement of how a model at peak demand can be represented, became questionable.

To address this issue, it was decided to base the linear regression models on the untransformed WDS models. After consulting an engineering statistical professional, this decision was encouraged as in complex systems, the derivation of a theoretical relationship may be practically impossible. Therefore, a more practical solution was to derive these relationships from empirical data.

It was again found that many of the network characteristics were for practical purposes, redundant. Four of the seven physical WDS characteristics were not contributing to or improving the quality of the models. These included: reservoir elevation above mean terrain elevation, reservoir distance from the centre of area, shape factor ratio and terrain index.

The final model comprised of a linear regression model that only considered the total pipeline length, total pipeline volume and area to make predictions of the total peak hour demand which a WDS could supply. The model also included a safety factor depending on the magnitude of the

Stellenbosch University https://scholar.sun.ac.za

total peak hour demand in order to improve the reliability of the model. At smaller total peak hour demand values, the regression model was less accurate and needed to be adjusted. To improve the usability of the model, different graphs outlining the relationship between the total pipeline length and total pipeline volume were included. These graphs provided a detailed outlay of what proportion of the total pipeline length each nominal pipe diameter should be. These graphs were designed specifically for different systems, depending on the topology, size and land use of the area.

The usefulness of this tool must be emphasised, as the model has the capability to provide a cost estimate per pipe diameter of a WDS network by simply considering the peak demand, or capacity of a system accompanied by the area of a system. This model holds great potential to assist planners to estimate pipeline construction costs during initial stages when no pipe layout information or designs are available.

Stellenbosch University https://scholar.sun.ac.za

SECTION 9

RECOMMENDATIONS

9.1. Recommendations and prospective future studies

The initial decision to use flow velocity as a fixed parameter became questionable, as pipes with large diameters can typically handle greater flow velocities than pipes of smaller diameters. Thus, using the energy gradient could potentially have resulted in a more reliable parameter for transforming the hydraulic models. Yet, as the untransformed models were ultimately used, this decision would likely have ended with the same conclusion, but would have been better practise.

The method of quantifying the area of the WDSs using an ellipse could also be questioned. An alternative approach would simply have been to remove the hydraulic models that had large sections with no pipes. This would have limited the amount of available hydraulic models, but would have eliminated the uncertainty attached to the assumptions of assuming each area is an ellipse.

The final regression model had the total user demand as the dependent variable of the equation, with the total pipeline length as one of the independent x-variables. The goal of the regression model was to compute the total pipeline length and then use the diameter distribution graphs to compute the lengths of the different pipe diameters. Considering this, the regression model could have been reinterpreted to have the total pipeline length as the dependent y-variable and the unknown variable that is solved for. As the exact workings and importance of the different variables were unknown initially, this decision was not made. But in hindsight would have made the regression model more user friendly.

Stellenbosch University https://scholar.sun.ac.za

Conclusion

Lastly, the model was developed for the initial cost estimation of urban WDSs. Hydraulic models for rural areas could also be researched in prospective future studies to develop a regression model specific to rural areas.

Stellenbosch University https://scholar.sun.ac.za

APPENDICES

10.1. Principal Component Analysis mathematical proofs

- **If a matrix's transpose and inverse are the same, the matrix is orthogonal**

An orthogonal matrix is a square matrix and its columns and rows are orthogonal (Mathworld, 2018). This means that if the dot product of any two columns is computed, the answer is 0. Or if described in a visual way, the column vectors are perpendicular to each other. Another characteristic of orthogonal matrices is that each row and column is a unit vector. Thus, the summation of each value squared along a row or column is equal to 1.

Let A equal an m × n orthogonal matrix.

$(A^T A)ij = ai^T aj = \{$1 if i = j
0 otherwise

Thus, because $A^T A = I$ $A^{-1} = A^T$

- **For any matrix, $A^T A$ and AA^T are symmetrical**

$$(AA^T)^T = A^{TT}A^T = AA^T$$

$$(A^T A)^T = A^T A^{TT} = A^T A$$

- **A matrix is symmetric only if it is orthogonally diagonalizable**

A symmetric matrix is orthogonally diagonalizable if there is a distinct orthogonal matrix U and diagonal matrix E so that A = EDE^{-1}.

Following the first proof outlines in Section 10.1, this equation becomes: A = EDE^T.

Stellenbosch University https://scholar.sun.ac.za

Appendix

Knowing that if A and A^T are the same the matrix is symmetric.

$$A^T = (EDE^T)^T = EE^T D^T E^T = EDE^T = A$$

- **A symmetric matrix is diagonalized by a matrix of its orthonormal eigenvectors**

This proof is broken down into two parts. The first shows that a matrix is only orthogonally diagonalizable if its eigenvectors are linearly independent. The second part shows that the eigenvectors of a symmetric matrix are both linearly independent and orthogonal.

1st part of proof

Let A be any matrix, E be a vector of the eigenvectors namely E = [e1 e2 …] and D be a diagonal matrix with the corresponding eigenvalues λi along the diagonal in the iith positions.

If AE = ED then Aei = λiei. This is validated with the eigenvalue equation, thus showing that AE = ED.

After rearranging the equation: A = EDE^{-1}

2nd part of proof

For the second part we must show that the eigenvectors of a symmetric matrix are orthogonal. This can be done using the dot product rule.

Let λ1 and λ2 be the corresponding eigenvalues for eigenvectors e1 and e2 for a symmetric matrix.

Stellenbosch University https://scholar.sun.ac.za

$$\lambda 1 e1 \cdot e2 = (\lambda 1 e1)^T e2$$

$$= (Ae1)^T e2$$

$$= e1^T A^T e2$$

$$= e1^T Ae2$$

$$= e1^T (\lambda 2 e2)$$

$$= \lambda 2 e1 \cdot e2$$

By rearranging: $(\lambda 1 - \lambda 2)e1 \cdot e2 = 0$

Since the eigenvalues are unique, the dot product of the eigenvectors must be zero $e1 \cdot e2 = 0$, thus showing that the eigenvectors are orthogonal.

Thus summarising, if A is a symmetric matrix with orthonormal eigenvectors, then E must be an orthogonal matrix which satisfies the rule: $E^T = E^{-1}$.

From this, equation $A = EDE^{-1}$ becomes $A = EDE^T$.

- **For any m × n matrix X, a symmetric matrix $\boldsymbol{X^T X}$ can be attained**.

This new symmetric matrix has a set of orthonormal eigenvectors and associated eigenvalues. The set of vectors $\{\widehat{\boldsymbol{Xv1}} \ldots \widehat{\boldsymbol{Xvn}}\}$ form an orthonormal basis, with each vector $\widehat{\boldsymbol{Xv\iota}}$ having a length of $\sqrt{\boldsymbol{\sigma i}}$.

Stellenbosch University https://scholar.sun.ac.za

Appendix

By considering the dot product of 2 vectors:

$(X\hat{v}i)\cdot(X\hat{v}j)$ $= (X\hat{v}i)^T(X\hat{v}j)$

$= \hat{v}i^T X^T X\hat{v}j$

$= \hat{v}i^T(\lambda j\hat{v}j)$

$= \lambda j\hat{v}i \cdot \hat{v}j$

$= \{\lambda j$ if i = j

0 otherwise

Thus, the vectors are orthogonal.

To show that each vector is of length $\sqrt{\lambda i}$, the vectors are squared.

$$||X\hat{v}i||^2 = (X\hat{v}i)\cdot(X\hat{v}i) = \lambda i$$

Stellenbosch University https://scholar.sun.ac.za

10.2. Raw data with outliers

Table 18: Raw data with outliers indicated in yellow

Count	Output (l/s)		Reservoir elevation above mean terrain elevation (m)	Terrain			Land use	Total pipeline length (m)	Total pipeline volume (m^3)		Reservoir distance from the centre of area (m)	Shape			Area (km^2)
	Transformed	Untransformed		Range (m)	Std deviation (m)	Terrain index			Transformed	Untransformed		Width (m)	Height (m)	Shape factor ratio	
1	33.3	18.5	59.0	84.9	13.4	3.0	Res	23281	248	248	2156	2814	1478	1.9	3.3
2	67.6	84.5	42.1	94.2	14.5	3.0	LCH	81170	1453	691	807	5865	4255	1.4	19.6
3	46.9	156.4	34.5	151.2	16.3	4.0	LCH	63824	693	664	883	5653	2306	2.5	10.2
4	25.7	84.2	77.4	203.7	38.0		LCH	63824	693	694	883	5653	2306	2.5	10.2
5	46.2	72.1	52.3	273.3	36.0		Res	31389	194	195	2853	4125	3917	1.1	12.7
6	23.2	33.1	17.4	88.7	17.1	3.0	Res	37746	520	524	686	7634	1383	5.5	8.3
7	1.0	1.0	37.4	57.3	13.6	2.5	Res	2012	16	16	172	1276	212	6.0	0.2
8	9.4	2.3	50.9	50.0	9.5	2.0	Res	2947	20	20	999	1955	202	9.7	0.3
9	15.6	13.0	32.3	57.1	15.1	2.5	Res	6604	171	171	2090	2221	1029	2.2	1.8
10	220.4	183.7	85.8	113.5	15.1	3.5	LCH	70888	1237	1236	4089	2965	1422	2.1	3.3
11	283.2	217.9	66.1	103.2	18.5	4.0	LCH	73892	1120	1098	1396	4020	2164	1.9	6.8
12	311.2	311.2	82.3	100.3	17.2	3.5	Res/LCH	120309	4422	4422	2789	6867	4244	1.6	22.9
13	2.8	55.8	23.9	18.2	4.8	1.0	LCH	13247	113	113	437	987	750	1.3	0.6
14	17.1	2.3	39.5	48.3	7.7	2.0	Res	8940	68	68	962	1105	802	1.4	0.7
15	21.3	21.3	27.1	20.0	4.7	1.0	Res	4627	48	48	138	1316	306	4.3	0.3
16	50.9	18.2	73.3	104.9	23.3	4.0	LCH	9780	97	95	1939	1121	737	1.5	0.6
17	17.1	19.0	32.8	35.2	9.9	1.5	LCH	6716	33	33	292	745	676	1.1	0.4
18	270.2	225.1	56.7	79.7	10.9	2.5	Res	51382	1014	1004	2428	3750	3449	1.1	10.2
19	191.9	225.8	66.6	104.1	16.4	3.5	LCH	83797	1209	1193	1167	2891	2835	1.0	6.4
20	99.7	153.3	39.4	61.7	8.3	2.0	Res	48298	687	651	1924	2617	1819	1.4	3.7
21	187.5	104.2	72.3	91.4	14.0	3.0	Res	29871	474	454	1520	3277	1352	2.4	3.5
22	754.9	419.4	95.4	119.5	17.6	3.5	Res	157017	3319	3188	3557	3921	3680	1.1	11.3
23	397.2	317.7	76.4	108.1	14.1	3.5	Res	116728	2189	2181	2087	4168	3718	1.1	12.2
24	532.5	409.6	79.1	114.4	21.5	4.0	Res	124865	3242	3122	1518	7347	3105	2.4	17.9
25	157.8	90.2	187.5	46.3	10.5	2.0	Res	34483	448	441	10908	3721	1728	2.2	5.1
26	112.9	75.3	66.5	94.9	20.9	3.5	Res	29824	374	373	692	2722	1817	1.5	3.9
27	740.7	493.8	77.4	114.1	20.0	4.0	Res	122795	3566	3542	1891	6542	4500	1.5	23.1
28	515.6	515.6	65.8	101.3	22.8	4.0	Res	109291	4473	4456	2073	3819	3570	1.1	10.7
29	743.1	495.4	52.2	79.3	13.2	3.0	CBD	97664	4373	4365	1980	4014	2545	1.6	8.0

Stellenbosch University https://scholar.sun.ac.za

30	214.9	195.3	53.5	96.9	16.8	3.0	Res	58297	1158	1155	1945	4878	1991	2.5	7.6
31	256.0	256.0	76.0	126.6	19.2	4.0	Res/BCI	65290	1479	1475	2582	3800	2820	1.3	8.4
32	748.4	575.7	79.1	105.5	14.9	3.5	Res/CBD	83794	3235	3094	2820	3822	3136	1.2	9.4
33	428.6	306.1	67.0	114.2	16.8	3.5	Res	95244	2100	2078	2717	4583	1872	2.4	6.7
34	145.2	107.5	60.8	90.7	12.7	3.0	Res	34750	628	628	1975	3606	1287	2.8	3.6
35	254.3	254.3	46.9	77.4	12.7	3.0	Res	69594	2287	2287	1576	4811	1900	2.5	7.2
36	161.0	247.8	64.5	96.5	19.5	3.5	Res	58554	1302	1302	609	3276	2600	1.3	6.7
37	361.6	328.7	71.3	105.8	19.8	4.0	Res	99619	2675	2675	981	6400	2500	2.6	12.6
38	164.7	205.9	66.5	112.8	23.0	4.0	Res	55570	1180	1177	1636	5001	2136	2.3	8.4
39	52.5	47.8	50.5	50.9	11.3	2.0	Res	12552	170	170	1162	1952	820	2.4	1.3
40	69.4	34.7	54.2	87.3	17.9	3.0	Res	13170	148	148	648	1608	1504	1.1	1.9
41	39.8	22.7	40.2	39.4	10.7	1.5	Res	6005	68	68	546	1067	1037	1.0	0.9
42	13.7	19.5	27.1	53.6	14.3	2.5	Res	5717	54	54	790	1400	618	2.3	0.7
43	37.9	47.4	56.5	63.2	16.5	2.5	Res	12734	100	99	339	3151	778	4.1	1.9
44	108.8	350.7	77.2	135.8	27.6	5.0	Res	92075	2477	2588	1608	5789	3119	1.9	14.2
45	391.7	373.0	58.7	111.9	19.4	4.0	Res	104953	2582	2580	0	6868	3213	2.1	17.3
46	40.2	22.3	67.5	97.9	21.0	3.5	Res	27828	233	233	1703	2648	2200	1.2	4.6
47	150.0	111.1	30.7	57.6	15.5	2.5	Res/BCI	22152	1156	2100	1059	2424	1100	2.2	2.1
48	66.0	23.6	41.5	41.8	9.7	2.0	Res	10105	107	106	557	1551	923	1.7	1.1
49	119.2	45.8	44.7	37.5	6.4	1.5	Res	25292	566	563	1360	2484	2247	1.1	4.4
50	42.0	15.0	34.4	55.9	13.0	2.5	Res	16569	553	553	1678	1908	764	2.5	1.1
51	307.6	439.4	75.4	132.5	23.2	4.5	LCH	192669	4334	4010	2541	6961	5902	1.2	32.3
52	4.3	225.6	44.8	87.0	19.5	3.5	LCH	101924	2111	2099	1790	4435	3018	1.5	10.5
53	72.5	241.7	60.2	77.7	16.6	3.0	LCH	93141	1714	1711	1857	3532	2505	1.4	6.9
54	124.9	249.9	45.0	69.0	12.9	2.5	LCH	113309	1636	1627	0	4522	3911	1.2	13.9
55	296.0	197.3	79.7	112.3	19.9	4.0	LCH	138054	2384	2374	1825	6263	2696	2.3	13.3
56	107.3	178.8	47.5	71.6	16.5	3.0	LCH	56042	705	705	979	4400	1100	4.0	3.8
57	114.1	57.0	65.1	81.0	12.4	3.0	LCH	17579	205	204	760	2145	860	2.5	1.4
58	162.1	69.0	171.9	81.2	20.4	3.5	LCH	23624	302	262	3776	1741	1400	1.2	1.9
59	414.0	295.7	57.8	83.2	11.2	2.5	Res/LCH	128224	2517	2517	3100	6267	3138	2.0	15.4
60	530.8	252.8	76.3	98.9	16.2	3.0	Res	105695	2567	2558	2297	5659	3918	1.4	17.4
61	171.2	171.2	72.8	108.4	22.9	4.0	Res	60179	2882	2878	3215	7600	900	8.4	5.4
62	758.8	1084.0	87.3	147.1	29.5	5.0	LCH	585650	9179	8441	4872	11791	6420	1.8	59.5
63	423.3	769.6	52.8	84.4	11.8	2.5	LCH	318399	6972	6854	2279	8297	3824	2.2	24.9
64	58.9	42.1	49.5	90.4	22.6	3.5	Res	35267	601	601	1364	4282	2370	1.8	8.0
65	212.5	249.8	70.4	114.3	22.1	4.0	Res	66204	1396	1393	1400	4311	2000	2.2	6.8
66	190.8	95.4	74.9	127.8	26.1	4.5	Res	42148	707	705	1728	2296	1481	1.6	2.7
67	506.4	506.4	63.2	133.2	22.2	4.5	Res	161571	2741	2718	588	6037	2815	2.1	13.3
68	95.9	70.3	40.3	37.6	9.2	1.5	Res	17967	155	155	500	1680	1548	1.1	2.0
69	130.3	325.0	45.2	93.1	20.4	3.5	Res	93091	2112	2111	0	5620	3552	1.6	15.7

Stellenbosch University https://scholar.sun.ac.za

70	326.4	163.2	46.1	45.2	12.2	2.5	Res/BCI	48649	783	780	280	3435	2100	1.6	5.7
71	410.3	256.4	62.9	116.4	22.5	4.0	Res	99857	2593	2591	1804	5716	2538	2.3	11.4
72	256.3	176.0	53.8	79.5	17.8	3.0	Res/BCI	39053	1284	1284	1490	3312	2258	1.5	5.9
73	113.4	90.7	44.3	43.0	9.6	2.0	Res/BCI	21305	294	292	627	1891	1400	1.4	2.1
74	76.4	152.8	49.7	71.9	12.0	2.5	Res	54672	898	898	1173	4366	3000	1.5	10.3
75	194.0	161.7	52.3	82.0	19.8	3.5	Res	43044	1011	1011	1574	3050	2800	1.1	6.7
76	875.1	460.6	79.3	142.0	22.2	4.5	Res	147019	3673	3301	1865	7202	3922	1.8	22.2
77	122.7	163.6	35.8	28.8	5.1	1.0	LCH	97393	941	936	1078	4305	1500	2.9	5.1
78	18.4	30.6	31.1	28.6	7.1	1.5	LCH	23716	231	231	1082	3143	1200	2.6	3.0
79	30.7	19.8	49.4	77.8	23.9	3.5	Res	7418	79	79	1147	1907	1700	1.1	2.5
80	15.9	17.7	35.3	48.3	11.0	2.0	Res	11960	118	118	1037	2052	970	2.1	1.6
81	326.4	384.0	33.4	18.7	4.1	1.0	LCH	107671	2369	2369	2092	4880	1512	3.2	5.8
82	141.0	176.2	52.7	40.7	9.3	2.0	LCH	56729	610	609	761	3632	1010	3.6	2.9
83	423.7	498.5	58.2	84.8	12.7	3.0	Res	273486	3735	3733	2345	8269	4770	1.7	31.0
84	482.5	740.1	51.9	66.6	13.6	2.5	Res	253781	13607	13598	3908	8944	3000	3.0	21.1
85	197.0	218.9	39.2	23.2	4.7	1.0	Res	62013	733	731	1701	2636	1419	1.9	2.9
86	272.8	303.1	40.2	26.1	5.3	1.0	Res	68885	946	937	829	3009	1600	1.9	3.8
87	33.0	660.3	29.5	30.0	7.2	1.5	Res	122357	1885	1814	796	4955	1900	2.6	7.4
88	178.7	223.4	42.6	41.0	6.1	2.0	Res	62268	1601	1595	2619	3211	1705	1.9	4.3
89	119.8	184.4	50.4	25.5	6.5	1.5	Res	23408	302	189	1824	1699	1086	1.6	1.4
90	166.0	110.6	40.4	10.6	2.0	1.0	Res	19441	137	123	274	1228	860	1.4	0.8
91	521.8	226.9	55.0	57.5	13.5	2.5	Res	112707	2212	2160	1787	3794	3600	1.1	10.7
92	126.7	31.7	100.6	10.2	2.5	1.0	Res	12903	203	196	1050	1685	570	3.0	0.8
93	565.3	706.0	50.6	70.5	9.9	2.5	Res	225842	4846	4839	2215	6513	3700	1.8	18.9
94	648.4	498.6	56.4	44.2	10.9	2.0	Res	131521	2013	1973	674	5577	1885	3.0	8.3
95	85.3	89.8	37.0	33.9	8.7	1.5	Res	36206	520	520	492	3155	2060	1.5	5.1
96	392.7	112.2	65.2	37.6	9.6	1.5	Res	36907	631	602	760	4045	1916	2.1	6.1
97	79.5	36.1	47.1	25.2	6.2	1.5	Res	17498	160	159	808	1575	801	2.0	1.0
98	99.4	76.5	33.2	20.3	5.1	1.0	Res	40875	260	260	369	1966	1476	1.3	2.3
99	181.5	201.7	33.2	34.5	7.2	1.5	Res	82116	726	723	1891	3338	1805	1.8	4.7
100	21.7	62.1	34.5	40.2	8.4	2.0	Res	35668	350	350	1469	3499	1001	3.5	2.8
101	20.8	37.9	34.5	21.0	4.8	1.0	Res	12648	64	64	64	1648	592	2.8	0.8
102	420.5	97.8	36.7	37.9	10.1	1.5	Res	52275	973	641	1709	2943	2270	1.3	5.2
103	284.9	335.2	54.5	50.5	12.5	2.5	LCH	123405	1830	1830	1778	5436	1748	3.1	7.5
104	539.0	718.6	47.6	62.4	14.4	2.5	LCH	227840	2876	2876	928	5651	2355	2.4	10.5
105	101.3	168.8	157.0	84.7	14.4	3.0	Res/BCI	62285	811	811	2872	2761	2600	1.1	5.6
106	118.1	131.3	71.7	106.8	16.2	3.5	Res/BCI	18933	745	745	2252	3146	2000	1.6	4.9
107	235.7	277.3	64.2	111.0	16.5	3.5	Res/BCI	86361	1233	1229	2368	4865	3363	1.4	12.8
108	526.9	619.9	42.4	34.7	8.5	1.5	LCH	133596	2068	2054	152	4628	1760	2.6	6.4

Stellenbosch University https://scholar.sun.ac.za

109	217.2	395.0	35.6	48.1	8.4	2.0	LCH	107152	3365	3168	2836	4143	1623	2.6	5.3
110	46.6	71.8	28.5	36.3	6.3	1.5	Res	23648	1778	1778	866	2053	580	3.5	0.9
111	87.0	54.4	35.9	25.5	5.8	1.0	Res	15570	239	239	383	1860	597	3.1	0.9
112	115.4	104.9	106.1	33.2	7.8	1.5	Res	41436	389	387	1000	2600	2067	1.3	4.2
113	324.7	72.5	64.5	28.4	6.2	1.5	Res/BCI	19052	186	145	126	1385	1104	1.3	1.2
114	167.3	41.9	75.0	61.4	12.8	2.5	Res/BCI	15826	135	116	499	1482	1100	1.3	1.3
115	177.4	101.1	62.5	113.8	15.1	3.5	Res	31969	557	553	1896	2436	1691	1.4	3.2
116	508.0	133.7	52.9	92.5	14.6	3.0	Res	67987	1749	1724	2303	3416	2235	1.5	6.0
117	364.0	728.0	60.9	86.3	15.8	3.0	LCH	110061	3330	3257	3438	4143	4007	1.0	13.0
118	60.0	32.5	62.6	53.0	11.9	2.0	Res/BCI	13687	245	245	1713	2413	1234	2.0	2.3
119	622.1	829.5	65.9	111.6	18.6	4.0	LCH	309940	5697	5691	2108	6344	3930	1.6	19.6
120	259.2	259.2	70.5	73.5	13.5	3.0	LCH	117699	2148	2145	1301	3217	2500	1.3	6.3
121	198.8	110.5	58.0	75.6	13.0	3.0	LCH	38588	590	588	1348	2460	1416	1.7	2.7
122	566.7	435.9	67.2	115.4	17.9	3.5	Res	177488	5976	5971	4355	5355	4900	1.1	20.6
123	238.2	93.4	65.0	82.5	11.9	2.5	Res	36102	761	699	2266	2514	1624	1.5	3.2
124	155.0	72.1	82.6	47.3	9.1	2.0	Res	35405	1239	1219	2972	4067	1682	2.4	5.4
125	137.5	110.0	40.3	62.5	10.0	2.0	Res/BCI	37970	1108	1107	1598	2918	1065	2.7	2.4
126	50.5	28.1	38.3	28.7	7.3	1.5	Res	20016	231	231	984	2107	1566	1.3	2.6
127	104.6	72.1	43.6	44.9	7.6	2.0	Res	43801	794	794	797	3386	1475	2.3	3.9
128	73.5	91.8	42.8	54.9	6.4	2.0	Res/LCH	22891	315	314	2372	2300	1165	2.0	2.1
129	495.4	660.5	43.9	39.0	7.7	1.5	Res/LCH	230764	3738	3732	2348	5517	3151	1.8	13.7
130	660.6	412.9	50.9	49.3	12.0	2.0	Res/BCI	161701	3246	3034	1431	6933	4384	1.6	23.9
131	135.8	61.7	52.1	35.0	6.6	1.5	Res	33092	338	321	1148	2141	1800	1.2	3.0
132	103.2	98.3	60.4	21.4	5.7	1.0	Res	37974	328	297	1830	2816	1227	2.3	2.7
133	119.3	125.6	43.1	56.1	9.5	2.0	Res	77840	1904	1904	2493	6443	1845	3.5	9.3
134	60.2	201.8	44.2	67.5	15.1	2.5	Res	96348	1926	1904	4062	7422	2162	3.4	12.6
135	74.4	67.7	36.3	40.8	10.9	2.0	Res	35375	509	508	319	3654	2027	1.8	5.8
136	82.2	27.4	29.0	45.9	11.7	2.0	Res/BCI	15678	453	450	2841	1560	920	1.7	1.1
137	57.8	26.9	43.9	80.8	21.0	3.5	Res	17461	291	291	607	1875	1107	1.7	1.6
138	44.4	17.7	36.5	45.8	10.6	2.0	Res	13386	298	298	510	1459	1438	1.0	1.6
139	4.9	5.5	57.4	95.3	26.0	4.5	Res	5193	36	36	425	823	491	1.7	0.3
140	16.0	26.6	66.9	138.4	29.5	1.0	LCH	19691	142	96	546	1592	1537	1.0	1.9
141	26.8	25.6	48.5	72.6	14.8	3.0	LCH	13279	140	140	534	2539	1008	2.5	2.0
142	14.0	12.2	76.8	100.0	18.3	3.5	LCH	6661	48	48	402	868	570	1.5	0.4
143	9.9	28.4	68.1	106.8	21.6	4.0	LCH	9723	53	53	783	2323	1407	1.7	2.6
144	47.7	12.2	72.8	93.3	14.3	3.0	LCH	6131	51	45	705	977	620	1.6	0.5

Stellenbosch University https://scholar.sun.ac.za

145	144.3	49.8	48.7	53.5	6.7	2.0	LCH	21988	246	245	989	1841	1272	1.4	1.8
146	0.0	4.8	34.1	75.3	11.3	2.5	LCH	11362	43	43	518	1633	1600	1.0	2.1
147	104.6	18.4	66.3	95.2	17.8	3.0	LCH	32886	340	339	210	2320	1900	1.2	3.5
148	111.5	30.1	48.7	82.1	14.2	3.0	LCH	35774	709	791	890	1953	1178	1.7	1.8
149	85.0	94.4	81.4	130.1	24.8	5.0	Res	31591	423	423	716	3773	1809	2.1	5.4
150	50.9	63.7	76.8	150.5	26.1	5.0	Res	22769	282	282	1131	2603	1207	2.2	2.5
151	85.0	75.1	75.1	219.1	41.3		Res	31591	423	425	716	3773	1809	2.1	5.4
152	50.9	64.2	47.2	179.2	25.4		Res	22769	282	284	1131	2603	1207	2.2	2.5
153	25.2	26.5	41.4	61.9	10.0	2.0	LCH	14331	118	118	1327	976	755	1.3	0.6
154	18.8	19.8	66.9	91.2	17.5	3.0	LCH	16503	71	71	672	1596	593	2.7	0.7
155	19.0	31.6	62.8	86.2	16.7	3.0	LCH	36818	289	284	484	3233	1458	2.2	3.7
156	9.2	16.7	50.4	70.1	12.6	3.0	LCH	8593	76	75	639	1983	807	2.5	1.3
157	28.8	20.6	54.6	100.2	22.1	4.0	LCH	13160	210	210	508	1825	1021	1.8	1.5
158	5.6	4.9	65.7	99.0	17.1	3.0	LCH	7432	67	63	1034	2391	1266	1.9	2.4
159	9.4	14.4	62.0	126.0	21.5	4.0	LCH	14807	308	301	2803	3005	828	3.6	2.0
160	2.0	10.0	51.4	79.7	16.2	3.0	LCH	16218	76	76	1629	2301	1041	2.2	1.9
161	15.7	5.4	56.1	93.7	17.2	3.0	LCH	6839	57	57	155	1158	674	1.7	0.6
162	30.8	24.7	73.0	115.9	24.5	4.5	LCH	19998	113	113	1554	2573	1139	2.3	2.3
163	39.1	39.1	73.8	124.4	26.6	4.5	LCH	49052	355	355	669	4313	1575	2.7	5.3
164	86.9	30.0	59.6	64.4	15.6	2.5	Res	18023	188	187	453	1488	1290	1.2	1.5
165	26.1	20.1	57.8	84.0	16.0	3.0	LCH	5325	43	43	378	1182	370	3.2	0.3

Stellenbosch University https://scholar.sun.ac.za

Appendix

10.3. Transformed data

10.3.1 Data summary

Table 19: Transformed data

Count	Output (l/s)	Reservoir elevation above mean terrain elevation (m)	Terrain index	Land use	Total pipeline length (m)	Total pipeline volume (m^3)	Reservoir distance from the centre of area (m)	Shape factor ratio	Area (km^2)
1	59.0	59.0	3.0	Res	23280.5	248.4	2156.0	1.9	3.3
2	42.1	42.1	3.0	LCH	81170.2	1452.5	807.0	1.4	19.6
3	34.5	34.5	4.0	LCH	63823.9	693.1	883.0	2.5	10.2
9	32.3	32.3	2.5	Res	6603.8	170.5	2090.0	2.2	1.8
10	85.8	85.8	3.5	LCH	70887.6	1236.8	4089.0	2.1	3.3
11	66.1	66.1	4.0	LCH	73892.4	1120.1	1396.0	1.9	6.8
12	82.3	82.3	3.5	Res/LCH	120309.0	4421.5	2789.0	1.6	22.9
14	39.5	39.5	2.0	Res	8939.9	67.8	962.0	1.4	0.7
15	27.1	27.1	1.0	Res	4626.7	47.9	138.0	4.3	0.3
16	73.3	73.3	4.0	LCH	9779.6	97.4	1939.0	1.5	0.6
17	32.8	32.8	1.5	LCH	6715.9	32.9	292.0	1.1	0.4
18	56.7	56.7	2.5	Res	51381.8	1014.0	2428.0	1.1	10.2
19	66.6	66.6	3.5	LCH	83796.5	1209.1	1167.0	1.0	6.4
20	39.4	39.4	2.0	Res	48297.5	686.8	1924.0	1.4	3.7
21	72.3	72.3	3.0	Res	29871.1	473.6	1520.0	2.4	3.5
22	95.4	95.4	3.5	Res	157017.1	3319.3	3557.0	1.1	11.3
23	76.4	76.4	3.5	Res	116727.5	2189.1	2087.0	1.1	12.2
24	79.1	79.1	4.0	Res	124865.4	3241.6	1518.0	2.4	17.9
26	66.5	66.5	3.5	Res	29823.7	374.0	692.0	1.5	3.9
27	77.4	77.4	4.0	Res	122794.9	3565.5	1891.0	1.5	23.1
28	65.8	65.8	4.0	Res	109290.5	4472.6	2073.0	1.1	10.7
29	52.2	52.2	3.0	CBD	97663.6	4372.5	1980.0	1.6	8.0
30	53.5	53.5	3.0	Res	58297.3	1157.9	1945.0	2.5	7.6
31	76.0	76.0	4.0	Res/BCI	65290.0	1478.5	2582.0	1.3	8.4
32	79.1	79.1	3.5	Res/CBD	83793.8	3234.5	2820.0	1.2	9.4
33	67.0	67.0	3.5	Res	95243.7	2100.4	2717.0	2.4	6.7
34	60.8	60.8	3.0	Res	34750.3	627.9	1975.0	2.8	3.6
35	46.9	46.9	3.0	Res	69593.7	2287.2	1576.0	2.5	7.2
36	64.5	64.5	3.5	Res	58553.9	1302.1	609.0	1.3	6.7
37	71.3	71.3	4.0	Res	99618.6	2674.9	981.0	2.6	12.6
38	66.5	66.5	4.0	Res	55570.3	1180.2	1636.0	2.3	8.4
39	50.5	50.5	2.0	Res	12552.3	170.0	1162.0	2.4	1.3
40	54.2	54.2	3.0	Res	13169.5	148.4	648.0	1.1	1.9

Stellenbosch University https://scholar.sun.ac.za

41	40.2	40.2	1.5	Res	6005.0	67.9	546.0	1.0	0.9
42	27.1	27.1	2.5	Res	5716.8	54.0	790.0	2.3	0.7
43	56.5	56.5	2.5	Res	12734.0	99.7	339.0	4.1	1.9
44	77.2	77.2	5.0	Res	92074.7	2476.5	1608.0	1.9	14.2
45	58.7	58.7	4.0	Res	104953.0	2581.9	0.0	2.1	17.3
46	67.5	67.5	3.5	Res	27828.0	232.5	1703.0	1.2	4.6
47	30.7	30.7	2.5	Res/BCI	22152.3	1156.0	1059.0	2.2	2.1
48	41.5	41.5	2.0	Res	10104.7	106.9	557.0	1.7	1.1
49	44.7	44.7	1.5	Res	25292.0	566.2	1360.0	1.1	4.4
50	34.4	34.4	2.5	Res	16569.0	553.0	1678.0	2.5	1.1
51	75.4	75.4	4.5	LCH	192668.8	4333.7	2541.0	1.2	32.3
53	60.2	60.2	3.0	LCH	93141.4	1713.5	1857.0	1.4	6.9
54	45.0	45.0	2.5	LCH	113308.8	1636.1	0.0	1.2	13.9
55	79.7	79.7	4.0	LCH	138053.8	2383.8	1825.0	2.3	13.3
56	47.5	47.5	3.0	LCH	56041.5	705.2	979.0	4.0	3.8
57	65.1	65.1	3.0	LCH	17578.5	205.2	760.0	2.5	1.4
59	57.8	57.8	2.5	Res/LCH	128223.9	2517.3	3100.0	2.0	15.4
60	76.3	76.3	3.0	Res	105694.8	2567.0	2297.0	1.4	17.4
64	49.5	49.5	3.5	Res	35267.2	601.2	1364.0	1.8	8.0
65	70.4	70.4	4.0	Res	66204.4	1395.8	1400.0	2.2	6.8
66	74.9	74.9	4.5	Res	42148.4	707.2	1728.0	1.6	2.7
67	63.2	63.2	4.5	Res	161570.5	2741.0	588.0	2.1	13.3
68	40.3	40.3	1.5	Res	17967.1	155.1	500.0	1.1	2.0
69	45.2	45.2	3.5	Res	93090.7	2112.0	0.0	1.6	15.7
70	46.1	46.1	2.5	Res/BCI	48649.4	783.4	280.0	1.6	5.7
71	62.9	62.9	4.0	Res	99856.8	2592.6	1804.0	2.3	11.4
72	53.8	53.8	3.0	Res/BCI	39053.2	1284.2	1490.0	1.5	5.9
73	44.3	44.3	2.0	Res/BCI	21305.4	293.9	627.0	1.4	2.1
74	49.7	49.7	2.5	Res	54671.6	898.1	1173.0	1.5	10.3
75	52.3	52.3	3.5	Res	43044.1	1011.2	1574.0	1.1	6.7
76	79.3	79.3	4.5	Res	147018.6	3672.5	1865.0	1.8	22.2
77	35.8	35.8	1.0	LCH	97392.7	941.3	1078.0	2.9	5.1
78	31.1	31.1	1.5	LCH	23716.2	231.3	1082.0	2.6	3.0
79	49.4	49.4	3.5	Res	7417.5	78.6	1147.0	1.1	2.5
80	35.3	35.3	2.0	Res	11960.0	118.4	1037.0	2.1	1.6
81	33.4	33.4	1.0	LCH	107671.4	2369.3	2092.0	3.2	5.8
82	52.7	52.7	2.0	LCH	56728.7	609.5	761.0	3.6	2.9
83	58.2	58.2	3.0	Res	273486.1	3734.5	2345.0	1.7	31.0
85	39.2	39.2	1.0	Res	62013.2	732.5	1701.0	1.9	2.9
86	40.2	40.2	1.0	Res	68884.7	946.4	829.0	1.9	3.8
88	42.6	42.6	2.0	Res	62268.0	1600.5	2619.0	1.9	4.3
89	50.4	50.4	1.5	Res	23408.3	301.6	1824.0	1.6	1.4
90	40.4	40.4	1.0	Res	19441.2	136.9	274.0	1.4	0.8
91	55.0	55.0	2.5	Res	112707.3	2211.5	1787.0	1.1	10.7

Stellenbosch University https://scholar.sun.ac.za

92	100.6	100.6	1.0	Res	12903.1	203.3	1050.0	3.0	0.8
93	50.6	50.6	2.5	Res	225841.8	4846.0	2215.0	1.8	18.9
94	56.4	56.4	2.0	Res	131521.2	2012.8	674.0	3.0	8.3
95	37.0	37.0	1.5	Res	36206.4	520.1	492.0	1.5	5.1
96	65.2	65.2	1.5	Res	36906.8	631.0	760.0	2.1	6.1
97	47.1	47.1	1.5	Res	17498.0	159.7	808.0	2.0	1.0
98	33.2	33.2	1.0	Res	40875.3	259.5	369.0	1.3	2.3
99	33.2	33.2	1.5	Res	82116.4	726.1	1891.0	1.8	4.7
100	34.5	34.5	2.0	Res	35667.8	350.0	1469.0	3.5	2.8
102	36.7	36.7	1.5	Res	52275.3	972.8	1709.0	1.3	5.2
103	54.5	54.5	2.5	LCH	123405.0	1830.3	1778.0	3.1	7.5
104	47.6	47.6	2.5	LCH	227840.1	2875.9	928.0	2.4	10.5
106	118.1	71.7	3.5	Res/BCI	18933.3	745.1	2252.0	1.6	4.9
107	235.7	64.2	3.5	Res/BCI	86360.5	1233.3	2368.0	1.4	12.8
108	526.9	42.4	1.5	LCH	133596.1	2068.3	152.0	2.6	6.4
109	217.2	35.6	2.0	LCH	107151.8	3364.7	2836.0	2.6	5.3
110	46.6	28.5	1.5	Res	23648.1	1778.4	866.0	3.5	0.9
111	87.0	35.9	1.0	Res	15569.6	239.4	383.0	3.1	0.9
112	115.4	106.1	1.5	Res	41436.4	389.0	1000.0	1.3	4.2
113	324.7	64.5	1.5	Res/BCI	19051.6	185.7	126.0	1.3	1.2
114	167.3	75.0	2.5	Res/BCI	15826.4	134.8	499.0	1.3	1.3
115	177.4	62.5	3.5	Res	31969.0	556.5	1896.0	1.4	3.2
116	508.0	52.9	3.0	Res	67987.1	1749.0	2303.0	1.5	6.0
117	364.0	60.9	3.0	LCH	110060.7	3329.9	3438.0	1.0	13.0
118	60.0	62.6	2.0	Res/BCI	13686.8	245.4	1713.0	2.0	2.3
120	259.2	70.5	3.0	LCH	117698.7	2148.0	1301.0	1.3	6.3
121	198.8	58.0	3.0	LCH	38588.2	589.5	1348.0	1.7	2.7
122	566.7	67.2	3.5	Res	177487.9	5976.1	4355.0	1.1	20.6
123	238.2	65.0	2.5	Res	36102.4	760.9	2266.0	1.5	3.2
124	155.0	82.6	2.0	Res	35404.9	1238.9	2972.0	2.4	5.4
125	137.5	40.3	2.0	Res/BCI	37970.2	1107.7	1598.0	2.7	2.4
126	50.5	38.3	1.5	Res	20016.1	231.2	984.0	1.3	2.6
127	104.6	43.6	2.0	Res	43801.2	793.7	797.0	2.3	3.9
128	73.5	42.8	2.0	Res/LCH	22891.2	314.5	2372.0	2.0	2.1
129	495.4	43.9	1.5	Res/LCH	230764.0	3738.1	2348.0	1.8	13.7
130	660.6	50.9	2.0	Res/BCI	161701.2	3246.3	1431.0	1.6	23.9
131	135.8	52.1	1.5	Res	33091.5	337.8	1148.0	1.2	3.0
132	103.2	60.4	1.0	Res	37974.4	328.1	1830.0	2.3	2.7
133	119.3	43.1	2.0	Res	77839.9	1904.4	2493.0	3.5	9.3
134	60.2	44.2	2.5	Res	96348.3	1925.6	4062.0	3.4	12.6
135	74.4	36.3	2.0	Res	35375.1	509.0	319.0	1.8	5.8
136	82.2	29.0	2.0	Res/BCI	15677.7	453.4	2841.0	1.7	1.1
137	43.9	43.9	3.5	Res	17460.9	290.9	607.0	1.7	1.6
138	36.5	36.5	2.0	Res	13386.1	298.4	510.0	1.0	1.6

Stellenbosch University https://scholar.sun.ac.za

139	57.4	57.4	4.5	Res	5192.6	36.3	425.0	1.7	0.3
140	66.9	66.9	1.0	LCH	19691.4	142.3	546.0	1.0	1.9
141	48.5	48.5	3.0	LCH	13279.2	139.6	534.0	2.5	2.0
142	76.8	76.8	3.5	LCH	6660.8	47.6	402.0	1.5	0.4
143	68.1	68.1	4.0	LCH	9722.7	53.2	783.0	1.7	2.6
144	72.8	72.8	3.0	LCH	6130.9	51.4	705.0	1.6	0.5
145	48.7	48.7	2.0	LCH	21988.4	245.6	989.0	1.4	1.8
147	66.3	66.3	3.0	LCH	32885.9	339.8	210.0	1.2	3.5
148	48.7	48.7	3.0	LCH	35774.1	708.6	890.0	1.7	1.8
149	81.4	81.4	5.0	Res	31590.9	423.4	716.0	2.1	5.4
150	76.8	76.8	5.0	Res	22769.2	282.3	1131.0	2.2	2.5
153	41.4	41.4	2.0	LCH	14330.7	118.1	1327.0	1.3	0.6
154	66.9	66.9	3.0	LCH	16503.1	70.7	672.0	2.7	0.7
155	62.8	62.8	3.0	LCH	36818.3	289.0	484.0	2.2	3.7
156	50.4	50.4	3.0	LCH	8593.1	76.2	639.0	2.5	1.3
157	54.6	54.6	4.0	LCH	13159.5	210.0	508.0	1.8	1.5
158	65.7	65.7	3.0	LCH	7432.1	66.6	1034.0	1.9	2.4
159	62.0	62.0	4.0	LCH	14807.2	307.6	2803.0	3.6	2.0
160	51.4	51.4	3.0	LCH	16217.8	75.8	1629.0	2.2	1.9
161	56.1	56.1	3.0	LCH	6838.7	56.5	155.0	1.7	0.6
162.0	73.0	73.0	4.5	LCH	19998.1	113.2	1554.0	2.3	2.3
163	73.8	73.8	4.5	LCH	49051.7	355.3	669.0	2.7	5.3
164	59.6	59.6	2.5	Res	18022.9	188.2	453.0	1.2	1.5
165	57.8	57.8	3.0	LCH	5324.6	43.4	378.0	3.2	0.3

Stellenbosch University https://scholar.sun.ac.za

Appendix

10.3.2 Multi Linear Regression

Table 20: Multi Linear Regression summary

7 x-variables								
General land use								
Method	**Formula**							
	Intercept	**Reservoir elevation**	**Terrain index**	**Pipeline length**	**Pipeline volume**	**Reservoir distance**	**Shape factor**	**Area**
Standard regression	22.176	2.590	33.497	0.001	0.147	-0.004	12.983	-8.995
Removing multicollinearity	-30.108	2.761	26.366	0.003		0.028	21.185	-0.502
Low p-values	-80.967	2.099		0.003				
Residential land use								
Method	**Formula**							
	Intercept	**Reservoir elevation**	**Terrain index**	**Pipeline length**	**Pipeline volume**	**Reservoir distance**	**Shape factor**	**Area**
Standard regression	-3.519	3.755	49.903	0.001	0.125	0.002	11.072	-0.730
Removing multicollinearity	4.081	3.740	53.145		0.138	0.004	10.053	1.214
Low p-values	-12.034	3.759	52.665		0.144			
Low-Cost Housing land use								
Method	**Formula**							
	Intercept	**Reservoir elevation**	**Terrain index**	**Pipeline length**	**Pipeline volume**	**Reservoir distance**	**Shape factor**	**Area**
Standard regression	-9.907	0.397	3.992	0.002	0.085	-0.004	-2.567	-12.450
Removing multicollinearity	-24.755	0.623	-2.223	0.003		0.011	-4.675	-8.185
Low p-values	4.764			0.003				-7.874
3 x-variables								
General land use								
Method	**Formula**							
	Intercept	**Reservoir elevation**	**Terrain index**	**Pipeline length**	**Pipeline volume**	**Reservoir distance**	**Shape factor**	**Area**
Standard regression	39.621			0.001	0.151			-9.632
Low p-values	47.915				0.170			-7.375
Residential land use								
Method	**Formula**							
	Intercept	**Reservoir elevation**	**Terrain index**	**Pipeline length**	**Pipeline volume**	**Reservoir distance**	**Shape factor**	**Area**
Standard regression	39.533			0.001	0.121			-2.456
Low p-values	46.429				0.145			
Low-Cost Housing land use								
Method	**Formula**							
	Intercept	**Reservoir elevation**	**Terrain index**	**Pipeline length**	**Pipeline volume**	**Reservoir distance**	**Shape factor**	**Area**
Standard regression	16.033			0.002	0.083			-11.679
Low p-values								

Stellenbosch University https://scholar.sun.ac.za

10.3.3 Principal Component Analysis

Table 21: Principal components summary

7 x-variables							
General land use							
Variable	**pc1**	**pc2**	**pc3**	**pc4**	**pc5**	**pc6**	**pc7**
Reservoir elevation above mean terrain elevation	-0.309	-0.623	0.024	-0.078	0.713	-0.037	-0.014
Terrain index	-0.470	0.338	0.091	0.161	0.136	0.291	0.727
Total pipeline length	-0.496	0.242	0.024	0.068	0.016	0.508	-0.657
Total pipeline volume	-0.335	0.001	0.086	-0.894	-0.250	-0.121	0.065
Reservoir distance from centre of area	0.087	0.020	0.989	0.060	0.024	-0.068	-0.068
Shape factor ratio	-0.483	0.203	-0.042	0.285	-0.043	-0.787	-0.142
Area	-0.290	-0.631	0.053	0.281	-0.639	0.131	0.104
Residential land use							
Variable	**pc1**	**pc2**	**pc3**	**pc4**	**pc5**	**pc6**	**pc7**
Reservoir elevation above mean terrain elevation	-0.341	-0.540	-0.248	-0.176	-0.707	0.008	0.007
Terrain index	-0.324	-0.577	-0.288	0.150	0.662	0.038	0.130
Total pipeline length	-0.470	0.262	0.247	0.136	-0.085	0.151	0.775
Total pipeline volume	-0.488	0.163	0.174	0.119	0.023	0.625	-0.547
Reservoir distance from centre of area	-0.294	0.245	-0.135	-0.880	0.220	-0.112	-0.018
Shape factor ratio	0.067	-0.449	0.850	-0.254	0.075	-0.038	-0.019
Area	-0.478	0.138	0.153	0.274	-0.009	-0.756	-0.287
Low-Cost Housing land use							
Variable	**pc1**	**pc2**	**pc3**	**pc4**	**pc5**	**pc6**	**pc7**
Reservoir elevation above mean terrain elevation	0.103	0.676	-0.012	-0.092	0.672	-0.268	0.033
Terrain index	0.152	0.624	0.017	0.515	-0.431	0.365	-0.056
Total pipeline length	0.530	-0.220	0.004	0.083	0.328	0.509	0.545
Total pipeline volume	0.562	-0.152	-0.017	-0.022	0.144	0.107	-0.792
Reservoir distance from centre of area	0.311	0.227	0.498	-0.674	-0.369	-0.009	0.117
Shape factor ratio	-0.033	-0.161	0.834	0.456	0.166	-0.201	-0.020
Area	0.520	-0.073	-0.233	0.239	-0.266	-0.696	0.239

Stellenbosch University https://scholar.sun.ac.za

Appendix

(Continued from above) 3 x-variables							
General land use							
Variable	**pc1**	**pc2**	**pc3**	**pc4**	**pc5**	**pc6**	**pc7**
Reservoir elevation above mean terrain elevation							
Terrain index							
Total pipeline length	0.580	-0.409	0.705				
Total pipeline volume	0.580	-0.401	-0.709				
Reservoir distance from centre of area							
Shape factor ratio							
Area	0.572	0.820	0.005				
Residential land use							
Variable	**pc1**	**pc2**	**pc3**	**pc4**	**pc5**	**pc6**	**pc7**
Reservoir elevation above mean terrain elevation							
Terrain index							
Total pipeline length	0.581	-0.273	0.767				
Total pipeline volume	0.578	-0.525	-0.624				
Reservoir distance from centre of area							
Shape factor ratio							
Area	0.573	0.806	-0.147				
Low-Cost Housing land use							
Variable	**pc1**	**pc2**	**pc3**	**pc4**	**pc5**	**pc6**	**pc7**
Reservoir elevation above mean terrain elevation							
Terrain index							
Total pipeline length	-0.579	0.550	0.602				
Total pipeline volume	-0.595	0.219	-0.773				
Reservoir distance from centre of area							
Shape factor ratio							
Area	-0.557	-0.806	0.200				

Stellenbosch University https://scholar.sun.ac.za

Table 22: Principal Component Analysis summary

7 x-variables								
General land use								
Method	**Formula**							
	Intercept	**pc1**	**pc2**	**pc3**	**pc4**	**pc5**	**pc6**	**pc7**
Standard regression	22.176	12.069	17.478	-14.172	-12.947	23.321	3.557	-1.448
Low p-values	-19.258	8.841	5.410		-7.257	10.793	6.993	
Residential land use								
Method	**Formula**							
	Intercept	**pc1**	**pc2**	**pc3**	**pc4**	**pc5**	**pc6**	**pc7**
Standard regression	-3.519	14.428	31.660	3.960	-5.509	-36.513	-0.842	-6.087
Low p-values	2.530	14.087	32.499		-4.497	-35.822		-5.579
Low-Cost Housing land use								
Method	**Formula**							
	Intercept	**pc1**	**pc2**	**pc3**	**pc4**	**pc5**	**pc6**	**pc7**
Standard regression	-9.907	-5.699	4.073	0.825	-2.131	1.450	10.538	-3.202
Low p-values	-7.795	-5.391	5.986		-1.220		12.067	-3.417
3 x-variables								
General land use								
Method	**Formula**							
	Intercept	**pc1**	**pc2**	**pc3**	**pc4**	**pc5**	**pc6**	**pc7**
Standard regression	39.621	-5.425	-7.959	-0.154				
Low p-values								
Residential land use								
Method	**Formula**							
	Intercept	**pc1**	**pc2**	**pc3**	**pc4**	**pc5**	**pc6**	**pc7**
Standard regression	39.533	-1.337	-2.043	0.286				
Low p-values	38.447	-0.215	-0.461					
Low-Cost Housing land use								
Method	**Formula**							
	Intercept	**pc1**	**pc2**	**pc3**	**pc4**	**pc5**	**pc6**	**pc7**
Standard regression	16.033	6.456	9.432	-2.403				
Low p-values								

Stellenbosch University https://scholar.sun.ac.za

10.3.4 Partial Least Squares

Table 23: Partial Least Squares summary

7 x-variables								
General land use								
Method	**Formula**							
	Intercept	**Reservoir elevation**	**Terrain index**	**Pipeline length**	**Pipeline volume**	**Reservoir distance**	**Shape factor**	**Area**
PLS	-69.804	1.319	9.961	0.001	0.041	0.034	-10.059	6.089
Residential land use								
Method	**Formula**							
	Intercept	**Reservoir elevation**	**Terrain index**	**Pipeline length**	**Pipeline volume**	**Reservoir distance**	**Shape factor**	**Area**
PLS	-97.920	1.763	14.269	0.001	0.041	0.035	-7.728	6.928
Low-Cost Housing land use								
Method	**Formula**							
	Intercept	**Reservoir elevation**	**Terrain index**	**Pipeline length**	**Pipeline volume**	**Reservoir distance**	**Shape factor**	**Area**
PLS	-9.463	0.089	-2.181	0.001	0.037	0.020	5.272	3.609
3 x-variables								
General land use								
Method	**Formula**							
	Intercept	**Reservoir elevation**	**Terrain index**	**Pipeline length**	**Pipeline volume**	**Reservoir distance**	**Shape factor**	**Area**
PLS	27.280			0.001	0.051			7.467
Residential land use								
Method	**Formula**							
	Intercept	**Reservoir elevation**	**Terrain index**	**Pipeline length**	**Pipeline volume**	**Reservoir distance**	**Shape factor**	**Area**
PLS	30.790			0.001	0.052			8.822
Low-Cost Housing land use								
Method	**Formula**							
	Intercept	**Reservoir elevation**	**Terrain index**	**Pipeline length**	**Pipeline volume**	**Reservoir distance**	**Shape factor**	**Area**
PLS	17.147			0.001	0.039			3.830

Stellenbosch University https://scholar.sun.ac.za

10.4. Untransformed data

10.4.1 Data summary

Table 24: Untransformed data

Count	Output (l/s)	Reservoir elevation above mean terrain elevation (m)	Terrain index	Land use	Total pipeline length (m)	Total pipeline volume (m^3)	Reservoir distance from the centre of area (m)	Shape factor ratio	Area (km^2)
1	18.5	59.0	3.0	Res	23280.5	248.4	2156.0	1.9	3.3
2	84.5	42.1	3.0	LCH	81170.2	690.7	807.0	1.4	19.6
3	156.4	34.5	4.0	LCH	63823.9	664.3	883.0	2.5	10.2
9	13.0	32.3	2.5	Res	6603.8	170.5	2090.0	2.2	1.8
10	183.7	85.8	3.5	LCH	70887.6	1236.4	4089.0	2.1	3.3
11	217.9	66.1	4.0	LCH	73892.4	1097.7	1396.0	1.9	6.8
12	311.2	82.3	3.5	Res/LCH	120309.0	4421.5	2789.0	1.6	22.9
14	2.3	39.5	2.0	Res	8939.9	67.8	962.0	1.4	0.7
15	21.3	27.1	1.0	Res	4626.7	47.9	138.0	4.3	0.3
16	18.2	73.3	4.0	LCH	9779.6	95.1	1939.0	1.5	0.6
17	19.0	32.8	1.5	LCH	6715.9	32.9	292.0	1.1	0.4
18	225.1	56.7	2.5	Res	51381.8	1004.2	2428.0	1.1	10.2
19	225.8	66.6	3.5	LCH	83796.5	1193.3	1167.0	1.0	6.4
20	153.3	39.4	2.0	Res	48297.5	651.4	1924.0	1.4	3.7
21	104.2	72.3	3.0	Res	29871.1	454.0	1520.0	2.4	3.5
22	419.4	95.4	3.5	Res	157017.1	3187.7	3557.0	1.1	11.3
23	317.7	76.4	3.5	Res	116727.5	2181.3	2087.0	1.1	12.2
24	409.6	79.1	4.0	Res	124865.4	3121.8	1518.0	2.4	17.9
26	75.3	66.5	3.5	Res	29823.7	372.6	692.0	1.5	3.9
27	493.8	77.4	4.0	Res	122794.9	3542.3	1891.0	1.5	23.1
28	515.6	65.8	4.0	Res	109290.5	4456.1	2073.0	1.1	10.7
29	495.4	52.2	3.0	CBD	97663.6	4365.0	1980.0	1.6	8.0
30	195.3	53.5	3.0	Res	58297.3	1155.3	1945.0	2.5	7.6
31	256.0	76.0	4.0	Res/BCI	65290.0	1475.3	2582.0	1.3	8.4
32	575.7	79.1	3.5	Res/CBD	83793.8	3093.8	2820.0	1.2	9.4
33	306.1	67.0	3.5	Res	95243.7	2078.5	2717.0	2.4	6.7
34	107.5	60.8	3.0	Res	34750.3	627.7	1975.0	2.8	3.6
35	254.3	46.9	3.0	Res	69593.7	2287.2	1576.0	2.5	7.2
36	247.8	64.5	3.5	Res	58553.9	1302.1	609.0	1.3	6.7
37	328.7	71.3	4.0	Res	99618.6	2674.6	981.0	2.6	12.6
38	205.9	66.5	4.0	Res	55570.3	1177.0	1636.0	2.3	8.4

Stellenbosch University https://scholar.sun.ac.za

39	47.8	50.5	2.0	Res	12552.3	170.0	1162.0	2.4	1.3
40	34.7	54.2	3.0	Res	13169.5	148.4	648.0	1.1	1.9
41	22.7	40.2	1.5	Res	6005.0	67.9	546.0	1.0	0.9
42	19.5	27.1	2.5	Res	5716.8	54.0	790.0	2.3	0.7
43	47.4	56.5	2.5	Res	12734.0	99.0	339.0	4.1	1.9
44	350.7	77.2	5.0	Res	92074.7	2587.9	1608.0	1.9	14.2
45	373.0	58.7	4.0	Res	104953.0	2580.3	0.0	2.1	17.3
46	22.3	67.5	3.5	Res	27828.0	232.5	1703.0	1.2	4.6
47	111.1	30.7	2.5	Res/BCI	22152.3	2100.3	1059.0	2.2	2.1
48	23.6	41.5	2.0	Res	10104.7	106.4	557.0	1.7	1.1
49	45.8	44.7	1.5	Res	25292.0	563.5	1360.0	1.1	4.4
50	15.0	34.4	2.5	Res	16569.0	553.0	1678.0	2.5	1.1
51	439.4	75.4	4.5	LCH	192668.8	4010.1	2541.0	1.2	32.3
53	241.7	60.2	3.0	LCH	93141.4	1710.9	1857.0	1.4	6.9
54	249.9	45.0	2.5	LCH	113308.8	1627.4	0.0	1.2	13.9
55	197.3	79.7	4.0	LCH	138053.8	2374.3	1825.0	2.3	13.3
56	178.8	47.5	3.0	LCH	56041.5	705.0	979.0	4.0	3.8
57	57.0	65.1	3.0	LCH	17578.5	204.2	760.0	2.5	1.4
59	295.7	57.8	2.5	Res/LCH	128223.9	2517.3	3100.0	2.0	15.4
60	252.8	76.3	3.0	Res	105694.8	2557.6	2297.0	1.4	17.4
64	42.1	49.5	3.5	Res	35267.2	601.2	1364.0	1.8	8.0
65	249.8	70.4	4.0	Res	66204.4	1392.5	1400.0	2.2	6.8
66	95.4	74.9	4.5	Res	42148.4	705.0	1728.0	1.6	2.7
67	506.4	63.2	4.5	Res	161570.5	2717.9	588.0	2.1	13.3
68	70.3	40.3	1.5	Res	17967.1	154.7	500.0	1.1	2.0
69	325.0	45.2	3.5	Res	93090.7	2111.0	0.0	1.6	15.7
70	163.2	46.1	2.5	Res/BCI	48649.4	780.4	280.0	1.6	5.7
71	256.4	62.9	4.0	Res	99856.8	2590.8	1804.0	2.3	11.4
72	176.0	53.8	3.0	Res/BCI	39053.2	1284.2	1490.0	1.5	5.9
73	90.7	44.3	2.0	Res/BCI	21305.4	292.5	627.0	1.4	2.1
74	152.8	49.7	2.5	Res	54671.6	898.1	1173.0	1.5	10.3
75	161.7	52.3	3.5	Res	43044.1	1011.2	1574.0	1.1	6.7
76	460.6	79.3	4.5	Res	147018.6	3301.4	1865.0	1.8	22.2
77	163.6	35.8	1.0	LCH	97392.7	935.9	1078.0	2.9	5.1
78	30.6	31.1	1.5	LCH	23716.2	231.3	1082.0	2.6	3.0
79	19.8	49.4	3.5	Res	7417.5	78.6	1147.0	1.1	2.5
80	17.7	35.3	2.0	Res	11960.0	118.4	1037.0	2.1	1.6
81	384.0	33.4	1.0	LCH	107671.4	2368.9	2092.0	3.2	5.8
82	176.2	52.7	2.0	LCH	56728.7	609.5	761.0	3.6	2.9
83	498.5	58.2	3.0	Res	273486.1	3733.2	2345.0	1.7	31.0
85	218.9	39.2	1.0	Res	62013.2	730.6	1701.0	1.9	2.9
86	303.1	40.2	1.0	Res	68884.7	936.6	829.0	1.9	3.8
88	223.4	42.6	2.0	Res	62268.0	1594.7	2619.0	1.9	4.3
89	184.4	50.4	1.5	Res	23408.3	188.6	1824.0	1.6	1.4

Stellenbosch University https://scholar.sun.ac.za

90	110.6	40.4	1.0	Res	19441.2	123.1	274.0	1.4	0.8
91	226.9	55.0	2.5	Res	112707.3	2160.0	1787.0	1.1	10.7
92	31.7	100.6	1.0	Res	12903.1	196.2	1050.0	3.0	0.8
93	706.0	50.6	2.5	Res	225841.8	4838.7	2215.0	1.8	18.9
94	498.6	56.4	2.0	Res	131521.2	1972.8	674.0	3.0	8.3
95	89.8	37.0	1.5	Res	36206.4	520.1	492.0	1.5	5.1
96	112.2	65.2	1.5	Res	36906.8	601.9	760.0	2.1	6.1
97	36.1	47.1	1.5	Res	17498.0	159.3	808.0	2.0	1.0
98	76.5	33.2	1.0	Res	40875.3	259.5	369.0	1.3	2.3
99	201.7	33.2	1.5	Res	82116.4	722.5	1891.0	1.8	4.7
100	62.1	34.5	2.0	Res	35667.8	350.0	1469.0	3.5	2.8
102	97.8	36.7	1.5	Res	52275.3	640.5	1709.0	1.3	5.2
103	335.2	54.5	2.5	LCH	123405.0	1830.3	1778.0	3.1	7.5
104	718.6	47.6	2.5	LCH	227840.1	2875.9	928.0	2.4	10.5
106	131.3	71.7	3.5	Res/BCI	18933.3	745.1	2252.0	1.6	4.9
107	277.3	64.2	3.5	Res/BCI	86360.5	1229.1	2368.0	1.4	12.8
108	619.9	42.4	1.5	LCH	133596.1	2053.9	152.0	2.6	6.4
109	395.0	35.6	2.0	LCH	107151.8	3167.9	2836.0	2.6	5.3
110	71.8	28.5	1.5	Res	23648.1	1778.4	866.0	3.5	0.9
111	54.4	35.9	1.0	Res	15569.6	239.4	383.0	3.1	0.9
112	104.9	106.1	1.5	Res	41436.4	387.2	1000.0	1.3	4.2
113	72.5	64.5	1.5	Res/BCI	19051.6	144.9	126.0	1.3	1.2
114	41.9	75.0	2.5	Res/BCI	15826.4	115.9	499.0	1.3	1.3
115	101.1	62.5	3.5	Res	31969.0	552.9	1896.0	1.4	3.2
116	133.7	52.9	3.0	Res	67987.1	1724.2	2303.0	1.5	6.0
117	728.0	60.9	3.0	LCH	110060.7	3256.9	3438.0	1.0	13.0
118	32.5	62.6	2.0	Res/BCI	13686.8	244.6	1713.0	2.0	2.3
120	259.2	70.5	3.0	LCH	117698.7	2145.3	1301.0	1.3	6.3
121	110.5	58.0	3.0	LCH	38588.2	588.1	1348.0	1.7	2.7
122	435.9	67.2	3.5	Res	177487.9	5970.6	4355.0	1.1	20.6
123	93.4	65.0	2.5	Res	36102.4	699.5	2266.0	1.5	3.2
124	72.1	82.6	2.0	Res	35404.9	1219.1	2972.0	2.4	5.4
125	110.0	40.3	2.0	Res/BCI	37970.2	1107.2	1598.0	2.7	2.4
126	28.1	38.3	1.5	Res	20016.1	231.2	984.0	1.3	2.6
127	72.1	43.6	2.0	Res	43801.2	793.7	797.0	2.3	3.9
128	91.8	42.8	2.0	Res/LCH	22891.2	314.4	2372.0	2.0	2.1
129	660.5	43.9	1.5	Res/LCH	230764.0	3732.0	2348.0	1.8	13.7
130	412.9	50.9	2.0	Res/BCI	161701.2	3033.8	1431.0	1.6	23.9
131	61.7	52.1	1.5	Res	33091.5	321.3	1148.0	1.2	3.0
132	98.3	60.4	1.0	Res	37974.4	296.6	1830.0	2.3	2.7
133	125.6	43.1	2.0	Res	77839.9	1904.4	2493.0	3.5	9.3
134	201.8	44.2	2.5	Res	96348.3	1903.5	4062.0	3.4	12.6
135	67.7	36.3	2.0	Res	35375.1	508.4	319.0	1.8	5.8
136	27.4	29.0	2.0	Res/BCI	15677.7	450.1	2841.0	1.7	1.1

Stellenbosch University https://scholar.sun.ac.za

137	26.9	43.9	3.5	Res	17460.9	290.9	607.0	1.7	1.6
138	17.7	36.5	2.0	Res	13386.1	298.4	510.0	1.0	1.6
139	5.5	57.4	4.5	Res	5192.6	36.3	425.0	1.7	0.3
140	26.6	66.9	1.0	LCH	19691.4	96.2	546.0	1.0	1.9
141	25.6	48.5	3.0	LCH	13279.2	139.5	534.0	2.5	2.0
142	12.2	76.8	3.5	LCH	6660.8	47.6	402.0	1.5	0.4
143	28.4	68.1	4.0	LCH	9722.7	53.2	783.0	1.7	2.6
144	12.2	72.8	3.0	LCH	6130.9	45.2	705.0	1.6	0.5
145	49.8	48.7	2.0	LCH	21988.4	245.0	989.0	1.4	1.8
147	18.4	66.3	3.0	LCH	32885.9	338.7	210.0	1.2	3.5
148	30.1	48.7	3.0	LCH	35774.1	790.8	890.0	1.7	1.8
149	94.4	81.4	5.0	Res	31590.9	423.4	716.0	2.1	5.4
150	63.7	76.8	5.0	Res	22769.2	282.3	1131.0	2.2	2.5
153	26.5	41.4	2.0	LCH	14330.7	118.1	1327.0	1.3	0.6
154	19.8	66.9	3.0	LCH	16503.1	70.7	672.0	2.7	0.7
155	31.6	62.8	3.0	LCH	36818.3	283.7	484.0	2.2	3.7
156	16.7	50.4	3.0	LCH	8593.1	75.3	639.0	2.5	1.3
157	20.6	54.6	4.0	LCH	13159.5	210.0	508.0	1.8	1.5
158	4.9	65.7	3.0	LCH	7432.1	62.8	1034.0	1.9	2.4
159	14.4	62.0	4.0	LCH	14807.2	301.3	2803.0	3.6	2.0
160	10.0	51.4	3.0	LCH	16217.8	75.8	1629.0	2.2	1.9
161	5.4	56.1	3.0	LCH	6838.7	56.5	155.0	1.7	0.6
162	24.7	73.0	4.5	LCH	19998.1	113.2	1554.0	2.3	2.3
163	39.1	73.8	4.5	LCH	49051.7	355.3	669.0	2.7	5.3
164	30.0	59.6	2.5	Res	18022.9	187.1	453.0	1.2	1.5
165	20.1	57.8	3.0	LCH	5324.6	43.3	378.0	3.2	0.3

Stellenbosch University https://scholar.sun.ac.za

10.4.2 Multi linear regression

Table 25: Multi-Linear Regression summary

7 x-variables								
General land use								
Method	**Formula**							
	Intercept	**Reservoir elevation**	**Terrain index**	**Pipeline length**	**Pipeline volume**	**Reservoir distance**	**Shape factor**	**Area**
Standard regression	22.790	0.155	-2.618	0.002	0.090	-0.013	-0.958	-7.397
Removing multicollinearity	42.923	0.242	-14.595		0.133	-0.012	10.524	-0.146
Low p-values	61.799		-13.106		0.128			
Residential land use								
Method	**Formula**							
	Intercept	**Reservoir elevation**	**Terrain index**	**Pipeline length**	**Pipeline volume**	**Reservoir distance**	**Shape factor**	**Area**
Standard regression	30.041	0.418	-7.126	0.002	0.095	-0.021	3.113	-4.656
Removing multicollinearity	52.392	0.394	-16.970		0.132	-0.014	5.830	1.019
Low p-values	30.123				0.128			
Low-Cost Housing land use								
Method	**Formula**							
	Intercept	**Reservoir elevation**	**Terrain index**	**Pipeline length**	**Pipeline volume**	**Reservoir distance**	**Shape factor**	**Area**
Standard regression	26.475	-1.039	8.092	0.003	0.053	-0.003	0.160	-8.862
Removing multicollinearity	62.870	-1.120	3.231		0.193	-0.022	12.876	-6.939
Low p-values	102.047	-1.490			0.184			-6.660
3 x-variables								
General land use								
Method	**Formula**							
	Intercept	**Reservoir elevation**	**Terrain index**	**Pipeline length**	**Pipeline volume**	**Reservoir distance**	**Shape factor**	**Area**
Standard regression	9.855			0.002	0.085			-7.253
Low p-values								
Residential land use								
Method	**Formula**							
	Intercept	**Reservoir elevation**	**Terrain index**	**Pipeline length**	**Pipeline volume**	**Reservoir distance**	**Shape factor**	**Area**
Standard regression	18.556			0.001	0.088			-4.333
Low p-values	16.107			0.001	0.079			
Low-Cost Housing land use								
Method	**Formula**							
	Intercept	**Reservoir elevation**	**Terrain index**	**Pipeline length**	**Pipeline volume**	**Reservoir distance**	**Shape factor**	**Area**
Standard regression	-13.978			0.003	0.039			-8.186
Low p-values	-18.911			0.003				-7.342

Stellenbosch University https://scholar.sun.ac.za

10.4.3 Principal Component Analysis

Table 26: Principal components summary

7 x-variables							
General land use							
Variable	**pc1**	**pc2**	**pc3**	**pc4**	**pc5**	**pc6**	**pc7**
Reservoir elevation above mean terrain elevation	-0.310	-0.622	0.029	-0.079	0.713	-0.018	0.032
Terrain index	-0.292	-0.629	0.060	0.281	-0.639	0.091	-0.140
Total pipeline length	-0.471	0.340	0.083	0.163	0.143	0.083	-0.776
Total pipeline volume	-0.492	0.246	0.030	0.066	0.001	0.668	0.495
Reservoir distance from centre of area	-0.337	0.004	0.081	-0.893	-0.248	-0.140	-0.034
Shape factor ratio	0.085	0.030	0.989	0.057	0.024	-0.057	0.078
Area	-0.483	0.201	-0.049	0.287	-0.035	-0.718	0.354
Residential land use							
Variable	**pc1**	**pc2**	**pc3**	**pc4**	**pc5**	**pc6**	**pc7**
Reservoir elevation above mean terrain elevation	-0.341	-0.538	-0.253	-0.176	-0.706	-0.015	-0.002
Terrain index	-0.325	-0.574	-0.291	0.152	0.662	-0.027	-0.137
Total pipeline length	-0.471	0.263	0.249	0.131	-0.090	-0.137	-0.777
Total pipeline volume	-0.488	0.158	0.177	0.124	0.032	-0.634	0.535
Reservoir distance from centre of area	-0.293	0.246	-0.139	-0.880	0.220	0.107	0.023
Shape factor ratio	0.066	-0.454	0.846	-0.258	0.074	0.041	0.017
Area	-0.478	0.140	0.154	0.270	-0.015	0.751	0.301
Low-Cost Housing land use							
Variable	**pc1**	**pc2**	**pc3**	**pc4**	**pc5**	**pc6**	**pc7**
Reservoir elevation above mean terrain elevation	0.108	0.674	-0.002	0.099	-0.656	0.307	-0.023
Terrain index	0.153	0.624	0.021	-0.516	0.402	-0.398	0.015
Total pipeline length	0.536	-0.224	-0.018	-0.089	-0.301	-0.361	-0.659
Total pipeline volume	0.558	-0.155	0.000	0.037	-0.210	-0.262	0.742
Reservoir distance from centre of area	0.320	0.220	0.491	0.665	0.391	0.023	-0.109
Shape factor ratio	-0.023	-0.170	0.833	-0.458	-0.154	0.207	0.034
Area	0.513	-0.070	-0.255	-0.251	0.311	0.711	-0.049

Stellenbosch University https://scholar.sun.ac.za

(Continued from above) 3 x-variables							
General land use							
Variable	**pc1**	**pc2**	**pc3**	**pc4**	**pc5**	**pc6**	**pc7**
Reservoir elevation above mean terrain elevation							
Terrain index							
Total pipeline length	0.583	-0.203	0.787				
Total pipeline volume	0.577	-0.578	-0.577				
Reservoir distance from centre of area							
Shape factor ratio							
Area	0.572	0.790	-0.219				
Residential land use							
Variable	**pc1**	**pc2**	**pc3**	**pc4**	**pc5**	**pc6**	**pc7**
Reservoir elevation above mean terrain elevation							
Terrain index							
Total pipeline length	0.581	-0.253	0.774				
Total pipeline volume	0.578	-0.541	-0.611				
Reservoir distance from centre of area							
Shape factor ratio							
Area	0.573	0.802	-0.168				
Low-Cost Housing land use							
Variable	**pc1**	**pc2**	**pc3**	**pc4**	**pc5**	**pc6**	**pc7**
Reservoir elevation above mean terrain elevation							
Terrain index							
Total pipeline length	-0.588	0.423	0.689				
Total pipeline volume	-0.592	0.356	-0.723				
Reservoir distance from centre of area							
Shape factor ratio							
Area	-0.551	-0.833	0.041				

Stellenbosch University https://scholar.sun.ac.za

Appendix

Table 27: Principal Component Analysis summary

7 x-variables								
General land use								
Method	**Formula**							
	Intercept	**pc1**	**pc2**	**pc3**	**pc4**	**pc5**	**pc6**	**pc7**
Standard regression	22.790	4.168	0.059	-0.733	-2.907	2.021	5.186	-2.274
Low p-values	21.183	4.181			-2.820	1.946	5.144	-2.230
Residential land use								
Method	**Formula**							
	Intercept	**pc1**	**pc2**	**pc3**	**pc4**	**pc5**	**pc6**	**pc7**
Standard regression	30.041	4.565	1.811	3.901	-3.186	-4.709	-3.251	-0.321
Low p-values	21.493	4.604		7.893	-4.191	-4.518	-2.529	
Low-Cost Housing land use								
Method	**Formula**							
	Intercept	**pc1**	**pc2**	**pc3**	**pc4**	**pc5**	**pc6**	**pc7**
Standard regression	26.475	-3.395	4.932	2.566	-2.124	1.144	-9.822	0.625
Low p-values	30.595	-3.770	3.203	1.542			-8.645	0.529
3 x-variables								
General land use								
Method	**Formula**							
	Intercept	**pc1**	**pc2**	**pc3**	**pc4**	**pc5**	**pc6**	**pc7**
Standard regression	9.855	-4.100	-5.781	1.544				
Low p-values								
Residential land use								
Method	**Formula**							
	Intercept	**pc1**	**pc2**	**pc3**	**pc4**	**pc5**	**pc6**	**pc7**
Standard regression	18.556	-2.432	-3.522	0.676				
Low p-values	16.293	-0.117	-0.274					
Low-Cost Housing land use								
Method	**Formula**							
	Intercept	**pc1**	**pc2**	**pc3**	**pc4**	**pc5**	**pc6**	**pc7**
Standard regression	-13.978	4.488	6.837	-0.359				
Low p-values								

Stellenbosch University https://scholar.sun.ac.za

10.4.4 Partial Least Squares

Table 28: Partial Least Squares summary

7 x-variables								
General land use								
Method	**Formula**							
	Intercept	**Reservoir elevation**	**Terrain index**	**Pipeline length**	**Pipeline volume**	**Reservoir distance**	**Shape factor**	**Area**
PLS	-66.901	0.810	10.171	0.001	0.039	0.027	-3.311	5.905
Residential land use								
Method	**Formula**							
	Intercept	**Reservoir elevation**	**Terrain index**	**Pipeline length**	**Pipeline volume**	**Reservoir distance**	**Shape factor**	**Area**
PLS	-80.657	1.120	14.382	0.001	0.037	0.024	-2.553	6.101
Low-Cost Housing land use								
Method	**Formula**							
	Intercept	**Reservoir elevation**	**Terrain index**	**Pipeline length**	**Pipeline volume**	**Reservoir distance**	**Shape factor**	**Area**
PLS	-1.971	-0.422	-4.706	0.001	0.053	0.022	8.346	5.436
3 x-variables								
General land use								
Method	**Formula**							
	Intercept	**Reservoir elevation**	**Terrain index**	**Pipeline length**	**Pipeline volume**	**Reservoir distance**	**Shape factor**	**Area**
PLS	9.226			0.001	0.047			7.115
Residential land use								
Method	**Formula**							
	Intercept	**Reservoir elevation**	**Terrain index**	**Pipeline length**	**Pipeline volume**	**Reservoir distance**	**Shape factor**	**Area**
PLS	13.967			0.001	0.046			7.686
Low-Cost Housing land use								
Method	**Formula**							
	Intercept	**Reservoir elevation**	**Terrain index**	**Pipeline length**	**Pipeline volume**	**Reservoir distance**	**Shape factor**	**Area**
PLS	-4.154			0.001	0.056			5.692

Stellenbosch University https://scholar.sun.ac.za

Appendix

10.5. Urban network Pipe diameter distributions

Small areas (< 10 km²)

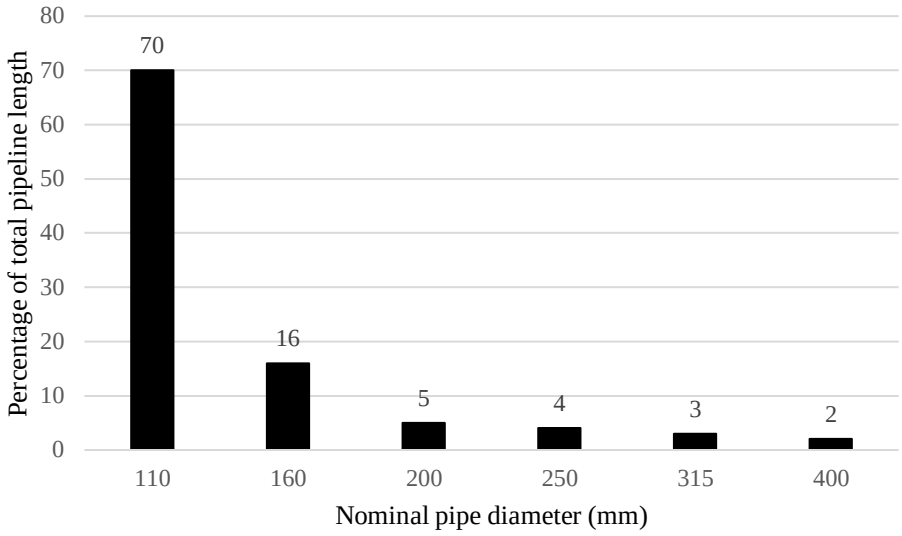


Figure 30: Small areas pipe diameter distribution

Moderately sized areas (between 10 and 20 km²)

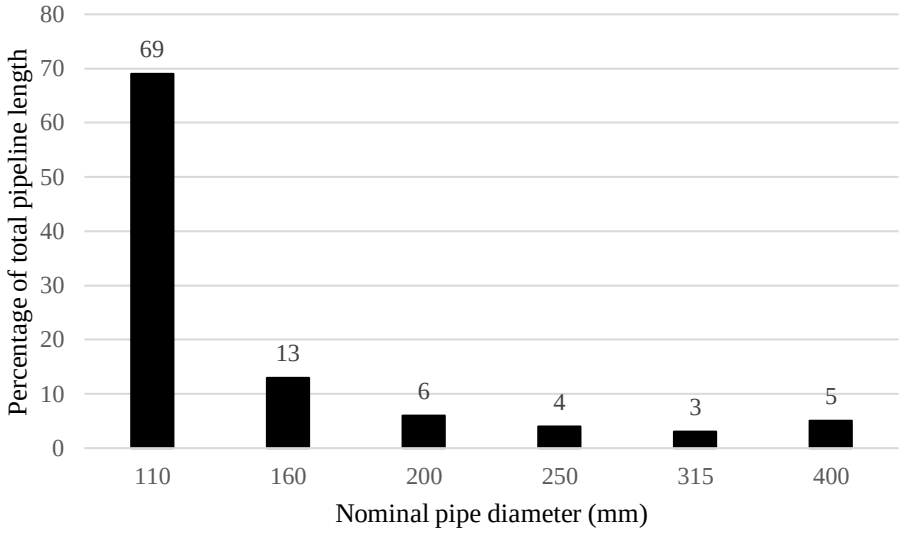


Figure 31: Moderately sized areas pipe diameter distribution

Stellenbosch University https://scholar.sun.ac.za

Large areas (> 20 km^2)

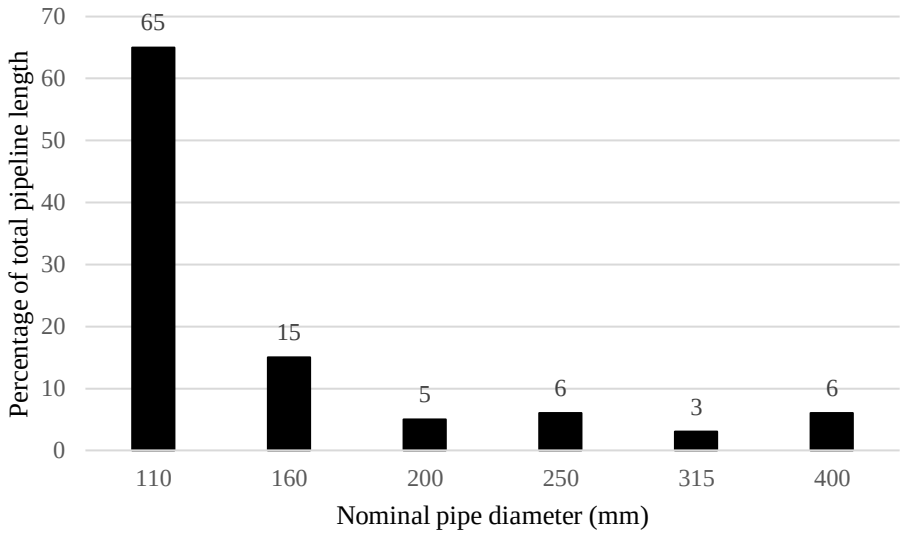


Figure 32: Large areas pipe diameter distribution

Flat areas

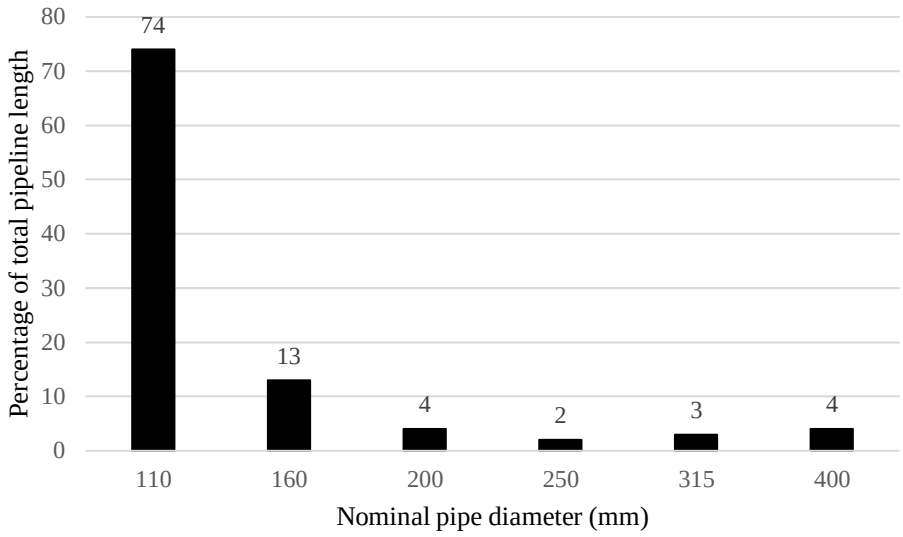


Figure 33: Flat areas pipe diameter distribution

Stellenbosch University https://scholar.sun.ac.za

Appendix

Partially hilly areas

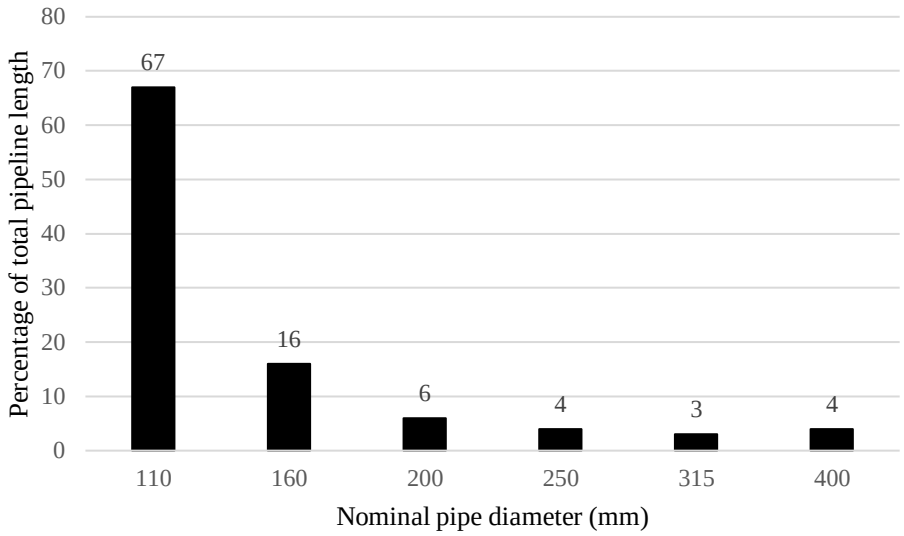


Figure 34: Partially hilly areas pipe diameter distribution

Hilly areas

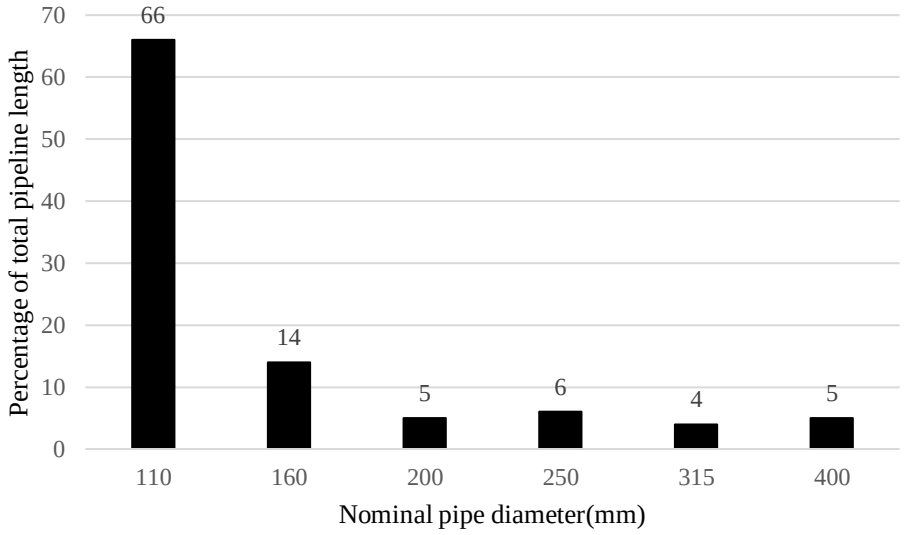


Figure 35: Hilly areas pipe diameter distribution

Stellenbosch University https://scholar.sun.ac.za

General areas

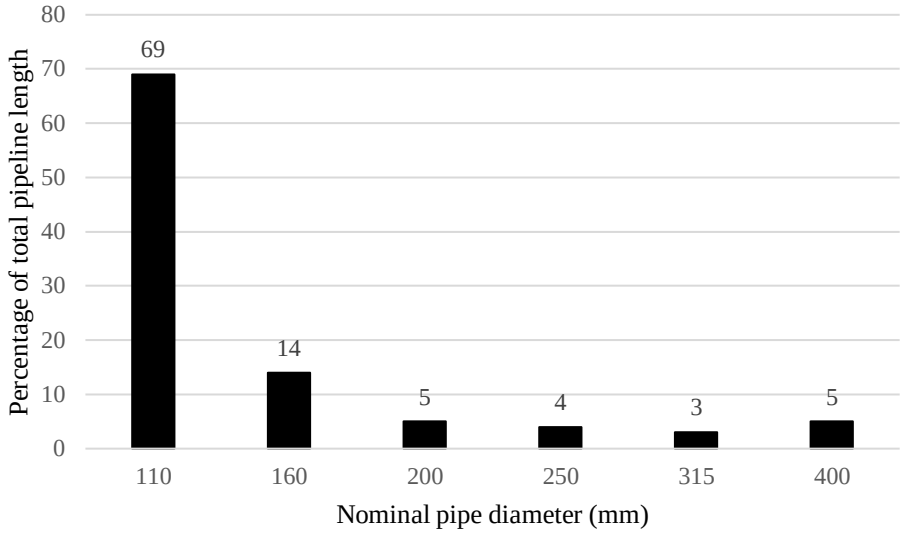


Figure 36: General areas pipe diameter distribution

Residential areas

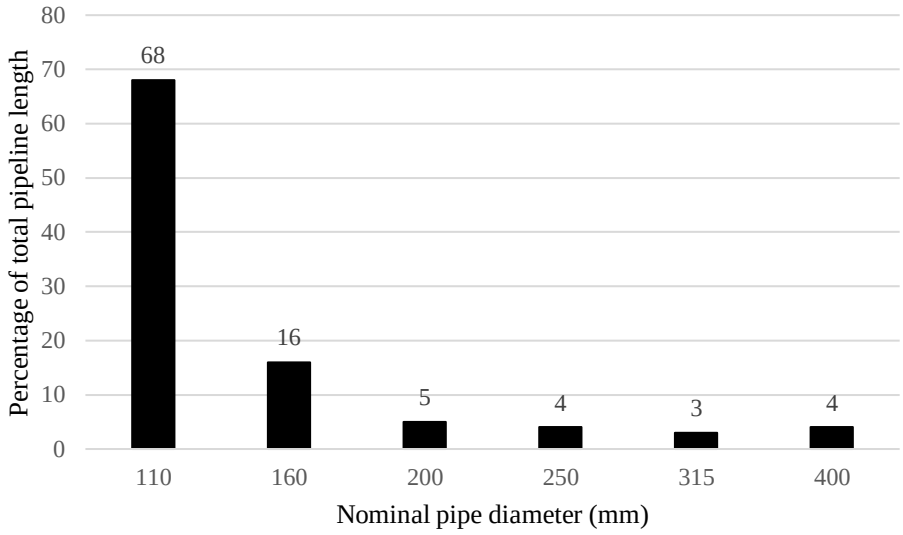


Figure 37: Residential areas pipe diameter distribution

Stellenbosch University https://scholar.sun.ac.za

Appendix

Low-Cost Housing areas

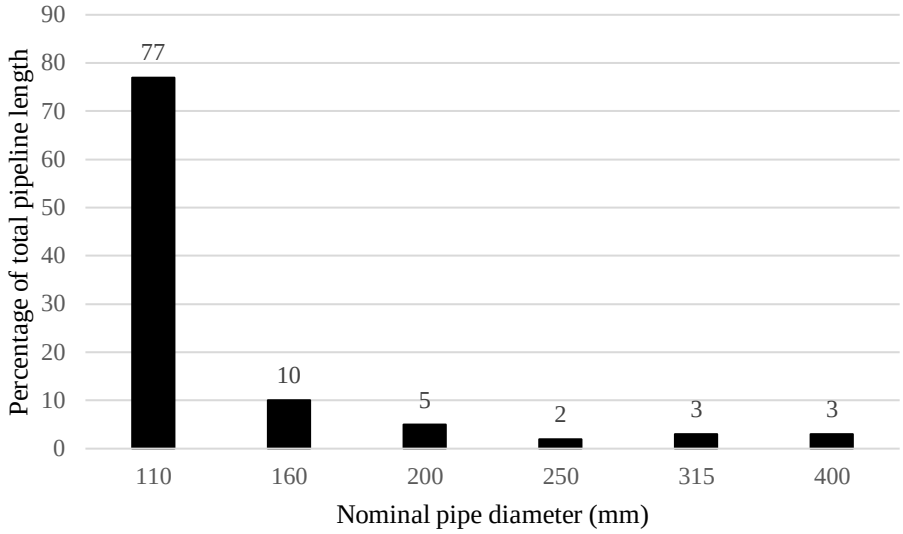


Figure 38: Low-Cost housing areas pipe diameter distribution

Stellenbosch University https://scholar.sun.ac.za

10.6. Example calculation

If a new medium density, medium sized plots residential area of 380 hectares is planned, the unit water demand can be determined from Table J.2 of *The Neighbourhood Planning and Design Guide* (Department of Human Settlements, 2019). This is found to be 9 kl/ha/d. With the unit water demand known, the annual average daily demand (AADD) can be calculated. Considering the area is 380 hectares, the AADD becomes 3420 kl/d. With the AADD known, the peak factor can be estimated using Table 1. Thereafter, the regression model can be used to calculate the approximate length of each pipe size required and provide an estimate of the construction costs for the water distribution network. The steps include:

1. Computing the AADD.

$$AADD = 9 * 380 = 3420\ kl/d$$

2. Using the peak factors from Table 1, the total peak hour demand can be computed.

$$Total\ peak\ hour\ demand = 4 * 3420 * \frac{1000}{24 * 60 * 60} = 158\ l/s$$

3. A safety factor can then be applied to the model. From Figure 20, a safety factor of 1.4 is applied, thus the new total peak hour demand becomes 222 l/s.
4. With the total peak hour demand and area known, the WDS capacity model can be used to calculate the total pipeline length. The final factor is attained from Table 16, knowing that it is a general type area, thus a final factor of 0.00424 is used.

$$Total\ peak\ hour\ demand = 9.855 + FF * Total\ pipeline\ length - 7.253 * Area$$

$$222 = 9.855 + 0.00424 * Total\ pipeline\ length - 7.253 * 3.8$$

$$Total\ pipeline\ length = 56534\ \text{m}$$

Stellenbosch University https://scholar.sun.ac.za

5. With the total pipeline length known, the pipe diameter distribution graph in Figure 27 can be used to compute the length of each pipe diameter. With the length of each pipe diameter known, the construction costs from Table 17 can be used to calculate the total approximate construction costs.

Table 29: Pipe length and construction cost computation

Pipe diameter (mm)	Proportion of total pipeline length (%)	Pipeline length (m)	Public open space unit cost (R/m)	Construction cost (R mil)
50	2	1131	452.00	0.51
63	2	1131	483.00	0.55
75	23	13003	518.00	6.74
90	10	5653	569.00	3.22
110	32	18091	610.00	11.04
160	14	7915	830.00	6.57
200	5	2827	1069.00	3.02
250	4	2261	1425.00	3.22
300	2	1131	1425.00	1.61
315	1	565	1937.00	1.10
355	1	565	2398.00	1.36
400	1	565	3009.00	1.70
450	3	1696	3383.00	5.74
Total				46.36

6. By applying the regression model, it is estimated that total construction cost for the water supply network associated with the planned medium density residential development would be approximately R46 million. Table 29 also provides the total required length for every pipe size.

Stellenbosch University https://scholar.sun.ac.za

Websites

Analyse-it. (2018). Analyse-it. [Online]. Available: https://analyse-it.com/ [Accessed 7 August 2018].

Bock, T. (2018). What are Variance Inflation Factors (VIFs). [Online]. Available: https://www.displayr.com/variance-inflation-factors-vifs/ [Accessed 7 August 2018].

FishXing. (2006). Continuity Equation. [Online]. Available: http://www.fsl.orst.edu/geowater/FX3/help/FX3_Help.html#8_Hydraulic_Reference/Continuity_Equation.htm [Accessed 8 March 2018].

GLS Consulting. (2018). GLS conslting. [Online]. Available: http://www.gls.co.za/about.html [Accessed 6 November 2018].

Lee, K. (2017). Surface & subsurface water sources. [Online]. Available: https://sciencing.com/surface-subsurface-water-resources-22528.html [Accessed 20 November 2018].

Mathworld. (2018). Orthogonal matrix. [Online]. Available: http://mathworld.wolfram.com/OrthogonalMatrix.html [Accessed 12 July 2018].

Rathore, D. (2015). How does the water distribution system work? [Online]. Available: https://www.quora.com/How-does-the-water-distribution-system-work [Accessed 6 March 2018].

Robor. (2018). Robor. [Online]. Available: http://www.robor.co.za/filebrowser/editorfiles/catalogs/03_product_catalogs/03_conveyance/2376_robor_ductile_iron_brochure5.pdf [Accessed 16 January 2019].

Rouse, M. (2010). Noisy data. [Online]. Available: https://searchbusinessanalytics.techtarget.com/definition/noisy-data [Accessed 12 July 2018].

Santoyo, S. (2017). A brief outline of outlier detection techniques. [Online]. Available: https://towardsdatascience.com/a-brief-overview-of-outlier-detection-techniques-1e0b2c19e561 [Accessed 12 July 2018].

Statistic Solutions. (2018). Homoscedasticity. [Online]. Available: https://www.statisticssolutions.com/homoscedasticity/ [Accessed 10 October 2018].

University of Iowa (2017). Percentage error formula. [Online]. Available: http://astro.physics.uiowa.edu/ITU/glossary/percent-error-formula/ [Accessed 18 November 2018].

Xlstat. (2018). Xlstat. [Online]. Available: https://www.xlstat.com/en/ [Accessed 5 Septhember 2018].

Stellenbosch University https://scholar.sun.ac.za

Journals

Bro, R. and Smilde, A. K. (2014). Principal component analysis. *Analytical Methods*, 6, 2812-2831.

De Souza, S. V. C. and Junqueira, R. G. (2005). A procedure to assess linearity by ordinary least squares method. *Analytica Chimica Acta*, 552, 25-35.

Department of Human Settlements. (2019). The Neighbourhood Planning and Design Guide. *Water Supply*, 2, 1-72.

Department of Water & Sanitation. (2005). Standard Specifications for Municipal civil engineering works. *City of Tshwane Metropolitan Municipality*, 3, 1-903.

Erickson et al. (2017). Water quality effects of intermittent water supply in Arraijan. *Water Res*, 114, 338-350.

Geladi, P. and Kowalski, B.R. (1985). Partial Least Squares Resgression: A Tutorial. *Analytica Chimica Acta*, 185, 1-17.

Ghorbanian et al. (2015). Minimum Pressure Criterion in Water Distribution Systems: Challenges and Consequences. *Urban Water Journal*, 14, 777-791.

Jacobs, H. and Strijdom, J. L. (2009). Evaluation of minimum residual pressure as design criterion for South African water distribution systems. *Water SA 2009*, 35, 183 - 191.

Mora-Rodriguez et al. (2014). An overview of leaks and intrusion for different. *Urban Water Journal*, 11, 1-10.

National Water Act. (1998). Guidelines for human settlement planning and design. National Water Act, 2, 1-23.

Nyende-Byakika et al. (2012). Comparative analysis of approaches to modelling water. *Civil Engineering and Environmental Systems*, 29, 79-89.

Romano, M. and Kapelan, Z. (2014). Adaptive water demand forecasting for near real-time management of smart water distribution systems. *Environmental Modelling & Software*, 60, 265-276.

Shlens, J. (2014). A tutorial on Principal component analysis. *Internation Journal of Remote Sensing*, 2, 1-13.

Siong, K. (2013). A simple Explanation of Partial Least Squares. *The Mathworks*, 1, 1-10.

Tanyimboh, T. and Templeman, A. (2000). A quantified assessment of the relationship between the reliability and entropy of water distribution systems. *Engineering Optimisation*, 33, 179-199.

Wang et al. (2014). Water hammer assessment techniques for water distribution systems. *Procedia Engineering*, 70, 1717–1725.

Wold et al. (1984). The collinearity problem in linear regression: The partial least squares approach to generalized inverses. *Society for Industrial and Applied Mathematics*, 5, 735-743.

Zeng et al. (2016). Extreme water-hammer pressure during one-after-another load shedding in pumped-storage stations. *Renewable Energy*, 99, 35-44.

Stellenbosch University https://scholar.sun.ac.za

Theses

Hirst, G. (2017). *Modelling of water reticulation networks under conditions of extrememly low pressures*. BEng. Stellenbosch University Department of Civil Engineering, 1, 1-58.

Strijdom, J. (2016). *Evaluation of minimum pressure head during peak flow as design criterion water distribution systems*. MEng. Stellenbosch University Department of Civil Engineering, 1, 1-136.

Books

Chadwick et al. (2013). *Hydraulics in Civil and Environmental Engineering.* Boca Raton: CRC Press, 648.

Montgomery et al. (2014). *Applied Statistics and Probability for Engineers.* New York: John Wiley & Sons, 200.

Savic, A. and Banyard, J. (2011). *Water Distribution Systems.* London: ICE Publishing, 180.

Interviews

Grotepass, F. and Auret, L. (2018). *Engineering Statistics.*

Grotepass, F. and Loubser, C. (2018). *Network Capacity Index.*