Bachelor’s Thesis
TVVR 12/4002

Cost functions for predicting capital expenditure of small town water systems

- A Minor Field Study in Ghana

Kristoffer Ristinmaa

Division of Water Resources Engineering
Department of Building and Environmental Technology
Lund University

Avdelningen för Teknisk Vattenresurslära

Cost functions for predicting capital expenditure of small town water systems

- *A Minor Field Study in Ghana*

Kristoffer Ristinmaa

Abstract

Using data gathered by researchers from WASHCost project in Ghana, this thesis examines how cost drivers can be used to predict capital expenditure for construction of piped water systems for small towns in Ghana. The data is collected from 45 small towns and communities in peri-urban areas in the Volta, Ashanti and Northern region. I have used correlation analysis and multiple regression analysis to derive cost functions for small town water systems and for the components in a water system. The aim is to derive functions for the total capital expenditure of a small town water system and for the fixed assets: mechanized borehole, water reservoir, pipe work and stand post. The resulting functions are created both for initial use before planning a new system and after having done the first design of a water system. The results are presented as eight functions, representing the total capital expenditure for a water system and capital expenditure for the assets water reservoirs, pipework and stand posts. The data was insufficient in deriving a function for mechanized boreholes. Key parameters for validating the functions are presented in relation to the functions. Explaining variables that were frequently used are design population, length of pipeline, volume of reservoir and dummy variables for region (location). The functions vary in accuracy with an explanation coefficient R^2 varying from 0.42 to 0.88. The functions that have explaining variables, requiring a first investigation at place of the intended water system, are particularly interesting to use in forecasting capital expenditure.

Key words: Ghana, small town water system, cost function, capital expenditure, multiple regression analysis

Sammanfattning

Rapporten är en sammanställning av ett kandidatarbete, huvudsakligen utförd som en ”Minor Field Study” i Kumasi, Ghana och i samarbete med WASHCost. WASHCost är ett pågående internationellt forskningsprojekt med holländska IRC som initiativtagare. Genom att använda data, insamlad av forskare från WASHCost i Ghana har jag undersökt vilka kostnadsdrivare som kan användas för att förutsäga kostnaderna vid nybyggnation av ett dricksvattensystem med grundvattenteknik för mindre orter med populationer under 20 000 invånare. Databasen täcker 45 tätorter ifrån tre ghananska regioner; Volta, Ashanti och Norra regionen. Huvudmålet med studien har varit att skapa en kostnadsfunktion för att i ett tidigt skede kunna få en uppfattning om de totala investeringskostnaderna vid nybyggnation av ett drickvattensystem. Sekundärt har funktioner skapats för att förutsäga kostnaderna för följande ingående komponenter i ett ghananskt vattennät: Borrhål med automatisk pump, vattenreservoarer, rörledningsnät och vattenposter. Till min hjälp för att skapa relevanta funktioner har jag använt mig av sambandsanalys och multipel regression analys med olika modellalternativ. Dataunderlaget innehåll inte tillräckligt med information för att kunna skapa någon funktion för borrhålen. I resultaten redovisas åtta olika funktioner, varav två beskriver den totala investeringskostnaden. Av funktionerna är fem skapade efter linjära modeller och tre efter exponentiella. De vanligaste förekommande kostnadsdrivarna, alltså förklarande variablerna, är: i vilken region vattensystemet är lokaliserat, den dimensionerande populationen, den totala rörlängden och vattenreservoarens volym. I samband med resultaten presenteras även nyckelvärden för att kunna bedöma funktionernas tillförlitlighet. Förklaringsgraden, R^2, för funktionerna varierar mellan 0.42-0.88, vilket ger en stor variation i tillförlitlighet. De mest intressanta funktionerna är de som kallas för ”first function” och kräver en första utredning för det tilltänkta vattensystemet, men ger istället en bättre precision i att förutsätta kostnaderna.

Nyckelord: Ghana, small town water system, cost function, capital expenditure, multiple regression analysis

Acknowledgment

Thanks to God for making all things possible and guidance through life.

Thanks to Swedish International Development Cooperation Agency, Sida, for granting my Minor Field Study with a scholarship. Without the scholarship I would not have been able to carry out the study in Ghana.

Special appreciation goes to my supervisor, Professor Magnus Persson, for his supervision and guidance. He has always been available and his constructive comments and suggestions have been helpful through the whole thesis.

Also, I would like to express my utmost gratitude to my co-supervisor, Dr. Kwabena B. Nyarko for all his supervision, guidance and hospitality at KNUST. He let me work in his office and take part of his very interesting work as country director for WASHCost. Without him, I would not have been able to do my study in Ghana.

My heartiest thanks to all my friends I met in Ghana, who supported me and made the journey in Ghana a wonderful experience and memory for life. My thankfulness goes especially to my dear friends Mr. Joshua Ofusuhene and Mr. Paa Kwesi Woode. Thank you for support, joy and company during two wonderful months.

My deepest appreciation to Dr. Lena Zetterqvist for her knowledge and support in mathematical statistics.

Finally a special thanks to my family, friends and particularly my wife Maria for love, encouragement and support.

Table of Contents

List of figures

List of tables

SECTION ONE

1 Introduction

Using data collected by researchers from WASHCost project in Ghana, this paper examines how cost drivers can be used to predict capital expenditure for construction of piped water systems for small towns in Ghana. I have used correlation analysis and multiple regression analysis to derive cost functions for small town water systems and for the components in a water system. The resulting functions are created both for initial use before planning a new system and after having done the first design of a water system.

1.1 Background: Water supply and management in Ghana

Ghana is a West African country where major political and economic changes and progress have taken place during the last couple of decades. Processes of democratisation and economic growth have occurred alongside with changes of the country's water supply and management (Dawuni 2009; The Presidency Republic of Ghana 2010). Especially the rural areas of Ghana have experienced major improvements in terms of water supply (WHO 2010).

Figure 1. Ghana in the world. Source: Wikimedia Commons

In many developing countries, including Ghana, water management is subject of major challenges. While Ghana is not yet in water stress, the country needs to handle its water resources effectively. The country is furthermore faced with problems in terms of both quantity and quality (Yeboah, 2008). Almost all urban water is based on surface water, unlike the rural water which comes from groundwater. Kankam-Yeboah et. al. (2010), from CSIR Water Research Institute in Ghana, report that Ghana, in 2020, might experience a reduction of up to 15-20% of the annual river flow. Apart from a reduction of hydropower generation, this will put pressure on Ghana's water supply. The possible water stress is enhanced by the high level of non-revenue water. According to Yeboah (2008), the non-revenue water can, under certain conditions and in some parts of Ghana, be up to 51% of the water supply.

The major part of Ghana's water consumption goes to domestic use, irrigation and to hold livestock. During the last two decades, almost 30% of Ghana's population have gotten access to improved water supply. In 2010, according to WHO's definition of improved water, 86% of Ghana's population is supplied by improved water (WHO 2010). In rural areas, the water coverage was only 59 % in 2009 and the next goal is a 76 % coverage in 2015 (Nkrumah et al 2010). The country is still highly dependent on foreign donors. Up to 95 % of the new capital investments in the water sector comes from developing partners (mostly through loans) (Moriarty et al 2010). Ghana's water supply scheme has also been struggling with large costs with insufficient revenues during the 90's. Despite the percentage of improved water, the non-governmental organisation WaterAid (2011) states that up to 80% of all diseases in Ghana are caused by insufficiently refined water and poor sanitation.

Although the majority of Ghana's population is provided with water from improved systems, there are still 14 % who are not (WHO 2010). In rural areas the relative number without improved water is 20 % in 2010, according to WHO, and 40 % in 2009 according to Moriarty et al (2010). The difference in relative number depends on how you define improved water in terms of quality, quantity, fetching time etc. In absolute numbers, the above mentioned percentages represent between two to four million people without access to improved water systems.

One common type of improved water system, so called rural piped systems, for small towns and communities in Ghana is a water system consisting of a secured mechanized borehole with connecting pipelines to storage tanks and stand pipes. While the usage of this water scheme is extensive, 451 systems in 2009, the costs vary a lot from one system to another (a.a.). This fact could, among other things, complicate the evaluation of different bids from building companies when projecting a new water system.

Present study has been initiated by Dr. K. B. Nyarko at, country director for WASHCost and senior lecturer at Kwame Nkrumah University of Science and Technology (KNUST) in Kumasi, Ghana. The university has, together with International Water and Sanitation Centre (IRC), an ongoing project called WASHCost. The research project is an international project and aims to increase the availability and use of cost information for both rural areas and small towns. In present study, the capital expenditure (CapEx) for current methods and schemes for water service delivery will be analyzed. With access to WASHCost's database, I will derive cost functions to enable estimations of future CapEx for small town water systems. Hopefully the study will contribute to the knowledge of the costs for water service delivery in Ghana for future decision-making processes.

1.2 Definitions

To reduce confusion and to encircle the objects discussed in the thesis, the physical components that have been investigated are defined in this section. Cost function and capital expenditure are also a central concepts in the present study and is defined below.

1.2.1 Cost function

A cost function is a function where the cost is given by input, explaining variables. The responding variable, y, is described by a function of explaining variables, X.

$$y = f(X)$$

In my study, the responding variable, y, will be capital expenditure, see section 1.2.2 below. The explaining variables, which are input to the function, affect the capital expenditure. Furthermore, I intend to derive the cost functions statistically. It means that only the explaining variables that can be shown having a statistical impact on the responding variable are accepted in the function. To see how the explaining variables are selected, see section 2 Method.

1.2.2 Capital expenditure

The capital expenditure (CapEx) is the investment cost in fixed assets. The assets are hardware such as pipes, pumps and storage tanks. It also includes software as all the one-off work for constructing the physical components. Characteristic for CapEx is that the costs are normally isolated, easy to relate to a physical object and often in lump sums. Even if the word 'capital' may lead you to think about the initial costs, the CapEx also includes major costs for extensions, enhancements and improvements (Fonesca et al 2011). In this study, CapEx for a certain object includes both the hardware and software; they have in other words not been separated. Instead, the study focuses on the aggregated CapEx for a type of object in a water system, e.g. a standpipe, and looks at the possible cost drivers, software and hardware.

1.2.3 Characteristics of a small town water system

A small town is defined by Ghana's Community Water and Sanitation Agency (CWSA) as a peri-urban settlement with a population between two thousand and five thousand inhabitants (CSWA 2007).

A water system is the sum of the physical equipment used to provide consumers with water within a certain area. It comprises of the whole process from production or collection and transmission to the distribution of the water. A water system can have different solutions of how to provide the consumers with water, e.g. water access through a mechanized borehole or transmission from Ghana Water Company Limited (GWCL), stand posts or connection to households, electricity through the national grid or a solar system etc. Several communities or small towns may be included in the same water system.

Mechanized borehole

A mechanized borehole is a common means to provide a community or a small town with improved water in Ghana. According to Dr. B. Ali (2012-03-02), at the Department of Geological Engineering, KNUST, Ghana, the location for a mechanized borehole should be determined by siting. Once that is done, the drilling, construction and development of the borehole starts. When all this is settled, the screens and plains, a pump, gravel and the important sanitary seals out of cement grout should be inserted into the hole. The output flow is determined by the yield, which has to be tested in situ. Usually the borehole is accompanied by a basic pump house to regulate the pump (Dwumfour-Asare 2009).

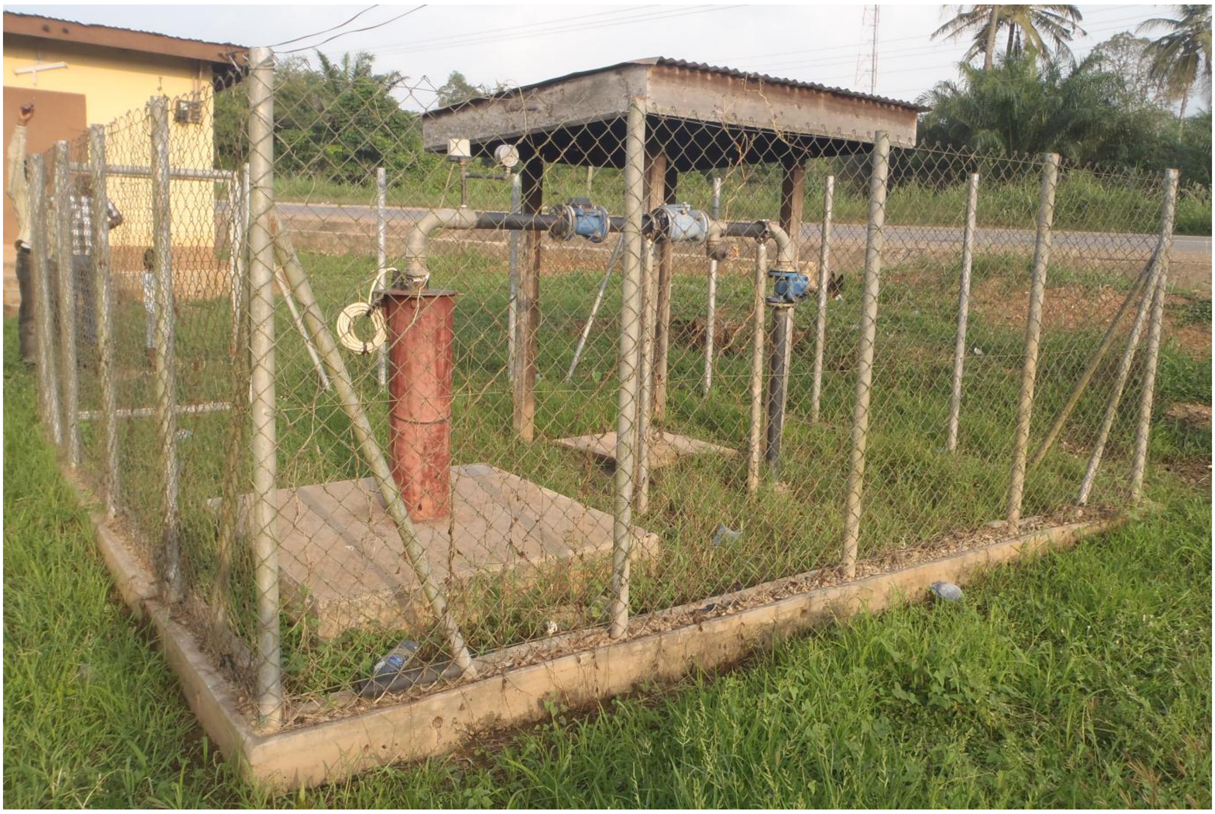


Figure 2. A mechanized borehole in Kuntanase with fence to prevent intrusion by animals. The yellow building in the background is the related pump house. Photo: Kristoffer Ristinmaa

Water reservoirs

Water reservoirs are stored water reserves of certain quantity, based on the design population's daily consumption alternatively based on the boreholes' yield if this falls below the consumption. There are commonly four types of reservoirs used for storing treated water in small towns in Ghana:

- Steel tank
- Excavated ground concrete reservoir
- Elevated concrete tank
- Plastic storage tank – commonly named Polytank after the brand with the same name

Pipework

Pipework is the scheme of pipelines within a water system. The pipework is the net which conveys the water from its source to the consumer. The network can be divided into transmission pipes and distribution pipes. Normally the transmission pipes have a larger diameter and serve as a link between source and reservoir. The distribution pipework distribute the water from the reservoirs to the households or standpipes (Nyarko 2007). Most of the pipelines are made from PVC or HDPE (Dwumfour-Asare 2009). In this study, pipework is defined as including both the transmission and the distribution in the system.

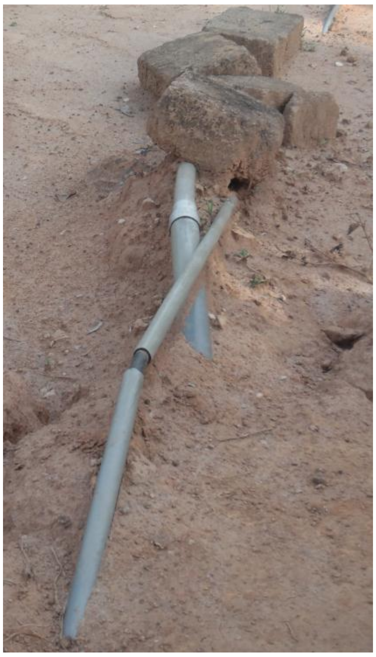


Figure 3. PVC-pipes above ground surface in Yaase. Photo: Kristoffer Ristinmaa

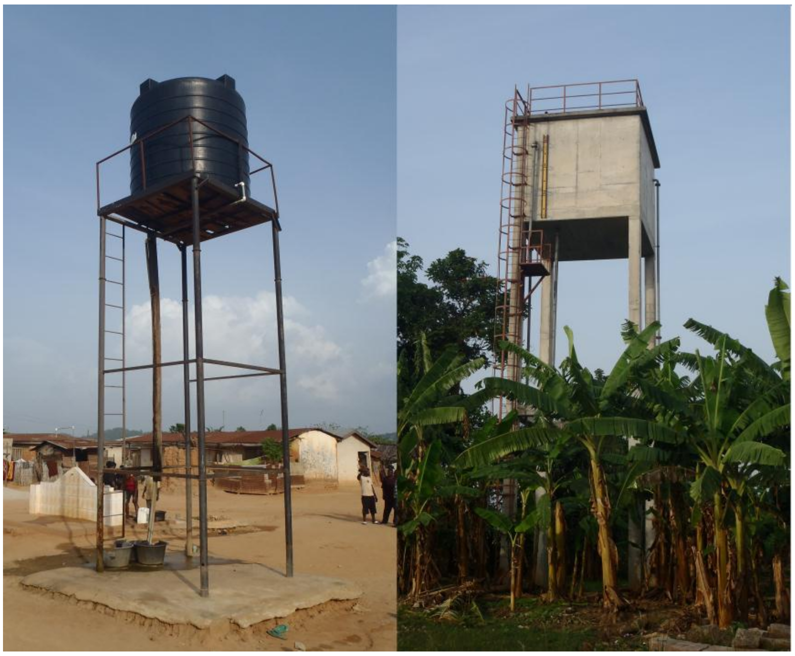


Figure 4. Plastic tank in Yaase to the left and elevated concrete reservoir in Kuntanase to the right. Photo: Kristoffer Ristinmaa

Stand post

A stand post, also called standpipe or tap stand, is a water point for consumers to fetch water. Stand posts usually consist of one or more taps for fetching water. Ghana has a standard requiring a stand post not be crowded with more than 300 consumers (Moriarty et al 2011). In other words, only 300 consumers should have a specific stand post as their main water point.

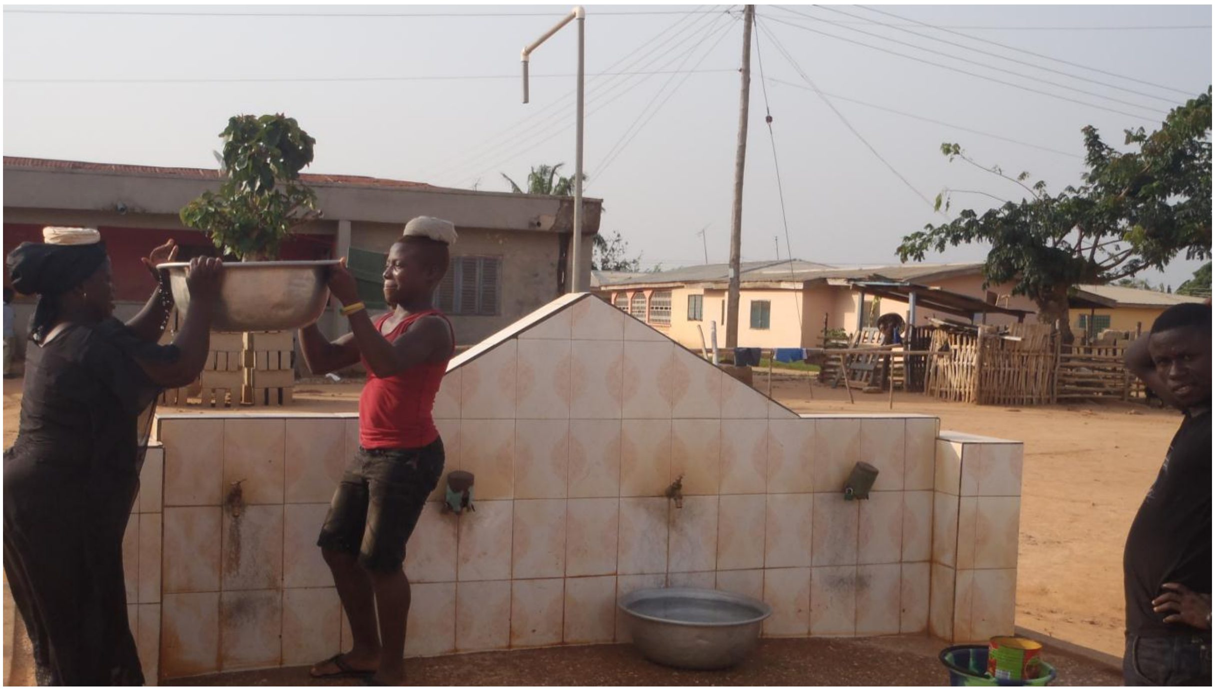


Figure 5. A stand post in Yaase with five taps including the PVC pipe that makes it possible to fetch water carrying a bowl on your head.

1.3 Previous studies

In order to get a picture of how a cost function could be derived, I consulted previous similar studies. My ambition was to see: a) what costs are investigated, b) what variables explain the cost of a water system and c) what models are used to derive the functions? The list of studies below is not a complete list of studies regarding water supply schemes and capital costs. Instead is an overview of studies that has worked as guidance and methodological framework for this study.

In 2011, Nkrumah et. al. (2011) presented the results of an investigation of the cost drivers for small water systems in Ghana. The researchers from that project collected data from three regions in Ghana and the gathered database is the same used for this thesis (see section 2.1 *Data collection*). The authors come to the conclusion that the main cost drivers for the evaluated water systems is technology, population density, hydrogeology of area and contract packaging.

Several studies have been made in the field of cost estimation and water supply. In 1977, Water Research Centre (WRc, 1977) in England published their technical report, *TR61 – Cost information of water supply and sewage -disposal*. The report presents cost functions for various components in water supply and sewage systems. The source of data used was contracts, such as bills of quantities (BoQs). The cost data from the BoQs were thereafter adjusted to the same year due to the inflation. WRc came to the conclusion that a multiplicative model (in this study called log-log model) of multiple linear regression analysis was the best way to derive their recommended functions.

Clark and Stevie (1981) have another approach to estimate the total cost for a water system in the United States. Unlike WRc (1977), which divided a water system into detailed components, Clark and Stevie divide the cost for a water system into acquisition/treatment cost and transmission/distribution cost. Assuming that the production quantity is the key explanatory variable for both parts of a water system, the authors derive a function for the total cost for a water system. The quantity is then estimated and depending on the population density. Normally the population density is described as person per area unit, but as a proxy. Nkrumah et. al. (2011) used pipe-length per capita as one of the variables for investigating the cost drivers for a water system. With an increasing pipe-length per capita it is assumed that less people live in a larger area, i.e. decreasing population density.

Eilers (1984), from the USA Environmental Protection Agency (EPA), estimate the costs for smaller water systems in the USA. He used the same type of log-log model for multiple regression analysis as was used by WRc in 1977. The study's main focus is on the capital and maintenance cost for the smaller water treatment systems. Eilers (a.a.) stresses the importance to make such cost analyses when designing a new small water system to determine the most cost effective design. Since the systems are small, Eilers (a.a.) claims that the systems often are in shortage of operating revenues and cannot make an economic advantage of large scale production.

Other studies in the field have been done more recently. Antonioli and Filippini (2002) estimated a multivariate variable cost function for annual cost for building and maintaining a water system in Italy. The purpose with the study is to refine the determination process of the tariffs and to estimate the optimal size for a water system distribution. As a model, they use a version of the Cobb-Douglas function, which is an often used multiplicative model in econometrics to derive the annual value of a certain production. With eight explanatory variables, Antoniolo and Filippini (a.a.) estimate the annual cost for a water supply scheme. The authors are however more interested in getting a satisfying determination coefficient than to have all of the explanatory variables significant different from zero.

Kirshen et. al. (2004) explored climate and regional aspects for driving the annual cost of water supply in the United States of America. The database used is voluminous and the results presented with a clear explanation degree. The functions where made as a log-log model with the delivered water quantity delivered per second as the single explanatory variable. No matter if it was a surface and water.

In *Rural cost functions for water supply and sanitation* (OECD 2005), functions for investment and operational expenditure are derived for EECCA countries. They present an extensive list of costs for different technical options of water supply for small towns and communities less than 5000 in population. The method used is not described, but by looking at the derived cost functions, I concluded that the models used are linear, log-log and polynomial models. The most common used model for deriving investment costs is the log-log model.

Tsegai et al (2009) estimated costs for water supply in Middle Oliphant sub-basin of South Africa to evaluate the sustainability by comparing the estimated marginal costs with the tariffs. He applied a more sophisticated statistical model, using the translog cost functions method. The translog function is basically a generalized form of the Cobb-Douglas function, used for deriving annual cost or yield.

Antoniolo and Filippini (2002) and Tsegai et al (2009) targeted the maintenance and annual cost for water supply. Present study differs from them by investigating only the CapEx for small town water systems. In that way my study is more like a field study with the same objectives as WRc (1977), Clark and Stevie (1981), Eilers (1984), Kirshen et.al. (2004) and OECD (2005). The modeling in this thesis will therefore be closer related to these studies theories.

1.4 Objectives and limitations

The major objective of my thesis is to derive cost functions for diverse parts of the water systems used for small towns in Ghana. The cost functions will hopefully give good estimations for predicting the CapEx in accordance to e.g. WRc (1977). The main target is a function to predict the total capital expenditure (TotCapEx) for a small town water system and secondly, functions for parts of the water system will be derived. The parts that will be analyzed are defined in section 1.3 and they are mechanized boreholes, water reservoirs, pipework and stand posts.

The objective is to derive functions to get an idea of the capital expenditure without having done a first investigation at place. If not possible or if the function will get parameters that are too uncertain, a second function will be derived with parameters that require a first investigation and design of the system.

Certain delimitations have been necessary. The data is limited to what WASHCost has already collected in Ghana and to additional literature such as reports and documents, mainly from Ghana's Community Water and Sanitation Agency (CSWA). The analysis has only been carried out to examine the capital expenditure for water systems using mechanized boreholes and for water systems with a design population beneath 20 000. The study period in Ghana was limited to seven weeks between 23rd of January to 12th of March 2012. Also, it is important to notice that the study makes no attempt to explain why a certain variable works as a cost driver and if the apparent variable truly is a cost driver.

1.5 Thesis structure

The thesis is structured as follows:

Section one, *Introduction*, positions the report into its context in Ghana and the sector of water supply. Moreover, central definitions are presented and described. Different parts of a small town water system are visualized and explained. The sub-section *Previous studies* examines and summarizes relevant literature to obtain a framework for the methodology. The objectives and limitations conclude the section.

Section two, *Method*, explains how the field area is selected and how the data is collected. The data is also summarized. All the models used for analyzing the data are illustrated, including the key parameters for validating the results. Finally the section discloses the modeling approach for each evaluated part of a water system.

In section three, *Results and Discussion*, the results are presented as recommended functions together with the validation parameters and a summary of used data. Plots illustrate the strength and relationship between the used variables.

Finally conclusions and recommendations are presented in section four, *Conclusions and recommendations*.

SECTION TWO

2 Method

The section starts with a description of how the data has been collected. Thereafter the tools for analyzing the data and the models tested for explaining the capital expenditures are presented. Key parameters to validate the derived function are also presented below. Finally, more detailed modeling approaches are presented for each analyzed components of a water system.

2.1 Data collection

The data collection has been carried out by the researchers within the WASHCost project in Ghana. When additional information was necessary to derive a function, I searched through published reports, mainly from Ghana's Community Water and Sanitation Agency (CSWA). Normally my additional research was without any success as the reports from CSWA frequently presented cost and technical data in lumps.

Criteria were set as guidance to select where the data should be collected. According to the researchers, Nkrumah et. al. (2011), the selection of regions was conducted so the data would fulfill the following:

- A diversity of donors and development partners with sufficient information regarding capital cost
- A diversity of hydro geological and hydro climatic zones
- Different approaches of how to implement a water system

The regions selected were Northern region, Ashanti region and Volta region, see figure 6. Northern region has a relatively dry tropical climate and is located in a zone of savannah land. Ashanti region is located in the midst of Ghana and is mostly covered by forest with a hot and humid tropical climate. Volta region is located in the western parts and has Volta Lake as its eastern border.

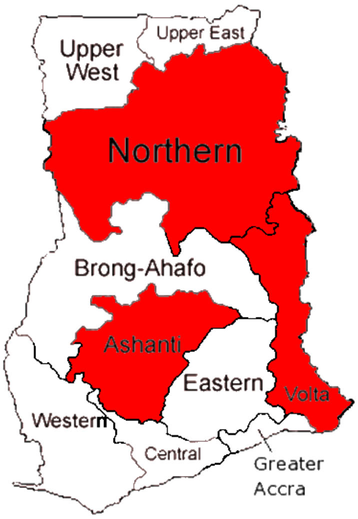


Figure 6. Map of Ghana with the three selected regions marked in red. Source: Wkimedia Commons, modified by author.

The database comprises of 45 water systems and all quantities and cost information have been collected from "contract documents, bills of quantities, payment certifications and completion reports" (Nkrumah et. al. 2011). Collected cost data are presented in the local currency, Ghana Cedi (GHs), and are summarized in Table 1.

Table 1. Minimum and maximum values of collected cost data in Ghana Cedis.

CapEx (GHs,2011)	Min	Max
Borehole site works	0	71 007
Borehole	0	308 657
Pipeline	14 873	520 837
Storage tank	967	226 180
Standpost	7 215	49 700
Total water system	58 109	960 748

In total, 13 variables are collected for all the data points, excluding CapEx in table 1. The minimum value of capital expenditure for borehole and borehole site works is zero. Three systems did not use boreholes, which explain the absence of capital expenditure. The numerical data is presented with minimum and maximum values in Table 2 below.

Table 2. Minimum and maximum values of collected numerical variables.

Data summary Variable	Min	Max	Unit
Number of communities	1	14	-
Design population	1 533	19 477	person
Number of mechanized boreholes	0	4	-
Length of pipeline	1 323	35 040	meter
Number of storage tanks	1	3	-
Volume of water reservoir	12	400	cubic meter
Number of standposts	4	29	-

Remaining data tells more about the character of a water system and is of a qualitative nature. Since they cannot be described in numbers, they are presented with a description in Table 3 below.

Table 3. Presentation and brief explanation of the collected quantitative data.

Variable	**Description**
Regions	Where the system is located. The regions are Volta, Ashanti or Northern region.
Year of construction	When the system is constructed.
Transmission and distribution pipe length	Percentage of the total length of pipelines in a water system that operates for transmission versus distribution of water.
Type of reservoir	A system could use ground concrete reservoir, high level concrete reservoir, steel or plastic tank. One water system can use several options.
Technichal option	Sorts out if mechanized borehole is used or the system is provided with bulk water from Ghana Water Company Limited's (GWCL)
Power source	Sorts out if power comes from the national grid or from a solar system
Contract packaging	Depending on the source of funding and could either be International Competitive Bidding (ICB) or National Competitive Bidding (NCB).

16 systems are located in Ashanti, 16 in Volta and 13 in Northern region. The systems were constructed and paid between 1998 and 2010. Steel tank is just used in Northern region and is also the only reservoir option used in the region. Three water systems use bulk water from GWCL and are the only systems using plastic tank. Only four water systems use solar system as power source and they are all located in Northern region. Contract packaging follows the variation of region. The only contract packaging presented in Volta region is NCB, whereas ICB is the only contract bidding system used in Ashanti and Northern region.

This study uses data from towns with populations between two thousand and twenty thousand, which makes this study investigating the costs for both peri-urban and small urban areas. Although the sizes of the observed towns divide the data into two categories of settlement, the characteristics of the water systems used are very similar for the settlements. The majority of the water systems in small towns or smaller urban areas observed in this study comprise of a mechanized borehole, pipework for transmission and distribution, some kind of reservoir and several stand posts.

All data used for modeling is presented in Appendix 1. For further reading about the data and how the data was selected and collected, read Nkrumah et.al. (2011)

2.2 Analytical tools

This section describes the analytical tools used for transforming and analyzing the data. The analysis used also come with both limitations and keys or parameters to determine whether or not a result is good enough. These limitations and strengths are discussed under each subtitle.

2.2.1 Deflated costs

The data points in WASHCost's database are from water systems built between 1998 and 2010. During this period, Ghana has gone through major changes in its economy. Because of this, the inflation is necessary to consider when analyzing the CapEx. A common way to handle the inflation is to adjust all the prices and relate them to the base year. The cost adjustments due to the inflation have been made through Prime Building Cost Index (PCBI), provided by Ghana Statistical Service. As the costs for labour and materials are presented in lump in the database used for the study, the annual growth in PCBI also is a combination of these two categories. The cost adjuster used due to the inflation is the annual growth of Ghana's Gross Domestic Product (GDP) deflator. The index is derived from dividing the current price for one year's GDP with the constant price for the same year's GDP. The deflator index is therefore depending on the chosen base year, but the annual growth of GDP deflator does not depend on the base year. The constant price is shortly described as the product of the current period's product quantity multiplied with the price for such a quantity during the base year. The current price is simply the price paid during the actual period for the period's quantity of products (OECD 2003). WASHCost has previously used a GDP deflator specifically made for costs in Ghana's water sector. WASHCost's GDP deflator is based on Ghana's PCBI but gives a lower inflation over time. The two deflators are presented in Figure 7. The difference will give lower costs if WASHCost deflator is used, which is done in present study. The costs will be particularly lower further away from year 2011, which is used as base year in this study. The deflator index provided from WASHCost's office is presented in Appendix 2.

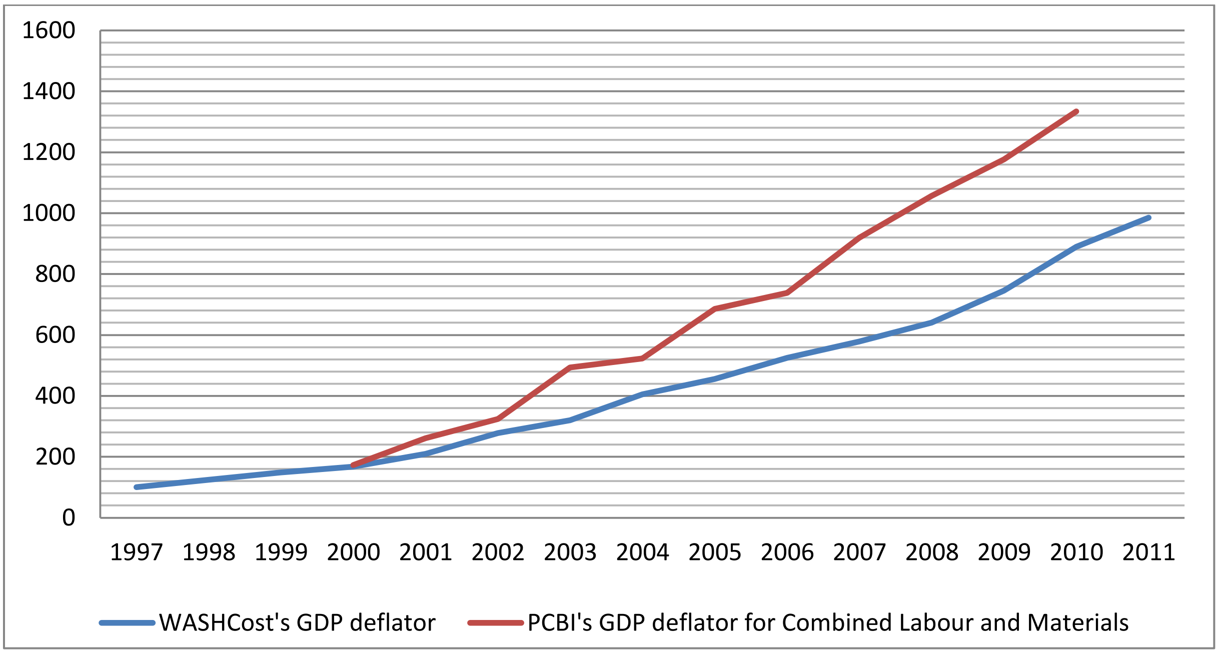


Figure 7. Comparison of the tow GDP deflators. The blue lower line are used in present study. 1997=100, base year. Source: Ghana Statistical Service (2011) and WASHCost (Appendix 2).

2.2.2 Correlation analysis

The correlation coefficient, ρ_{xy}, is a measurement for evaluating the correlation between variables. The coefficient is a normed measure derived from dividing covariance with the product of the standard deviation of the tested variables (Matematisk Statistik 2010). The coefficient is defined as

$$\rho_{xy} = \frac{C(X,Y)}{D(X) \cdot D(Y)}$$

$C(X,Y)$ is the covariance between the tested variables
$D(X)$ is the standard deviation for variable X and $D(Y)$ for variable Y

The coefficient is always in the range -1 > ρ_{xy} > 1. If ρ_{xy} is positive, then there is a positive correlation between X and Y. If ρ_{xy} is negative, then the correlation between X and Y is negative. If $\rho_{xy} = 0$, then no correlation exist. As in the case of this study, an approximation of the correlation coefficient can be estimated. Pairs of variables are taken from *n* data points and are placed in a series of pairs. Then the estimation of ρ_{xy}, here called r_{xy}, can be calculated

$$\rho^*_{xy} = r_{xy} = \frac{c_{xy}}{s_x \cdot s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Since calculating r_{xy} is nothing else than estimations, some tools are required to approach and to examine the results from the correlation analysis. Also because of the uncertainty in the estimations, a level of probability has to be ensured. In general, the closer r_{xy} gets to either -1 or 1, the higher probability of a correlation between variable X and Y. To ensure that the correlation is of a certain probability, a t-test can be done (MathWorks 2012). If the variables are normally varied, then the t-value can be computed at n-2 degrees of freedom as shown below:

$$t = r_{xy}\sqrt{\frac{n-2}{1-r_{xy}^2}}$$

A chart, see appendix 3, can then be used to find out the probability that ρ_{xy} is significant different from zero. The computed t-value corresponds to a certain level of probability. A level that is lower than 0.05 means that the probability for two variables to not have any correlation is lower than five percent. As a rule, all the independent variables which correlated with the dependent variable at a level < 0.05 where selected for further analysis. All the parameters had a proxy degree of freedom equal to 40. It means that to obtain a level lower than 0.05; r_{xy} had to be greater than $|0.31|$.

If selection of interesting independent variables is the first use of a correlation analysis; the second use is to find out which variables really are independent. If the correlation is very high between two independent variables, then probably they are not really independent and will be hard to use in the same multiple regression analysis.

There are both some limitations and some deceptive scenarios with a correlation analysis. Once again it is worth mentioning that a correlation analysis only gives a correlation between two variables; it tells nothing about causes and effects. The correlation is only measured as a linear correlation. It means that a perfectly quadratic curve will give no correlation at all in the analysis, where the case actually is the opposite. On the other hand, the analysis may give a very strong linear correlation with just a couple of outliers, where there would have been no correlation without the outliers. These misinterpretations can easily be prevented by plotting each possible independent variable against the dependent (Matematisk Statistik 2010). The results from the correlation analysis are presented under section 2.3

Modeling approach.

2.2.3 Multiple regression analysis

A regression analysis is a method to estimate a responding variable, also named dependent variable, on the basis of explaining variables, also named independent or predictor variables. Through a regression analysis, the variety of an independent variable are sought to be described by explaining variables. Multiple regression analysis is simply a regression analysis with multiple explaining variables. According to Kinney (2002), "Multiple linear regression models help experimenters perform complex analyses in very practical situations". A general model for regression analysis consists of a responding variable, y, a function consisting of explaining variables, $f(X)$, and a random error or deviation, ε.

$$y = f(X) + \varepsilon$$

The selection of explaining variables included in the functions varies from one part of the small town water system to another. Each selection is described and discussed under section 0. The other issue of modeling is to choose the most accurate function to describe each capital expenditure. According to section 1.3, *Previous studies*, the most common function used to derive functions for capital costs is the so-called log-log model. Beside the log-log model, a regular linear model is tested and in the cases where one explaining variable seems to be highly correlated to the capital expenditure in a polynomial way a polynomial model is tested.

The question that is the most complicated to answer in the analysis is: When is a model accurate enough? There are plenty of other models that might fit better with the data, so when is the model reliable and when to stop the analysis? This is one of the thesis' weaknesses and the question can only be answered with comparative key parameters. The parameters I used to analyze a model in present thesis are described under section *Key parameters for validation* and are as follow:

- Significance level
- Residual Analysis
- Coefficient of determination (R^2)
- Root Mean Square Error (RMSE)

Linear model

A general linear model with multiple explaining, independent variables is written

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_n x_{ni} + \varepsilon_i$$

$y_i =$ dependent variable from data point i, $i = 1,2,\dots,n$
$x_{1i}, x_{2i}, \dots, x_{ni} =$ explaining variables from data point i
$\alpha =$ constant to be estimated
$\beta_1, \beta_2, \dots, \beta_n =$ coefficients to be estimated, relating to explaining variables $x_1, x_2, \dots, x_n$
$\varepsilon_i =$ random error for data point i, $\varepsilon_i \in N(0, \sigma)$

The linear model is a basic model, serving as a base for many other models, including the other one used in this thesis. The model is used by OECD (2005).

Log-log model

The log-log model, also called multiplicative model, is basically a version of the linear model, see *Linear model*. Several studies present results with a log-log model, e.g. WRc (1977), Clark and Stevie (1981), Eilers (1984), Kirshen et.al. (2004) and OECD (2005). In a general form the log-log model is written

$$y_i = \alpha \cdot x_{1i}^{\beta_1} \cdot x_{2i}^{\beta_2} \cdot \dots \cdot x_{ni}^{\beta_n} \cdot \varepsilon_i$$

$y_i =$ dependent variable from data point i, $i = 1,2,\dots,n$
$x_{1i}, x_{2i}, \dots, x_{ni} =$ explaining variables from data point i
$\alpha =$ constant to be estimated
$\beta_1, \beta_2, \dots, \beta_n =$ coefficients to be estimated, relating to explaining variables $x_1, x_2, \dots, x_n$
$\varepsilon_i =$ random error for data point i, $\log(\varepsilon_i) \in N(0, \sigma)$

To compute the parameters of interest, α and β, the model is turned into natural logarithms on both side of the equal sign. The parameters can then be computed in the same way as for a linear model. This is why the model is called “log-log”. The log-log form of the multiplicative model is written

$$log(y_i) = \log(\alpha) + \beta_1 \log(x_{1i}) + \beta_2 \log(x_{2i}) + \cdots + \beta_n \log(x_{ni}) + \log(\varepsilon_i)$$

Polynomial model

The polynomial model is, like the log-log model, developed from the linear model. In this study, the model is only used when a single variable is highly correlated to the responding variable and the correlation does not seem to be linear. Polynomial model is for instance used by OECD (2005). The model is written

$$y_i = \alpha + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_n x_i^n + \varepsilon_i$$

y_i = dependent variable from data point i, $i = 1,2,\ldots,n$
x_i = explaining variable from data point i
α = constant to be estimated
$\beta_1, \beta_2, \ldots, \beta_n$ = coefficients to be estimated, relating to each polynomial level of x_i
ε_i = random error for data point i, $\varepsilon_i \in N(0,\sigma)$

Key parameters for validation

Significance level

In this report, the significance level describes at what level all the computed factors (β_i) and constants (α) in the model are significant apart from zero. Derived from the data, the computed factors get a most probable value, but because of variation and uncertainty in data it is a very little chance that this is the exact value. Instead, it is more interesting to know the confidence interval of a factor at a certain level of probability. It is particularly interesting to know at what level the interval for a value does not include zero. For instance, if the significance level is 0.01 for the following model

$$Cost = \alpha + \beta x; \ \alpha = 1 \ ; \ \beta = 2$$

it is only one percent risk that the real value of α and β are zero or negative. The security of the model increases when the significance level becomes higher. It is considered as “a very low risk” that zero is included if the significance level is below 0.05 (Vännman 2010). On the other hand, it is arbitrary to use level 0.05 in statistics when arguing for a low risk (Stigler 2008).

Residual Analysis

The models described above all relies on the assumption that the errors, ε_i, are normal distributed and independent from each other. This leads to the assumption that the y-values are normal distributed and independent from each other too. To verify, a residual analysis is good method. A residual, Δy, is the deviation from the observed y-value, y_i, to the same point on the estimated line, y_i^* (Matematisk Statistik 2010). The residual is defined as

$$\Delta y_i = y_i - y_i^*, \qquad i = 1,2,3,\ldots,n$$

y_i = The real value for y in data point i according to the collected data
y_i^* = The responding value of y computed by the function $f(x_i)$
n = Number of data points

The residuals are then plotted next to each other for analysis of patterns, see Figure 88 below. If no patterns are observed, then it is justified to draw the conclusion that the residuals are independent. The residuals should also be plotted in a normal distribution graph to see that they roughly follow a normal distribution curve (Kinney 2002). If a model does not satisfactory pass the residual analysis, it has to be rejected.

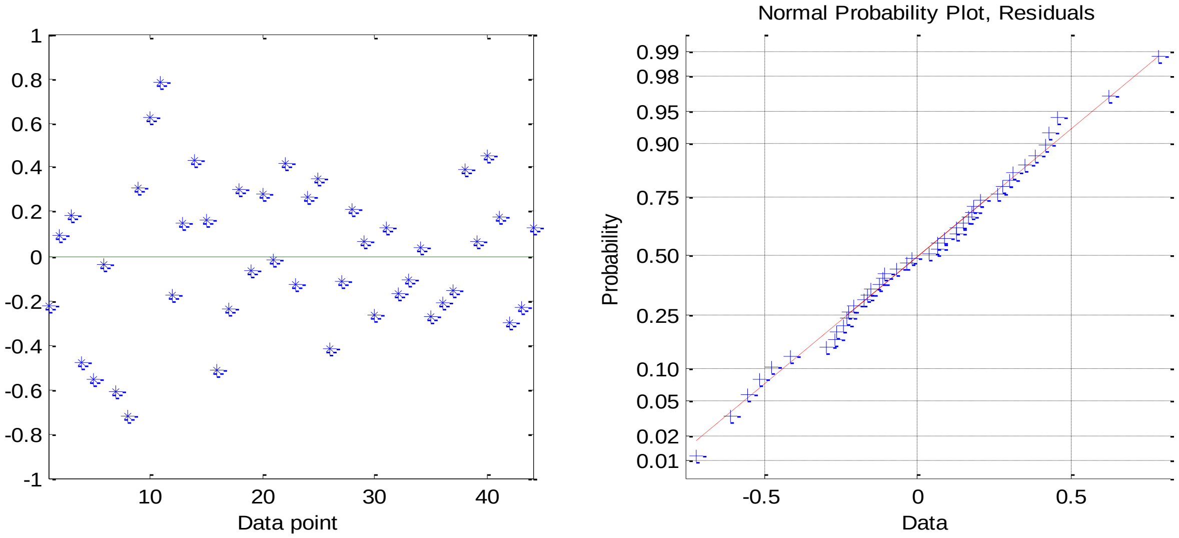


Figure 8. Example taken from residual analysis of CapEx for building stand posts. Left chart: Residuals plotted next each other without patterns. Right chart: Residuals plotted in a normal distribution chart following a normal distribution in an acceptable way.

Coefficient of determination (R²)

Statistically, the coefficient of determination (R^2) is a squared r_{xy}, see section 2.2.2. As r_{xy} shows the correlation between two variables, R^2 describes the correlation between observed data and the estimated line. Since R^2 is squared, the value goes from zero to one, where zero means that the model explains nothing of the observed data and one means that the model explains 100 % of the observed data. All added explaining parameters will increase R^2, so common sense has to validate if the increase shows an improvement (Kinney 2002). Together with e.g. the confidence interval for each explaining variable, the increase of R^2 can tell if a model is an improvement compared to a similar model. To get a second opinion regarding the correlation, it is worth plotting both the observed data and the predicted values in the same graph and compare.

Root Mean Square Error (RMSE)

The RMSE is a measurement of how the data points are scattered around the computed function. It gives a picture of how well a model follows the collected data. The RMSE is calculated as following

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} \Delta y^2}{n}}$$

Δy = See equation under section *Residual Analysis*
n = Number of data points

Since the residuals are squared, the RMSE weighs larger residuals more than smaller. It makes the method very sensible for outliers. This is an advantage, as the RMSE reveals if a strong correlation only exist due to outliers. To enable comparison between studies, the coefficient of variation of the RMSE, CV(RMSE), is presented in the results together with RMSE. It is simply the RMSE normalized to the mean of the observed CapEx and is presented in percentage.

$$CV(RMSE) = \frac{RMSE}{\bar{x}}$$

$\bar{x}$ = Mean of data

2.2.4 Dummy variables

All the above mentioned models for examining variables require numerical input data. Almost half of the variables for the study are of a qualitative nature. Dummy variable is a numeric stand-in for these qualitative variables. A dummy variable is usually either 0 or 1 and indicates if a statistical occurrence is absent (0) or present (1). Consider the variable region, which has three qualitative options. Each data point is located in a specific region. If data point X is located in Ashanti region, the dummy variable "Ashanti region" is 1 and the variables "Northern region" and "Volta region" are 0 for data point X. This makes it possible to consider quantitative data in numerical methods.

For mathematical reasons the normally used values 0 and 1 has to be changed to 1 and 2 when a log-log model is applied. If an input data is 1 in a log-log model, the outcome from the function will not change just like if the input data would be 0 in a normal linear model.

2.3 Modeling approach

This chapter describes the modeling approach for each object of the analysis. I used the following two programs as tools for modeling:

- Microsoft Excel 2010
- Mathwork's Matlab R2011a

Excel has been a platform for basic calculations, e.g. computing the deflated costs. The best advantage of excel has been to store all data in a good visual way. With excel it is also easy to get an overview of charts and compare a lot of data and values, such as the correlation analysis.

Matlab is software developed to handle effectively matrixes. Since the program is specialized for matrixes, statistical analysis with multiple variables is well suited. Especially two of Matlab's functions are used throughout the analysis. The first is "corrcoef", which computes correlation coefficients between the input variables and presents the relationships in a matrix. The second function is "regress", which estimates the best values for α and β_i for the examined model. The function also gives the confidence interval for each parameter, the significance level and coefficient of determination.

Charts from the correlation analysis are presented under this section since the correlation is just part of the method towards the results and not part of the thesis' objectives. Charts plotting each explaining variable against CapEx are not presented in order to reduce the volume of the report. In case a plot indicates a non-linear correlation; it is mentioned under each subtitle.

2.3.1 Small town water system

To not neglect any possible explanatory variables, I put all variables into relationship with total capital expenditure (TotCapEx) for a water system. In total, all 13 variables were examined to find out those of the available variables which could explain the TotCapEx. Because of the greater amount of variables, the first sorting was to remove all the variables which did not have a significance level lower than 0.05. None of the explaining variable indicated a non-linear correlation; hence a polynomial model is not of interest. Left from this first analysis are the variables showed with their relative correlation in following Table 4.

Table 4. Results from the correlation analysis with TotCapEx as responding variable.

	TotCapEx	Volta region	Northern region	Population	Length of pipeline	Total volume of tanks
TotCapEx	1	-	-	-	-	-
Volta region	- 0,69	1	-	-	-	-
Northern region	0,44	-0,47	1	-	-	-
Population	0,40	0,02	0,05	1	-	-
Length of pipeline	0,50	-0,13	0,02	0,72	1	-
Total volume of tanks	0,59	-0,37	0,12	0,64	0,58	1

The variables with strongest correlation to TotCapEx are 1) if the system is located in Volta region and 2) how large volumes that are stored within the water system. Also the length of pipelines seems to explain the cost. Population has a stronger correlation to the length of pipeline than the total cost, which indicates that they might be unusable together in a multiple linear regression analysis. It is also the case for population and total volume of tanks. Thereafter I tested the explaining variables in varied combinations in both normal and log-log models. After comparing the different models, the final recommended functions were derived and are presented under *section 3 Results and discussion*.

2.3.2 Mechanized borehole

In an interview with Dr. B. Ali and Dr. K.B. Nyarko (2012-03-02), Department of Civil Engineering, Kwame Nkrumah University of Science and Technology, Ghana, I learned that the cost drivers when building a mechanized borehole should be number of boreholes, chosen diameter and depth. The casing length and quantity of gravel and cement might be factors too. This is in line with the cost drivers used for examining the investment costs for boreholes in *TR61 – Cost information of water supply and sewage –disposal* (WRC 1977). The only available variable for this study was the number of boreholes. Without success, I have carried out a review of previous reports from CSWA regarding cost and technical information from piped water systems to fill the gaps of data. All the information was reported in lumps, which made it useless for the study. Other variables, such as regions and design population, were tested as variables as well but did not give any satisfying results in the correlation analysis. None of the variables used gave a sufficient correlation to CapEx for either borehole site works or for building the boreholes. For this reason I have not reported any function for mechanized borehole under *section 3, Results & Discussion*.

2.3.3 Water reservoirs

Possible independent variables to CapEx of water reservoirs were chosen from the collected raw data. These variables are regions, number of storage tanks, design population, volume, height and type of reservoir. During the analysis I found that steel tanks are the only used option within the water systems in Northern region, wherefore the steel tanks are treated individually. The data consists of just three water systems using Polytank, which is far too few to draw statistical conclusions. The Polytanks were removed from the analysis to be able to focus the cost function on reservoirs made out of concrete. The volume and the height of the reservoirs were kept in a reasonable range, why it felt unnecessary to distinct smaller from larger storage tanks.

For the water reservoirs, two cost functions were finally derived for estimating CapEx:

1. Steel tank – normally only used in the northern region
2. Reservoirs made of concrete – both excavated ground reservoirs and high level concrete tanks are evaluated in the same function.

Steel tank

As mentioned above, the steel tank is only used in Northern region; hence region is not an interesting parameter. Although one must bear in mind that the result only reflects the cost for a steel tank in Northern region. The water systems using steel tank have just one tank each, which makes the number of storage tanks unusable as a variable. The volume, height and design population were all highly correlated to the cost, but they also correlated strongly amongst themselves. The last three variables were tested in a multiple linear regression in all possible combinations and in both normal and log-log models. No non-linear correlation was found by plotting the explaining variables against CapEx.

Table 5 below presents the best correlated variables.

Table 5. Results from the correlation analysis with CapEx for building a steel tank as responding variable.

	CapEx Steel tank	Height	Population	Volume
CapEx Steel tank	1	-	-	-
Height	0,72	1	-	-
Population	0,76	0,88	1	-
Volume	0,94	0,81	0,79	1

Concrete reservoirs

The construction of concrete reservoirs can be done in two different ways. The two ways are either to bury them down into the ground, here named ground concrete tank (GCT), or to use a high level concrete tank, here solely named concrete tank (CT). Except for three water systems using polytanks they all used CT or GCT in Volta and Ashanti region. Only five of 27 communities use GCT, which is a bit sparse in number of data points. Almost all the communities or water systems used one reservoir. In the cases where two or three reservoirs were used, the data indicates that the cost increases with higher number of reservoirs. However, they were too few and the cost for just one reservoir varied too much to make statistically significant assumptions. A cost function for reservoirs made of concrete was harder to derive. None of the possible independent variables from the raw data strongly correlated and no non-linear correlation was found to explain CapEx for concrete reservoirs. The variables were tested in a multiple linear regression in all possible combinations and in both a normal and a log-log model. Table 6 below presents the best correlated variables.

Table 6. Results from the correlation analysis with CapEx for building a concrete tank as responding variable.

	CapEx Concrete	Volume	CT	Volta region
CapEx Concrete	1	-	-	-
Volume	0,24	1	-	-
CT	0,30	-0,25	1	-
Volta Region	-0,46	-0,17	0,08	1

2.3.4 Pipework

The first step in the analysis was to distinguish possible independent variables to explain the CapEx for the pipework in a piped water system. All available and possibly independent variables were picked out from the raw data. These variables are region, number of communities, design population, length of pipeline used for transmission and for distribution, number of stand post and selected technical option.(i.e. mechanical borehole or water provided from GWCL). Some variables which were not available can easily be assumed as interesting for the cost of a pipework. Examples that have been used as cost drivers in *TR61 – Cost information of water supply and sewage –disposal* (WRC 1977) are material of the pipeline and the most common diameter used in the net.Of the existing data, Volta region, design population, length of transmission, length of distribution and total length of pipelines correlated at an acceptable level. As shown in Table 7 below, the correlation is high between population, total, transmission and distribution length of pipelines.

Table 7. Results from the correlation analysis with CapEx for building pipework as responding variable.

	CapEx Pipework	Volta region	Design Pop.	Total length (m)	Trans.pipe length (m)	Distr. pipe length (m)
CapEx Pipework	1	-	-	-	-	-
Volta region	-0,43	1	-	-	-	-
Design Pop.	0,53	0,14	1	-	-	-
Total length (m)	0,82	-0,11	0,62	1	-	-
Trans. pipe length (m)	0,81	-0,15	0,61	0,90	1	-
Distr. pipe length (m)	0,44	-0,01	0,30	0,68	0,30	1

No non-linear correlation was found by plotting the explaining variables against CapEx. Due to the strong relation between the supposed independent variable, they were analyzed in various constellations. When two parameters with too strong correlation where used as explaining variables, then the model showed several parameters with its confidence interval covering zero, i.e. not significant. The best result where achieved using the total length of pipelines and the dummy variable for Volta region.

2.3.5 Standpost

The possible independent variables to explain the responding CapEx for stand posts are region, number of communities, design population and number of stand posts. The first step after having decided the possible variables is to analyze the correlation between all the parameters. The analysis showed that region did not matter at all. The three other parameters, presented in Table 8, showed enough reasonable significant levels (<0.05) to be worth evaluating in a multiple linear regression. No non-linear correlation of interest was found by plotting the explaining variables against CapEx.

Table 8. Results from the correlation analysis with CapEx for building a single stand post as responding variable.

	CapEx Stand Post	# of Com.	Design Pop.	# of Stand Posts
CapEx Stand Post	1	-	-	-
# of Communities	0,38	1	-	-
Design Population	0,79	0,08	1	-
# of Stand Posts	0,68	0,52	0,74	1

As the only case, the cost function for stand posts did not give any significant constant, α, in the normal multiple linear regression models, no matter what independent variable were used. Instead I chose the log-log model to derive the cost function.

SECTION THREE

3 Results & Discussion

This section describes the estimated results from the analysis. The results are built on the empirical data presented under section 2.1, *Data collection.* Since the analysis failed to derive a function for predicting CapEx for mechanized borehole, there is no such function presented in this section.

Almost all the results are presented with two recommended cost functions for each analyzed object. The *first* function has a better accuracy, but has such variables that a first design investigation must take place before having any use of the function. The *second* function is therefore derived to get a picture of the cost before even an investigation of the location for the forthcoming water system has taken place. In one particular case, water reservoir made of concrete, only one recommended function is presented. Due to the inferior reliability even of the first function, I found no reason to present an even less accurate function. All the functions are presented in thousand (1000) Ghana Cedis (GHs). This is showed in the function as ‘000 GHs. It means that the user should multiply the output result with thousand.

3.1 Small town water system

3.1.1 Data summary

The data used for deriving the two cost functions to explain TotCapEx is summarized and presented in Table 9 below.

Table 9. Data summary for the variables used in the functions to predict TotCapEx.

Variable	Label	Unit	Min.	Max.	Mean	Median	Std.Dev.
Total Capital expenditure for a Water system	COST	GHs	58 109	960 748	494 150	499 217	235 352
Dummy variable for Volta region. 1 if Volta, 0 if other	REG	-	0	1	0,33	0	0,48
Dummy variable for Northern region. 1 if Northern, 0 if other	REG	-	0	1	0,31	0	0,47
Design population	POP	person	1 533	11 493	5 288	4 746	2 738
Total lenght of pipelines	PIPE	meter	1 323	17 574	6 328	5 188	4 192
Total volume of storage tanks in a water system	VOL	m3	12,0	200,0	77,5	75,0	39,2

3.1.2 Recommended functions

Both the first and second function are expressed in ‘000 GHs and are derived as normal linear models. ‘000 means that the user should multiply the result from the function with thousand (1000).

First function

The function to explain TotCapEx with best accuracy contains the explaining variables length of pipeline, volume of tank and dummy variables for Volta region and Northern region. The recommended function to estimate TotCapEx for a small town water system is

$$\boldsymbol{TotCapEx\ (\text{‘}000\ GHs) = 338 + REG + 0.0178 \cdot PIPE + 1.26 \cdot VOL}$$

REG: If Volta region, then REG = -243, if Northern region, then REG = 86., if other region, then REG = 0
PIPE= Total length of pipelines
VOL= Total volume of storage tanks

Table 10. Key parameters for validation of the first function of TotCapEx

Observations	Significance level	R^2	R	RMSE	CV(RMSE)
42	< 0.10	0,70	0,84	127 067 GHs	25.7 %

As presented in Table 10, the level of significance does not reach the commonly used level 0.05. 70 % of TotCapEx are explained by the first function with above mentioned input parameters. The CV(RMSE) is 25.7 % of 494 150 GHs, the mean of observed TotCapEx. The values are, compared with the other functions in my study, in the upper range. As displayed in Figure 9, the predicted costs are slightly overestimated for most of the first data points. These are data points from Ashanti region, but the deviation where not significant enough to explain the predicted cost in the function. I did not detect any patterns by analyzing the residual plots.

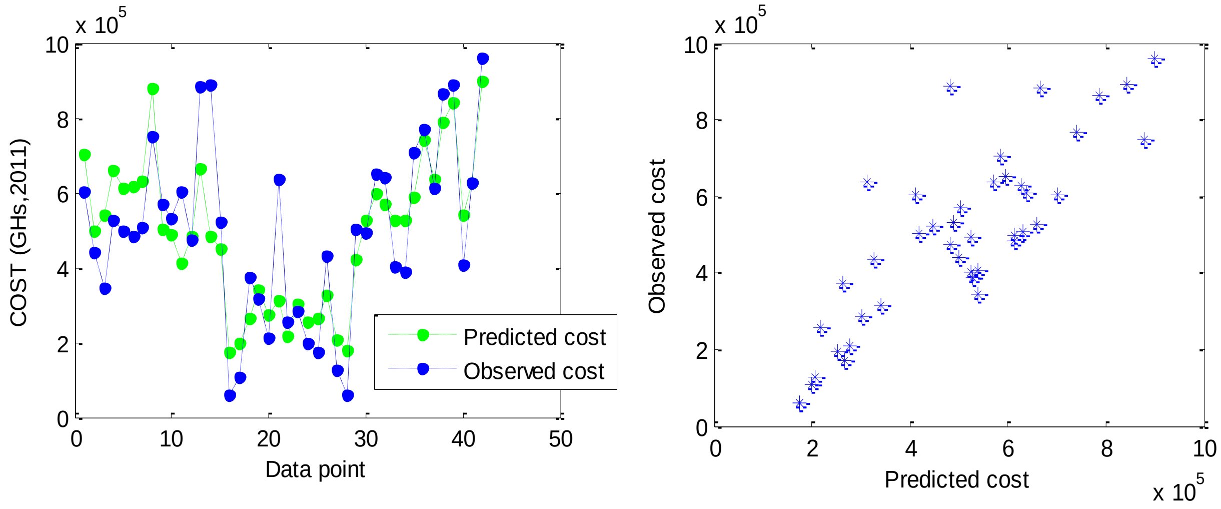


Figure 9. Predicted cost is the output from the function with the explaining variables from data point x as input and observed cost is the cost observed in data point x.

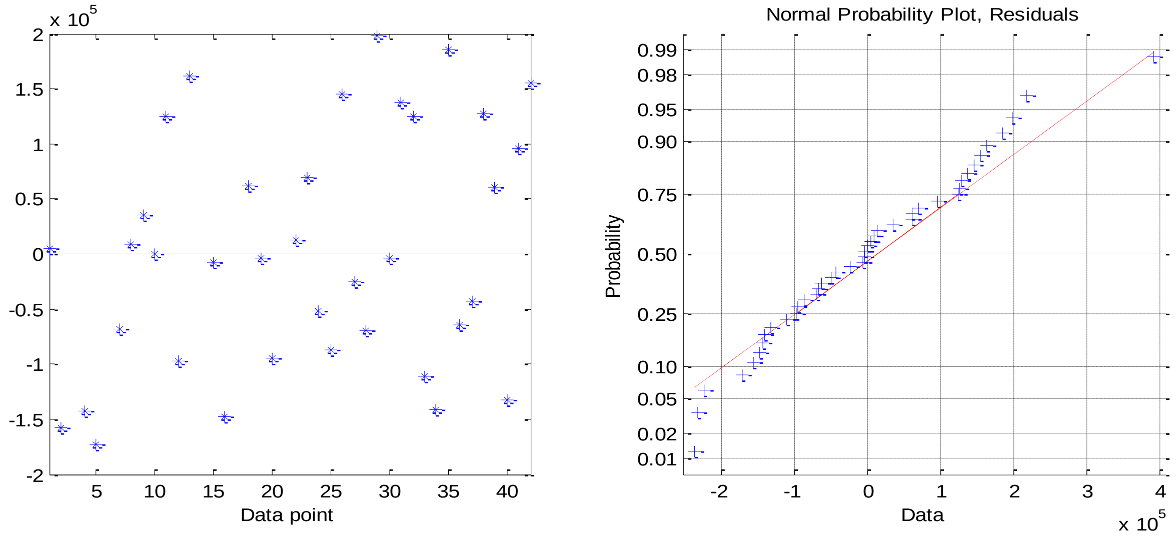


Figure 10. Left: The residuals plotted in order with their distance from the function line. Right: The residuals plotted against a normal probability plot with the red line representing a perfect normal distribution.

Second function

The second cost function, which can be used initially in planning a new water system, depends on the explanatory variables for Volta region and design population. The second, less precise function is:

$$CapEx\ (\text{‘}000\ GHs) \ = \ 420 - 345 \cdot REG \ + \ 0.0359 \cdot POP$$

REG: If Volta region, then REG = 1

If other region, then REG = 0

POP= Design population in persons

Table 11. Key parameters for validation of the second function of TotCapEx

Observations	Significance level	R^2	R	RMSE	CV(RMSE)
42	<0.01	0,65	0,81	137 111 GHs	27.7 %

Compared to the first function, this second function has the advantage to be useful before even starting to design a water system. On the other hand, the second function uses explaining variables which give less accuracy as showed in Table 11. R^2 decreases while the RMSE increases. The only favorable parameter is the lower significance level.

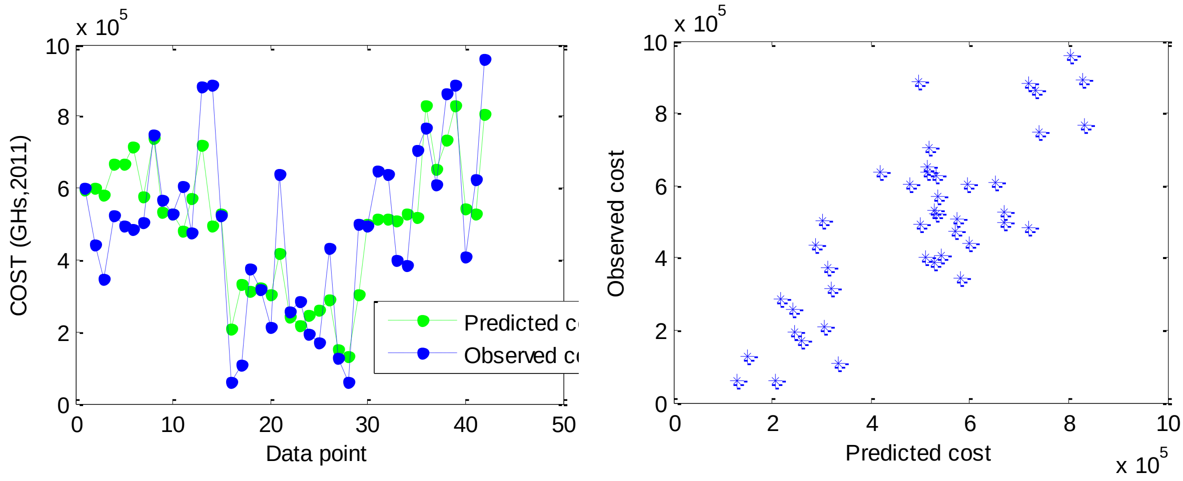


Figure 11. Predicted cost is the output from the function with the explaining variables from data point x as input and observed cost is the cost observed in data point x.

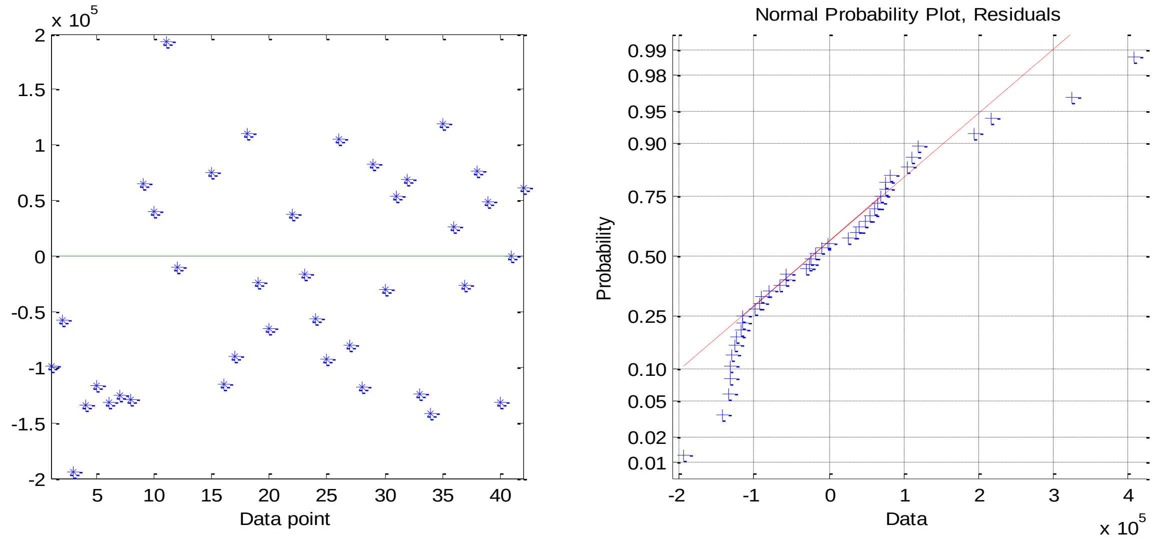


Figure 12. Left: The residuals plotted in order with their distance from the function line. Right: The residuals plotted against a normal probability plot with the red line representing a perfect normal distribution.

3.2 Water reservoirs

The results from analyzing the costs of water reservoirs are split up and presented separately for steel tanks and for reservoirs made of concrete. No function is presented for plastic tanks. To see the discussion behind these decisions, see the sub section 2.3.3, *Water reservoirs,* in *Modeling Approach*.

3.2.1 Steel tank

Data summary

The data used for deriving the two cost functions to explain CapEx for building steel tanks is summarized and presented in Table 12 below.

Table 12. Data summary for the variables used in the functions predicting CapEx for steel tank.

Variable	Label	Unit	Min.	Max.	Mean	Median	Std.Dev.
CapEx for steel tank (corrected to 2011)	COST	GHs	150 872	226 180	176 522	161 482	27 943
Volume of steel tank	VOL	m^3	50	150	85	80	34
Design population for the water system	POP	person	2 207	11 493	5 482	3 098	3 772

Recommended functions

Both first and second function are expressed in '000 GHs and are derived as normal linear models. '000 means that the user should multiply the result from the function with thousand (1000).

First function

The first recommended function use the designed total volume of a storage tank. This function can only be used after a pilot study in the study area where the need of storage capacity is investigated. The function is written

$$\boldsymbol{CapEx}\ (\text{'}\mathbf{000}\ \boldsymbol{GHs})\ =\ \mathbf{110.28}\ +\ \mathbf{0.78}\cdot\boldsymbol{VOL}$$

VOL = Designed volume of steel tank in m^3.

Table 13. Key parameters for validation of the first function predicting CapEx for steel tank.

Observations	Significance level	R^2	R	RMSE	CV(RMSE)
13	<0.01	0,88	0,94	9 123 GHs	5.1 %

Apart from the small number of observations, which is a weakness, this function has the strongest values of all functions I have derived in my study. This is especially observable by looking at the CV(RMSE) and the R^2. The residuals in Figure 14 do not have any specific pattern and the plotted figures in Figure 15 reveal the coherence between predicted and observed costs.

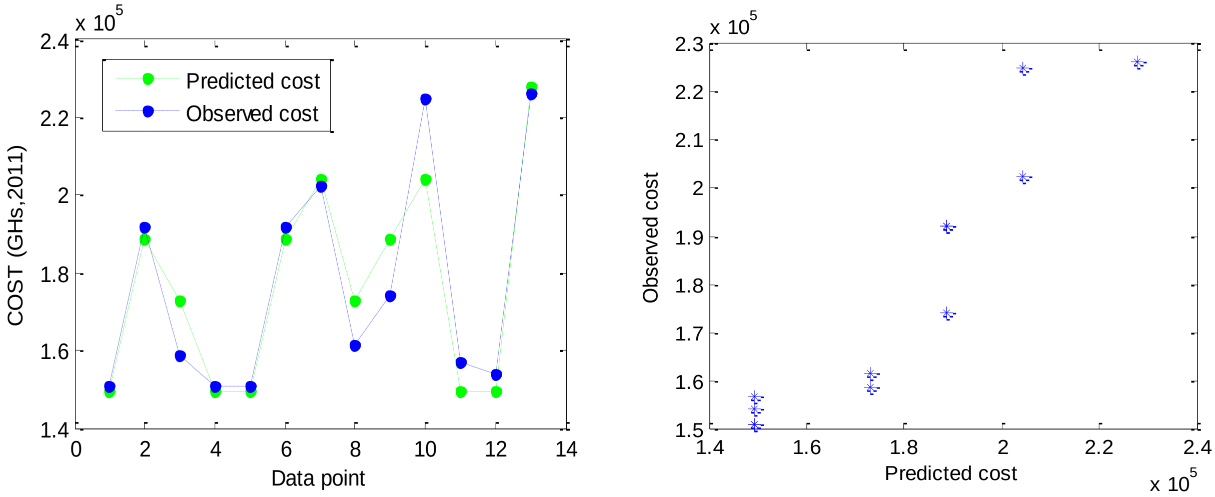


Figure 13. Predicted cost is the output from the function with the explaining variables from data point x as input and observed cost is the cost observed in data point x.

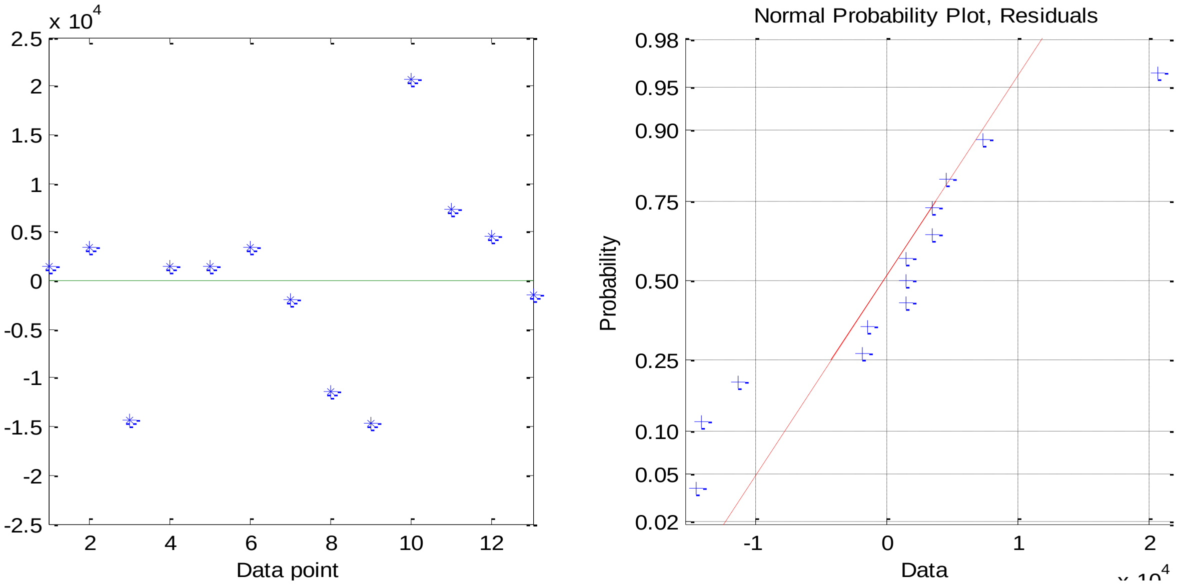


Figure 14. Left: The residuals plotted in order with their distance from the function line. Right: The residuals plotted against a normal probability plot with the red line representing a perfect normal distribution.

Second function

To estimate the cost for a steel tank before starting an investigation to design the storage capacity, the recommended function is:

$$\boldsymbol{CapEx\ ('000\ GHs)\ =\ 145.57 + \ 0.00564 \cdot POP}$$

POP = Design population in persons

Table 14. Key parameters for validation of the second function predicting CapEx for steel tank.

Observations	Significance level	R^2	R	RMSE	CV(RMSE)
13	<0.01	0.58	0.76	17 382 GHs	9.8 %

The second function, with only the design population as explaining variable, is far from the first function regarding accuracy and explanation degree. Table 14 lists that the R^2 drops to 0.58, but using design population as the only variable still gives a very low significance level. Figure 15 illustrates how the predicted costs do not correspond well to the observed cost. The user of the derived functions should consider possibilities to estimate the total need of storage volume in the study area to enable usage of the first rather than the second function.

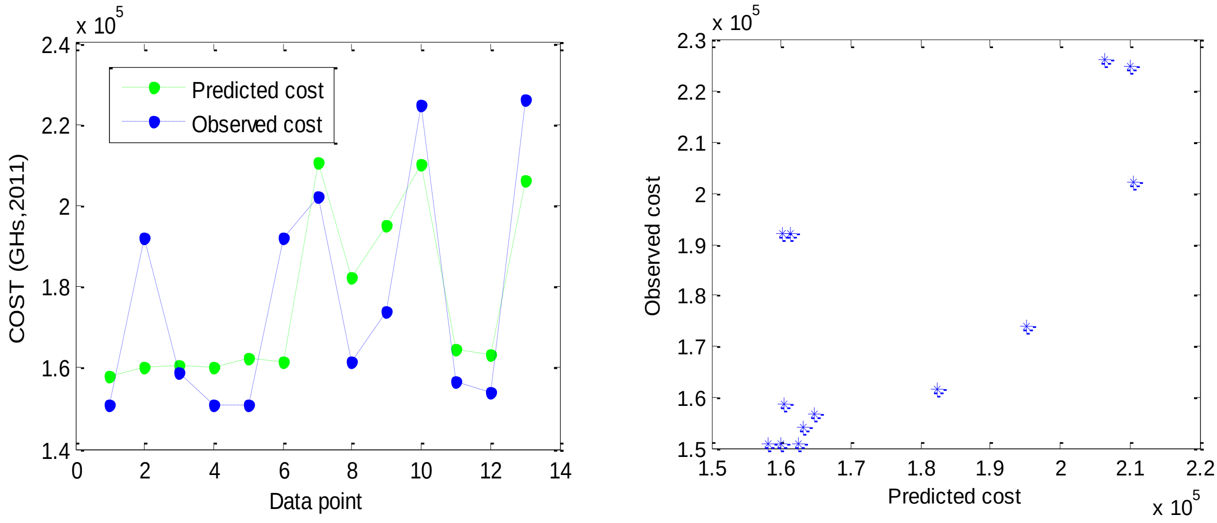


Figure 15. Predicted cost is the output from the function with the explaining variables from data point x as input and observed cost is the cost observed in data point x.

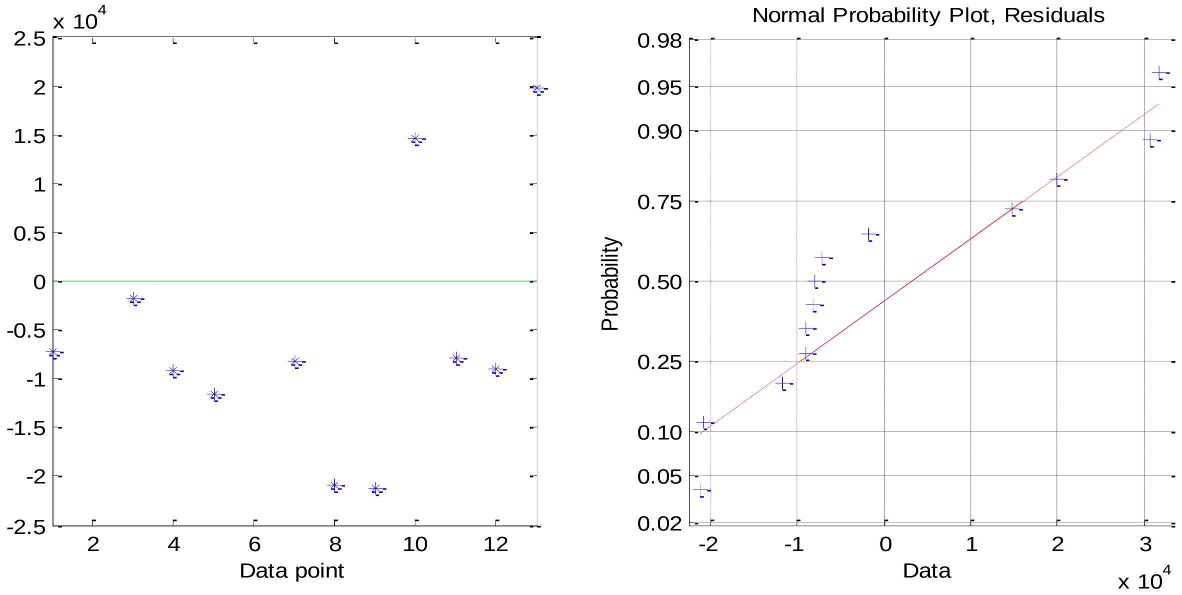


Figure 16. Left: The residuals plotted in order with their distance from the function line. Right: The residuals plotted against a normal probability plot with the red line representing a perfect normal distribution.

3.2.2 Reservoirs made of concrete

Data summary

The data used for deriving the two cost functions to explain CapEx for building reservoirs made of concrete is summarized and presented in Table 15 below.

Table 15. Data summary for the variables used in the function predicting CapEx for reservoirs made of concrete.

Variable	Label	Unit	Min.	Max.	Mean	Median	Std.Dev.
Capital expenditure for a reservoir made of concrete	COST	GHs	15 829	148 222	74 289	65 865	32 746
Size of the reservoir	VOL	m3	12,0	200,0	78,8	75,0	41,9
Dummy variable for type of reservoir. 2 if CT, 1 if GCT	TYPE	-	1,00	2,00	1,89	2,00	0,32
Dummy variable for region. 2 if Volta region, 1 if other region	REG	-	1,00	2,00	1,44	1,00	0,51

Recommended function

The final recommended function for deriving cost for a concrete reservoir is built on the explaining variables: dummy for Volta region (reduces the cost), the volume of the reservoir and whether it is a normally levitated concrete tank (CT) or it is a ground concrete tank (GCT). The function is built on a log-log model and computes the cost in thousand Ghana Cedis. The recommended function for reservoirs made of concrete is useful after a pilot study in the community and is articulated as:

$$\boldsymbol{CapEx\ ('000\ GHs) = 8519 \cdot VOL^{0.37} \cdot TYPE^{0.05} \cdot REG^{-0.66}}$$

VOL = Designed volume of concrete reservoir in m^3
TYPE = 2 if concrete tank, 1 if concrete ground reservoir
REG = 2 if Volta region, 1 if other region

Table 16. Key parameters for validation of the function predicting CapEx for reservoirs made of concrete.

Observations	Significance level	R2	R	RMSE	CV(RMSE)
27	<0.10	0,43	0,65	26 750 GHs	36.0 %

The function to derive cost for a concrete reservoir has the poorest key values of all investigated assets. The function needs a pre-investigation of the study area to be useful and still, the output from the function is very uncertain. Therefore the function is only recommended in the sense of being the best possible from the input data. It does not mean that I recommend using the function to determine the potential costs. Figure 17 reveals the lack of coherence between observed and predicted cost and underlines the uncertainty of the function.

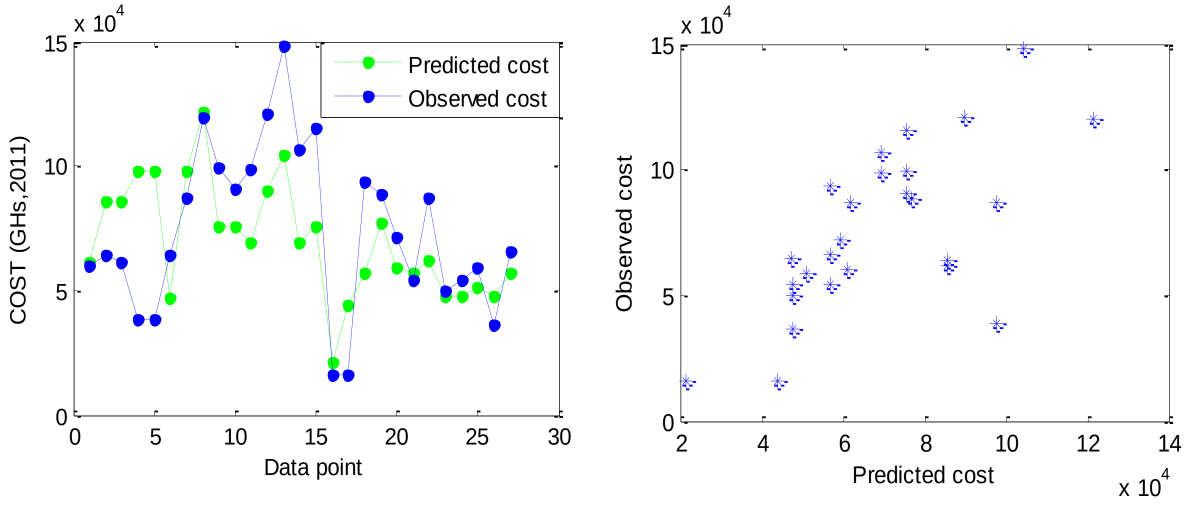


Figure 17. Predicted cost is the output from the function with the explaining variables from data point x as input and observed cost is the cost observed in data point x.

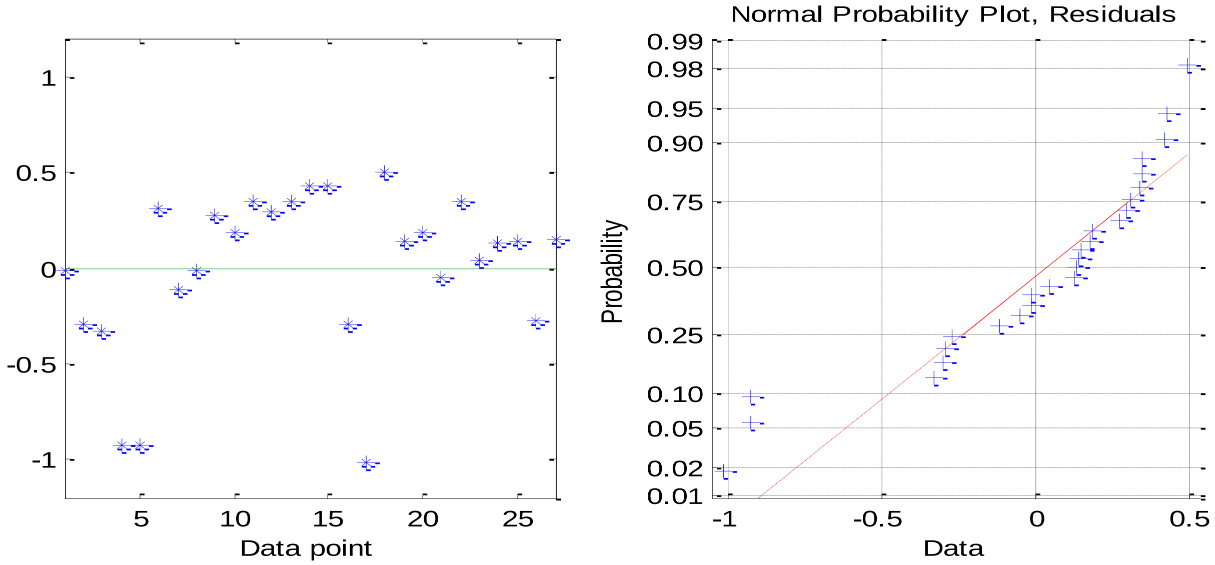


Figure 18. Left: The residuals plotted in order with their distance from the function line. Right: The residuals plotted against a normal probability plot with the red line representing a perfect normal distribution.

3.3 Pipework

3.3.1 Data summary

The data used for deriving the two cost functions to explain CapEx for building a pipeline network for water service in a small town is summarized and presented in Table 17 below.

Table 17. Data summary for the variables used in the functions predicting CapEx for pipeworks.

Variable	Label	Unit	Min.	Max.	Mean	Median	Std.Dev.
Capital expenditure for pipework	COST	GHs	14 870	455 360	156 740	137 880	103 060
Dummy variable for region. 1 if Volta region, 0 if other region	REG	-	0	1	0,36	0	0,48
Total length of pipelines	PIPE	km	1.32	17.57	6.25	5.19	4.16
Design population	POP	person	1 533	15 942	5 469	4 746	3 162

3.3.2 Recommended functions

The first function is derived as a normal linear model, while the second function is written in a log-log model.

First function

CapEx for pipework in a small town water system is best explained by the total length of pipelines in the system and if the study area is located in Volta region. The first recommended function for pipework is useful after a pilot study in the study area and is written

$$\boldsymbol{CapEx\ ('000\ GHs)\ =\ 61.21\ +19.5\cdot PIPE\ -74.50\cdot REG}$$

PIPE= Total length of pipeline in km
REG= 1 if Volta region, 0 if other region

Table 18. Key parameters for validation of the first function predicting CapEx for pipeworks.

Observations	Significance level	R^2	R	RMSE	CV(RMSE)
42	<0.01	0,80	0,89	45 455 GHs	29.0 %

The number of observations is similar to all the other functions, except for the water reservoirs'. The significance level is sufficiently low and the cost is well explained by the two input variables with R^2=0.80. An objection to use the function is the scattered residuals, especially some of the outliers that bring the values down for all the key parameters. The scattered residuals are showed with a CV(RMSE) = 29 % and in the left plot of Figure 20.

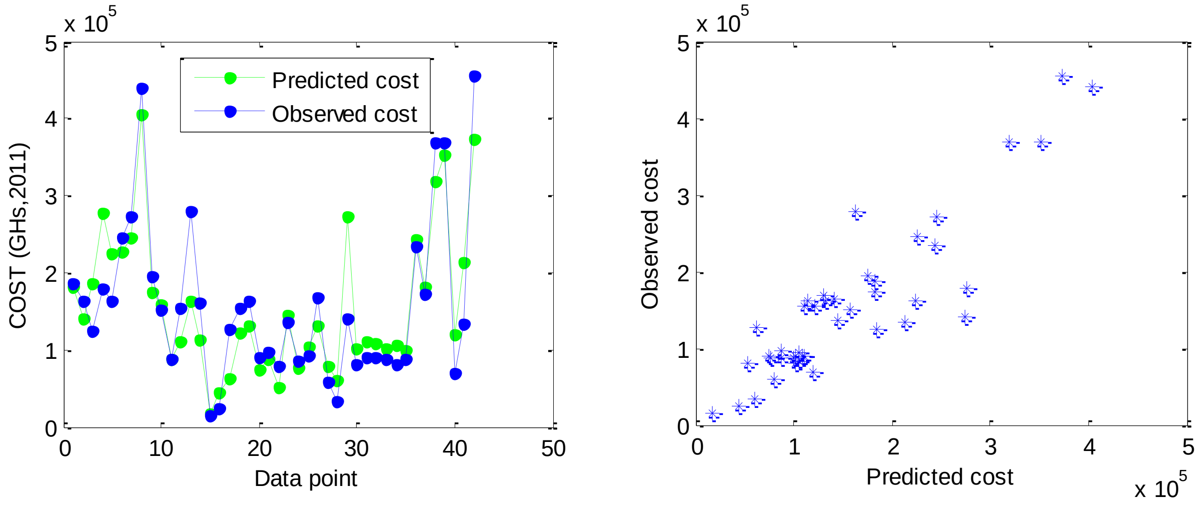


Figure 19. Predicted cost is the output from the function with the explaining variables from data point x as input and observed cost is the cost observed in data point x.

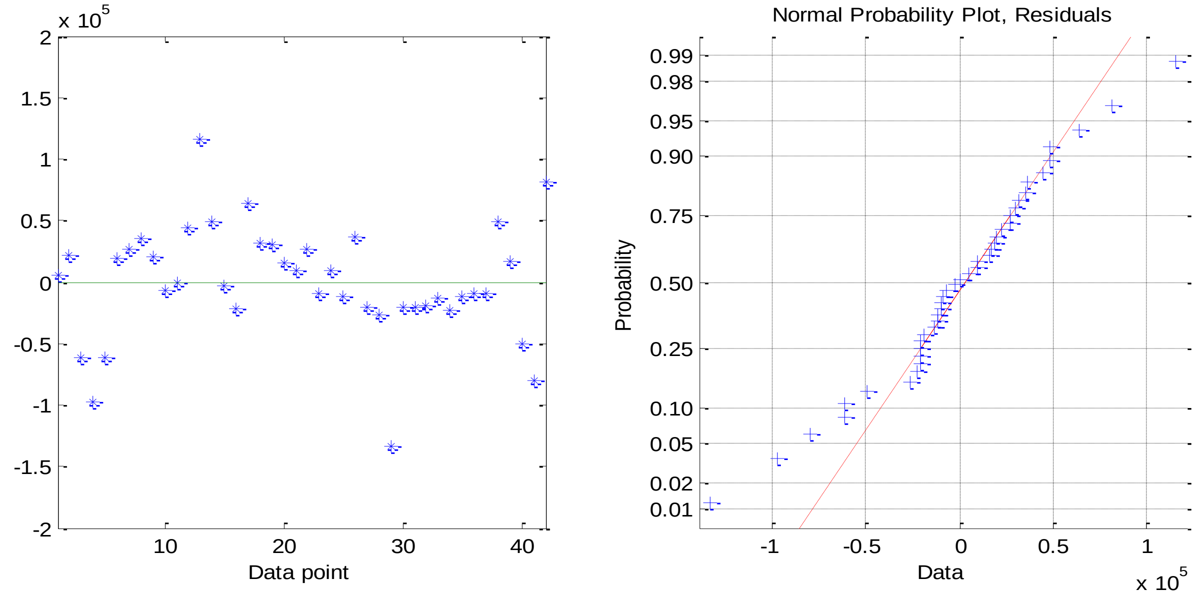


Figure 20. Left: The residuals plotted in order with their distance from the function line. Right: The residuals plotted against a normal probability plot with the red line representing a perfect normal distribution.

Second function

The second function can be used without knowing more than the design population and where the study area is located and is written

$$CapEx\ ('000) = 472.15 \cdot POP^{0.70} \cdot REG^{-1.20}$$

POP= Design population in person

REG= 2 if Volta region, 1 if other region

Table 19. Key parameters for validation of the second function predicting CapEx for reservoirs made of concrete.

Observations	Significance level	R^2	R	RMSE	CV(RMSE)
42	<0.01	0,54	0,74	63 812 GHs	40.7 %

As for all system characteristics above, the second function for pipework does not reach the same accuracy as the first function does. Although the significance level is decent and the R^2 might be considered acceptable, the RMSE shows that the residuals are very scattered. This will give a very wide interval for the predicted cost. The real prediction interval would be larger than the derived cost plus/minus the RMSE and already the RMSE interval will give a multiplicative interval between 0.59 and 1.41 of the predicted cost. The right plot in figure XX shows how most of the data points are concentrated beneath twenty-five thousand (25’000) Ghana Cedis and the residual plot in figure YY reveals how the outliers are scattered.

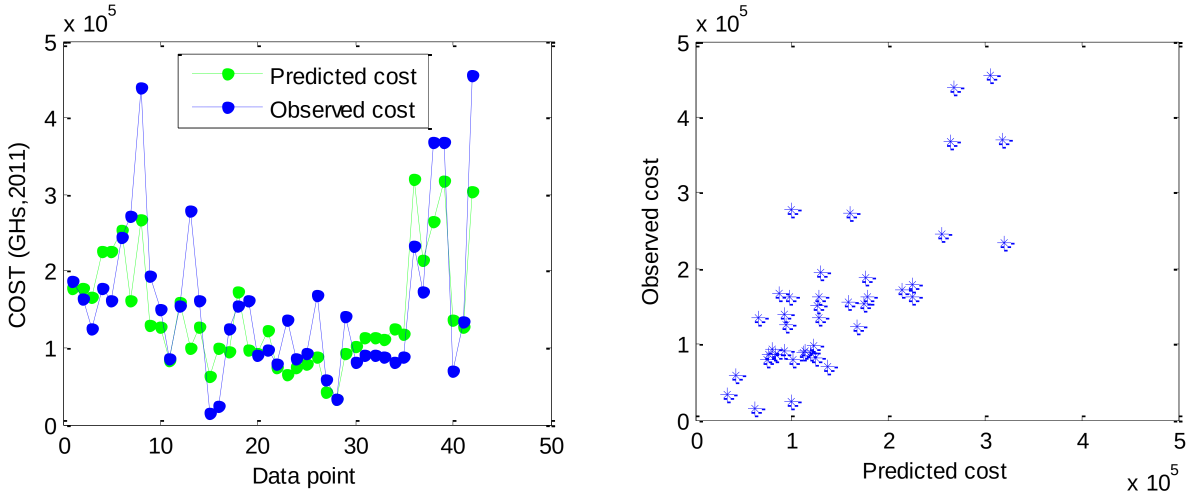


Figure 21. Predicted cost is the output from the function with the explaining variables from data point x as input and observed cost is the cost observed in data point x.

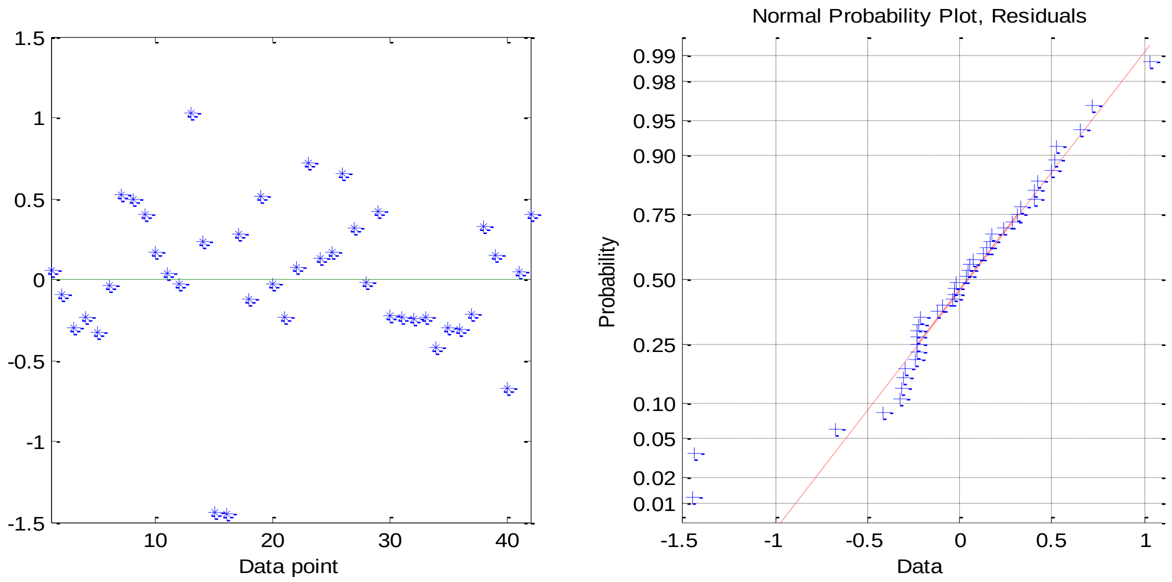


Figure 22. Left: The residuals plotted in order with their distance from the function line. Right: The residuals plotted against a normal probability plot with the red line representing a perfect normal distribution.

3.4 Stand post

Surprisingly, the design population gave better values than the number of stand posts in the multiple regression analysis and because of the high correlation between them the model just got significant with one at a time, see section 2.3.5 *Standpost*, in *Modeling approach* for correlation analysis. Therefore the final cost function depends on the variables number of communities and design population. As the function is the best possible, considering the data given and the models tested, and at the same time uses variables which do not depend on a pilot study, it is presented as the only function in *Recommended function*.

3.4.1 Data summary

The data used for deriving the cost function to explain CapEx for building the stand posts within a small town water system is summarized and presented in Table 20 below.

Table 20. Data summary for the variables used in the functions predicting CapEx for stand posts.

Variable	Label	Unit	Min.	Max.	Mean	Median	Std.Dev.
Capital expenditure for stand posts	COST	GHs	7 215	49 700	21 457	17 677	12 386
Number of communities included in the water system	COM	-	1	14	1,75	1	2,28
Design population	POP	person	1 533	15 942	5 577	4 882	3 130

3.4.2 Recommended function

The recommended function can be used without knowing more than the design population and the number of communities included in the water system. The function is written

$$\boldsymbol{CapEx\ ('000\ GHs)\ =\ 33.85 \cdot COM^{0.20} \cdot POP^{0.74}}$$

COM= Number of communities
POP= Design population in person

Table 21. Key parameters for validation of the function predicting CapEx for stand posts.

Observations	Significance level	R2	R	RMSE	CV(RMSE)
44	<0.05	0,64	0,80	6 991 GHs	32.6 %

The key parameters reveal that the function estimates the CapEx in a reasonable way. The values are not optimal but as a function for prediction, the function may give a first indication of the cost range. Again, the CV(RMSE) shows fairly scattered residuals which can also be seen in Figure 24.

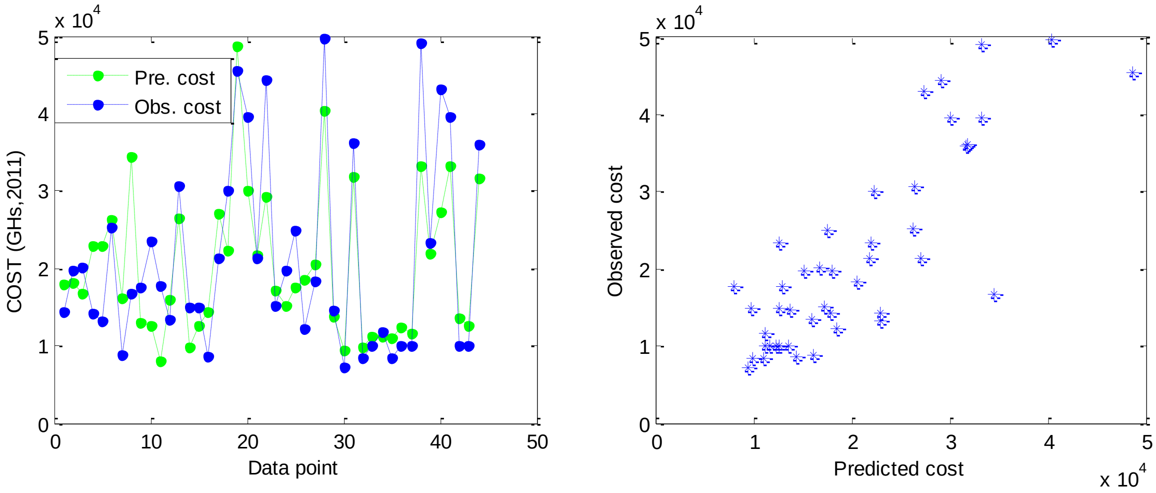


Figure 23. Predicted cost is the output from the function with the explaining variables from data point x as input and observed cost is the cost observed in data point x.

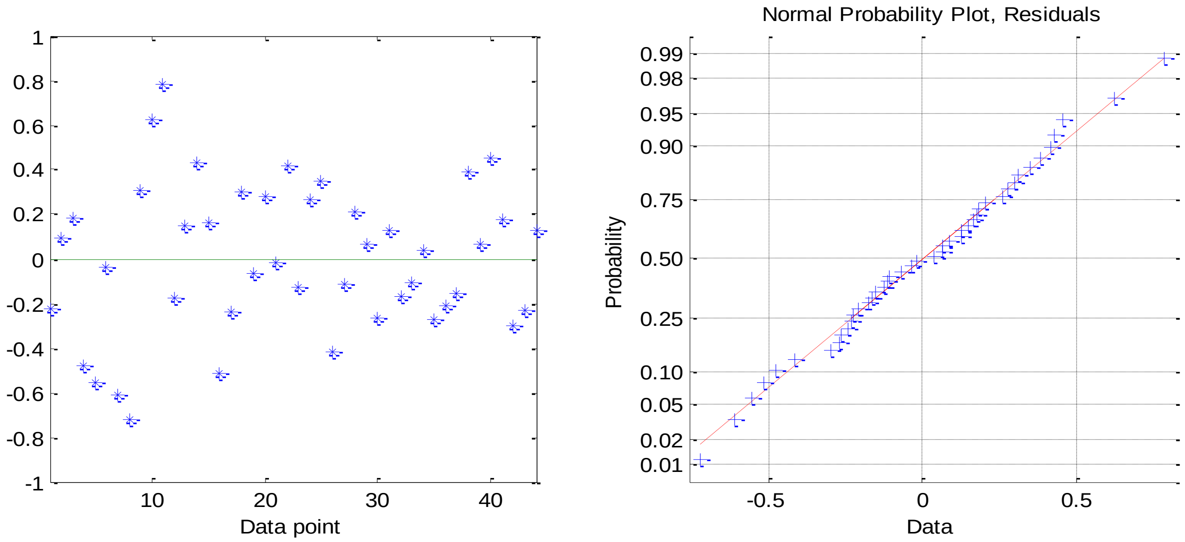


Figure 24. Left: The residuals plotted in order with their distance from the function line. Right: The residuals plotted against a normal probability plot with the red line representing a perfect normal distribution.

SECTION FOUR

4 Conclusions and suggestions for further work

In this section the results are questioned and discussed. The results are compared and discussed in relation to previous studies. Conclusions will be drawn from the discussion and suggestions will be given for further work to deepen the knowledge within the field of capital cost for small town water systems in Ghana.

4.1 Final discussion with conclusions

Overall, it has been possible to derive cost functions for a small town water system in Ghana in accordance to the thesis' objectives. Especially the first functions for TotCapEx and for the CapExs may be interesting to use, due to their better key values. These functions normally require a first investigation and design of the intended building area, but will on the other hand give more reliable cost predictions. However, in two cases it has not been possible to create functions adequate to predict the cost. In one of the cases, the key values of the function for concrete reservoirs are not good enough and the other case is the lack of a function for mechanized boreholes. The data was insufficient to derive a function to predict the cost for mechanized boreholes. These parts of the water system are discussed further down together with recommendations. The second functions, that are intended to operate as a forecaster in the early stage of a new construction project, should be used sparsely and critically. I would say, arguing from the key parameters, that the second function for TotCapEx, CapEx for steel tank and the function for stand posts are the function with enough accuracy to use. One might also question if 13 observations are enough for the functions to predict CapEx for steel tank, but then again the number of observations has taken into account when the models has been derived and when calculating the key parameters.

Most of the finally derived functions used data that fits best into a normal linear model instead of the multiplicative log-log model. This deviates from most of the previous studies, where the log-log model or versions of the log-log model are used, e.g. WRc (1977), Clark & Stevie (1981), Kirshen et.al. (2004) and Eilers (1984). Three out of eight functions in my study are best fit in a log-log model. The coefficient of determination, R^2, ranges from 0.42 to 0.88. This is in accordance with the results in the previous studies by WRc (1977) and Clark & Stevie (1981). Kirshen et. al. had over 1 300 data points in their study and got an R^2 above 0.90 for all their functions. I was not able to find any R^2 in Eilers' (1984) or OECD's (2005) reports, which makes them difficult to use for comparison. The level of significance is consistently lower than 0.10 (often <0.01) in present study which is reasonable. Although the conventional level of significance to form firm arguments is 0.05, this is still arbitrary, which brings me to the conclusion that a function with level 0.10 should not be immediately rejected (Stigler 2008). Instead it is important to evaluate the significance level together with all other key parameters, such as the RMSE and the R^2. The residual analysis where used to reject function where the errors did not fulfill the criteria to be independent each other and normal distributed. All the functions presented are considered to fulfill the criteria.

The design population is used in all the second functions to determine CapEx without having done a first design of the water system. Population is one of few variables that are available before planning a new water system and is normally, together with region, the one with most correlation to capital cost. As could have been presumed, the design population does not explain all the variability of CapEx. Nkrumah et. al. (2011) suggests that the population density of an intended construction area probably could explain the cost in a good way, but data for population density is not yet collected for the observed data points. Moreover, the authors suggested pipe length/capita as a proxy for the population density, but then again it requires knowing the length of the designed pipework and cannot be used before planning.

As the results shows, the regions often occur as explaining variables. Particularly Volta region is part of most functions and consequently reduces the cost prediction. Here it is important to stress once again that explaining versus responding variables do not imply a clear causal relationship. It is easy to understand that a region in itself cannot drive the cost; instead it is the specific circumstances for a region which drive the cost. The problem is to determine what circumstances that might inflate CapEx. The differences between the regions, based on the criteria for collecting the data, are possible answers to the question of what drives the cost. However, this cannot be concluded in my study. Contract packaging is an example of a possible cost driver, already concluded by Nkrumah et. al. (2011). For example, NCB is the only contract packaging used in Volta region, which means that it would have worked identical with a dummy variable for Volta region in a function. Still, I chose Volta region as explaining variable due to the more general character of a region. NCB as a contract package is not represented in the other two regions, Northern and Ashanti, which prevents a comparison across the regions which implies that other factors cannot be excluded as cost drivers. Moreover, there is one more problem with using region as an explaining variable; it gets more complicated to predict the cost for a water system in another region than the observed. If the functions are used for another region, the circumstances have to be well examined to ensure that the region has similar conditions as one of the regions in this study.

The GDP deflator has an impact on the results as well. Due to the used GDP deflator index, CapEx for older construction is estimated lower than if PCBI from Ghanas statistical service would have been used without modifications.

4.2 Suggestions for further work and studies

The suggestions I present below are developed from the question of how to proceed with the results. What more has to be done to refine the data to be able to adjust the functions in order to get more precise results? What could give a broader picture of the cost building a new small town water system in Ghana? The suggestions are as follows:

1. Collect more relevant physical data regarding mechanized boreholes in order to find the right cost drivers for this major asset in a water system. Data such as depth, diameter and casing length are used in WRc (1977).
2. Collect more data points to increase the number of observations for concrete reservoirs. Moreover, the quantity of used material, such as concrete and rebar would be interesting to test as explaining variables.
3. Conduct more qualitative analyses to determine what really causes an inflation of CapEx instead of the regions.
4. Conduct a life-cycle cost analysis where functions are derived while both CapEx and operational and maintenance costs are considered. Tsegai's (2009) and Antoniolo and Fillipini's (2002) studies are examples of that kind of modeling approach which processes the data to annual costs.

5 References

Ali, Dr. B. (2012-03-02). "Interview with Dr. B. Ali regarding mechanized boreholes". Department of geological engineering, KNUST, Ghana.

Antonioli, B. & Filippini, M. (2002). "The use of a variable cost function in the regulation of the Italian water industry". University of Lugano, Switzerland. Link 2012-06-07: http://doc.rero.ch/lm.php?url=1000%2C42%2C6%2C20051021110309-QZ%2F1_wp0201.pdf

Armstrong, J.S. & Collopy, F. (1992). "Error Measures for Generalizing About Forecasting Methods: Empirical Comparisons". *International Journal of Forecasting*, Volume 8, Issue 1, pages 69-80. Link 2012-05-06: http://repository.upenn.edu/cgi/viewcontent.cgi?article=1075&context=marketing_papers

CWSA (2007). The Community Water and Sanitation Agency: Corporate Brochure.

Clark, R.M. & Stevie, R.G. (1981)."A Water Supply Cost Model Incorporating Spatial Variables". *Land Economics*. Vol. 57. No. 1, pp. 18-32. University of Wisconsin. Link 2012-06-10: http://www.jstor.org/stable/3145749

Dawuni, J. (2009-02-24). "The Ghana Women's Movement and Democratic Changes". Association for Women's Rights in Development. Link 2011-09-14: http://staging.awid.org/eng/Issues-and-Analysis/Library/The-Ghana-Women-s-Movement-and-Democratic-Changes

Dwumfour-Asare, B. (2009). "Investment Cost of Small Town Water Supply Schemes in the Greater Accra Region". Msc thesis, KNUST, Ghana. Link 2012-06-07: http://www.washcost.info/page/629

Eilers, R.G. (1984). "Cost equations for small drinking water systems". US Environmental Protection Agency, EPA. Link 2012-06-07: http://nepis.epa.gov search words: "Eilers 1984"

Fonseca, Franceys, Batchelor, McIntyre, Klutse, Komives, Moriarty, Naafs, Nyarko, Pezon, Potter, Reddy, Snehalatha (2011) "Briefing Note 1a – Life-cycle costs approach". IRC – International Water and Sanitation Centre. Link 2012-06-07: http://www.washcost.info/page/1557

Ghana Statistical Service (GSS) (2011). "Prime Building Cost Index (CPI)". Link 2012-05-07: http://www.statsghana.gov.gh/docfiles/news/prime_building_costs_index(pbci)_jan2000-sept2011_P3_2011nov30.pdf

Kankam-Yeboah, K., Amisigio, B. & Obuobi, E. (2010). "Climate change impacts on water resources in Ghana". *Ghana and UNESCO*. Link 2012-05-07: http://www.natcomreport.com/ghana/livre/climate-change.pdf

Kinney, J.J. (2002). "Statistics for science and engineering". Boston: Addison-Wesley

Kirshen, P.H., Larsen, A.L., Vogel, R.M. & Moomaw, W. (2004). “Lack of influence of climate on present cost of water supply in the USA”. *Water Policy 6 ss 269-279*. Link 2012-06-07: http://engineering.tufts.edu/cee/people/vogel/publications/lack-influence.pdf

Matematisk statistik (2010). “Sambandsanalys – Regression och Korrelation”. Lund: Matematikcentrum, Lunds universitet. Link 2012-05-09: http://www.maths.lth.se/matstat/kurser/fms032/sambandVL_10.pdf

MathWorks (2012). ”corrcoef – Correlation coefficients”. Product documentation. Link 2012-05-09: http://www.mathworks.se/help/techdoc/ref/corrcoef.html

Moriarty, Naafs, Pezon, Fonesca, Uandela, Potter, Batchelor, Reddy, Mekala. (2010). “Working paper 1 – WASHCost’s theory of change: reforms in the water sector and what they mean for the use of unit costs”. IRC – International Water and Sanitation Centre. Link 2012-06-07: http://www.washcost.info/page/753

Moriarty, Batchelor, Fonesca, Klutse, Naafs, Nyarko, Pezon, Potter, Reddy, Mekala. (2011). “Working paper 2 – Ladders for assessing and costing water service delivery”. IRC – International Water and Sanitation Centre. Link 2012-06-07: http://www.washcost.info/page/753

Nkrumah, E., Nyarko, K.B., Dwumfour-Asare, B., Oduro-Kwarteng, S. & Moriarty, P (2011). “Drivers of capital expenditure of rural piped water systems in Ghana: The Volta, Ashanti and Northern region”. Link 2012-06-07: http://rwsnforum.files.wordpress.com/2011/11/194-dwumfour-ghana-long-paper-docx.pdf

Nyarko, Dr. K.B. (2007). ”Drinking water sector in Ghana – Drivers for performance”. Netherlands: Taylor & Francis Group.

OECD. (2003). *Glossary of Statistical Terms*. Retrieved from OECD's Statistics portal. Link 2012-06-07: http://stats.oecd.org/glossary/detail.asp?ID=1164

OECD EAP Task Force Secretariat (2005) “ Rural cost functions for water supply and sanitation”. Link 2012-06-07: http://www.oecd.org/dataoecd/18/12/36228167.pdf

Stigler, S. (2008). "Fisher and the 5% level". *Chance* 21 (4): 12. Link 2012-06-05: http://www.springerlink.com/content/p546581236kw3g67/

The Presidency Republic of Ghana. (2011-01-04).“Ghana: The World's Fastest Growing Economy in 2011 “. Link 2011-10-02: http://www.presidency.gov.gh/press-centre/general-news/ghana-worlds-fastest-growing-economy-2011

Tsegai, D.W., Linz, T. & Kloos, J. (2009). “Economic analysis of water supply cost structure in the Middle Olifants sub-basin of South Africa”. Center for Development Research ZEF Bonn, Germany. Link 2012-06-07: http://www.zef.de/fileadmin/webfiles/downloads/zef_dp/zef_dp_129.pdf

Vännman, K. (2010). ”Matematisk Statistik”. Lund: Studentlitteratur AB

WaterAid. (2011) “Ghana – context”. Link 2010-10-02: http://www.wateraid.org/uk/what_we_do/where_we_work/ghana/default.asp

WHO – World Health Organization. (2010). “WHO / UNICEF Joint Monitoring Programme (JMP) for Water Supply and Sanitation”. Link 2011-10-02: http://www.wssinfo.org/data-estimates/table/

WRc – Water Research Centre. (1977). “TR61 – Cost information of water supply and sewage disposal”.

Yeboah, P. A. (2008). “Management of non-revenue water: A case study of the water supply in Accra, Ghana”. Longbourough University. Link 2011-10-02: http://www.switchurbanwater.eu/outputs/pdfs/W3-1_CACC_PHD_Management_of_Non-Revenue_Water_-_Case_study.pdf

APPENDIX 1

Data base with raw material from WASHCost, slightly modified by me:

NR= Northern region VR= Volta region AR= Ashanti region

CT= Concrete tank ST= Steel Tank PT= Poly Tank

NG= National Grid SS= Solar system

MB= Mechanized borehole GWCL= Bulk water through Ghana Water Company Limited

Region	dum_AR	dum_VR	dum_NR	Water system	# of Com.	Design Pop.	Year of constr.	Year numeric, start
NR	0	0	1	Buipe	1	10762	2006	9
NR	0	0	1	Yoggu	1	3098	2006	9
NR	0	0	1	Gulpi	1	3387	2006	9
NR	0	0	1	Kpandai	1	11441	2006	9
NR	0	0	1	Sawla	1	8793	2006	9
NR	0	0	1	Nanton	1	6512	2006	9
NR	0	0	1	Diare	1	11493	2006	9
NR	0	0	1	Bakamba	1	2780	2007	10
NR	0	0	1	Buya	1	3000	2007	10
NR	0	0	1	Katijeli	1	2550	2007	10
NR	0	0	1	Mankarig	1	2640	2007	10
NR	0	0	1	Langbinsi	1	2606	2007	10
NR	0	0	1	Busunu	1	2207	2007	10
VR	0	1	0	Tsiame	7	6396	2007	10
VR	0	1	0	3	3	1533	2006	9
VR	0	1	0	6	6	2122	2006	9
VR	0	1	0	14	14	7329	2006	9
VR	0	1	0	Borae No	1	5940	2006	9
VR	0	1	0	Nkonya	1	5156	2006	9
VR	0	1	0	Chenderi	1	4820	2006	9
VR	0	1	0	Ave	1	3965	2006	9
VR	0	1	0	Akpafu	1	4671	2006	9
VR	0	1	0	Kpando	1	9600	2005	8

# of Mech Boreholes	Pipeline (m)	Transmission pipe length	Distribution pipe length	% of Trans. on CapEx per	% of dist on CapEx per capita	# of Storage tanks	Type Storage tank	dum_CT	dum_ST	dum_PT	Tank size
1	16025	9635	6390	28,07828	18,62172	1	ST	0	1	0	150
1	7819	5032	2787	13,38606	7,41394	1	ST	0	1	0	50
1	2967	700	2267	3,869228	12,53077	1	ST	0	1	0	50
2	14885	8491	6394	22,87465	17,22535	1	ST	0	1	0	120
2	13230	7592	5638	23,64251	17,55749	1	ST	0	1	0	100
1	6204	3147	3057	14,05092	13,64908	1	ST	0	1	0	80
1	9312	6710	2602	21,54521	8,35479	1	ST	0	1	0	120
2	1978	890	1088	5,354398	6,545602	1	ST	0	1	0	100
1	2250	200	2050	1,822222	18,67778	1	ST	0	1	0	50
1	2073	603	1470	6,195803	15,1042	1	ST	0	1	0	50
2	2455	622	1833	3,420367	10,07963	1	ST	0	1	0	80
2	2540	580	1960	3,037008	10,26299	1	ST	0	1	0	100
2	2068	700	1368	5,212766	10,18723	1	ST	0	1	0	50
1	14700	4161	10539	7,784184	19,71582	1	CT	1	0	0	50
0	3744	0	3744	0	0	1	PT	0	0	1	12
0	4737	0	4737	0	0	1	PT	0	0	1	21
0	35040	0	35040	0	0	1	PT	0	0	1	33
2	7393	5506	1887	28,0774	9,622602	1	CT	1	0	0	80
1	6036	3088	2948	25,68217	24,51783	1	CT	1	0	0	50
1	4605	4200	405	37,48534	3,614658	1	CT	1	0	0	60
2	8106	6243	1863	33,27135	9,928645	1	CT	1	0	0	50
2	3370	1549	1821	12,91599	15,18401	1	CT	1	0	0	50
1	5159	248	3505	2,286384	32,31362	1	CT	1	0	0	100

Tanksize total volume	# of Standpost	Power source (NG, SS)	dum_SS	Tech.Option (MB, SW,GWCL	dum_gwcl	Cost Borehole site works	Cost Borehole	Cost pipeline	Cost storage tank	Cost standpost	Total cost (Incl. all extra
150	8	NG	0	MB	0	5546,924	34857,32	455360,9	226180,4	35997,98	960748,2
50	6	NG	0	MB	0	5453,105	26266,36	133706,4	153979,6	10040,34	626750,6
50	15	NG	0	MB	0	5453,105	29394,27	69379,29	156756,6	10040,34	408384,3
120	14	NG	0	MB	0	8624,545	47597,87	368939,3	224850,1	39667,66	889823,3
100	8	NG	0	MB	0	11370,35	48220,8	368295,4	173986	43093,53	862878,6
80	15	NG	0	MB	0	7468,664	31501,28	172792,6	161481,8	23376,45	609585,6
120	15	NG	0	MB	0	7468,664	29044,61	233938,8	202317,5	49059,88	767696
100	6	SS	1	MB	0	0	50997,95	87743,27	191981,6	10038,27	704897
50	6	NG	0	MB	0	0	27602,74	82189,87	150872,2	10038,27	386050
50	5	NG	0	MB	0	0	27602,74	88399,53	150872,2	8365,228	400387,9
80	7	SS	1	MB	0	0	84194,83	90107,22	158654,8	11711,32	638767,3
100	6	SS	1	MB	0	0	47250,36	90112,64	191981,6	10038,27	650386,6
50	5	SS	1	MB	0	0	51679,33	80576,71	150872,2	8365,228	494788,1
50	20	NG	0	MB	0	0	9683,478	140267,5	138766,6	36162,93	502254,9
12	5	NG	0	GWCL	1	0	0	33392,16	9471,764	7215,486	59339,81
21	10	NG	0	GWCL	1	0	0	58441,92	43166,28	14601,4	125697,6
33	26	NG	0	GWCL	1	0	0	520837	154981,2	49700,25	778764,3
80	12	NG	0	MB	0	0	25533,45	167919,3	65865,08	18278,84	432703,4
50	10	NG	0	MB	0	0	11645,93	93314,4	36348,64	12176,14	172157,6
60	11	NG	0	MB	0	0	13325,26	86180,1	59017,6	24951,78	195962,2
50	10	NG	0	MB	0	0	27748,96	135496,3	54388,51	19780,5	285523,8
50	8	NG	0	MB	0	0	12984,07	79563,03	49947,27	15112,58	254990,8
100	16	NG	0	MB	0	0	9378,288	96932,09	87171,77	44427,42	636849,8

Region	dum_AR	dum_VR	dum_NR	Water system	# of Com.	Design Pop.	Year of constr.	Year numeric, start	# of Mech Boreholes	Pipeline (m)	Transmission pipe length
AR	1	0	0	Boanim	1	4943	2006	9	2	6177	3063
VR	0	1	0	Fesi	1	6458	2005	8	2	4488	3539
VR	0	1	0	VE GHAD	4	6877	2004	7	1	7443	4972
VR	0	1	0	Wusuta	2	15942	2003	6	3	6938	4671
VR	0	1	0	Osramani	1	6665	2001	4	1	3840	3057
VR	0	1	0	Nolopi	2	7200	1999	2	1	2976	574
VR	0	1	0	Amanta	1	3679	1998	1	1	1588	649
AR	1	0	0	Pepease	1	3100	2010	13	2	2674	431
AR	1	0	0	No 3	1	2165	2010	13	2	5217	2380
AR	1	0	0	Kofiase	1	8378	2010	13	2	9909	2520
AR	1	0	0	wu	1	4231	2010	13	2	2513	256
AR	1	0	0	Naama	1	1683	2010	13	2	1323	456
AR	1	0	0	Apaah	1	3066	2010	13	2	4954	1078
AR	1	0	0	Yonso	1	3182	2010	13	2	5816	2060
AR	1	0	0	pobi	3	8943	2009	12	3	17574	14496
AR	1	0	0	Dampong	1	4312	2009	12	3	9410	7190
AR	1	0	0	Fomena	1	8305	2008	11	3	8450	6000
AR	1	0	0	Juabeng	1	19477	2006	9	4	17580	11070
AR	1	0	0	Kwaso	1	6950	2006	9	2	8330	6262
AR	1	0	0	Onwe	1	6950	2006	9	4	11028	7339
AR	1	0	0	Bompata	1	4546	2006	9	2	6345	4745
AR	1	0	0	Atwedie	1	5008	2006	9	2	4066	2231
AR	1	0	0	Boanim	1	4943	2006	9	2	6177	3063

Distribution pipe length	% of Trans. on CapEx per	% of dist on CapEx per capita	# of Storage tanks	Type Storage tank	dum_CT	dum_ST	dum_PT	Tank size	Tanksize total volume	# of Standpost	Power source (NG, SS)
3114	14,62822	14,87178	2	CT	1	0	0	100100	200	11	NG
949	29,9648	8,035205	1	CT	1	0	0	80	80	11	NG
2471	32,6657	16,2343	2	CT	1	0	0	60, 30	90	18	NG
2267	19,99549	9,704511	3	CT	1	0	0	100	180	29	NG
783	24,67891	6,321094	1	CT	1	0	0	80	80	11	NG
2402	4,860484	20,33952	1	CT	1	0	0	40	40	14	NG
939	12,34244	17,85756	1	CT	1	0	0	40	40	7	NG
2243	4,706507	24,49349	1	CT	1	0	0	50	50	6	NG
2837	13,82289	16,47711	1	CT	1	0	0	40	40	6	SS
7389	14,41962	42,28038	1	CT	1	0	0	120	120	16	NG
2257	3,127417	27,57258	1	CT	1	0	0	80	80	7	NG
867	4,825397	9,174603	1	CT	1	0	0	40	40	4	NG
3876	5,853492	21,04651	1	CT	1	0	0	50	50	8	NG
3756	11,47593	20,92407	1	CT	1	0	0	50	50	6	NG
3078	45,4495	9,650495	1	CT	1	0	0	180	180	15	NG
2220	37,51637	11,58363	1	CT	1	0	0	100	100	7	NG
2450	32,52071	13,27929	1	CT	1	0	0	100	100	14	NG
6510	20,71689	12,18311	2	CT, ST	1	1	0	150, 250	400	15	NG
2068	23,07844	7,621561	1	CT	1	0	0	100	100	15	NG
3689	20,03118	10,06882	1	CT	1	0	0	100	100	16	NG
1600	24,52892	8,27108	1	CT	1	0	0	70	70	8	NG
1835	18,87516	15,52484	1	CT	1	0	0	70	70	8	NG
3114	14,62822	14,87178	2	CT	1	0	0	100100	200	11	NG

dum_SS	Tech.Option (MB, SW,GWCL	dum_gwcl	Cost Borehole site works	Cost Borehole	Cost pipeline	Cost storage tank	Cost standpost	Total cost (Incl. all extra
0	MB	0	10387,58	51780,3	187046,3	60284,42	14297,94	601547,4
0	MB	0	0	23013,14	90248,91	54025,5	21410,46	210911,5
0	MB	0	0	17490	162255,4	71756,8	39657,19	316780,5
0	MB	0	0	51707,74	154210	88349,25	45495,27	479171,4
0	MB	0	0	46508,28	125715,1	93628,08	30033,91	374905,4
0	MB	0	0	31480,78	23534,61	15917,88	21311,63	107585,2
0	MB	0	0	9148,611	14873,15	15829,42	8608,188	58109,2
0	MB	0	49137,64	68349,63	161934,3	115397,2	14938,74	523402,5
1	MB	0	19439,95	308656,9	278744,9	106476,7	14938,74	889033,2
0	MB	0	6607,335	40445,04	517552,8	148222,1	30662,14	881942,1
0	MB	0	6633,019	46785,9	154982,2	120651,3	13414,69	473696
0	MB	0	27787,41	52128,17	86977,16	98464,7	17725,51	604741,7
0	MB	0	18549,76	148366	150689,2	90492,98	23479,22	529372,6
0	MB	0	21538,05	151744,9	194813,9	99532,38	17628,68	569284,9
0	MB	0	9430,698	38840,75	439689,3	119777,9	16757,83	749160
0	MB	0	6574,147	27531,78	271889,8	87122,54	8795,608	505751,3
0	MB	0	17237,12	30028,44	244965,6	64438,07	25303,57	483416,4
0	MB	0	0	21967,71	163105,7	966,7432	16481,22	429406
0	MB	0	0	17860,88	162595,3	38614,77	13184,99	496179,2
0	MB	0	0	17675,87	178451,4	38590,84	14283,73	525888
0	MB	0	8289,801	31658,11	123856,4	61582,64	20177,73	344688
0	MB	0	71007,46	19493,6	162938,8	63919,66	19772,43	441274,9
0	MB	0	10387,58	51780,3	187046,3	60284,42	14297,94	601547,4

APPENDIX 2

GDP Deflator from WASHCost:

			1980	1981	1982	1983	1984	1985	1986	1987
		year	1980	1981	1982	1983	1984	1985	1986	1987
Ghana	GHA	Inflation, GDP deflator (annual %)	50,005	116,504	22,296	122,875	39,665	10,305	24,565	39,815
Ghana	GHA	Index	100	150,005	324,668252	397,1768	885,2079	1236,326	1363,729	1698,729
Ghana GDP Deflator multiplier to convert past costs to current (2011) prices			*2464,28*	*1642,80*	*758,78*	*620,45*	*278,38*	*199,32*	*180,70*	*145,07*
Exchange rate	GH/US$	1,6								

1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998
31,359	25,224	37,259	18,031	10,056	24,96	24,87	59,462	44,357	24,838	19,215
2375,078	3119,879	3906,837	5362,485	6329,395	6965,879	8704,562	10869,39	17332,54	25020,74	31235,39
103,76	*78,99*	*63,08*	*45,95*	*38,93*	*35,38*	*28,31*	*22,67*	*14,22*	*9,85*	*7,89*

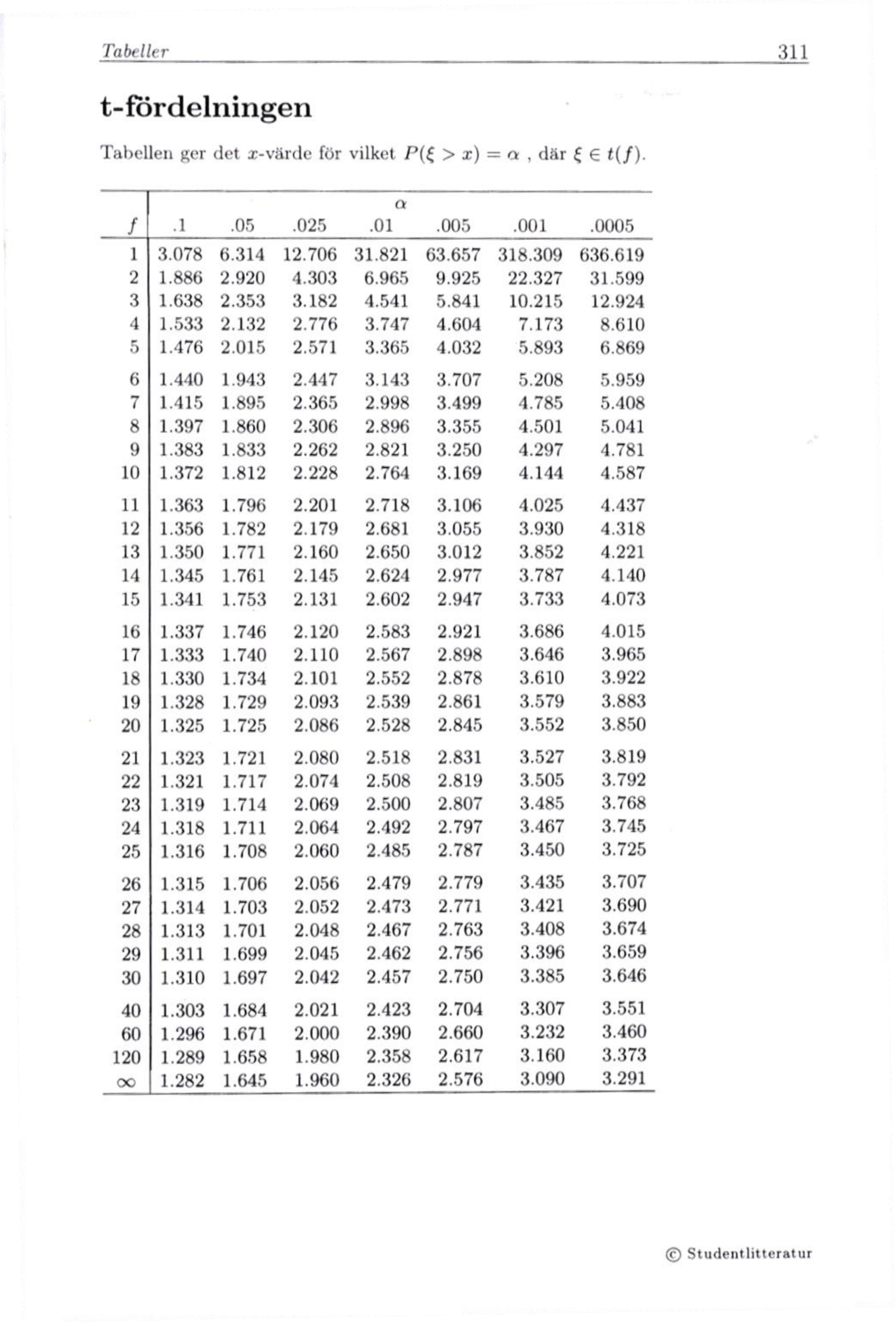
2009	2010	2011
19,251	10,709	8,55
186657,4	222590,8	246428
1,32	*1,11*	*1,00*

APPENDIX 3

Table for t-test. Source: Vännman, K. (2010). ”Matematisk Statistik”. Lund: Studentlitteratur AB

Translation: **t-distribution,** The table gives the x-value which fulfills $P(\xi > x) = \alpha$, with $\xi \in t(f)$.

f = degrees of freedom. $\alpha =$ level of significance.



Tabeller 311

t-fördelningen

Tabellen ger det x-värde för vilket $P(\xi > x) = \alpha$, där $\xi \in t(f)$.

	α						
f	.1	.05	.025	.01	.005	.001	.0005
1	3.078	6.314	12.706	31.821	63.657	318.309	636.619
2	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	1.296	1.671	2.000	2.390	2.660	3.232	3.460
120	1.289	1.658	1.980	2.358	2.617	3.160	3.373
∞	1.282	1.645	1.960	2.326	2.576	3.090	3.291

APPENDIX 4

MATLAB files, created to be used in my thesis. Exemplified by matlab file for stand posts.

```
%% Deriving a cost function for standpost/standpipes in a water system
% show min, max, mean, median and std deviation
x=spcost;    % CapEx for stand posts
y=numcom;    % Number of communities
z=pop;       % Design population
standfact=[min(x) max(x) mean(x) median(x) std(x);
    min(y) max(y) mean(y) median(y) std(y);
    min(z) max(z) mean(z) median(z) std(z)];
exp(3.522)

%% Histfit and variation plots
figure()
subplot(211)
histfit(spcost)
subplot(212)
histfit(log(spcost))

figure()
subplot(221)
normplot(spcost)
subplot(222)
normplot(log(spcost))
subplot(223)
wgumbplot(spcost)
subplot(224)
wblplot(spcost) % Does not really fit well to any of them, maybe gumbel is
the best one.

%% Corr.analysis for standposts #1 - All independent variables
standcorr1=[spcost dum_regions numcom pop standpost];
[stand1,level]=corrcoef(standcorr1) % Region doesn't matter, maybe numcom
and def pop and standpost

%% Corr.analysis for standposts #2 - with numcom pop and standpost
standcorr2=[spcost numcom pop standpost];
[stand2,level]=corrcoef(standcorr2)

%% Corr.analysis for standposts #3 - All independent variables log-log
standcorr3=[log(spcost) log(dum_logregions) log(numcom) log(pop)
log(standpost)];
[stand3,level]=corrcoef(standcorr3)

%% Regress normal #1 - with numcom pop and standpost
standreg1=[ones(size(standpost)) numcom pop standpost];
alpha=0.10  % significance level
[b,bint,r,rint,stats]=regress(spcost,standreg1,alpha);

rmse=sqrt(sum(r.^2)/length(r));      % Root mean square error for real value
beta=[b bint]                        % Shows beta-koefficient and the
intverval
R2oRMSE=[stats(:,1) sqrt(stats(:,4)) rmse] % Shows R2 sqrt(variance) rmse
```

```
% It does not matter which variables I use, the constant is not
significant!

%% Regress log-log #1 - with numcom pop and standpost
standlog1=[ones(size(pop)) log(numcom) log(pop)];
alpha=0.05  % significance level
[b,bint,r,rint,stats]=regress(log(spcost),standlog1,alpha);

y0=standlog1*b; % Giving the predicted data

% this looks like the best one of all the comb I've tried
% Best both in R2=0.64 sigma=0.1174 RMSE=6991 GHs

x=standlog1;
rmse=sqrt(sum(r.^2)/length(r));     % Root mean square error for real value
R=spcost-exp(y0);                     % Calc residuals in real value
rmsereal=sqrt(sum(R.^2)/length(R));  % Root mean square error for real
value

beta=[b bint]                    % Shows beta-coefficient and the intverval
R2oRMSE=[length(x) 1-alpha stats(:,1) sqrt(stats(:,1)) rmsereal]; % Shows
n, R2, R and rmse


x=[1:44];

figure()
plot(x,exp(y0),'--g.',x,spcost,'--b.','MarkerSize',15) % Plots both the
obeserved and predicted data
ylabel('COST (GHs,2011)')
xlabel('Data point')
legend('Predicted cost','Observed cost','location','best')
figure()
plot(exp(y0),spcost,'*') % Plots observed against predicted
ylabel('Observed cost')
xlabel('Predicted cost')
figure()
subplot(121)
plot(x,r,'*',x,zeros(44,1)) % Plots residuals
axis([1 44 -1 1])
xlabel('Data point')
subplot(122)
normplot(r) % Shows how well the errors fits into normal variation
title('Normal Probability Plot, Residuals')
```