The normalisation of resource efficiency measures in healthcare facilities: the case of energy and water

by

Abimelek Shikongo Amunjela

Thesis presented in fulfilment of the requirements for the degree of Master of Engineering (Industrial Engineering) in the Faculty of Engineering at Stellenbosch University

Supervisor: Mrs. I. H. de Kock

Co-supervisor: Prof. A. C. Brent

April 2019

Declaration

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third-party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Date: April 2019

Abstract

Energy and water consumption performance comparisons are used at a healthcare management and policy formulation level when formulating benchmarks and energy improvement targets. Normalising for the differences between hospitals is a key part of hospital consumption performance comparisons. It ensures that the measures used in these comparisons are commensurate, thereby increasing the reliability and robustness of the comparisons. Currently, the measures used in these comparisons are only normalised for the size of a hospital and are not adjusted to account for the inter-hospital differences in the level of medical service provision.

This study investigated the feasibility of including normalisation factors that are representative of the level of medical service provision in the normalisation model used to compare the energy and water performance of hospitals. The complexity and level of specialisation of the composition of a hospital's diagnostic caseload and the output of a hospital were used to quantify the level of medical service provision at a hospital. Measures were also formulated to quantify the size of a hospital.

Statistically-based modelling methods were used to conduct an exhaustive analysis of the relationships between combinations of the normalising factors in the analysis and the energy and water consumption of the respective hospitals. Multiple linear regression (MLR) models were developed for all the possible combinations of normalising factors. These models were used to assess and rank the explanatory power provided by each combination of normalising factors in explaining the variations in the energy and water consumption of hospitals.

Based on these MLR analyses and the rankings of the explanatory power provided by the respective models, it was concluded that the level of medical service provision of a hospital (as represented by its output, complexity and level of specialisation), does not significantly contribute to increasing the reliability or robustness of the current normalisation model. Furthermore, accounting for the level of medical service

provision in the normalisation model would complicate the model without providing any significant additional explanatory power or increasing the objectivity of hospital consumption performance comparisons.

Uittreksel

Energie- en waterverbruik-prestasievergelykings word gebruik op 'n gesondheidsorgbestuurs- en beleidsformuleringsvlak wanneer maatstawwe en doelwitte vir energieverbetering geformuleer word. Normalisering vir die verskille tussen hospitale is 'n belangrike aspek wanneer hospitaalverbruikverrigting vergelyk word. Normalisering verseker dat die maatstawwe wat in hierdie vergelykings gebruik word regverdig is, en verhoog dus die betroubaarheid en robuustheid van die vergelykings. Tans word daar in die maatstawwe wat in hierdie vergelykings gebruik word slegs genormaliseer vir die grootte van 'n hospitaal, en word dit nie aangepas om byvoorbeeld die verskille tussen hospitale in die vlak van mediese diensverskaffing nie in berekening te bring nie.

Hierdie studie het ondersoek ingestel na die uitvoerbaarheid van die insluiting van normaliseringfaktore wat verteenwoordigend is van die vlak van mediese diensverskaffing in die normalisasiemodel wat gebruik word om die energie- en waterprestasie van hospitale te vergelyk. Die kompleksiteit en vlak van spesialisering van die samestelling van 'n hospitaal se diagnostiese gevallelading en die uitset van 'n hospitaal is gebruik om die vlak van mediese diensverskaffing by 'n hospitaal te kwantifiseer. Maatstawwe is ook geformuleer om die grootte van 'n hospitaal te kwantifiseer.

Statistiek-gebaseerde modelleringsmetodes is gebruik om 'n volledige ontleding van die verhoudings tussen kombinasies van die normaliseringfaktore in die analise en die energie- en waterverbruik van die onderskeie hospitale te doen. Meervoudige lineêre regressie (MLR) modelle is ontwikkel vir al die moontlike kombinasies van die normaliseringfaktore. Hierdie modelle is gebruik om die verklarende krag wat deur elke kombinasie van normaliseringfaktore verskaf word te assesseer en te rangskik om die variasies in die energie- en waterverbruik van hospitale te verduidelik.

Op grond van hierdie MLR-ontledings en die ranglys van die verklarende krag wat deur die onderskeie modelle uitgewys word, is daar tot die gevolgtrekking gekom dat die vlak van mediese dienslewering van 'n hospitaal (soos mee gebring deur die lewering, kompleksiteit en spesialiseringsvlak) nie beduidend bydra tot die verbetering van die betroubaarheid of robuustheid van die huidige normaliseringsmodel nie. Om die vlak van mediese diensverskaffing in die normaliseringsmodel in ag te neem sou die model onnodig kompliseer sonder om enige beduidende addisionele verklarende krag te gee, of om die objektiwiteit van die vergelykings van hospitaal-verbruikverrigting te verhoog.

Acknowledgements

With the completion of this project, I would like to acknowledge and express my sincere gratitude to the following people and organisations for their contribution and assistance throughout this study:

- My supervisor: Mrs. Imke de Kock and co-supervisor: Prof. Alan Brent, for their patience and guidance throughout my master's studies.
- GlaxoSmithKline plc, for providing the financial assistance required to pursue this study.
- The Health Systems Engineering and Innovation research group for believing in my potential and affording me both the opportunity and means to conduct this study.
- My family and friends for their love, encouragement, understanding and unwavering support.

Contents

List of tables

List of figures

Nomenclature

ACRONYMS

ABH	Ladismith (Alan Blyth) Hospital
AEC	Total annual electricity consumption of a hospital
AIC	Akaike information criterion
AWC	Total annual water consumption of a hospital.
BE	Building equipment
BED	Number of beds metric
BWH	Beaufort West Hospital
CLD	Caledon Hospital
CLH	Clanwilliam Hospital
CMPX	Case mix complexity metric
CRS	Ceres Hospital
CWHC	Cold water for human consumption
DEA	Data envelope analysis
DHW	Domestic hot water
EIG	Expected Information Gain
GDP	Gross domestic product
HER	Hermanus Hospital
HVAC	Heating ventilation and air conditioning
ICD-10	10th Revision of the International Classification of Diseases Master
MIT	Industry Table
ICT	Information communication technology
IQR	Interquartile range
JRC	Joint Research Centre
KNY	Knysna Hospital
LAP	LAPA Munnik Hospital
LBH	Laingsburg Hospital

MATLAB	Matrix Laboratory software
MLR	multiple Linear Regression
MON	Montagu Hospital
MTE	Medical technical equipment
MVLR	Multivariate linear regression
MWRA	Massachusetts Water Resources Authority
N/A	Not Applicable
OECD	Organisation for Economic Co-operation and Development
PC	Principal Component
PCA	Principal Component Analysis
PDE	Patient Day Equivalent metric
PRH	Prince Albert Hospital
RIV	Riversdale Hospital
RKH	Radie Kotze Hospital
ROB	Robertson Hospital
RSADoH	Republic of South Africa Department of Health
SANDoH	South African National Department of Health
SMTE	Small medical technical equipment
SPEC	Level of specialisation of case mix metric
SSE	Residual sum of squares
STB	Stellenbosch Hospital
SWE	Swellendam Hospital
TFA	Total floor area metric
UDH	Uniondale Hospital
VDoH	Victorian Government Department of Health
VIF	Variance inflation factor
VRE	Vredendal Hospital
WCDEADP	Western Cape Department of Environmental Affairs and Development Planning
WCDoH	Western Cape Department of Health
WOC	Worcester Hospital

SYMBOLS

α	Significance level
∞	Infinity
C	Total number of cases in the analysis
C_i	Total number of cases treated by each hospital in the analysis
C_j	Total number of cases in the analysis in to each diagnostic category
H_0	Null hypothesis
H_1	Alternative hypothesis
kL	Kilolitres
kWh	Kilowatt-hours
L	Litres
m	Meters
m^2	Square metres
m^3	Cubic metres
N	Number of hospitals in the analysis
P	The proportion of overall cases of diagnostic type j treated at each hospital
P_i	The proportion of total cases in the analysis treated by each hospital
PC_i	*i-th* principal component
Q	The proportion of each hospital's caseload that is of diagnostic type *j*
Q_j	The proportion of total cases in the analysis that are of diagnostic type j
r	Correlation coefficient
R^2	Coefficient of determination
R_a^2	The adjusted coefficient of determination
s_j	The standard deviation of the set of variables associated with the hospitals.

Chapter 1 Introduction

This thesis investigated the feasibility of normalising for both the size and function[1] of a hospital when benchmarking the energy and water performance of hospitals. Furthermore, the thesis investigated whether this would increase the objectivity of hospital consumption performance comparisons. In this chapter, the study is introduced by describing the context that created the need for the research. Furthermore, the research objectives and methodology are stated, and a chapter outline for the thesis is provided.

1.1 Study background

The drive for sustainability and efficiency when considering energy and water usage in the building sector has resulted in the increased importance of the effective management of energy and water resources. Hospitals are complex resource-intensive facilities, resulting in increased complexities when managing their energy and water usage. In addition, differences in the characteristics of hospital buildings and variations in the medical services provided by different hospitals have resulted in significant differences in energy and water usage between hospitals (MWRA n.d.; Szklo et al. 2004).

Buildings are one of the largest consumers of energy, especially in developing countries; for example, they account for 50 percent of the energy used in Brazil, and 42 percent of the energy used in Botswana (Abu Bakar et al. 2015). In comparison to commercial offices, hospitals are twice as energy intensive, and six times as water intensive (Rajagopalan & Elkadi 2014). Furthermore, hospital buildings must adhere to strict hygiene and air quality requirements and ensure patient comfort. Meeting these requirements requires the continuous operation of hospital building services,

[1]The function of a hospital is represented by its level of medical service provision.

which, when coupled with the specialised services that are provided by hospitals, has led to high energy and water consumption.

Energy and water-related expenses are major expenditures for hospitals; third after medicine and staff wages (Hu et al. 2004). The resource-intensive nature of hospitals; the rising global energy demand, which is projected to be 45 percent higher by 2025 than it was in 2002 (Abu Bakar et al. 2015); and South Africa's limited freshwater reserves (Thopil & Pouris 2016), have led to increasing economic pressure on the health sector to be more sustainable and efficient with its energy and water consumption.

Energy is used in both medical operations and by support services to ensure the smooth and seamless operation of all medical activities in a hospital and to maintain the health and comfort of both patients and staff. Hospital buildings typically comprise of patient wards, operating theatres, an X-ray department, administration offices, and an array of support, or auxiliary services, such as a boiler house, workshops, laundry rooms, a kitchen, and a dining hall (Gupta et al. 2007). In the Western Cape province, heating, ventilation and air conditioning (HVAC) and water heating account for 50 percent of the energy consumption in hospitals (WCDEADP 2008). Furthermore, lighting and medical equipment (general and specialised) are also significant consumers of energy, as shown in Figure 1.1.

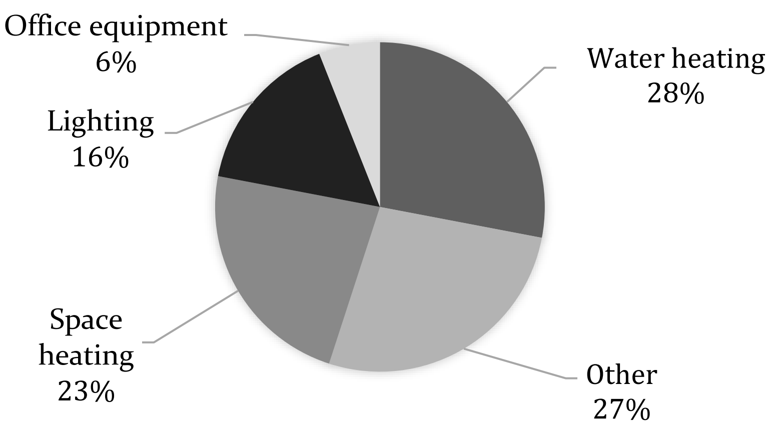


Figure 1.1: Breakdown of the energy load of a typical hospital (adapted from WCDEADP 2008)

In hospitals, water is important for maintaining the health and hygiene of both the patients and hospital staff. Water use can be classified into five categories: ablution, ingestion, irrigation, process, and sanitation. Figure 1.2 shows a breakdown of the water load of a typical hospital. Consumption by each category varies depending on the size and function of the hospitals. However, ablution (water use in basins, sinks and showers) and process (water used for sterilisation, cleaning, heating and cooling) are responsible for 60 to 80 percent of water usage (VDoH 2009). The water usage by rehabilitation swimming pools and other support services, such as laundries, is also significant; however, the presence of these services is generally limited to large hospitals.

The first step in improving inefficient energy and water usage is to understand how and why the resource consumption patterns of hospital facilities vary. By studying the flows of energy and water in hospitals and performing system and subsystem-level comparisons, inefficiencies can be identified (Abu Bakar et al. 2015). These comparisons may be skewed by inter-hospital variations in confounding factors that affect the energy and water utilisation in hospitals, thus rendering these comparisons inaccurate and unreliable.

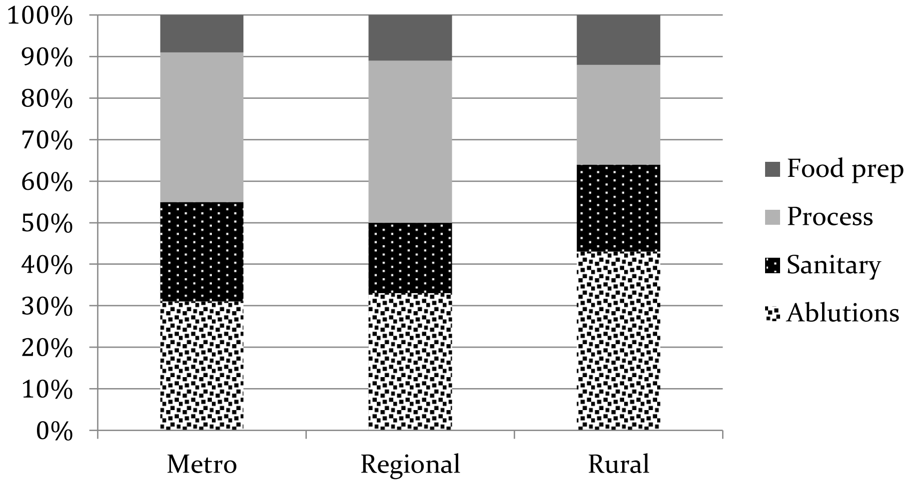


Figure 1.2: Breakdown of the water load of a typical hospital (adapted from VDoH 2009)

The accuracy and reliability of the comparisons can be improved by normalising the energy and water consumption data used in these comparisons, for the effects exerted by the characteristics of the hospitals and the medical services they provide. Since these comparisons are used at a health system-level in a policy-making environment as part of decision-support tools, obtaining more objective and robust comparisons will aid decision-makers in developing policies that are based on data that reflects the actual performance of the respective hospitals.

This highlights the need for the development of an approach for the normalisation of energy and water consumption data against the effects of these factors. In this study, this was achieved through identifying confounding factors that result in inter-hospital variations in energy and water consumption; devising or obtaining a method of determining objective measures of these confounding factors in hospitals when analysing their performance; and then accounting for the effect of these confounding factors in the energy and water consumption comparisons of hospitals.

1.2 Problem statement

Currently the benchmarks used for energy and water consumption performance analysis are predominantly standardised for hospital size. Consumption is quoted with respect to the number of beds (i.e. kilowatt hours per bed per day [kWh/bed/day]) or with respect to floor area (i.e. litres per square metres per day [L/m^2/day]). These approaches do not account for the effect of a hospital's function[2] on its energy and water consumption performance. Thus, there is a need to identify and define measures that are representative of the function of a hospital, and to investigate whether accounting for these measures in the formulation of benchmarks for energy and water utilisation in hospitals will result in more robust[3] comparisons.

1.3 Research objectives

The primary aim of this study was to test the feasibility of using a normalisation approach that accounts for both the size and function of a hospital when assessing energy and water consumption performance. The research objectives were:

1. To establish a relationship between the inter-hospital variations in the characteristics and function of a hospital, and the inter-hospital variations in their energy and water consumption performance.

[2] The function of a hospital is represented by its level of medical service provision.

[3] In this study robustness is defined as the ability to account for a more statistically significant share of the variation in the energy and water consumption of hospitals.

2. To identify and formulate methods and measures for quantitatively capturing the inter-hospital variations in the function and characteristics of a hospital.
3. To evaluate the feasibility of using combinations of these measures as normalising factors when evaluating energy and water consumption performance in hospitals at a health system-level.

1.4 Delineations

The following delineations outline the scope of the problem investigated by this study:

1. The study was limited to assessing the explanatory power of the normalisation factors with respect to the consumption of two types of resources, namely: electricity and water. No other forms of energy, such as liquid fuels, were considered.
2. The normalisation factors investigated were limited to only representing the characteristics of a hospital and its function as defined by the factors identified in the study.
3. The normalising factors representing the function of a hospital were limited to the activities that are directly involved in patient care and are outside the direct control of hospital management.
4. The explanatory power provided by the respective normalisation factors was assessed with respect to the quantitative significance of each factor in a statistical model.
5. The study was conducted in a Western Cape health system context. The conclusions formed in this study are based on a data analysis that was predominantly focused on district hospitals in the Western Cape. Thus, the generalisation of the study's findings and conclusions to other regions is not claimed nor implied.

1.5 Research approach

This study applied a quantitative research approach. Quantitative research is deductive in nature and it focuses on the empirical testing of theory. Thus, the first phase of the six-phase research approach, as shown in Figure 1.3, defined and contextualised the real-world problem based on information obtained from the analyses of literature. This consisted of identifying and summarising the attributes and factors that affect energy and water consumption in hospitals; and understanding the formulation of normalised benchmarks.

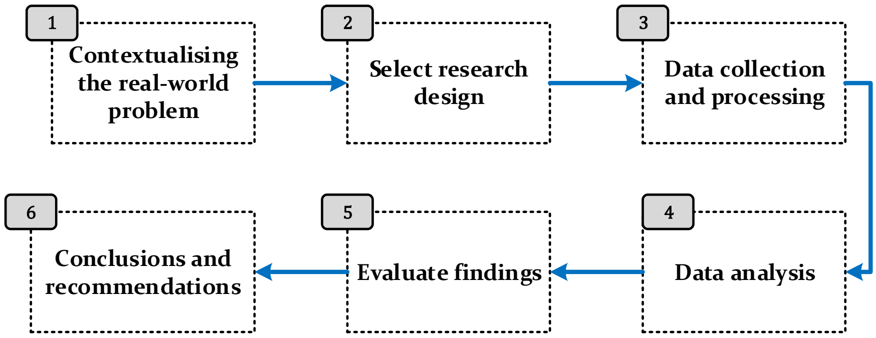


Figure 1.3: Overview of the research approach

In phase 2, an appropriate research design was selected or devised for solving the research problem. This involved the definition of concepts around which the research was conducted, and the definition of the measurements and indicators used to quantify these concepts. In the third phase, data was collected and processed in accordance with the research instruments specified in the research design. In the fourth phase, data analysis techniques were applied to evaluate the relationships between the variables in the collected data. In phases 5 and 6, the findings from the analyses were evaluated with respect to the research question and conclusions were drawn on the implications of the findings on the real-world problem that was investigated.

1.6 Document outline

Figure 1.4 provides an outline of the thesis layout and a description of the key aspects of the respective chapters. The chapters are structured such that they chronologically address the objectives of the research study as discussed in Section 1.3.

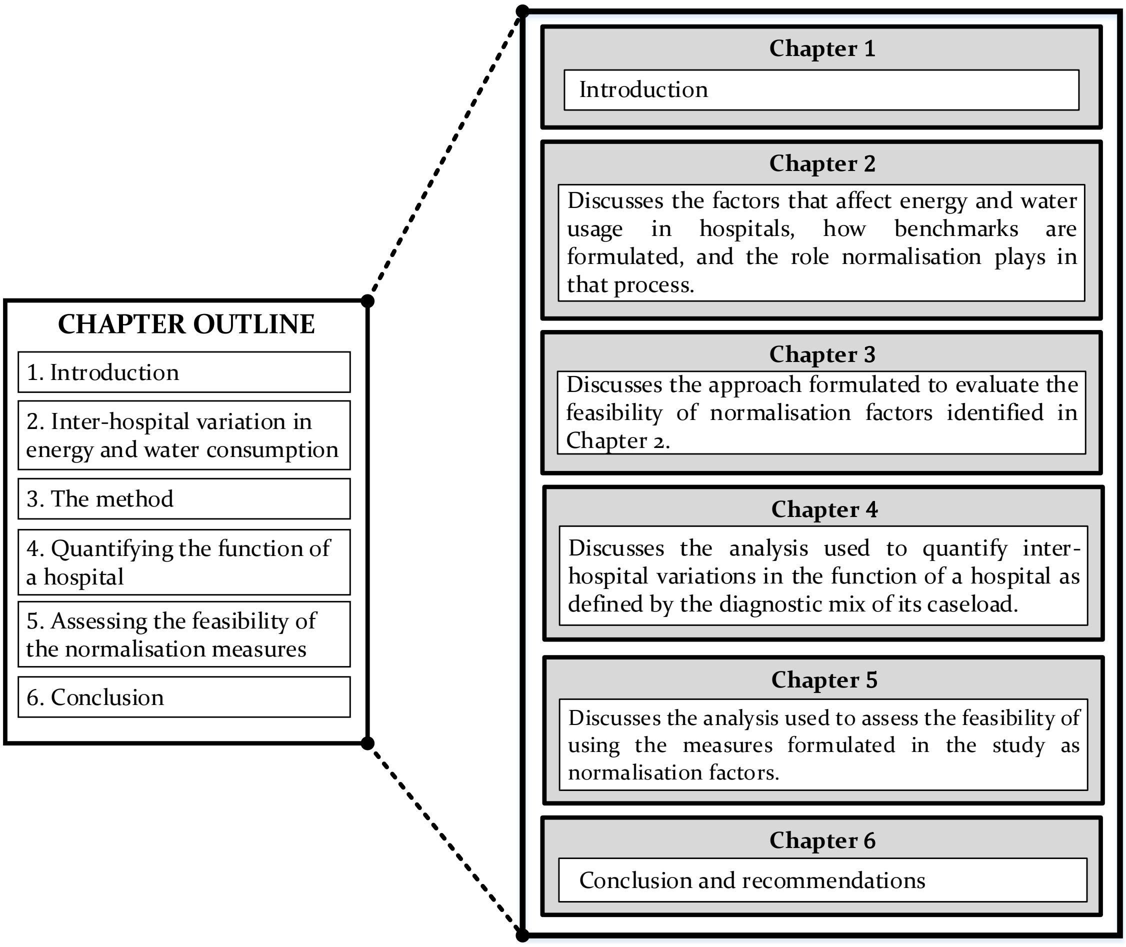


Figure 1.4: Thesis outline

Chapter 2 Inter-hospital variations in energy and water consumption

This chapter studies the energy and water flow within hospitals. It discusses the results of a literature analysis that identified the factors affecting energy and water consumption in hospitals and studied the relationship between these factors and resource consumption. Figure 2.1 outlines where this chapter fits into the research process. This chapter presents and builds on work that was published in a conference paper titled '*Developing normalised metrics for comparing the energy use of hospitals*' at the International Association for Management of Technology 2017 in Vienna Austria, see Amunjela et al. (2017).

The first phase of the research approach builds the theoretical foundation on which the study is based. It warranted a review of literature for the identification of factors that affect energy and water consumption at hospitals and the characteristics that cause variations in the performance of different facilities. This literature review aided in outlining the theoretical foundation of the chosen indicators and the definitions of their metrics by providing an understanding of the different aspects of the phenomenon being studied.

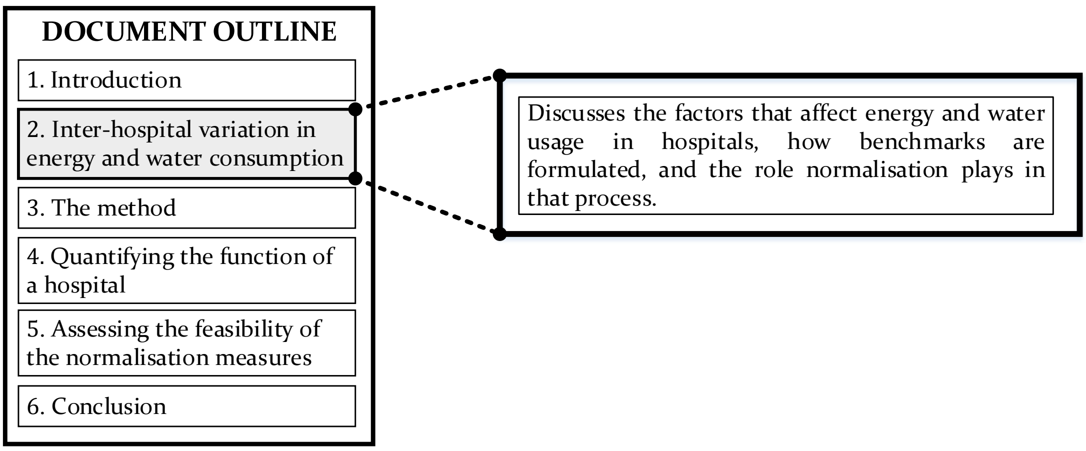


Figure 2.1: Thesis document outline: Chapter 2 contextualised

2.1 Factors that affect the energy use of hospitals

Hospitals have two main types of energy sources: electrical energy, used to power most hospital systems, and thermal energy which is used for heating (Szklo et al. 2004). Electrical energy is supplied via the grid and thermal energy is provided by combusting fossil fuels such as oil, gas, and coal. Various research inquiries have studied the utilisation of energy by non-domestic buildings, such as office buildings, factories and hospitals, and how it relates to benchmarking and quantifying the energy performance of buildings.

An analysis of both published and grey literature on the energy consumption in hospitals was conducted to gain an understanding of the factors that affect energy consumption in hospitals. This analysis and its results are summarised in Appendix A.1.

The study by Ma et al. (2017) focused on public buildings and studied the energy consumption of 119 buildings in China (office buildings, schools, and hospitals). From their analysis, factors that affect the energy consumption of these buildings were classified into two categories: internal factors, and external factors. Internal factors refer to the building's characteristics and condition such as its year of construction, its orientation and the structure of its envelope. External factors are defined as the events and activities that influence the indoor thermal environment of a building such as weather factors, heat gains from occupants, equipment and lighting.

Rajagopalan & Elkadi (2014), in their study of the energy performance of medium-sized healthcare facilities in Australia, defined three categories of factors affecting consumption at hospitals: physical characteristics, occupancy characteristics, and operational characteristics. Physical characteristics provide an understanding of how the architectural design and form of the hospital building, construction material, and certain building systems affect the hospital's energy consumption. The operational and occupancy categories account for the effect of the behaviour of the hospital's occupants on its energy load profile.

Deru et al. (2011) studied the most common types of commercial buildings in the USA and developed energy reference building models for use in the study of building energy efficiency. In their simulation, Deru et al. (2011) divided the factors that are influential to a building's energy behaviour into 4 groups: ‘building programme’, ‘form’, ‘fabric’, and ‘equipment’. The parameters that fall into the building programme category model the effects of the building’s activities, end uses, schedule, location and occupancy on its energy load profile. The form parameters account for the building's geometry and its effect on consumption. The parameters in the fabric category account for the type of construction material used to build the building and the thermal characteristics of these materials. Finally, equipment accounts for the parameters that govern the consumption of some of the major equipment involved in the building systems.

From a normalisation point of view, the interest lies in those factors that affect energy consumption in hospitals but are independent of the efficiency and administration of the hospital. Singer et al. (2009) stated that because of the nature of the service provided by hospitals the factors that affect its energy consumption can be grouped into two general categories. The first general category accounts for the effect of the hospital's characteristics, its contextual setting, and the configurations of its different building systems. The second general category accounts for the effect of the medical services provided at the hospital.

Singer et al. (2009) also stated that the energy consumption of a hospital scales with the extent and capacity of its medical service provision; the presence and extent of certain clinical specialities and energy-intensive medical services, such as critical care units and large medical technical equipment, drive the overall energy consumption of hospitals.

The authors differentiate between two categories of determinant factors: characteristics that are fixed and do not depend on the behaviour of the hospital's occupants but differ from hospital to hospital; and factors and activities in the building that scale with the type of building use and the level of medical service provision at hospitals. There is a need to understand the level of influence the factors in these categories have on the energy consumption of a hospital, determine how these factors vary from hospital to hospital and account for this variation when normalising the energy performance of hospitals.

Table A.2 in Appendix A.1 shows a summary of the results of the literature analysis. From the literature analysis, the factors that affect energy consumption were grouped into three major categories, namely, hospital characteristics and construction, weather and climate factors, and the effect of the hospital's function. Figure 2.2 shows how each of these groupings is related to the respective subsections that discuss the effect of the factors found in the literature review.

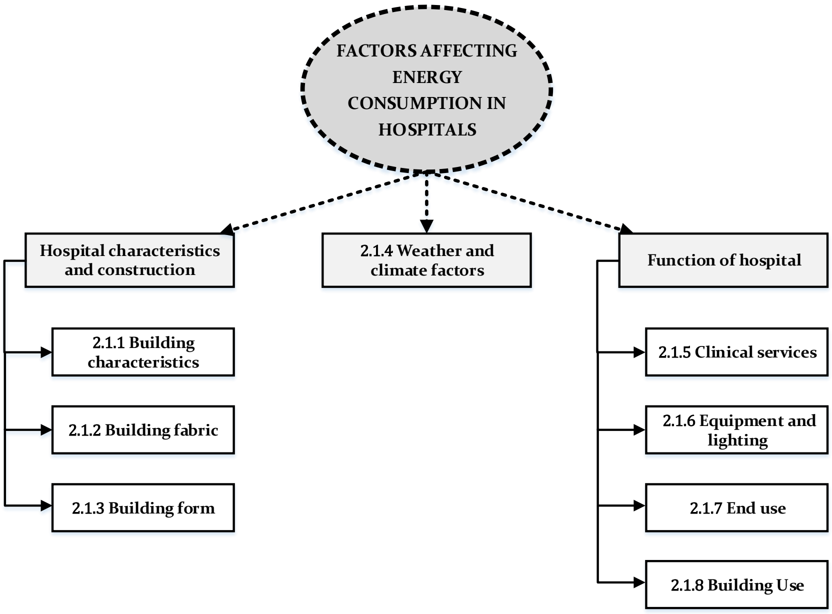


Figure 2.2: Organisation of subsections discussing the results of the literature analysis for energy consumption

2.1.1 Building characteristics

This group houses the factors that represent the effects of the size, layout, and age of a hospital building on its energy consumption. The size of a building is one of the most important determinants of its energy consumption (Singer et al. 2009). Size is the parameter that most of the studies normalised for when analysing the energy consumption of a building. Building size refers to the physical size of the building and is quantified using parameters such as total roofed area, total floor area, and total heated volume.

The general consensus within literature is that the energy consumption of a building or hospital scales with its size. The research done by de Fátima Castro et al. (2015) found that resource consumption and cost in buildings are directly proportional to changes in the net floor area of a building. Rajagopalan & Elkadi (2014) also found a good correlation between both the floor area and building volumes of hospitals, and the hospital's electricity consumption, whereas Murray et al. (2008) found no correlation between the electricity consumption of a building and its treated volume. However, this was attributed to the fact that the buildings in the study by Murray et al. (2008) were mainly heated using natural gas.

The year of construction of a building determines the material used in its construction and the way it was constructed, namely: its wall thickness (Caldera et al. 2008). Thus, the building envelopes of hospitals from different eras may vary in terms of heat transmittance and thermal capacity due to the use of different design configurations and materials. Consequently, this can cause variations in the cooling and heating loads of hospitals from different eras.

Ma et al. (2016) and Ma et al. (2017) both studied the energy consumption in Chinese public buildings (hospitals, schools and offices) and found that for buildings that performed the same function, energy consumption per unit area was higher in older buildings than in newer buildings. This was attributed to improvements in the thermal performance of the doors and windows, the improved thermal performance of modern building envelopes due to better insulation practices, and the improved energy efficiency of building HVAC equipment, lights and elevators. Murray et al. (2008) reported a weak correlation between heat energy consumption and the age of the facilities in the study. It was noted that five of the eight authors that mentioned the age or year of construction of a building as a factor did not elaborate on the relationship between age and consumption.

Singer et al. (2009) state that it is important to normalise for the effect of the location of a hospital as it is one of the factors that impacts its energy consumption and is outside the control of hospital administrators. Singer et al. (2009) further argue that the local climate and weather, which affect the thermal environment and lighting requirements of a building, are dependent on the location of the building.

Interestingly, this contradicts the findings of Murray et al. (2008) who found a poor correlation between heat energy consumption and physical location.

2.1.2 Building fabric

The building's envelope is the thermal barrier between the external environment and the thermal conditions inside the building. The thermo-physical properties of the walls, roof and windows determine the amount and rate of heat transfer through the building's envelope (Catalina et al. 2013; Pacheco et al. 2012). Heat transferred through the façade is one of the major contributors to the HVAC load of a hospital building (Rajagopalan & Elkadi 2014).

The buildings HVAC loads are sensitive to changes in the overall heat transfer coefficient of the building. Ma et al. (2017) found that an increase in the overall heat transfer coefficient of a building results in an increase in the building's total energy consumption. The heat transfer coefficient of a building depends significantly on the thermo-physical properties of its envelope i.e. the thermal transmissivity of its walls, roofs and windows. The use of external wall insulation, green roofs and cool roofs, or energy-saving window glass and frames reduces the overall heat transfer coefficient (Ma et al. 2016). Taleb (2016) observed a 24.7 percent reduction in solar heat gains in simulations studying the envelope performance of a hospital in Abu Dhabi when the hospital model used a green roof instead of a normal roof.

Window glazing is one of the features used to regulate the thermal environment of a building (Pacheco et al. 2012). The size, orientation, type and amount of window glazing influence the heating and cooling load of a building. The use of heat-absorbing glass, heat-reflecting glass and low-radiation glass can reduce the amount of energy needed by a cooling system or a heating system, in the respective heating and cooling seasons, based on the spectral properties of the glass (Pacheco et al. 2012). The extent of these gains is such that Catalina et al. (2013) argued that by optimising the use of energy-efficient windows the building envelope's thermal performance would outperform a highly insulated wall.

2.1.3 Building form

The aspect ratio, orientation, window-to-wall ratio, shading, and the building's relative compactness were identified as the factors used to account for the impact of the hospital's form on its energy consumption (Fumo et al. 2010). The aspect ratio represents the ratio between the length of the building's east-west facing façade divided by the overall length of its north-south facing façade (Deru et al. 2011). This factor is usually studied together with the orientation as they both influence the amount of solar radiation received by the building, which can increase the building's cooling load by 25 percent (Pacheco et al. 2012).

Alshayeb et al. (2015) analysed the impact of a hospital building's orientation and use of shading on its energy consumption. The optimal orientation of a building reduces solar heat gains through the building's façade, and thus the amount of active control required to ensure an optimal internal thermal environment (Alshayeb et al. 2015). However, the study found that the impact of the orientation of the building was not significant. A 0.38 percent difference in energy consumption was observed between the two buildings with the greatest difference in orientation. This difference was attributed to the square-shaped profile of the buildings, and thus a small difference in the area of the east-west facing façade and the area of the north-south facing façade. The east-west facing façade has a higher solar heat gain coefficient than the north-south facing façade.

The shading performance of the building's façade regulates the amount of solar heat gains through the façade and can reduce peak cooling loads and lighting loads (Pacheco et al. 2012). Taleb (2016) observed a 1.5 percent reduction in heat gains through the façade in building models using sunshade-designed façades.

The window-to-wall ratio of a building also has a significant impact on heat gains through the façade. Windows are more permeable to solar radiation than walls and are thus more susceptible to direct solar heat gains. Both Korolija et al. (2011) and Radwan et al. (2016) found a strong correlation between total building energy use and a building's window-to-wall ratio. Radwan et al. (2016) found that a 10 percent change in this ratio results in a 4.3 percent change in the energy consumption of a hospital’s HVAC system.

The compactness ratio, also known as the shape factor or building’s surface area to volume ratio, is related to a building’s ability to retain heat and avoid heat losses through the façade (Pacheco et al. 2012). These losses need to be compensated for by the buildings HVAC system. Buildings with large compactness ratios have higher heat losses through their façades than a similar building with the same type of façade and smaller compactness ratio (Rajagopalan & Elkadi 2014). Korolija et al. (2013) found a strong correlation between total annual energy use and the compactness ratio.

2.1.4 Weather and climate factors

Climate factors, such as the outdoor temperature, solar radiation and the humidity of the area in which a hospital is located, affect the amount of solar heat gained and/or lost through the building's envelope. Thus, the local climate influences the cooling and heating load of the hospital, which directly affects the energy consumption of the HVAC system.

Ma et al. (2016) studied energy consumption in public buildings in China’s cold Tianjin region and found that energy consumption for heating was responsible for a significant proportion of building energy use in winter. Chung & Park (2015) also attributed variations in the energy consumption for heating of Japanese and Korean buildings to differences in the climates of the two countries. Chung & Park (2015), Bagnasco et al. (2015) and Szklo et al. (2004) found that the electrical energy consumption of buildings also varied seasonally. Electrical energy consumption was significantly higher in summer than in winter. Bagnasco et al. (2015) found that electrical energy consumption of buildings in summer was 30 percent higher than in winter. This was attributed to the high and continuous demand placed on HVAC systems in large hospitals during the summer months.

2.1.5 Clinical services

The Subsections 2.1.1 to 2.1.4 discuss how the characteristics of physical building that constitutes the hospital and the local climate affect the energy consumption of a hospital. The energy consumption affected by these factors was mainly concentrated on the performance of the buildings HVAC system. This subsection and the ones to follow focus on how the medical services provided at a hospital affect the hospital's energy consumption. The effects of these factors are responsible for the energy use of the medical aspect of a hospital and are experienced by an array of building systems.

Chung (2011) mentions the importance of building use as a determinant of energy consumption in hospitals. Due to the heterogeneous nature of hospitals, energy consumption varies significantly with the type of services provided at a hospital. Careful consideration is therefore required to account for the change in the energy consumption of hospitals due to changes in the types of medical services provided by a hospital. As mentioned earlier in Section 2.1, the energy consumption of hospitals scales with the extent and capacity of its medical service provision (Singer et al. 2009).

The clinical specialities of a hospital are a major determinant of its energy use and the presence of different specialities has varying effects on the hospital's load profile. For example, the high power per unit consumption of large medical technical equipment is associated with medical imaging departments. Rohde & Martinez (2015) found large variations in the energy usage of the different hospital departments.

The respective clinical specialities have different energy requirements. This variance in energy requirements affects their energy load profile to different degrees. Clinical specialities affect the internal load of a hospital in three ways: they govern the type of services provided by a hospital, the type of equipment in use at the hospital and the capacity of a hospital (Szklo et al. 2004). Thus, clinical specialities also have an effect on the internal loading and internal heat gain of a hospital due to occupancy, equipment, and lighting. Buonomano et al. (2014) found that the functionality of a hospital correlates to its electricity use and HVAC load. Furthermore, larger hospitals are more complex and also provide a more diverse set of services (Szklo et al. 2004).

2.1.6 Equipment and lighting

Equipment use, and lighting, are two of the major drivers of energy consumption in hospitals. The energy consumed by equipment and lighting has two components, namely: the direct equipment use induced electric load component from the operation of the equipment, and the thermal load component from the waste heat generated by the equipment (Rohde & Martinez 2015).

The contribution of these components is governed largely by three factors: the energy intensity of each device type, the variations in the usage level and activity patterns of the respective devices, and the prevalence of each device type in the hospital (Rohde & Martinez 2015). Equipment and lighting density is a measure of device prevalence. Ma et al. (2017) found that the energy consumption of buildings increases with an increase in its equipment density and lighting density. This is because of the direct energy consumption associated with medical and building equipment and the increase in cooling energy consumption due to the extraction of the waste heat attributed to equipment use.

Three types of equipment are used in a hospital: large and small medical technical equipment (MTE); building equipment (BE), and information communication technology (ICT) equipment. Rohde & Martinez (2015) studied the energy usage of medical equipment in large teaching hospitals in Norway. The study found that medical equipment contributes significantly to both the electrical and heating loads of a hospital.

Furthermore, the medical equipment associated with each clinical speciality have different power ratings. For example, medical imaging equipment has higher power ratings than small medical technical equipment (SMTE) such as point-of-care devices (Rohde & Martinez 2015). However, although SMTE has a low per unit energy consumption rating, these devices are present in large enough numbers in hospitals and/or are often operated continuously for extended periods of time, thus making their collective energy consumption significant. This further emphasises the important role of the size and capacity of the hospital on its energy consumption.

2.1.7 End use

The energy consumption of a building is captured by the energy demand of six sublevel building systems (see Figure 2.3): the cooling system, space heating system, water heating system, lighting system, mechanical ventilation system, and plug and process loads (Leipziger 2013). It is important to account for the effect that the characteristics of these building systems have on the energy performance of a hospital building when determining the actual performance potential of a hospital. The system type and system characteristics parameters encompass both the micro and macro inter-building system differences.

Macro differences refer to the large-scale differences between system configurations i.e. using a centralised variable air volume HVAC system vs. using a constant air volume HVAC system. The collective energy consumption varies depending on the system, the energy consumption of fans, pumps, motors and other significant electricity-consuming devices depending on the type of system used (Korolija et al. 2011). Micro differences refer to the differences between components that make up the system, their electric loading, and efficiency characteristics.

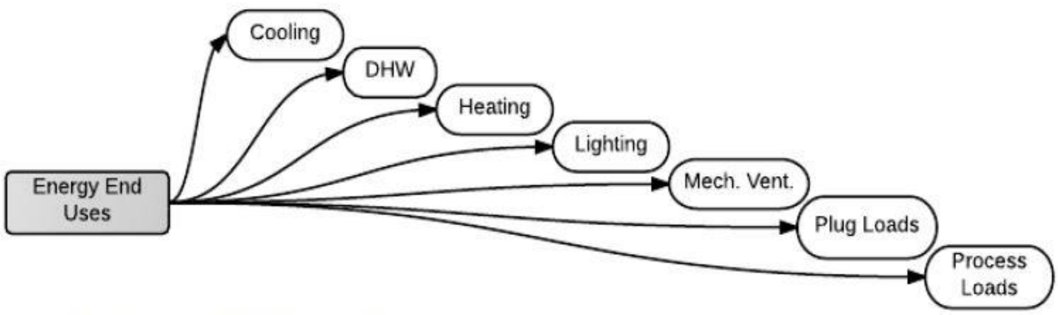


Figure 2.3: Energy end uses in hospitals (Leipziger 2013)

The varying configurations and combinations that are used to aggregate the respective system components to perform the function of the building service have different efficiencies. These efficiencies can be divided into two categories: macro-level efficiency; the overall efficiency of the system, and micro-level efficiency; the efficiency of the components that make up the system. Macro-level efficiency is the most important measure of efficient energy consumption because efficient components can be combined in an inefficient manner resulting in wasteful energy consumption (Singer et al. 2009). The building is mostly operated in one of two modes, at design-load conditions or at part-load conditions, thus both the macro- and micro- level efficiencies of these modes are of interest (Zhu 2006).

2.1.8 Building use

The space within a hospital building is divided into different zones with similar thermal comfort requirements. These climatic zones are referred to as thermal zones. The respective functional spaces within the hospital such as the waiting rooms in the emergency department, administrative offices, critical and intensive care, and operating rooms have different thermal comfort and outside air requirements. The heating and cooling demand profiles of these thermal zones and their fresh air requirements have a significant impact on the energy efficiency and performance of the buildings HVAC system (Korolija et al. 2011). The sizing of the HVAC system's respective components are also largely dependent on the size and amount of thermal zones within the building (Zhu 2006).

Rajagopalan & Elkadi (2014) studied the energy performance of medium-sized Australian hospitals. They divided the hospital buildings into six functional areas, according to the different functions performed in each space type. Each functional area was allocated a percentage that represents its energy consumption as a fraction of the total energy consumption of the hospitals. The differences in the functions of the respective functional areas resulted in different percentages allocated to each functional unit. These differences are reflected in an array of parameters such as the occupancy rates of the different zones in terms of occupant density and operational patterns, different ventilation requirements (flow rates and outside air requirements), different equipment and lighting specifications.

2.2 Factors that affect the water use of hospitals

In this section, the factors affecting the water consumption of hospitals were grouped into two general categories according to commonality. The first set of categories houses the factors identified from literature that account for the effect of the hospital's characteristics and context on its water consumption. The second set of categories houses the factors that account for the effect of the hospital's function on its water consumption. The literature analysis process is discussed in Appendix A.2 and a summary of the results of the literature analysis is presented in Table A.4. Figure 2.4 outlines the relationship between the two general categories and the following subsections discussing the results of the literature analysis.

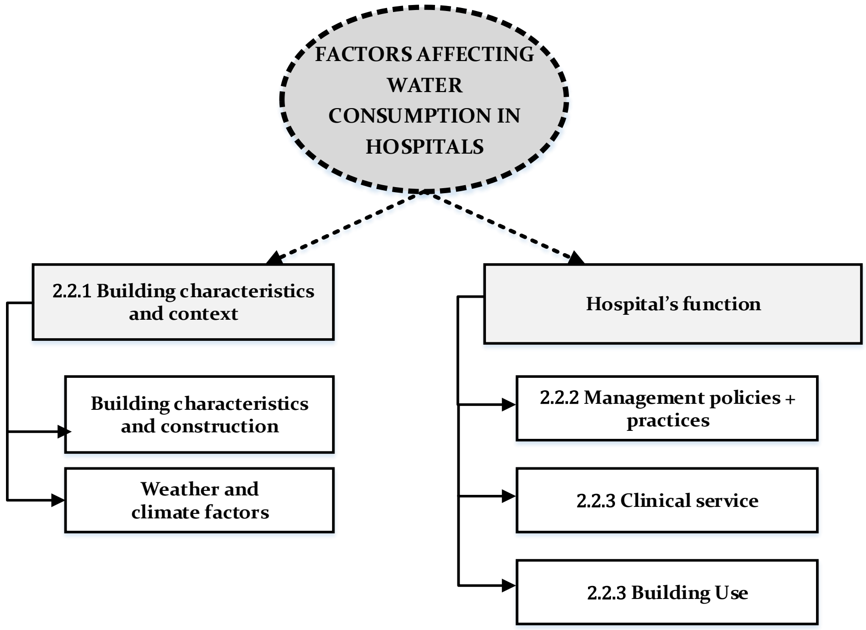


Figure 2.4: Organisation of subsections discussing the results of the literature analysis for water consumption

2.2.1 **Building characteristics and context**

Building characteristics are the factors most authors accounted for when assessing the water consumption of hospitals. Emphasis was placed on accounting for the size of the facility by stating water consumption performance with respect to the hospital's floor area (m^3 per m^2 floor area).

Four publications discussed the impact of a hospital's size on its water consumption. Garcia-Sanz-Calcedo et al. (2017) found a strong and statistically significant correlation between the built floor area of a hospital and its cold water for human consumption (CWHC) ($R^2 = 0.9645$) and domestic hot water (DHW) consumption ($R^2 = 0.9447$). González et al. (2016) also found a statistically significant correlation between the built floor area of a hospital and its CWHC ($R^2 = 0.8417$). Wong & Mui

(2008) state that the size of a hospital is one of the factors the Hong Kong Building Environmental Assessment Method normalises for in their model. Taken together, these highlight the significant impact of a hospital size on its water consumption.

The effect of a hospital's year of construction on its water consumption was also mentioned by both González et al. (2016) and MWRA (n.d.). These publications agreed that water consumption varies between hospitals constructed in different years. Furthermore, geographically dependent climate factors such as the temperature and humidity also affect water consumption in hospitals. For example, warmer climates result in higher water consumption. In addition, Verlicchi et al. (2010) found that water consumption also varies seasonally and observed higher monthly water consumption in the warmer months.

BIS & Cranfield University (2009) attributed this increase in consumption to an increase in the amount of water consumed for ingestion, hygiene and cooling. Furthermore, outdoor water consumption was also observed to be higher in warmer climates. The heating degree days' metric can be used to track the effect of climate on water consumption of hospitals. González et al. (2016) established a relationship between the CWHC and annual heating degree days in the region where the hospital is situated.

2.2.2 Management policies and practices

The hospital's water management policies and the level of awareness within the institution of the hospital's water footprint play a significant role in the attitude people have towards water usage within hospitals (D'Alessandro et al. 2016). The policies adopted by hospital management and their relation to the aim of reducing the amount of water consumed by hospitals or increasing the hospital's consumption efficiency impact its water consumption. This refers to the use of innovative water-saving systems such as flow control on taps, water-saving shower heads and the reuse of grey water and/or rainwater for irrigation and fire services (D'Alessandro et al. 2016a). Furthermore, increased water consumption awareness and minor investments can reduce water consumption. The study by MWRA (n.d.) observed a

19 percent reduction in annual water consumption following increased water consumption awareness.

Incentives for implementing water sustainability methods provide encouragement in terms of funds and recognition to hospital administrators (Faezipour et al. 2013). The implementation of policies in favour of the recycling and reuse of non-potable waste water can lead to significant reductions in the use of water from municipal mains (VDoH 2009). However caution needs to be exercised in deciding how this water is reused as it may have health implications for building occupants and fouling implications on the mechanical components of the building's systems (González et al. 2016).

2.2.3 Clinical service

Water use in hospitals can be split into three main categories: water for human consumption, water used for medical purposes and water for support services. The nature and extent of these services varies from hospital to hospital. The effects of a service on the water consumption of a hospital can be captured by accounting for three aspects of the service: its presence, the nature of the service and the extent of the service provided. This is done by determining how water use at the respective facilities scales with the presence, nature and extent of these services.

Water for human consumption represents the water that is consumed by both patients and staff for ingestion, hygiene and sanitation. The effect of human consumption on the hospital's total water demand was discussed in Subsection 2.2.1.

The portion of a hospital's water needs that is due to its medical service provision consists of water needed to satisfy water-intensive medical procedures such as haemodialysis, hemofiltration and hydrotherapy (Collett et al. 2016). Other aspects of the water needed for medical services include the sterilisation and decontamination of surgical and medical instruments in the central sterile department, the pre-surgery surgical scrubbing, and water used in medical laboratories. The nature of these services differs from hospital to hospital and scales with the extent and capacity of these services at hospitals.

Process water consumption is mainly due to kitchen water use, laundry, housekeeping activities, HVAC and cooling tower consumption, and irrigation (Collett et al. 2016). Kitchen water use is mainly for food preparation and cleaning purposes. The water used in the laundry is for the care of the medical staff and patient gowns, and the bedding and clothes used in both the operating theatre and the patient wards. HVAC systems are consumers of large amounts of water in commercial buildings. The water-cooled chillers used in the HVAC systems of most commercial buildings are significantly more energy-efficient than the air-cooled chiller alternatives (Weimar & Browning 2010). However, these systems consume large quantities of water due to the evaporation in the cooling towers that is responsible for heat rejection to the atmosphere.

González et al. (2016) acknowledge that the services provided by a hospital are directly related to its water consumption. However, in their analysis, they did not include them in the quantitative aspect of their study. A part of their study focused on determining whether a relationship exists between the microeconomic climate of the region in which a hospital is situated, and its water consumption. The author argues that the microeconomic climate determines the level of services that the hospital can provide and thus by extension its water consumption. The study used the gross domestic product (GDP) of a region as a measure of its microeconomic climate and found that there is no link between GDP and the water consumption of hospitals.

Another method that can be used to account for the effect of the type of services provided by a hospital on its water consumption is to use its mix of area ratios (Wong & Mui 2008). The logic behind the mix of area ratios is that by using ratios of the floor areas of significant departments in hospitals relative to the total floor area of a hospital, one can capture the demand of the specific areas on the total demand of the hospital's water consumption. Thus, as these ratios change at the respective hospitals the water consumption changes in a similar manner.

Verlicchi et al. (2010) also identified the importance of a hospital’s mix of areas. The study identifies the need to account for the number and types of wards and units in a hospital. The number and type of wards are factors that affect how much waste water is produced by the hospital and are proportional to the quantities of water needed to satisfy the different needs of the hospital. In the study by Verlicchi et al. (2010) they also highlight the importance of the general types of services provided in a hospital: kitchen, laundry and air conditioning.

The capacity of a hospital's medical service provision is another important factor affecting its water consumption. Verlicchi et al. (2010) analysed and aggregated the water consumption figures stated in various publications dated 1994 to 2009 from different countries and in different contextual settings. The analysis found no clear correlation between the average daily water consumption of a hospital and its capacity in terms of the number of beds in the hospital. This is inconsistent with the findings of D'Alessandro et al. (2016a), González et al. (2016), and BIS & Cranfield University (2009). D'Alessandro et al. (2016a) found a strong correlation between water consumption and the number of beds. Unlike the study of Verlicchi et al. (2010), the hospitals in this study are all situated in one region in Spain and not from different countries. This standardises the effect of contextual factors such as climate and microeconomic climate.

Both González et al. (2016) and BIS & Cranfield University (2009) also found a strong and statistically significant correlation between water consumption and the number of beds in a hospital. González et al. (2016) analysed the cold water use for human consumption (CWHC) in Spanish hospitals and found that CWHC was strongly correlated (R^2=0.9046) to the number of beds in a hospital. Garcia-Sanz-Calcedo et al. (2017) found a correlation between the number of beds in a hospital and both CWCH (R^2 = 0.8245) and DHW consumption (R^2 = 0.8484) respectively. This supports the findings of D'Alessandro et al. (2016a) and shows that water consumption does scale with respect to the number of beds in a hospital.

2.2.4 **Building use**

The number of occupants is a key driver of water use in hospitals (BIS & Cranfield University 2009). It is a measure of the level of demand that the need for services and resources places on hospital water resources (Faezipour et al. 2013). The level of demand for water resources is proportional to the number of hospital occupants.

Larger hospitals require more water for both human consumption (ingestion and hygiene) and process activities such as cooling and sanitation.

The second aspect of occupancy is the occupancy attributes. Occupancy attributes refers to the needs of the respective building occupants and is related to the hospital's inpatient to outpatient ratio (MWRA n.d.). Inpatients consume more water for human consumption than outpatients because the length of their stay in hospital warrants the provision of certain services that outpatients do not need. Furthermore, the water needed for medical purposes is dependent on the treatments and services provided to the different types of patients.

The third significant aspect of occupancy is the occupancy patterns. Garcia-Sanz-Calcedo et al. (2017) found that the peak time for water consumption in hospitals is between 12:00 and 18:00. This coincides with the periods of peak occupancy. Verlicchi et al. (2010) have similar findings, the water consumption in hospitals between 08:00 and 16:00 was 20 percent higher than the hospital's average water consumption and 30 percent lower than the average water consumption between 01:00 and 08:00.

Water is essential in the management of energy within hospital buildings (D'Alessandro et al. 2016a). The type of HVAC system used by the hospital is the second significant aspect of a hospital building use that has a significant impact on the hospital's water consumption. HVAC systems that use water-cooled chillers are more energy efficient than air-cooled HVAC systems, but are significantly more water intensive (Weimar & Browning 2010).

In water-cooled HVAC systems, the waste heat is exhausted into the atmosphere in the cooling tower via evaporative cooling. On average, for every degree Fahrenheit of temperature cooled, approximately 0.1 percent of the water that circulates through the cooling tower is lost to the atmosphere (Weimar & Browning 2010). Consequently, large amounts of make-up water are required to replace the water lost due to evaporation.

Furthermore, as part of the preventative maintenance practices, cooling towers must be bled regularly. As the water evaporates it leaves behind dissolved minerals. The bleeding process prevents the build-up of dissolved minerals in the building's hydrological network (MWRA n.d.). This process is water-intensive as large amounts of water are discharged from the cooling tower and replaced with new water (González et al. 2016). By replacing the water-cooled units with modern energy-efficient air-cooled units, hospitals can achieve a balance between energy and water sustainability. 58 percent of the 21 hospitals in the study by D'Alessandro et al. (2016a) planned to implement such interventions.

2.3 Quantifying consumption performance

2.3.1 Benchmarks

Benchmarking is a performance evaluation technique in which predefined performance-related metrics measured or estimated at a facility are compared to measurements at other facilities or to performance targets (Singer et al. 2009). This approach allows for the identification of performance gaps[4] and facilitates the development of management policies and practices that improve the performance of the facility.

In the resource consumption context, benchmarks are the reference performance levels against which resource consumption parameters that correspond to different facilities are compared and assessed (de Fátima Castro et al. 2015). Benchmarks act as a predefined quantitative baseline against which the performance of different facilities can be compared. The performance of each facility is quantified by metrics measured or estimated at that facility. These metrics are compared to those of the benchmark. The benchmark corresponds to a reference facility or specific performance target (Singer et al. 2009).

The appropriate application of robust benchmarks provides an accurate evaluation of a facility's operational performance (Hong, Burman, et al. 2014). In a healthcare context benchmarks have been used for:

[4]Defined as the difference between the actual performance of a facility and the performance of the reference against which it is benchmarked (Morgenstern et al. 2016).

1. Planning healthcare service provision, determining which type of services are appropriate for a certain context and the extent to which those services must be provided (see Böhme et al. (2013)).
2. Assessing the efficiency of facilities with respect to healthcare outcomes or performance targets in an operational context (see Araújo et al. (2013) and Nayar & Ozcan (2008)).
3. Sizing infrastructure requirements for healthcare facilities (see Fumo et al. (2010)).

Although benchmarks are not formulated in this study, it is important to understand what benchmarks are, how they are formulated and how normalisation is applied in the formulation process; because they are the main method used to evaluate the energy and water performance of hospitals in the Western Cape.

2.3.2 Benchmarking approaches

Benchmarking approaches are divided into two macro-level categories: top-down methods and bottom-up methods. Hong et al. (2014) and (Burman et al. 2014) conducted comparative studies of the benchmarking approaches for non-domestic buildings. The methodologies are classified into these categories based on the level of detail represented by the information used to derive the benchmarks in each category.

Top-down performance benchmarking methods derive their benchmarks by studying the consumption at a building-level and refining the level of analysis to a subsystem-level if more detailed consumption information is available (Hong et al. 2014). Top-down methods are dependent on the statistical analysis of a sample of buildings with similar but not identical characteristics to the building. This reduces the risk of outliers distorting the derived benchmark.

In top-down methods the benchmarks are based on building-level energy performance figures, the benchmarks are usually expressed as a single metric indicating how buildings with similar demand use energy. These benchmarking methodologies are better suited for policy formulation applications as they allow for the comparative analysis of energy use by a building against other existing buildings.

The bottom-up methods provide a means of quantifying a buildings consumption efficiency based on a theoretical analysis of the building which accounts for the effect of the building's unique architectural and system characteristics (Burman et al. 2014). The building is studied as a system: the lower levels of the system are clearly specified, and their performance is analysed generating metrics that represent the consumption of that system (see Figure 2.5). Subsystem-level information is then aggregated to generate system-level performance figures. This system-level information is used to generate system/building-level benchmarks.

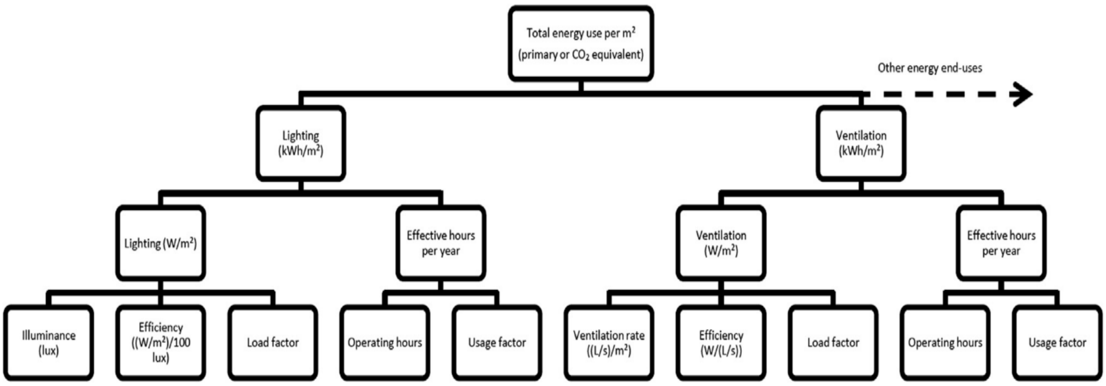


Figure 2.5: An example of the application of a bottom-up benchmarking approach (from Burman et al. (2014))

These methods are suitable for generating contextualised benchmarks that can be used to identify efficiency improvement opportunities within a hospital by studying the performance of individual systems such as HVAC and DHW. The system-level consumption figures would then be aggregated together into a single figure that represents the hypothetical whole building energy and water performance level. Bottom-up methods are more suited to identifying areas for performance improvements and how performance improvements can be realised (Burman et al. 2014).

2.3.3 Formulating benchmarks

A diverse array of statistical methods have been used in literature to compare the resource consumption performance of buildings. These methods provide a structured and systematic approach for objectively evaluating the performance of a building. Benchmarks are formulated using one of four different approaches:

1. Estimated by sector-specific experts from historic performance data of a facility. These serve as good practice averages based on the retrospective performance of the facility (see Cunninghame (2015)).
2. Estimated using computerised simulations such as EnergyPlus to develop a reference building model against which the performance of other facilities is compared. This requires a strong understanding of the interactions between the attributes of a facility and its consumption (see Fumo et al. (2010)). This approach is expensive, complex, and time- and data-intensive (Caldera et al. 2008).
3. Estimated from regression analysis[5] (see Chung, Hui & Y. Miu Lam (2006)). Regression is applied to a set of data made up of the consumption performance of an array of hospitals. The regression analysis is used to estimate a benchmark for the dataset by fitting an estimate to the data such that the estimate minimises the difference between the estimated value and the actual observed performance of each hospital.
4. Estimated using data envelopment analysis[6] (DEA) (see Nayar & Ozcan (2008)). Data envelopment analysis (DEA) is used to formulate an efficiency frontier. This frontier is used as the reference level against which the performance of facilities are benchmarked (OECD & JRC 2008).

The choice of benchmarking methods depends on (Hong, Paterson, et al. 2014):

- the objective of the analysis,
- the availability of data needed to develop and normalise benchmarks,
- the granularity of the data involved in developing benchmarking, and
- the accuracy and robustness of the desired model.

---

[5] See Appendix B.1 for a more detailed explanation.

[6] See Appendix B.2 for a more detailed explanation.

2.3.4 Selected approach to benchmarking

The confounding factors identified in Sections 2.1 and 2.2 that affect the degree of energy and water consumption in a hospital were summarised into two high-level categories: the hospital's building structural composition (characteristics, construction and context); and its function (utilisation of capacity and case mix composition).

Most of the factors in the hospital building's structural composition category were identified in literature that used a bottom-up approach. In these publications, physical computer models were used in simulations to assess the consumption behaviour of hospital buildings. Parameters that represent the characteristics and structure of a hospital building were used to develop a model of the hospital or an area of interest in the hospital. These models were used to perform simulations that studied the consumption behaviour of the modelled hospital building or to develop a reference model to benchmark against.

In the context of South African hospitals using a bottom-up approach is infeasible, because of the high granularity of the data required. A bottom-up benchmarking approach is both data- and resource-intensive, as well as computationally expensive as it requires the collection of information pertaining to a wide range of characteristics for all the hospitals being comparatively evaluated.

The data on most of these factors is not available for most facilities or able to be quantified in such a way as to provide a consistent picture over time. For example, many of these hospitals were built in different eras, and have been continuously extended and refurbished over their service life, therefore the fabric of the hospital is not homogenous. Thus, the level of confidence that can be placed in metrics developed using these factors will be limited since the data used to construct them is of a poor quality.

Furthermore, when one starts accounting for too many factors, the metric developed becomes difficult for the intended end user to conceptualise. This reduces the usefulness of the benchmarks developed by this normalisation process. Another problem is that it has too many factors that potentially do not significantly affect the energy and water consumption at hospitals. Thus, there was a need to find a balance between the level of significance of the factors accounted for in the normalisation process and the relative usefulness of the benchmarks developed with respect to the trends they can be used to study.

The following criteria were used to identify and select the factors and characteristics for inclusion (adapted from OECD (2009)):

- The factor must be of a general or total building-level in nature;
- The factor must be quantifiable and practically definable;
- The factor must have some statistical utility;
- The availability of data on the factor; and
- The factor is outside the direct control of hospital administrators.

2.3.5 Levels of the hospital at which consumption is compared

A more suitable approach would be to use a top-down approach to study the hospitals in the province collectively and try to explain the variance in the data structure of their energy and water consumption. This is done by identifying factors that vary between hospitals and result in significant differences in their consumption patterns. This is a more feasible approach to take as the granularity of data required to account for the individual factors identified in the preceding sections is high.

The guideline proposed by Singer et al. (2009) for benchmarking hospitals identifies three distinct levels at which consumption performance can be evaluated, namely: total building-level, building system-level and departmental-level. The total building-level defines macro metrics which facilitate the capturing or comparison of overall building consumption performance. The trade-off for this approach is that some accuracy is sacrificed due to the simplifying assumptions made to the macro-level model, but a more holistic picture of performance is gained especially in situations where the available data is limited.

The building system-level targets building system-level improvements. Metrics are defined to capture the consumption by significant building systems and their components. This analysis allows for the identification of poor-performing systems, inefficient equipment, inefficient operational parameters, and prioritising the improvement to localised building systems such as: HVAC systems, DHW systems, steam systems, and lighting systems (Singer et al. 2009).

The departmental-level facilitates department vs department comparisons, benchmarking and improvements. This approach is subject to limitations, though, as the demand and consumption of some resources by various departments is interdependent and cannot always be isolated to specific departments. For example, the HVAC load is generally distributed at a building-level and not isolated to a specific department. Like the building system-level approach, the departmental-level approach also requires the measurement of specific characteristics and consumption measurements.

The chosen scope determines the metrics that can be used in the analysis. This choice is reflected in the variables used to represent the identified factors. In this study the total building-level approach was used. The focus will be primarily split between the two first-order categories of factors: medical service capacity and building characteristics. These represent the most significant drivers of resource consumption at a total building-level (Singer et al. 2009).

The level at which consumption is evaluated plays a vital role in the selection of the factors that are used to describe and characterise the hospitals in the normalisation model. The chosen factors are characteristic of the level at which consumption is evaluated. In this study consumption is evaluated at a total building-level. Thus, the level of abstraction of the normalisation factors is much higher than for a departmental level or a building system-level approach. Thus, only macro-level variables will be suitable for use in the approach selected for this study.

2.4 Concepts and indicators evaluated

In quantitative research, concepts are quantifiable theoretical points around which research is conducted. They are used to describe or evaluate causal relationships, and either serve as the real-world phenomenon being studied (dependent variables) or provide a causal explanation of aspects of the real-world phenomenon (independent variables) (Bryman et al. 2014). In this study the interest was in evaluating the relationships between the concepts of resource consumption performance in hospitals, and the concepts discussed in Sections 2.1 and 2.2.

In literature, studies that focus on capturing the effect of the hospital's function on its energy and water consumption are less common. However, the effect of a hospital's function on its resource consumption with respect to hospital cost analysis is a widely researched field in health economics. Thus, by studying the techniques used to model the hospital's function for cost analysis, it is possible to identify suitable metrics and techniques that can be adapted for use in the analysis of energy and water consumption in hospitals.

From the current applications of normalisation in benchmarking and from the literature analysis, it follows that the size of a hospital and its function are significant drivers of energy and water consumption. Two factors are of interest when studying the function of a hospital: the diagnostic mix of its patient population, and the hospital's output. These two parameters capture the types of services provided by each hospital and the capacity of that hospital to provide those services, where, capacity refers to the extent and complexity of the services offered.

Furthermore, there is a need to test the feasibility of changing the way benchmarks are normalised, and to determine whether a combination of these factors is a better normalising factor than the size of the hospital alone. It is proposed that comparisons can be made more comprehensive by using benchmarks that are normalised for the size of the hospital, the output of the hospital and the hospital's case mix. These concepts are measurable, and the data is readily available. This makes it both feasible and cost-effective for implementation in the South African context, as it excludes concepts that are difficult or unrealistic to measure which may not significantly affect consumption. Each concept is expanded on in the following subsections.

2.4.1 **The size of a hospital**

This concept represents and captures the inter-hospital variation in the size of healthcare facilities. As discussed in Section 2.1.1, the two most commonly used measures to quantify the size of a hospital are its building footprint in terms of total floor area, and its capacity as represented by its number of available beds. The benchmarks currently being used to evaluate and compare energy and water consumption in the public healthcare sector in the Western Cape are predominantly normalised for the scale of the healthcare facility (Cunninghame 2015).

The capacity focuses on the scaling effect due to the demand placed on resources by patient needs. The potential to treat a patient is used as the unit of analysis. This measure is often specified in the forms $KWh/bed/day$ for electricity consumption and $KL/bed/day$ for water consumption. The building footprint focuses on the scaling effect associated with the operation of the hospital's systems. A unit of floor area is defined as the unit of analysis and the changes in the performance of these systems are evaluated with respect to this unit. This measure is often specified in the forms $KWh/m^2/year$ for electricity consumption and $KL/m^2/year$ for water consumption.

2.4.2 The output of a hospital

A patient-day is a healthcare time unit used to represent the demand placed on a healthcare facility's resources by hospitalising a patient for one day (Gray et al. 2011). Each patient-day represents a unit of time during which the services of a hospital are being used by a patient. For example, '20 patient-days' is equivalent to 20 patients being hospitalised for one day each. This concept was used as a measure of hospital output to gauge how many patients were treated or admitted and their average length of stay.

A patient-day-equivalent (PDE) is a form of the patient-day adjusted to the inpatient-to-outpatient ratios of hospitals. The PDE is defined as:

$$\begin{aligned} \text{PDE} = {} & \text{inpatient day} + \frac{1}{2}\,\text{\# of day patients} + \frac{1}{3}\,\text{\# of outpatients} \\ & + \frac{1}{3}\,\text{emergency headcount} \end{aligned} \tag{2.1}$$

By its definition, the PDE measure assumes that the demand on a hospital's services due to 1 inpatient visit is equivalent to that of 2 inpatients or 3 outpatient visits or 3 emergency visits (EHMI 2017). Thus, it facilitates the evaluation of resource utilisation efficiency at hospitals with respect to service-related data (Massyn et al. 2017).

2.4.3 The composition of a hospital's case mix

The compositions of the caseloads of hospitals differ significantly in terms of the number, type and extend of the diagnosis of the patients treated at the hospital. The different diseases that make up the caseload of a hospital, and their corresponding procedures have different levels of resource consumption (Gray et al. 2011). This results in inter-hospital variations in the demand placed on the resources of hospitals due to the diagnostic makeup of their patient populations. This concept evaluates the complexity of a hospital's caseload and the level of specialisation of that caseload, with respect to the range of diagnoses treated at a hospital and the number of cases corresponding to each treated diagnosis.

2.5 Conclusion

This chapter identified and provided an insight into the characteristics and factors that vary between hospitals and have a significant impact on their energy and water performance. The literature analysis identified multiple interconnected characteristics and factors that should be accounted for when normalising the energy and water consumption of hospitals. These characteristics and factors provided a theoretical foundation for the selection of inputs and indicators for the normalisation model.

It was found that resource consumption scales with the size, capacity, and level of medical service provision at a hospital. Three concepts that represent these factors were selected as potential normalisation factors in the formulation of benchmarks, namely: size, hospital output and case mix composition. These concepts were selected because they are quantifiable, practically definable, outside the direct control of hospital administrators, and the data needed to quantify them was available. The feasibility of using these concepts as normalising factors will be evaluated in subsequent chapters.

Chapter 3 The method

Chapter 2 discussed the factors and characteristics affecting energy and water consumption in hospitals, and the role of normalised benchmarks in the evaluation and comparison of the resource consumption performance of hospitals. It proposed that these comparisons can be made more comprehensive by using benchmarks that are normalised for: the size, output and case mix composition of a hospital. This chapter outlines the research methodology used to test the feasibility of changing the way benchmarks are normalised to determine whether a combination of these factors is a better normalising factor than the scale of the facility or output alone.

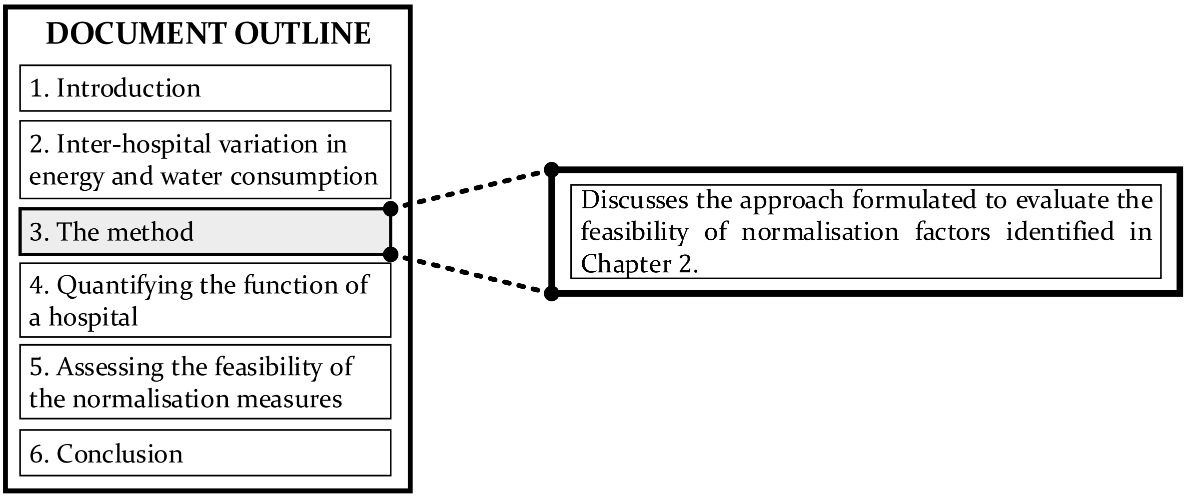


Figure 3.1: Thesis document outline: Chapter 3 contextualised

3.1 Research design

The research problem addressed in this study is descriptive in nature. The aim of the research was to describe how the normalising factors are related to the energy and water consumption of a hospital and to quantify these relationships. These relationships were then used to propose a normalisation model.

Using the research design type classification framework of Mouton (2005), a statistical modelling research design was selected for this study. This research design is empirical in nature and focuses on the analysis of primary, existing or hybrid (combination of primary and existing) data. The research design consists of studying a process or system, identifying important variables, and then capturing, describing and validating a representation of that process or system (Hofstee 2006).

Statistical analysis uses a sample that is representative of a population to make inferences about the behaviour or characteristics of the population using the results obtained from the analysis. The analysis process consists of:

- collecting sample data from a population,
- organising the data,
- analysing the data,
- representing the results, and
- interpreting the results to make inferences about the population.

In this study, statistical analysis was used to develop a model that estimates a function that relates inter-hospital variations in total annual energy and water consumptions to patterns in different combinations of normalising factors. Multiple analyses were conducted with different combinations of normalising factors as independent variables to determine which combination statistically accounts for the largest amount of variance in the energy and water consumption dataset.

One of the benefits associated with this research design is its ability to capture large-scale processes and systems accurately and to facilitate the simplification of the relationships in the process or system into a model (Mouton 2005). This leads to a greater understanding of that process or system as the model can be used to describe, explain or predict the process or system's behaviour under varying conditions.

However, this research design has some limitations that need to be considered. For example, data collection issues, such as obtaining a large enough and complete enough dataset to construct and test the model. Other limitations are associated with accounting for unexpected variables, the potential impacts of simplifying assumptions used to allow the model to practically capture reality, and the potential impact of errors in the modelling process (Hofstee 2006).

3.2 Research methodology

This section details the research methodology used to develop normalisation models that allow for the commensurable comparison of the energy and water consumption of a set of hospitals with respect to combinations of normalising factors. This methodology was used to assess the significance of the explanatory power associated with each combination of normalising factors and to identify the factors with the most statistically significant explanatory power. The research methodology used in this study and the chapters that correspond to each of the steps in the methodology are shown in Figure 3.2.

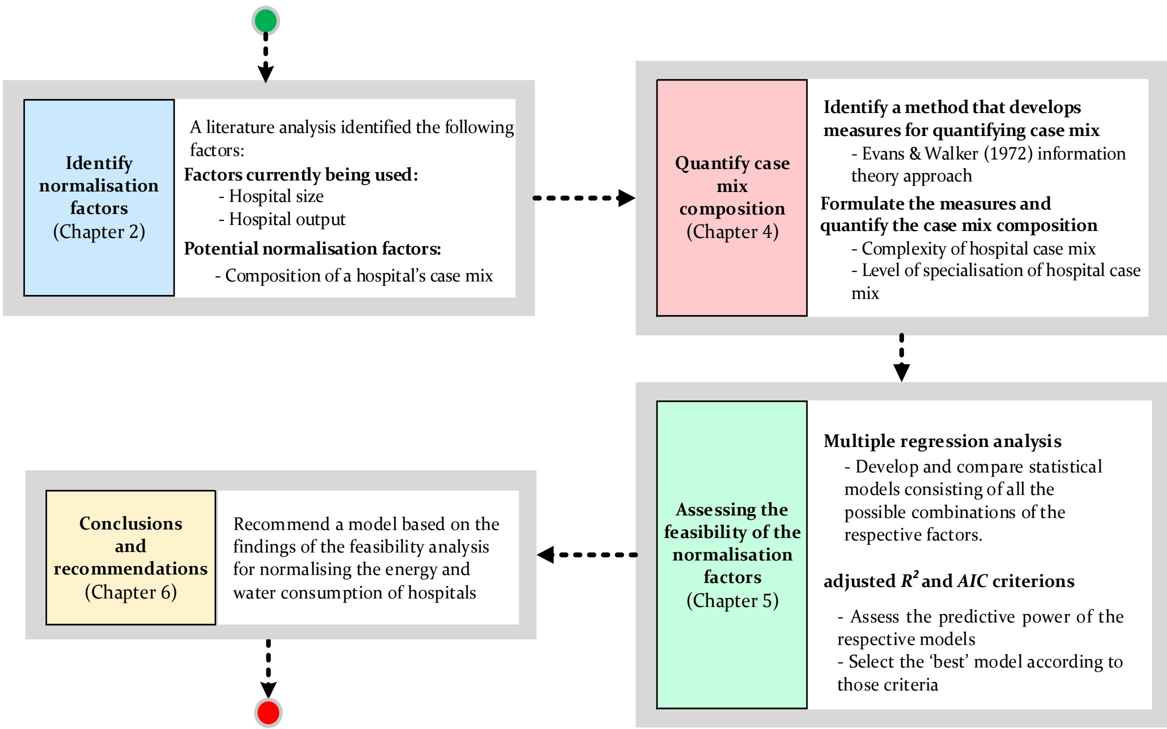


Figure 3.2: Research methodology

The first part of the methodology was discussed in Chapter 2. It consisted of the literature analysis that contextualised the real-world problem and identified the normalisation factors. From this analysis two sets of normalisation factors were identified: the factors currently being used to conduct normalisation (hospital size and output), and the potential factors for inclusion into the normalisation model (case mix composition).

The aim of the second part of the research methodology is to develop quantifiable measures to capture the case mix composition of a hospital. An information theory-based approach developed by Evans & Walker (1972) was used to model the complexity and level of specialisation of a hospital's caseload. This will be discussed in Chapter 4. Measures already exist for quantifying the factors currently being used to normalise the energy and water consumption of hospitals. The number of beds and the total floor area are used for the size of a hospital. The patient-day-equivalent is used for the output of a hospital.

In the third part of the research methodology, as discussed in Chapter 5, a statistical analysis was conducted using different combinations of the measures of the normalising factors. Multiple linear regression analyses were conducted to determine the amount of variation in the energy and water consumption data of hospitals that is explained by variations in combinations of the normalising factors.

The aim of the multiple regression analyses was to identify the combination of normalising factors with the most statistically significant explanatory power. This is achieved by determining how much of the variation in the hospital's consumption data is explained by variations in the various combinations of the normalising factors. The amount of inter-hospital variation in resource consumption accounted for by the models were compared to evaluate the explanatory power of the respective models. The best model was selected according to the R_a^2 and AIC criterions. These criteria facilitate the selection of the model with the best trade-off between simplicity and fit in the models evaluated.

In part 4 of the research methodology, the results of the regression analyses are used to draw conclusions on the feasibility of adding complexity and level of specialisation to the normalisation model to account for the case mix composition of a hospital. Furthermore, a normalisation model for the commensurable comparison of the energy and water consumption of a set of hospitals is recommended.

3.3 The information theory approach

This section discusses the information theory approach that was used to formulate measures for quantifying the case mix composition of a hospital. This information theory approach was proposed by Evans & Walker (1972) as a framework for standardising hospital output when conducting comparative analyses of hospital costs. Barer (1982) empirically tested the framework and refined the case mix standardisation method developed by Evans & Walker (1972).

Both Evans & Walker (1972) and Barer (1982) focused on formulating functions for estimating and comparing the costs associated with hospital services. This was a popular field of study during the 1970s to 1990s, which focused on identifying and developing equations for the relationships between the inter-hospital variations in hospital costs and sets of predefined factors.

The approach proposed in the framework of Evans & Walker (1972) used the diagnostic proportions of a hospital's caseload to formulate measures for its complexity and level of specialisation. The framework proposed that the treatment of complex cases requires extensive facilities and specialised staff. Thus, complex cases are responsible for a larger demand on the resources of a hospital and thus result in an increase in hospital expenditure. By studying the distribution of medical cases across a healthcare network it is possible to develop a measure of the complexity and level of specialisation of each hospital in the network.

The Evans & Walker (1972) framework proposed that the degree of concentration of medical cases across hospitals can be used as a measure of the complexity of a medical case type. It argued that complex medical case types tend to be concentrated in a few hospitals that have the specialised facilities, equipment and staff needed to treat them, whereas comparatively straightforward medical case types can be treated by most hospitals and are thus distributed across the healthcare network. Thus, the complexity of a hospital's caseload is calculated using the complexity of each medical case type in the hospital's caseload and the number of cases of that medical case type treated by the hospital.

The level of specialisation is also calculated using the diagnostic proportions of a hospital's caseload. It is a measure of the breadth of medical case types treated by a hospital. Specialised hospitals treat a small portfolio of diagnostic cases while more generalist hospitals cater to a larger set of medical case types. Subsection 3.3.1 discusses the expected information gain measure which was used to calculate complexity and level of specialisation as outlined in Subsections 3.3.2 and 3.3.3 respectively.

3.3.1 Expected information gain

The information theory approach developed by Evans & Walker (1972) uses two sets of probabilities (the prior probabilities and posterior probabilities) to define expected information measures. The treatment associated with each medical case is defined as an event, and each event has an associated prior probability and an associated posterior probability. The prior probabilities are based on the information available about the hospitals in the healthcare network before data collection or analysis. The posterior probabilities are based on each hospital's share of the overall provincial caseload treated for the diagnostic case type of interest.

After data collection and analysis, the actual distribution of medical cases across the hospitals in the healthcare network is known. Thus, the posterior probabilities can be calculated. The difference between the prior probabilities and posterior probabilities associated with each diagnostic case type at each hospital is called the information gain. It represents the knowledge gained from learning the actual distribution of medical cases at each of the hospitals in the healthcare network.

The information gain associated with the prior probabilities ($P_{prior,i}$) of an event i is calculated using:

$$IG_{prior,i} = ln\left(\frac{1}{P_{prior,i}}\right) \tag{3.1}$$

Similarly, the information gain associated with the posterior probabilities ($P_{posterior,i}$) of an event i is given by:

$$IG_{posterior,i} = ln\left(\frac{1}{P_{posterior,i}}\right) \tag{3.2}$$

The size of the information gain (IG_i) associated with the occurrence of each i-th event is given by the difference between the information gains associated with the prior and posterior probabilities of its occurrence. This relationship is defined by equation (3.3). The difference between $P_{prior,i}$ and $P_{posterior,i}$ for events with a high likelihood of occurrence is small, and thus the information gain associated with the occurrence of this event is also small. However, if an unlikely event occurs, the difference between $P_{prior,i}$ and $P_{posterior,i}$ will be large, resulting in a large information gain.

$$\begin{aligned} IG_i = IG_{prior,i} - IG_{posterior,i} &= ln\left(\frac{1}{P_{prior,i}}\right) - ln\left(\frac{1}{P_{posterior,i}}\right) \\ &= ln\left(\frac{P_{posterior,i}}{P_{prior,i}}\right) \end{aligned} \tag{3.3}$$

The framework defines an expected information gain measure (EIG_i) as the weighted sum of the information gains of all the possible events that are classified into a category of interest. Each information gain is weighted by its posterior probability of occurrence. The expected information gain from the occurrence of the i-th event is given by:

$$EIG_i = \sum_{i=1}^{n} P_{posterior,i} \cdot IG_i = \sum_{i=1}^{n} P_{posterior,i} \cdot ln\left(\frac{P_{posterior,i}}{P_{prior,i}}\right) \tag{3.4}$$

This process was used to formulate the EIG_i measures that were used to define measures for the complexity and level of specialisation of a hospital's caseload. The formulation of measures for complexity is discussed in Section 3.3.2 and the formulation of measures for level of specialisation is discussed in Section 3.3.3.

3.3.2 Complexity of caseload

The Evans & Walker (1972) framework is used to formulate measures for quantifying the complexity of the case mix composition of a hospital. The medical cases treated in the healthcare network are grouped according to their diagnostic case type. Each diagnostic case type grouping is studied and the distribution of its medical cases amongst the hospitals in the healthcare network is evaluated. In these evaluations each medical case treated by a hospital in the healthcare network is defined as an event.

Before data collection and analysis, only the number of hospitals in the network (N) is known. Based on the available information, it is assumed that the cases corresponding to the respective diagnostic case types have an equal likelihood of being treated at any of the hospitals in the network. Thus, there are N possible events for every medical case in need of treatment at a hospital in the healthcare network. This is shown in Figure 3.3, where $P_1, P_2, P_3, ..., P_n$ are the probabilities associated with the n possible events for each diagnostic case type.

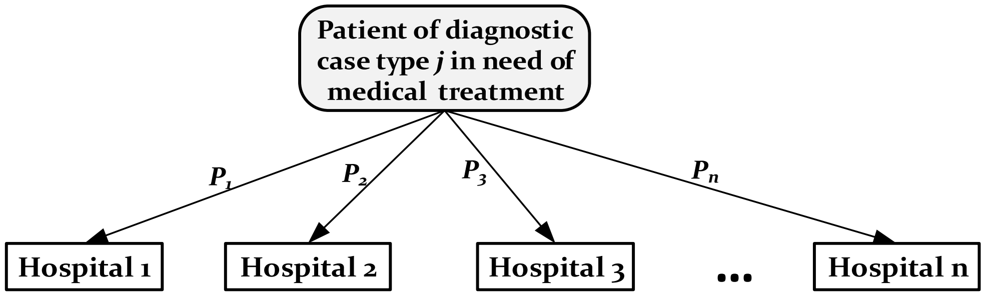


Figure 3.3: Focus of probabilities used to develop the complexity of caseload metric

Thus, when determining the complexity of a hospital's caseload, the prior probabilities correspond to the case where the treatment of medical cases of diagnostic type j is distributed evenly across the hospitals in the analysis. An even distribution is defined as the case where: the proportion of the overall number of cases of diagnostic type j treated in the network that is treated at hospital i $\left(\frac{c_{ij}}{C_j}\right)$ is equal to the likelihood of the medical case being treated at hospital i $\left(\frac{1}{N}\right)$. Therefore, the prior probability is given by:

$$prior\ probability = \frac{1}{N} \tag{3.5}$$

Where N is the number of hospitals in the healthcare network.

After data collection the actual distribution of cases of diagnostic type j treated at each hospital in the healthcare network is known. Thus, the posterior probability is based on each hospital's share of the overall healthcare network's patient load treated for the diagnostic case type of interest. This probability is denoted as:

$$posterior\ probability = q_{ij} = {}^{c_{ij}}/{}_{C_j} \tag{3.6}$$

Where

- c_{ij} is the number of diagnostic case type *j* cases treated at hospital *i* in a given year, and
- C_j is the overall number of diagnostic case type *j* cases treated at all the hospitals in the network in a given year.

The expected information gain associated with learning the actual distribution of the total cases of diagnostic type *j* treated at each hospital is given by:

$$G_j = \sum_{i=1}^{n} q_{ij} \cdot ln\left(N q_{ij}\right) \tag{3.7}$$

The expected information gain measure has an upward bias and tends to overestimate complexity. A standardisation is applied to correct for that bias. The $\bar{G}_j$ metric is the standardised version of the expected information gain measure. Where $\bar{G}_j$ is standardised to an average value of one. This metric is given by:

$$\bar{G}_j = \frac{G_j}{\sum_{j=1}^{m} G_j Q_j} \tag{3.8}$$

Where $Q_j = C_j/C$.

The complexity of the diagnostic case mix metric for hospital *i* is a weighted sum of the individual complexity values of all the diagnostic case types treated at the hospital. The weights represent the proportion of the overall provincial patient load for each diagnostic case type treated at hospital *i*. The complexity of the diagnostic case mix of hospital *i* is given by:

$$CMPX_i = \sum_{\mathrm{j}=1}^{m} \bar{G}_j p_{ij} \tag{3.9}$$

Where the weights (p_{ij}) are given by:

$$p_{ij} = {}^{c_{ij}}/{}_{C_i} \tag{3.10}$$

For diagnostic case type *j*, the expected information gain from learning the actual distribution of patients at each hospital was used as a measure of the complexity of that case type. A high information gain is indicative of a concentration of the treatment of diagnostic case type *j* at a hospital. The degree of concentration is used as a measure of the complexity of the caseload of a hospital. If a large proportion of the overall treatment of diagnostic case type *j* in the province is concentrated in a few hospitals., these hospitals are assumed to specialise in this diagnosis.

3.3.3 **Level of specialisation**

The level of specialisation measures formulated using the Evans & Walker (1972) framework evaluate the diversity of the diagnostic case types in the caseload of a hospital relative to other hospitals in the healthcare network. It measures whether diagnostic effort at a hospital is concentrated on a limited range of diagnostic case types or widely distributed across a large set of diagnostic case types. The focus is on the diagnostic case type classification of each medical case treated at hospital *i*. This is shown in Figure 3.4, where $P_1, P_2, P_3, \ldots, P_q$ are the probabilities associated with each case type.

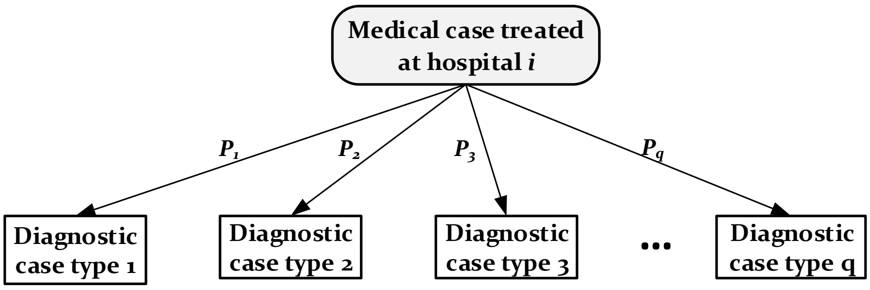


Figure 3.4: Focus of probabilities used to develop the level of specialisation metric

It is assumed that before data collection and analysis only the overall proportions of each diagnostic case type treated in the healthcare network is known. Thus, the prior probability of a medical case treated at hospital *i* being of diagnostic case type *j* is equal to the overall proportion of cases treated in the healthcare network that belong to diagnostic case type *j* (Q_j). This is given by:

$$prior\ probability = \frac{1}{Q_j} = \frac{1}{C_j/_C} = \frac{C}{C_j} \tag{3.11}$$

Where,

- C_j is the total number of diagnostic case type *j* cases treated at all the hospitals in the network in a given year, and
- C is the total number of cases treated at all the hospitals in the healthcare network in a given year.

The posterior probability is determined after data collection and analysis and is based on knowing the actual distribution of medical cases at each of the hospitals in the healthcare network. Thus, the posterior probability is given by:

$$posterior\ probability = p_{ij} = {}^{c_{ij}}/_{C_i} \tag{3.12}$$

Where,

- c_{ij} is the number of diagnostic case type *j* cases treated at hospital *i* in a given year, and
- C_i is the total number of cases treated at hospital *i* in a given year.

Thus, the information gain is given by:

$$H_i = \sum_{j=1}^{m} p_{ij}\, ln\left(p_{ij}/Q_j\right) \tag{3.13}$$

The expected information gain of hospital *i*, and by extension its $SPCL_i$ measure, are proportional to the size of the difference between the distribution of patients across the diagnostic case types within hospital *i,* and the overall distribution of patients across the diagnostic case types treated at all the hospitals in the province.

As with the $\bar{G}_j$ metric, $\bar{H}_i$ is the standardised version of the expected information gain measure for the level of specialisation of a hospital (H_i). It is also standardised to an average value of one. Furthermore, $\bar{H}_i$ is also equal to the $SPCL_i$ measure for the level of specialisation of the diagnostic caseload of a hospital. This measure is given by:

$$SPCL_i = \bar{H}_i = \frac{H_i}{\sum_{i=1}^{n} H_i P_i} \tag{3.14}$$

Where $P_i = C_i/C$.

3.4 Linear regression analysis

According to Chatterjee & Simonoff (2013) the three most common uses of regression analysis are:

1. to model the relationship between dependent variable(s) and independent variable(s);
2. to forecast the behaviour of independent variable(s); and
3. for hypothesis testing.

In this study, regression analysis was used to characterise the relationship between the dependent variables ($\boldsymbol{y}$) and independent variables ($\mathbf{x}$) by quantifying the dependence relationship ($y = f(\mathbf{x})$) between the variables. The dependent variables are the energy consumption (y_1) and water consumption (y_2) of hospitals. The independent variables are the size of the facilities (x_1), their output (x_2) the complexity of their case mix (x_3), and the level of specialisation of their case mix (x_4).

3.4.1 Multivariate situation

In statistics, a multivariate situation is one where each observation is characterised by multiple variables. An example of a multivariate situation is shown in Table 3.1, where there are n observations that are characterised by p variables. The term 'variate' refers to the fact that the situation can be represented as a weighted linear combination of variables with empirically determined weights (see equation (3.15)).

These weights (β-coefficients) are determined empirically from the observations. A univariate situation involves only one variable ($p = 1$). A multivariate situation involves more than one variable ($p \geq 2$). Thus, a multivariate situation can be defined as a linear combination of p variables where the weights are determined empirically from n observations that are collected from the population being studied.

Table 3.1: An example of a multivariate situation

Observations	X_1	X_2	X_3	…	X_p
1	x_{11}	x_{12}	x_{13}	…	x_{1p}
2	x_{21}	x_{22}	x_{23}	…	x_{2p}
⋮	⋮	⋮	⋮		⋮
n	x_{n1}	x_{n2}	x_{n3}	…	x_{np}

$$Y = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p \tag{3.15}$$

The behaviour of a system or process that is characterised by multiple variables should be studied by taking into consideration all the variables characterising it that are important to the design, development or improvement of the system or process being analysed. This is done to capture and account for the covariance structure that exists between variables that characterise the system or process. Using univariate statistics for a multivariate problem would result in a significant loss of information due to not accounting for this covariance structure.

3.4.2 Multiple linear regression

The primary interest of the regression analysis in this study is to evaluate the feasibility of including metrics that account for the complexity and level of specialisation of a hospital's case mix in the normalisation process. This is achieved through evaluating the variability in the dependent variables (the energy and water consumption of hospitals) that is associated with different combinations of independent variables (the normalisation factors).

It is assumed that the problem at hand's variability structure is characterised by linear relationships between n sets of observations of a dependent (y) and p independent ($\mathbf{x}$) variables. Although, the energy and water consumption are assessed relative to the same set of independent variables, it is assumed that there is no causal relationship between the energy consumption and the water consumption. It is also assumed that the independent variables do not covary. Thus, the multiple regression dependence model was deemed most suitable for characterising the relationship ($y = f(\mathbf{x})$) between the dependent and independent variables.

In multiple regression, the linear relationship between one dependent variable (y) and multiple independent variables ($\mathbf{x}$) is represented in a single model as illustrated in Figure 3.5. The arrowheads indicate the causal relationship in the model, $y = f(x) + \varepsilon$. The β-coefficients are unknown parameters that represent the degree of influence of the independent variable on the dependent variable. For example, for every unit change in x_1 there is a corresponding β_1 change in y assuming that all other independent variables are fixed (Chatterjee & Simonoff 2013).

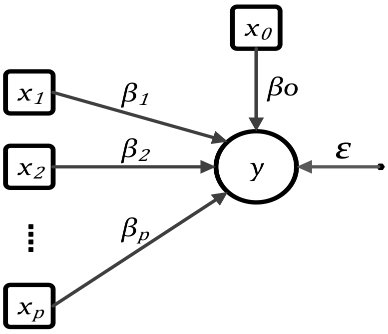


Figure 3.5: Pictorial representation of a multiple regression model

The mathematical representation of the general multiple regression model shown in Figure 3.5 is given by:

$$y = \beta_0 x_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon \tag{3.16}$$

Since this is a multivariate situation each of the individual x_i and y consists of sets of individual data points. Assuming that a dataset of n observations that are characterised by p variables was collected, then the independent variable and the dependent variables are represented by the following data vector and matrix:

$$y_{n\times 1} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{bmatrix} \quad \text{and} \quad x_{n\times p} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_i \\ \vdots \\ \mathbf{x}_n \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ip} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}_a \tag{3.17}$$

The general regression equation is applied to each observation in the dataset, as shown in equation (3.18) below for the i-th observation.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i \tag{3.18}$$

As shown in equation (3.19), y_i is variate and consists of two main parts the predicted value ($\hat{y}_i$) and the residual (ε_i).

$$y_i = \hat{y}_i + \varepsilon_i \tag{3.19}$$

The predicted value ($\hat{y}_i$) represents the expected value of y_i of the regression model given data collected in the i-th observation $\mathbf{x}_i$.

$$\hat{y}_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} = E(y_i | \mathbf{x}_i) \tag{3.20}$$

Thus, the multiple linear regression model for all observation in the collected dataset in matrix form is given by:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}_{n\times 1} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}_{n\times(p+1)} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}_{(p+1)\times 1} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}_{n\times 1} \tag{3.21}$$

This can be summarised in matrix form as:

$$\boldsymbol{Y} = \boldsymbol{X} \cdot \boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{3.22}$$

This equation represents the estimated underlying relationship between the dependent variable and independent variables in the data analysed. The regression coefficients ($\boldsymbol{\beta}$) are the estimation parameters that approximate this relationship. The residual components ($\boldsymbol{\varepsilon}$) represent the variability of the dependent variable that is due to factors (controllable or uncontrollable) that are not accounted for in the analysis. $\boldsymbol{X}$ and $\boldsymbol{Y}$ represent the dataset used to develop the regression model.

3.5 Model selection strategies

The multiple linear regression analysis method discussed in Section 3.4.2 was used to propose a set of candidate normalisation models. This section introduces the model comparison and selection approach that was used to objectively determine and select the 'best' normalisation model from this set of candidate models. This approach was proposed by Chatterjee & Simonoff (2013) and is based on the principle of parsimony. This principle emphasises the importance of maintaining a balance between the simplicity of a model and the strength of the model with respect to its ability to represent important relationships in the data.

Identifying and selecting the 'best' model involves comparing several regression models consisting of varying combinations of independent variables. These models are non-nested in nature, that is the models being compared consist of different combinations of a set of factors. For example, when comparing regression models with independent variables $\{x_2, x_3\}$ and $\{x_1, x_4\}$ respectively. The first model is not a subset of the second or vice versa. The following criterions were used to identify and select the 'best' model:

- the Adjusted coefficient of determination (R_a^2), and
- the Akaike information criterion (AIC).

These criteria assess the trade-off between the simplicity and strength of fit of a model (Chatterjee & Simonoff 2013).

3.5.1 The adjusted coefficient of determination (R_a^2)

The coefficient of determination (R^2) is an overall measure of the statistical significance of the regression. It is a measure of how much of the total variance in the dependent variable is associated with the set of independent variables in the regression analysis (Chatterjee & Simonoff 2013). The values of R^2 range from 0 to 1, where values close to 0 indicate a low similarity between the fitted target value ($\hat{y}_i$) and the observed value (y) and low predictive power, whereas values close to 1 indicate high similarity and predictive power. Thus, higher values for R^2 are desirable.

The R^2 measure has an upward bias and tends to overestimate the statistical significance of the predictive power of a regression model (Chatterjee & Simonoff 2013). This is corrected for by using the adjusted coefficient of determination (R_a^2) measure. As given by:

$$R_a^2 = R^2 - \frac{p}{n-p-1}(1 - R^2) \tag{3.23}$$

It adjusts the R^2 using the $\frac{p}{n-p-1}$ multiplier to incorporate the complexity of the regression model into the measure as shown below (Chatterjee & Simonoff 2013). Thus, if predictors that are not statistically significant are added to the regression model, the number of predictor variables (p) will increase and thus the complexity of the model increases. However, the predictive power of the model stays the same and thus the R_a^2 value for the model decreases. Thus, when comparing a set of models, the model with the highest R_a^2 value is desirable.

3.5.2 The Akaike information criterion (AIC)

The Akaike information criterion (AIC) estimates the quality of a regression model by using the loss of information due to the difference between the regression model and the true nature of the relationship between the predictor and target variables (Chatterjee & Simonoff 2013). This criterion is given by:

$$AIC = n\log(\hat{\sigma}^2) + n\log[(n-p-1)/n] + 2p + 4 \tag{3.24}$$

When comparing a set of models, the model with the lowest AIC statistic is desirable. In accordance with the principle of parsimony, the AIC statistic rewards models that have a good strength of fit (as represented by $\hat{\sigma}^2$) and penalises models with a high amount of predictor variables. A large amount of predictor variables is associated with high model complexity. This may lead to overfitting as increasing the number of predictors improves the model's strength of fit.

3.5.3 Model selection approach

Chatterjee & Simonoff (2013) suggested the following approach for selecting the 'best' model from a set of non-nested regression models:

1. Run a multiple regression analysis for all the models being compared (all possible combinations of predictor variables) and collect all the relevant statistics for each analysis.
2. When comparing the models, arrange the models chronologically; first, the models with the single predictor variables, then the two predictor variable models, and so on.
3. Keep increasing the predictor variables in the model until the value of R^2 stabilises (the R^2 statistic for successive consecutive models tends to a constant value). This implies that further increasing the number of predictor variables in the model only increases the complexity of the model and does not provide additional predictive power.
4. From this set of models choose the model that maximises R_a^2 and minimises AIC.

3.6 Conclusion

This chapter discussed the research methods that were used in this study. It provided an overview of the techniques and methods used to develop metrics from real-world data, and the models used to assess the normalisation potential of these metrics were also provided. In Chapter 4 and Chapter 5, this methodology is applied to real-world data to develop and assess several normalisation models.

Chapter 4 Quantifying the function of a hospital

This chapter is part of the second phase of the research methodology illustrated in Figure 3.2. It describes the application of the information theory approach outlined in Section 3.3 to real-world data. It also discusses the data analysis used to formulate measures for the complexity and level of specialisation of a hospital's caseload. The figure below, Figure 4.1: Thesis document outline: Chapter 4 contextualised , contextualises the current chapter in the thesis document.

A detailed overview of the variables used to formulate these measures is presented in Section 4.1. The data collection process and the dataset collected for analysis are discussed in Section 4.3. Section 4.4 discusses the data analysis process and the application of the information theory approach. In Section 4.5, the results of the data analysis are presented, and the formulated measures are discussed.

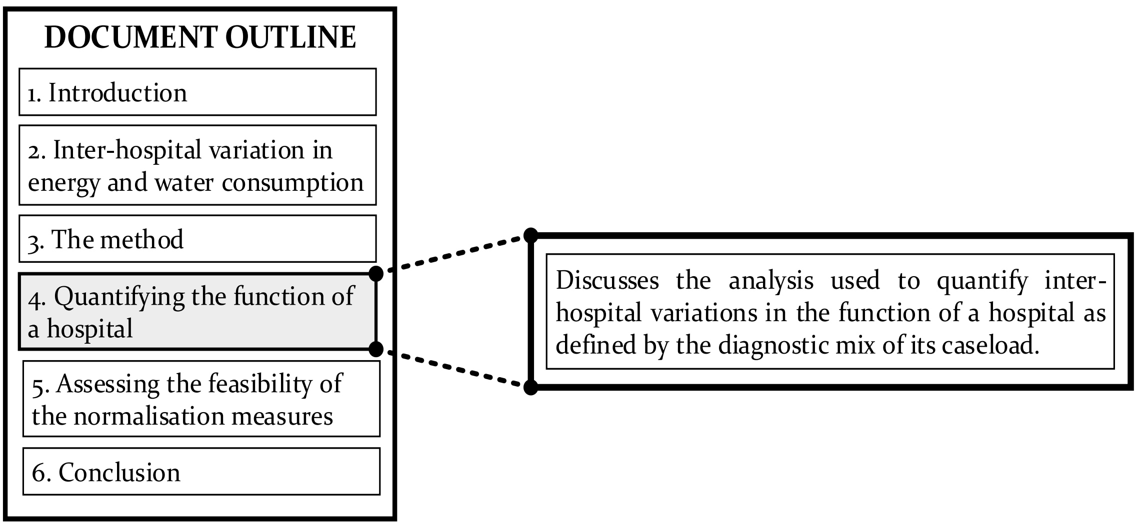


Figure 4.1: Thesis document outline: Chapter 4 contextualised

4.1 Classifying the caseload of a hospital

This section discusses and defines the variables used to formulate metrics for the complexity and level of specialisation of a hospital's medical caseload. The information theory approach proposed by Evans & Walker (1972) was used to develop the set of measures for quantifying and capturing the inter-hospital variations in case mix composition.

As outlined in Section 3.3, this approach uses the distribution of the number of cases treated per diagnostic case type across a set of hospitals to develop measures that are representative of the complexity and level of specialisation of the medical caseload of each of the hospitals in the analysis. The caseloads are classified in terms of the range and types of diagnosis treated by the medical services provided by a hospital (Barer 1982). A treated case, so to say a patient that leaves the hospital (through discharge, death or referral) was used as a measure of a hospital's output.

This classification required a comprehensive specification of the caseload of all the hospitals being analysed. Emphasis was placed on two sets of considerations when formulating and selecting indicators for this specification:

- Generalisability, comprehensiveness and objectivity: This ensured that the indicators are applicable at any district hospital, and that the results achieved using them is accurate and will facilitate a more robust comparison of resource consumption in hospitals.
- Data availability and accessibility: Data on the indicators selected must be available and accessible to allow for the quantitative testing of the said sub-indicator.

The individual treated cases that make up the caseload of a hospital vary significantly across hospitals with respect to the type of diagnosis treated and the severity of these diagnoses. Thus, a classification scheme was needed that was capable of accurately capturing this variability in the diagnostic proportions of hospitals. The South African version of the 10th Revision of the International Classification of Diseases Master Industry Table (ICD-10 MIT) was used to define the categories used to identify and group the different cases treated by the hospital.

The ICD-10 MIT medical coding system is the health industry standard for recording and reporting patient diagnoses and medical records at all public and private medical facilities in South Africa (SANDoH 2012b). The ICD-10 MIT is a tabular list that assigns an alphanumeric code to each treated case. This code describes the patient's primary diagnoses[7] to its maximum level of specificity in accordance with the rules and conventions of the World Health Organisation (SANDoH 2014).

The guidelines for the use of this medical coding system in the South African context are specified in SANDoH (2014). The ICD-10 MIT is divided into 23 chapters based on the potential diagnostic condition of a patient. The chapters describe a large grouping of medical conditions based on clinical diagnosis and cause of morbidity. The level of specificity of the classification scheme increases as one moves across the scheme from chapters to subdivision until the patient's primary diagnosis is defined to the highest level of specificity. Each chapter is subdivided into sections and these sections are further subdivided into subsections as shown in the example in Figure 4.2. The sections and subsections are representative of groupings of the nature of the causative organisms and location of illness within each chapter.

This classification scheme was used to classify patients according to their diagnosis. This provided a comprehensive specification of the diagnostic compositions of the caseloads of each of the hospitals in the analysis. The ICD-10 MIT codes assigned to each case in a hospital's caseload were used as the input variables in the information theory analysis. For each ICD-10 MIT code, data was collected pertaining to the total number of cases treated annually for that diagnosis under each respective clinical speciality, at each hospital in the analysis.

The ICD-10 MIT caseload data recorded at hospitals is in primary diagnosis form (see Figure 4.2). This level of specificity was too high for the applications of this study. At this level there are 40,546 potential ICD-10 MIT classifications for each diagnostic case treated by each clinical speciality provided at a hospital. The result obtained from

[7] The condition diagnosed by a physician as the main reason the patient needed treatment (SANDoH 2012b).

applying the information theory approach to data at this level of specificity has a low practical utility.

Thus, instead of applying the information theory approach at the primary diagnosis level, the cases in the caseloads of each hospital in the analysis were classified according to their clinical diagnosis and cause of morbidity as described at the ICD-10 MIT chapter level. At this level there are 23 potential ICD-10 MIT classifications associated with each case treated by a clinical speciality. Since these subsections house conditions with very similar treatment requirements the trade-off between information loss and practical utility gained was acceptable.

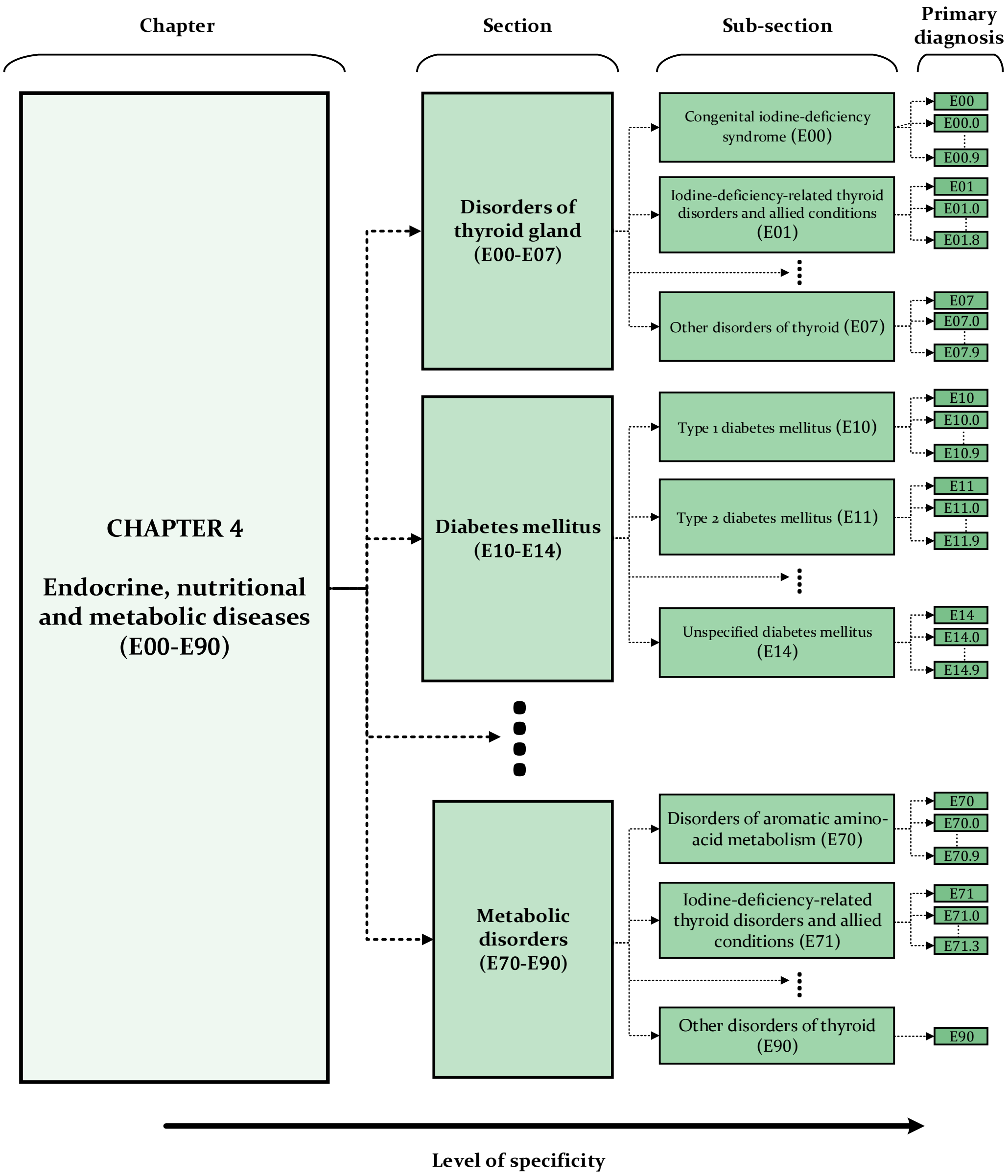


Figure 4.2: Example different levels of specification in the ICD-10 MIT

4.2 Healthcare in the Western Cape

A prerequisite for facility performance benchmarking is that the buildings being compared should have similar characteristics. Minimising the differences between the facilities being compared ensures more meaningful comparisons. This is done by categorising the hospitals into a set of groups that have similar characteristics to reduce the effects of the non-homogeneity in the characteristics of the hospitals at each level of the health system. This ensures that comparisons within categories and across adjacent categories are similar enough to produce meaningful comparisons.

To ensure comparability, the hospitals analysed and compared to each other need to be of similar size and have similar capacities. Thus, it is important to segment the hospitals into a set of appropriate categories that reflect the size and capacity of each hospital in this grouping. Since this approach is designed in the South African context and the data to be used to extend and develop the normalisation approach is based on the energy consumption of public hospitals in the Western Cape Province in South Africa, an understanding of the Western Cape healthcare context was needed.

Public healthcare in the province is delivered via a four-tier hierarchical structure as shown in Figure 4.3. The level of complexity and specialisation of the health service increases towards the top of the structure. The first tier houses the province's largest healthcare facilities. These hospitals offer tertiary and central referral health services in specialised units by highly and uniquely skilled medical personnel that perform advanced diagnostic procedures and treatments (SANDoH 2012a). These hospitals are also attached to a medical school and serve as training institutions for healthcare providers.

The second tier houses the province's acute regional and specialised hospitals. These healthcare facilities offer general surgery, internal medicine, paediatrics, obstetrics and gynaecological care (SANDoH 2012a). The regional hospitals are responsible for the treatment of referral patients from the district hospitals in their catchment area. Specialised hospitals attend to referral patients from primary healthcare clinics and district hospitals who need more sophisticated treatments. Unlike central hospitals, the catchment area of regional and specialised hospitals is limited to the population of the province in which the hospital is situated.

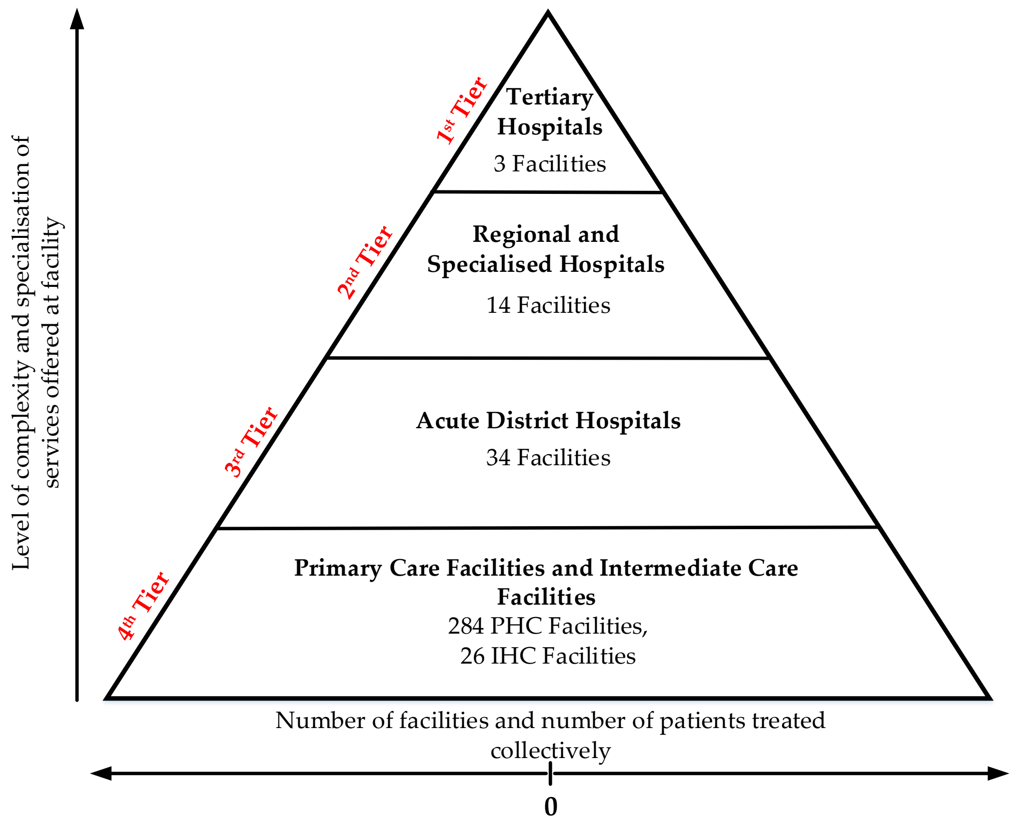


Figure 4.3: Healthcare facilities hierarchy and referral structure in the Western Cape

In the third tier of the health systems hierarchy are the province's acute district hospitals. These facilities provide inpatient, outpatient and emergency care on a 24-hour basis to the population of their respective health districts (SANDoH 2012a). District hospitals also attend to patients who were referred by primary healthcare clinics and need more sophisticated treatment. Primary healthcare clinics make up the fourth tier of the hierarchy. These facilities provide basic medical services and serve as the first point of contact between the patients and the health system. They have a low energy and water footprint because of their small size and capacity. However, the large number of clinics in the province means that their aggregated energy and water footprint is significant.

4.3 The dataset

The proposed research methodology uses data recorded by the Western Cape Department of Health (WCDoH) as a basis for investigating the feasibility of accounting for the case mix of a hospital when comparing the energy and water consumption of hospitals. This section discusses the collection of the dataset used in the analysis and the process by which this dataset was cleaned and transformed into a dataset suitable for use in conjunction with the Evans & Walker (1972) information theory approach.

4.3.1 The data collection

The data used in the study was provided by the Western Cape Department of Health (WCDoH) and is secondary in nature (existing hospital data). The WCDoH collects a diverse array of data at the various healthcare facilities in the province for planning, management, research, statistics, and policy formulation. Data used for statistical, management and planning purposes is collected and recorded on a centralised database. More specific data is collected by the directorates within the WCDoH that are interested in the content of this data.

For research purposes the WCDoH provides data from the centralised database and other sources upon request from researchers. Researchers are required to apply for access to the data by submitting a written request for data. This is done via the 'Annexure A: application for health data' form (see Appendix D.1). The application form must be accompanied by a comprehensive research proposal outlining the aim of the study and the intended purpose of the requested data. Ethical clearance from the research institution associated with the study must also be provided (see Appendix D.2).

Upon approval of their request the researcher was put into contact with the various directorates within the WCDoH that are relevant to the study. In the case of the analysis discussed in this chapter, the required data was provided by the Directorate of Information Management. The patient statistics data is based on the medical records of the patients that are treated by a hospital, as recorded by the attending physician. These records are captured at each of the individual hospitals and collected in a centralised database by the WCDoH's Directorate for Information Management.

4.3.2 The initial dataset

The Directorate of Information Management at the WCDoH provided data pertaining to the inpatient caseloads of 32 public district hospitals in the Western Cape. A list of the district hospitals is provided in Table E.1 in Appendix E.1.

The inpatient population was selected for analysis because it is the largest and most resource-intensive medical service category in a hospital (Barer 1982). There is a direct linkage between the care provided to inpatients during their stay at a hospital and the resource consumption of a hospital due to its inpatient load as inpatients are housed, fed, treated, and cared for using the resources of the hospital.

District hospitals were selected for the analysis because they are the largest set of hospital type in the Western Cape (see Figure 4.3). Furthermore, the structure of the patient referral system in the province is such that a type of treatment is provided at a level that is appropriate for it. Patients report to the lowest level of care in the system for treatment first and are only referred up the referral structure if care cannot be provided at that level (SANDoH 2002). Thus, the patient population of these hospitals was treated as a subset of the province's overall patient population that could be studied in isolation.

The South African National Department of Health's norms and standards for a district hospital's service package (SANDoH 2002) and the WCDoH's definitions for acute hospital packages of care WCDoH (2009) were used as the basis for defining the types of inpatient clinical specialities available at district hospitals.[8] These inpatient clinical specialities were categorised into the macro-functional groups found in district hospitals, as shown in Figure 4.4.

[8] These documents outline the clinical treatments and procedures provided, the competency and skills requirement of the staff at the hospital, and the type of equipment needed to provide treatment at district hospitals.

Emergency medicine	- Emergency medicine - Non-specialist as per DoH
Orthopaedics	- Orthopaedics - Non-specialist as per DoH
Psychiatry	- General psychiatry - Non-specialist as per DoH
Internal medicine	- Family practice - General medicine - Non-specialist as per DoH
Obstetrics & gynaecology	- Gynaecology - Obstetrics - Non-specialist as per DoH
Paediatrics	- General paediatrics - Paediatrics emergency medicine - Non-specialist as per DoH
Surgery	- Dental Surgery - Ear, Nose and Throat - General Surgery - Ophthalmology - Plastic Reconstructive Surgery - Trauma - Urology - Non-specialist as per DoH

Figure 4.4: Categories of clinical inpatient activities at district hospitals

The initial dataset was in the form of a Microsoft Excel workbook containing:

- the ICD-10 MIT description of each diagnosis that was treated at each district hospital;
- the ICD-10 MIT code corresponding to said description; and
- the total number of patients treated with that diagnosis for the year 2016.

The ICD-10 MIT descriptions of this initial dataset were organised according to the macro-category and clinical disciplines presented in Figure 4.4. An excerpt from the initial dataset is shown in Figure 4.5.

Hosp Code	Speciality	Description	ICD 10 Description	ICD 10 Code	2016	Grand Total
MON	Gynaecology	Gynaecology	Abnormal uterine and vaginal bleeding, unspecified	N93.9	28	28
			Contraceptive management, unspecified	Z30.9	1	1
			Diabetes mellitus in pregnancy, unspecified	O24.9	1	1
			Dyspnoea	R06.0	2	2
			Ectopic pregnancy, unspecified	O00.9	6	6
			Endometriosis, unspecified	N80.9	1	1
			False labour before 37 completed weeks of gestation	O47.0	1	1
			Haemorrhoids in pregnancy	O22.4	1	1
			Low back pain, sacral and sacrococcygeal region	M54.58	1	1
			Low back pain, site unspecified	M54.59	1	1
			Medical abortion, complete or unspecified, without complica	O04.9	1	1
			Migraine, unspecified	G43.9	1	1
			Missed abortion	O02.1	41	41
			Nausea and vomiting	R11.X	2	2
			Nonobstructive reflux-associated chronic pyelonephritis	N11.0	1	1
			Other and unspecified abdominal pain	R10.4	3	3
			Other specified noninflammatory disorders of vagina	N89.8	1	1
			Pain localized to other parts of lower abdomen	R10.3	6	6
			Pre-eclampsia, unspecified	O14.9	3	3
			Rheumatism, unspecified, site unspecified	M79.09	1	1
			Spontaneous abortion, complete or unspecified, without con	O03.9	2	2
			Spontaneous abortion, incomplete, complicated by genital tr	O03.0	2	2
			Spontaneous abortion, incomplete, with other and unspecifi	O03.3	1	1
			Sterilization	Z30.2	22	22
			Threatened abortion	O20.0	1	1
			Unspecified abortion, incomplete, without complication	O06.4	1	1
			Urinary tract infection, site not specified	N39.0	3	3
			(blank)	11	1	1
				12	1	1
				I10	2	2
				(blank)	1	1
		Gynaecology Total			140	140
	Gynaecology Total				140	140

Figure 4.5: Excerpt from initial dataset for the formulation of measures for the complexity and level of specialisation

4.3.3 **Data cleaning**

The data cleaning process was undertaken to prepare the collected data for analysis. This data pertains to the diagnostic proportions of the patient populations of the hospitals in the analyses and is dependent on the ICD-10 coding done at each of the individual hospitals. The first step of the cleaning process was to identify any errors in the collected dataset. This was done in conjunction with the diagnostic coding guidelines specified in (SANDoH 2014). Two types of errors were identified and corrected. In some cases, for example, patients were seen under a speciality but not coded on the system, whereas in other cases patients were coded incorrectly. An example of these errors can be seen in the highlighted area in Figure 4.5.

Table 4.1 details the specific types of errors that were identified. The incorrect entries under each clinical speciality sub-grouping for each hospital were aggregated into one entry named ‘*Error*’ for that clinical speciality at each hospital. An example of this transformation can be seen in Figure 4.6. For the year analysed (2016), the Electronic Discharge Summary system was not implemented at all the hospitals, thus the quality of the ICD-10 coding varied between the respective district hospitals.

Table 4.1: Medical coding errors identified in initial dataset

Error	Example
No ICD-10 Code provided.	(Blank)
The ICD-10 code is not alphanumeric.	51
Three-character ICD-10 code ends on a full stop.	J22.
Three-character ICD-10 code does not contain a full stop.	O02 1

Figure 4.6: Transformation of the error entries

The second phase of the data cleaning process involved identifying hospitals in the dataset whose caseloads contain a high proportion of '*Error*' entries. The '*Error*' entry percentage with respect to the total cases treated at a hospital was calculated for all the hospitals in the analysis. Figure 4.7 shows the proportion of a hospital's '*Error*' ICD-10 entries vs its total patients treated. The average for the entire dataset is also shown.

The average of these percentages was calculated and used as an exclusion criterion. Hospitals whose '*Error*' entry percentage was higher than the sample average were excluded from the analysis as their caseload composition may potentially skew the results of the information theory analysis. For this reason, only 18 of the initial 32 hospitals were included in the analysis. These hospitals are listed in Table E.2 in Appendix E.2.

Thirdly, as discussed in Section 4.1, the caseload data in the initial dataset was provided to the maximum level of specificity in accordance with the ICD-10 code description of the conditions of the patients in each hospital's patient population. This inpatient data collected was too specific for use in the analysis. At this level of specificity, the total provincial patient population consisted of 40,547 potentially different types of medical diagnosis treated under each clinical speciality at each of the district hospitals in the province as specified by the respective ICD-10 MIT codes.

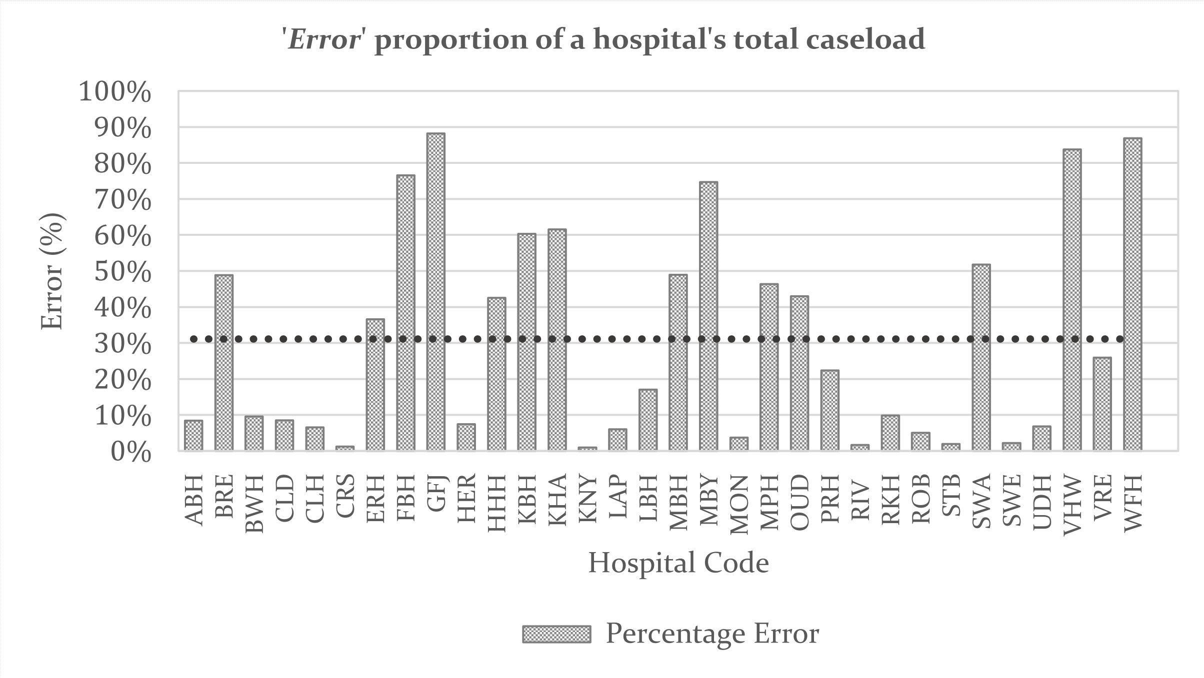


Figure 4.7: Proportion of a hospital's caseload that was an '*Error*' ICD-10 entry in the initial dataset

Thus, the caseload of each macro inpatient clinical speciality of the hospitals in the analysis, as described in the initial dataset, was categorised and condensed into the macro groups of medical classifications defined in the ICD-10 MIT as introduced in Section 4.1 (see Table C.1 in Appendix C for a full list of the groups). This generated finer groupings for each clinical speciality. The number of patients for each of the consolidated category groupings is the sum of the conditions that are housed in that grouping at that hospital (see Figure 4.8). This reduced the overall possible classifications and thus reduced the number of potential input variables in the analysis from 40,547 to 23 for each of the clinical specialities.

Hosp Code	Speciality	Description	ICD 10 Code	2016
ABH	Medicine	General Paediatrics	A09.0	6
ABH	Medicine	General Medicine	A09.0	11
ABH	Medicine	General Paediatrics	A09.9	64
ABH	Maternity	Obstetrics	A09.9	3
ABH	Surgery	General Surgery	A09.9	1
ABH	Medicine	General Medicine	A09.9	49

Hosp Code	Speciality	Description	ICD-10 Code	2016
ABH	Medicine	General Medicine	(A00-B99)	154
ABH	Medicine	General Medicine	(C00-D48)	33
ABH	Medicine	General Medicine	(D50-D89)	15
ABH	Medicine	General Medicine	(E00-E90)	125
ABH	Medicine	General Medicine	(F00-F99)	66

Figure 4.8: Consolidating ICD-10 codes into ICD-10 groupings

4.3.4 The final dataset

The final dataset consisted of 4,122 data entries distributed in an 18 by 229 matrix that represented the caseloads of 18 hospitals classified into to 229 diagnostic groupings across 23 clinical specialities. A total of 278,577 inpatient cases were treated at the district hospital level in the Western Cape in 2016. Of these cases, 91,939 were treated at the 18 hospitals analysed in this study. Figure 4.9 displays an excerpt from the final dataset. The matrix entry in cell *ij* represents the total number of cases treated at hospital *i* in 2016, that belong to diagnostic case category *j*.

Hospital code	Medicine, General Medicine, (A00-B99	Medicine, General Medicine, Error	Medicine, General Medicine, (C00-D48	Medicine, General Medicine, (D50-D89	Medicine, General Medicine, (E00-E90)	Medicine, General Medicine, (F00-F99)	Medicine, General Medicine, (G00-G99	Medicine, General Medicine, (H00-H59	Medicine, General Medicine, (H60-H95	Medicine, General Medicine, (I00-I99)	Medicine, General Medicine, (J00-J99)
ABH	154	87	33	15	125	66	61	3	2	138	259
BWH	223	217	45	30	99	19	45	2	3	200	263
CLD	196	158	57	42	91	57	120	2	1	304	257
CLH	223	158	45	25	117	40	78	10	9	273	284
CRS	718	65	105	139	394	126	224	2	2	770	940
HER	271	330	65	46	117	53	175	2	7	325	347
KNY	455	33	130	72	177	160	168	6	11	491	380
LAP	34	46	5	8	57	30	44	1	0	93	103
LBH	39	138	2	16	48	18	27	1	4	126	90
MON	120	69	28	50	60	24	90	3	1	185	265
PRH	48	309	18	8	34	31	49	10	2	52	104
RIV	196	54	92	45	140	55	86	13	5	278	366
RKH	122	188	32	26	83	46	77	5	1	148	267
ROB	195	88	42	46	129	8	132	1	1	269	311
STB	540	59	14	112	534	20	101	0	0	454	116
SWE	181	63	101	41	115	29	87	2	2	184	382
UDH	35	53	13	5	40	7	39	0	4	44	91
VRE	311	1565	58	47	228	116	170	39	14	333	392

Figure 4.9: Excerpt from the final dataset

4.4 The data analysis

The objective of this analysis was to apply the information theory approach proposed by Evans & Walker (1972) to the final dataset, thereby computing measures for the complexity and level of specialisation of the caseloads of each of the hospitals in the analysis.

4.4.1 The data analysis strategy

As introduced in Section 3.3, this approach studies the proportions of the case mix compositions of the respective hospitals and uses them to develop measures of the complexity and level of specialisation associated with the caseloads of the hospitals. These measures are potential normalisation factors that will be used when comparing the energy and water consumption performance of hospitals. Figure 4.10 provides an outline of the data analysis process and the application of the Evans & Walker (1972) approach to the final dataset.

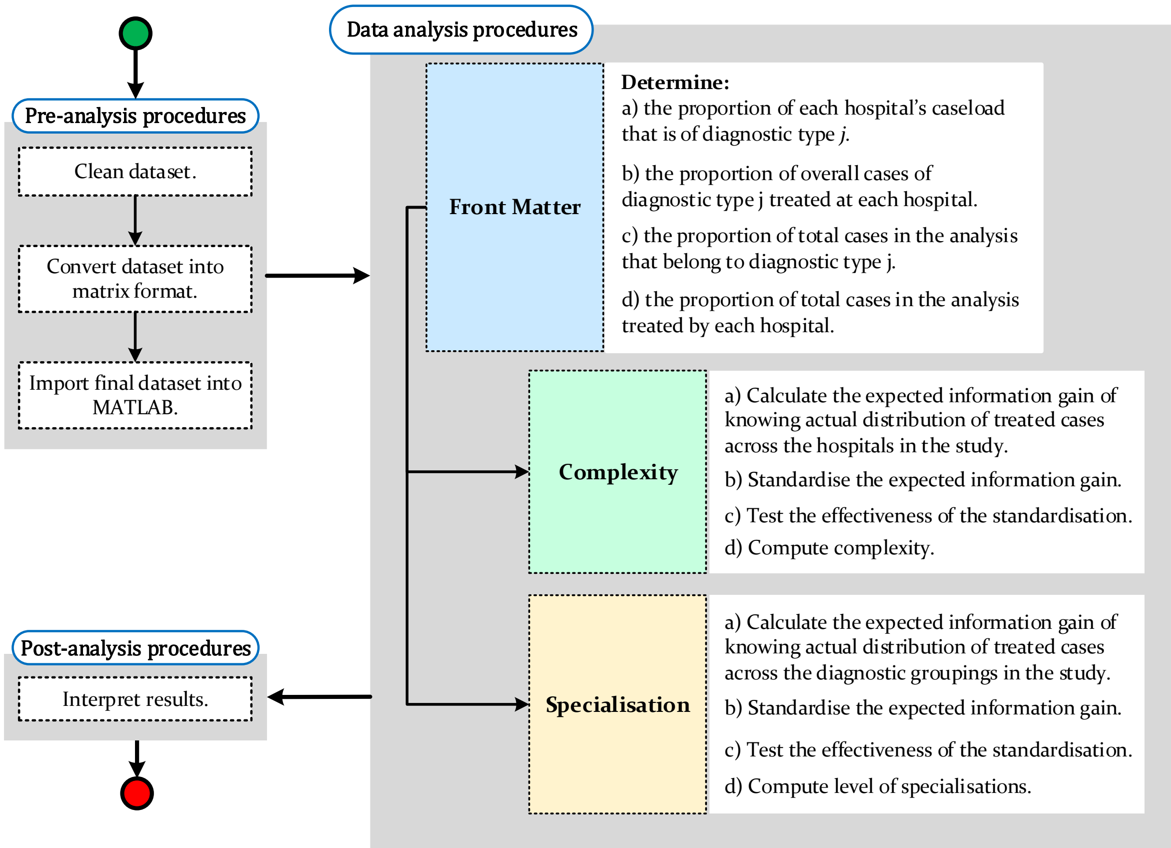


Figure 4.10: Data analysis process for the application of the Evans & Walker (1972) approach

4.4.2 Expected information gain

The data analysis process was centred around calculating the expected information gain associated with each hospital's caseload composition. The approach compared what was known about the entire system (the set of hospitals being analysed and their diagnostic proportions) before data collection, to what was known after data collection and analysis. It then measured the 'information gain' associated with the new knowledge generated about the system through data collection and analysis.

The prior probability represents the state of information before data collection. It is assumed that prior to data collection, only the number of hospitals in the analysis was known. Thus, patients are assumed to be equally distributed across the hospitals in the analysis. That is, the proportion of the provincial patient population for each diagnostic category treated at the district hospital level in the Western Cape Province[9] is distributed equally across each of the hospitals in the analysis. Thus, the reciprocal of the number of hospitals $\left(\frac{1}{N}\right)$ was used as the prior probability when calculating the expected information gain.

After data on the case mix compositions of the hospitals was collected and analysed, the actual distribution of the diagnosis at each of the hospitals is known. The effects of each hospital's capabilities and characteristics is reflected in the information that is known about each hospital's patient population. Thus, the proportions of the diagnostic distribution are used to calculate the posterior probabilities associated with each diagnostic grouping.

There are differences between the prior probabilities and the posterior probabilities of each diagnostic grouping at the respective hospitals. These differences are a result of variations in the actual distribution of patients across the different diagnostic groupings at the hospitals in the analysis. The expected information gain metrics measure the differences between the prior probability (calculated using the assumed

[9] The study initially considered all the district hospitals in the province. However, due to data quality challenges some of the hospitals had to be omitted from the analysis. Going forward, the assumptions related to studying the entire patient population of the district hospitals in the province are maintained.

distribution of patients) and the posterior probability (calculated using the actual distribution of patients) at each hospital in the analysis.

In this analysis the idea was to measure the degree of concentration of each case type at each hospital. Concentration was used as a measure of complexity and specialisation. The more concentrated a case type, the more its posterior probability differs from its prior probability, and thus the higher its expected information gain. The expected information gain associated with the case distribution of each diagnosis in the analysis was used to quantify complexity and level of specialisation.

The standardised version of the expected information gain measure, corrected for any potential upward bias, was used when computing complexity and level of specialisation. The approach used the standardised expected information gain measures for each of the respective diagnostic categories to compute complexity measures for the caseload of each hospital in the analysis. The complexity metric (*CMPX*) for a hospital's case mix is calculated as the weighted sum of the number of patients treated at the hospital that fall into each diagnostic category. For each diagnostic category, the weights are the standardised expected information gain measure associated with that diagnostic category.

For the level of specialisation metric, the focus was on the distribution of each hospital's patient population amongst the range of possible diagnoses that the hospital could treat. To study the distribution of patients amongst the respective diagnostic categories at each hospital, the prior probability changed to the proportion of the overall provincial patient population that is treated under diagnostic group j, while the posterior probability is the actual number of patients treated under diagnostic group j at each of the respective hospitals.

The expected information gain was calculated and standardised in the same way as for the complexity metric. However, the level of specialisation (*SPEC*) metric for each hospital's caseload was the weighted sum of the number of patients treated under each diagnostic category at each hospital i. For each diagnostic category, the weights were the standardised expected information gain measure associated with that diagnostic category.

4.4.3 MATLAB analysis

The data analysis was conducted in MATLAB R2018a. This is a programming language for solving complex equations and analysing data structures. MATLAB was selected because it is suitable for doing an array of complex analysis on large datasets, especially when that data is in matrix form. Furthermore, Stellenbosch University has a Total Academic Headcount MATLAB license and thus the program was readily available to the author.

The final dataset was imported into MATLAB and a script file was written to conduct the matrix calculations required for the data analysis (see Appendix F). The MATLAB script file was divided into 5 sections, in accordance with the data analysis procedure outlined in Figure 4.10. The first phase of the analysis imported the content of the final dataset's 18 x 229 matrix into MATLAB from a Microsoft Excel workbook. It then computed the parameters pertaining to the dataset that were used in the subsequent phases of the analysis to assess the distribution of the diagnostic cases across the respective diagnostic groupings at each hospital. These parameters are listed in Table 4.2.

Table 4.2: Parameters associated with the dataset matrix

Parameter	Definition
N	Number of hospitals in the analysis
$\boldsymbol{C_i}$	18 x 1 vector detailing the total number of cases treated by each hospital in the analysis
$\boldsymbol{C_j}$	1 x 229 vector detailing the total number of cases in the analysis that belong to each diagnostic category
C	Total number of cases in the analysis

In the second phase, the parameters calculated in phase 1 and the dataset matrix were used to define the diagnostic proportions associated with the respective aspects of the caseloads of the hospitals in the analysis. Table 4.3 presents the ratios associated with each of the respective caseload proportions used when calculating the expected information gains. These proportions are contained in the: 18 x 229 **P** matrix, 18 x 229 **Q** matrix, the 18 x 1 $\boldsymbol{P_i}$ vector, and the 1 x 229 $\boldsymbol{Q_j}$ vector. The entries of these matrices and vectors detail the ratios described in Table 4.3 that are associated with each parameter for each hospital or diagnostic category.

Table 4.3: Caseload proportion ratios

Q	***Caseload proportion:*** The proportion of each hospital's caseload that is of diagnostic type *j*. ***Associated ratio:*** $\frac{\text{Diagnostic proportions of hospital i's caseload}}{\text{Total number diagnosis j cases treated at all hospitals}}$
P	***Caseload proportion:*** The proportion of overall cases of diagnostic type j treated at each hospital. ***Associated ratio:*** $\frac{\text{Diagnostic proportions of hospital i's caseload}}{\text{Total number cases treated at hospital i}}$
$\boldsymbol{Q_j}$	***Caseload proportion:*** The proportion of total cases in the analysis that belong to diagnostic type j. ***Associated ratio***: $\frac{\text{Total number diagnosis j cases treated at all hospitals}}{\text{Total number cases treated by hospitals in the analysis}}$
$\boldsymbol{P_i}$	***Caseload proportion:*** The proportion of total cases in the analysis treated by each hospital. ***Associated ratio:*** $\frac{\text{Total number cases treated at hospital i}}{\text{Total number cases treated by hospitals in the analysis}}$

As discussed in Section 3.3.2, the complexity metric (CMPX) of a hospital's caseload was calculated using the standardised expected information gain measure associated with each diagnostic case type and the proportions of the hospital's caseload falling into that diagnostic category. In phase 3, this was achieved by substituting the **Q** matrix and $\boldsymbol{Q_j}$ vector into equation (4.1) to determine the expected information gain and then standardising the expected information gain by applying equation (4.2).

$$\mathrm{EIG}_{CMPX} = \sum_{n=1}^{18} \boldsymbol{Q}\, ln(N\mathbf{Q}) \tag{4.1}$$

$$\overline{EIG}_{CMPX} = \frac{\mathrm{EIG}_{CMPX}}{\sum_{m=1}^{229} \mathrm{EIG}_{CMPX} \cdot Q_j} \tag{4.2}$$

These equations are applied in their current form as MATLAB allows for element by element operations when doing matrix calculations. This feature of the software conducts independent scalar calculations between the respective entries of matrices and vectors. This allows one to conduct calculations with vectors and matrices whose dimensions would be problematic under conventional linear algebraic manipulations. Thus, the expected information gains associated with each of the 229 diagnostic groupings are calculated both simultaneously and independently by applying equation (4.1).

In phase 4, the **P** matrix, $\boldsymbol{P_i}$ vector, and $\boldsymbol{Q_j}$ vector were substituted into equations (4.3) to (4.4) to calculate and standardise the expected information gains which were used in the calculation of the specialisation metric (SPEC) for the caseloads of the respective hospitals.

$$\mathrm{EIG}_{SPEC} = \sum_{m=1}^{229} \mathbf{P}\, ln\left(\mathbf{P}/\boldsymbol{Q_j}\right) \tag{4.3}$$

$$\overline{EIG}_{SPEC} = \frac{\mathrm{EIG}_{SPEC}}{\sum_{n=1}^{18} \mathrm{EIG}_{SPEC} \cdot \boldsymbol{P_i}} \tag{4.4}$$

The fifth phase of the data analysis as described in the MATLAB script file exports the results of the analysis to a Microsoft Excel workbook. The exported CMPX and SPEC metrics associated with the caseload of each hospital in the analysis are recorded into a workbook for use in the next phase of the research process.

4.5 Results

This section discusses the results of the MATLAB analysis that was used to formulate the complexity and specialisation measures for the caseloads of the respective hospitals in the analysis. The MATLAB analysis outputted two 1 x 18 vectors detailing the complexity and level of specialisation measures attributed to the caseload of each of the hospitals in the analysis. These results are presented in Table 4.4. These measures are relative measures and represent the complexity and specialisation of the caseload of each hospital relative to the caseloads of the other 18 hospitals in the analysis.

Table 4.4: The complexity and level of specialisation metric associated with the caseload of each hospital

Hospital	Hospital Code	CMPX	SPEC
Ladismith (Alan Blyth) Hospital	ABH	21.78	1.03
Beaufort West Hospital	BWH	48.01	1.62
Caledon Hospital	CLD	39.94	0.91
Clanwilliam Hospital	CLH	31.31	0.77
Ceres Hospital	CRS	114.70	0.97
Hermanus Hospital	HER	56.56	0.59
Knysna Hospital	KNY	73.18	0.62
LAPA Munnik Hospital	LAP	9.01	1.30
Laingsburg Hospital	LBH	10.80	0.94
Montagu Hospital	MON	23.73	0.83
Prince Albert Hospital	PRH	16.60	1.40
Riversdale Hospital	RIV	38.93	0.77
Radie Kotze Hospital	RKH	23.60	0.99
Robertson Hospital	ROB	36.37	0.82
Stellenbosch Hospital	STB	59.30	1.28
Swellendam Hospital	SWE	29.55	0.86
Uniondale Hospital	UDH	8.35	3.73
Vredendal Hospital	VRE	73.29	1.15

4.5.1 **Complexity of caseloads**

This section discusses the complexity measures calculated in the MATLAB analysis using the distributions of the respective diagnostic case type groupings defined in Table C.1 in Appendix C across the hospitals studied. The MATLAB analysis studied the expected information gain (EIG) associated with each of the cases treated at each hospital to assign complexity measures to the caseloads of the hospitals in the analysis. The analysis calculates a standardised EIG measure for each of the potential 23 diagnostic groupings that can be treated under each clinical speciality. These standardised EIG measures represent the complexity of their respective diagnostic case types.

An EIG measure was calculated and assigned to each diagnostic case type in the study based on the degree of concentration of each case type across the range of hospitals studied. The degree of concentration was calculated using the difference between the prior and posterior probabilities attributed to each diagnostic case type. These probabilities represent the likelihood that a patient of diagnostic case type *j* has of being treated at hospital *i* according to what is known before data collection and analysis (prior probability), and after data collection and analysis (posterior probability).

The lowest possible information gain measure for each diagnostic case type *j* at each hospital is zero (0). It represents the case where the prior probability and posterior probabilities of diagnostic case type *j* at hospital *i* are equal. If the difference between the prior and posterior probability of diagnostic case type *j* at hospital *i* is small, then consequently the information gain measure associated with case type *j* at hospital *i* will also be small. In contrast, if the difference between the two probabilities is large, then the information gain measure will be large.

The expected information gain (EIG) for diagnostic case type *j* is the weighted sum of all the information gain measures of that case type at all the hospitals in the analysis, where the weights are the proportion of overall cases of diagnostic type *j* treated at each hospital. Straightforward diagnostic case types are associated with a low EIG measure as they can be treated at any of the hospitals in the analysis. Complex diagnostic case types are concentrated at a few hospitals that specialise in their treatment. This concentration results in a high EIG measure due to a large difference between the prior and posterior probabilities of the case type at these hospitals.

The box and whiskers plot provided in Figure 4.11 presents a visual representation of the distribution of the standardised EIG associated with each of the ICD-10 diagnostic groupings for each of the clinical specialities present at the hospitals in the analysis. The whiskers represent the minimum and maximum standardised EIG attributed to an ICD-10 diagnostic category in that clinical speciality. The boxes represent the 25^{th} percentile, median and 75^{th} percentile of the standardised EIG of the ICD-10 diagnostic category in that clinical speciality. The boxes are grouped, and colour-coded by clinical speciality grouping. The colour coordination scheme is explained in Table 4.5.

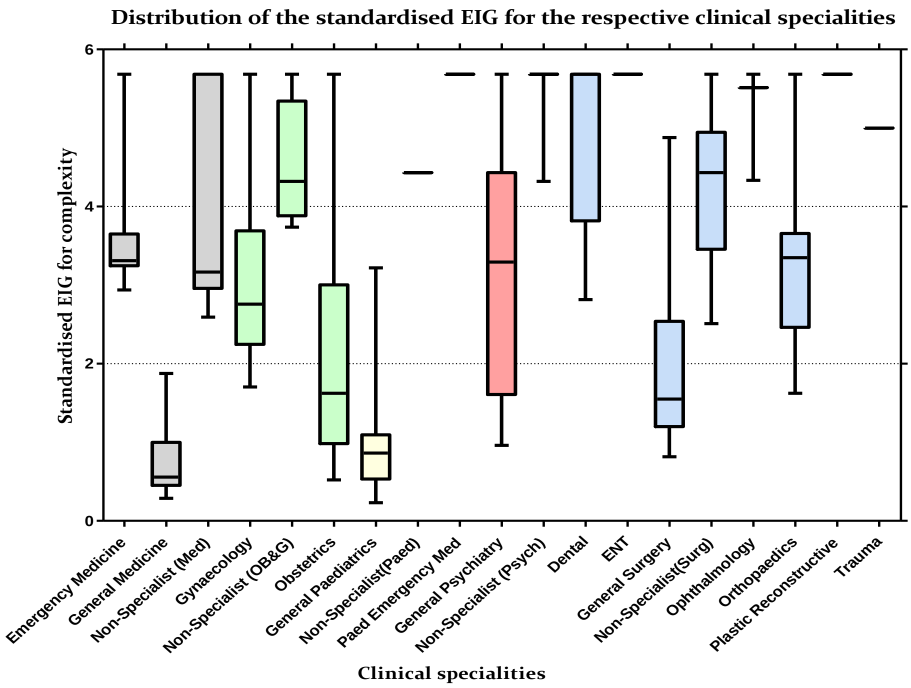


Figure 4.11: Box and whiskers plot of the standardised EIG measure of the clinical specialities found at the hospitals in the study

A comparison of Figure 4.11 and Figure 4.12 shows that the clinical specialities that are provided at a large number of hospitals as shown (see Figure 4.12) correspond to having low standardised EIG values (see Figure 4.11). In the analysis, only 4 clinical specialties are offered at an average of more than 10 hospitals, namely: General Medicine, Obstetrics, General Paediatrics, and General Surgery. These four clinical specialities also have the lowest standardised EIG measures of the clinical specialities in the analysis as represented by their interquartile range, as illustrated in the box and whiskers plot shown in Figure 4.11.

Table 4.5: Legend for colour scheme used Figure 4.11 and Figure 4.12.

Clinical speciality grouping	Colour
Medicine	Grey
Obstetrics & gynaecology	Green
Paediatrics	Yellow
Psychiatry	Red
Surgery	Blue

The three clinical specialities that are offered by the least number of hospitals in the analysis (an average of only one hospital) are: Paediatric Emergency Medicine, Ear Nose and Throat Surgery, and Plastic Reconstructive Surgery. These clinical specialities have the highest standardised EIG measures in the analysis. In accordance with the assumptions of the Evans & Walker (1972) approach, these diagnostic case type groupings are deemed to contain complex medical diagnosis because of their low distribution amongst the hospitals in the analysis. The treatment of the diagnoses that fall into these groupings is concentrated at a single hospital.

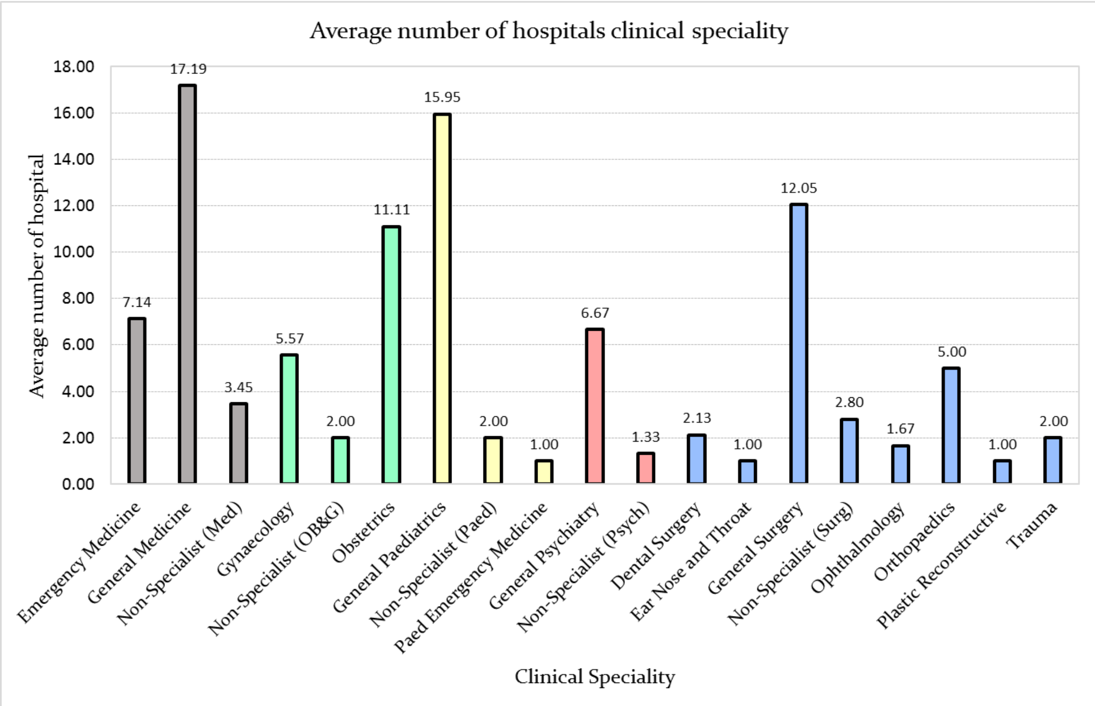


Figure 4.12: Average number of hospitals that provided treatment in each clinical speciality

A correlation analysis of the standardised EIG measure attributed to each ICD-10 diagnostic category under each clinical speciality vs the average number of hospitals that treated cases in that diagnostic category was also conducted. The results of the correlation analysis are shown in Figure 4.13. The results of the analysis confirm that as observed in the comparison of Figure 4.11 and Figure 4.12, there exists a negative and significant relationship between the standardised EIG measures and the average number of hospitals treating the respective diagnostic case type (R^2= 0.8511, and P value <0.0001).

The complexity measure for the caseloads of each of the hospitals in the analysis is the weighted sum of the cases in the caseload of each hospital, where the weights are the complexity metric associated with each ICD-10 diagnostic group for each clinical speciality in the caseload of the hospitals as defined by their standardised EIG measures. The complexity score of the caseload of each hospital is characterised by the proportion of the hospital's caseload that is distributed into the respective 'complex' diagnostic category's vs the proportion in the 'straightforward' diagnostic categories.

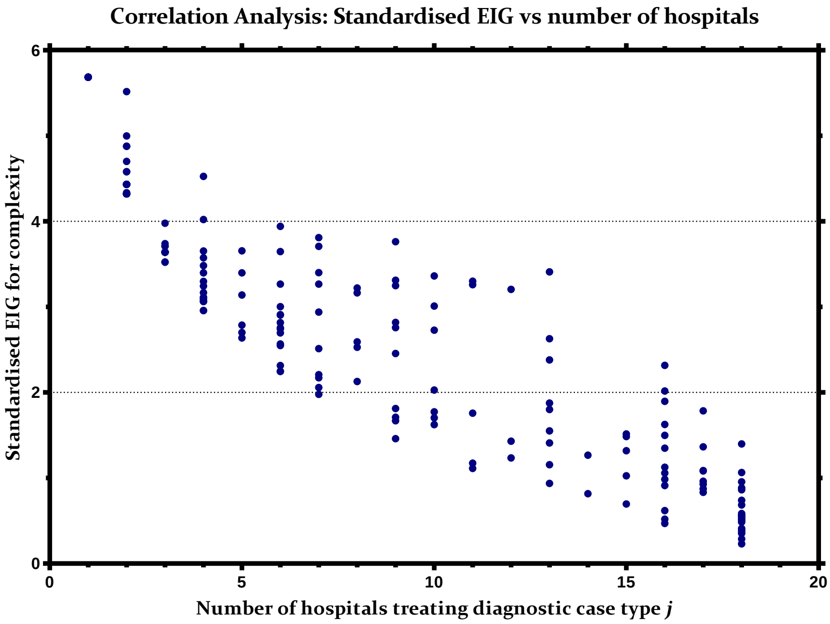


Figure 4.13: Correlation analysis: standardised EIG vs number of hospitals treating diagnostic case type *j*

A comparison of the complexity measures of each hospital to the number of diagnostic case types treated by each hospital was also conducted. The comparison found that the hospital with the highest complexity measure (Ceres Hospital: CMPX = 114.70) treated the second most-diagnostically diverse array of cases. Ceres Hospital (CRS) treated 11 less diagnostically different types of cases than Knysna Hospital (KNY). However, overall CRS treated 36.19 percent more cases than KNY. Also, it treated a larger proportion of the rarer diagnostic case types than KNY. Hence, its CMPX score is higher than that of KNY.

Furthermore, the hospital that had the lowest complexity measure (Uniondale Hospital: CMPX = 8.35) treated the least-diagnostically diverse set of cases, and as expected diagnostic cases that were common amongst the other hospitals in the study. Appendix G.2.2 details the result of a correlation analysis between the complexity measures assigned to the caseload of the respective hospitals and the number of diagnostically different cases treated by each of the hospital in the analysis. The correlation analysis found a positive and significant relationship (R^2= 0.7091, and P value <0.0001) between them. This confirmed that the complexity of a hospital's caseload increases as the number of diagnostically different cases treated by a hospital increase.

Furthermore, As shown in Table G.1 in Appendix G.1, the complexity measures of the hospitals in the analysis have a mean of 39.72 and a standard deviation of 27.57. The scatter amongst the CMPX measures is low, with 17 of the 18 hospitals CMPX values clustered around the mean (see Figure 4.14). The CMPX value for the caseload of Ceres Hospital differs significantly from those of the other hospitals in the analysis. This is due to the size (number of cases treated) and diagnostic diversity of its caseloads.

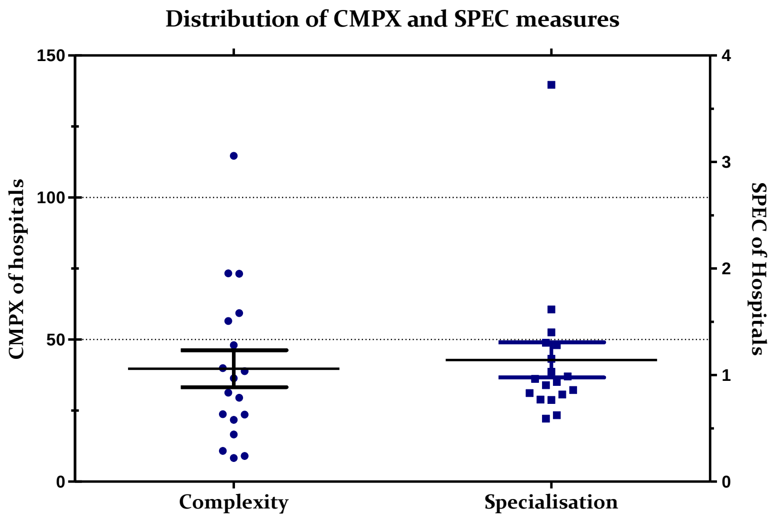


Figure 4.14: Scatter diagram showing the distribution of the complexity and specialisation measures calculated for the hospitals in the analysis

4.5.2 Level of Specialisation

This subsection discusses the results of the MATLAB analysis that formulated the level of specialisation measures (SPEC) associated with the caseloads of the respective hospitals in the analysis. The SPEC measure captures the degree of concentration of diagnostic effort at a hospital (Barer 1982). The measure gauges whether a hospital specialises in treating a select range of diagnoses from a limited set of clinical specialities, or whether the hospital is more generalist and thus caters for a diverse portfolio of diagnosis across an array of clinical specialities.

As in the case of the complexity measures discussed in Section 4.5.1, the level of specialisation measure assigned to a hospital's caseload is a relative metric. It gauges the level of specialisation observed in the caseload of hospital *i* relative to the caseloads of the other hospitals in the analysis. Furthermore, this measure is also defined using prior and posterior probabilities. For the level of specialisation measure, the focus was on whether the proportion of cases of diagnostic type *j* treated at hospital *i* (posterior probability) varies from the overall proportion of cases of diagnostic type *j* treated by all the hospitals in the analysis (prior probability).

An expected information gain (EIG) measure was calculated and standardised for the inpatient caseload of each hospital in the MATLAB analysis from the final dataset matrix. The level of specialisation measure calculated for each hospital is equal to the standardised EIG associated with the hospital's level of specialisation. The specialisation measures calculated for each hospital are presented in Table 4.1 above.

The minimum possible value for the level of specialisation measure is zero (0). It represents the case where the posterior and prior probabilities are equal. In this case, no new information is gained from the formulation of the posterior probabilities. As in the case of the complexity measure, the deviations of the proportion of the caseload composition of hospital *i* from the proportion of the overall caseload composition of all the hospitals in the analysis results in changes in the standardised EIG for specialisation, and consequently the SPEC measure associated with hospital *i*'s caseload.

The scatter diagram shown in Figure 4.14 presents a visual representation of the distribution of the level of specialisation measures associated with the inpatient case load of each hospital in the analysis. This distribution of SPEC measures has a mean of 1.143 and a median of 0.9505. The specialisation measures for the caseloads of 17 of the 18 hospitals are clustered around SPEC = 1. The standard deviation is 0.6.99 for the SPEC measures of the 18 hospitals in the analysis.

This increased variability is because of the level of specialisation of Uniondale Hospital (SPEC = 3.73). This SPEC measure is 2.3 times higher than the second-highest specialisation measure in the analysis, Beaufort West Hospital (SPEC = 1.62). Uniondale Hospital has the smallest and least diagnostically diverse caseload of the hospitals in the analysis. The hospital's caseload consists of 1,074 cases classified into 44 different diagnostic case types. However, these two factors are not individually responsible for the high specialisation measure.

Appendix G.2.3 and Appendix G.2.4 show the results of the correlation analyses which studied the relationships between the SPEC measure vs the number of different diagnosis treated by a hospital , and the SPEC measure vs the number of cases treated by a hospital. Both analyses found that the relationship between the variables was not statistically significant, (R^2= 0.1938, and p = 0.0675) and (R^2= 0.09701, and p = 0.2084) for the respective analysis.

There are two potential reasons why the caseload of a hospital would have a large level of specialisation measure (Evans & Walker 1972):

- The hospital is small and does not have the capacity to service a diagnostically diverse portfolio of cases.
- The hospital is large but designed to specialise in the treatment of a limited range of specialised diagnoses.

In the case of Uniondale Hospital, the former reason is applicable. An examination of the caseload composition of Uniondale Hospital found that 89.66 percent of the 1 074 cases treated by the hospital belong to only 20 diagnostic case types from 4 clinical specialities. These case types are also common amongst the caseload composition of the other hospitals in the analysis.

4.6 Conclusion

The complexity and specialisation measures developed in this chapter represent the case mix composition concept discussed in Section 2.4.3. These two measures represent the demand placed on a hospital's energy and water resources by the type and severity of the diagnoses in its patient population. They served as independent variables in the multiple regression analyses that develop normalisation models for the energy and water consumption of hospitals. These analyses are discussed in Chapter 5. They were used to assess the feasibility of using CMPX and SPEC as normalisation factors.

Chapter 5 Assessing the feasibility of the normalisation measures

This chapter discusses the analysis that evaluated the relationship between the prospective normalisation factors, and the energy and water consumption of hospitals. Statistical models were developed to capture the effect of the normalising factors on the consumption performance of hospitals. These models were used to evaluate the feasibility of using the complexity and specialisation measures developed for a hospital's caseload as potential normalising factors when comparing the energy and water consumption of hospitals.

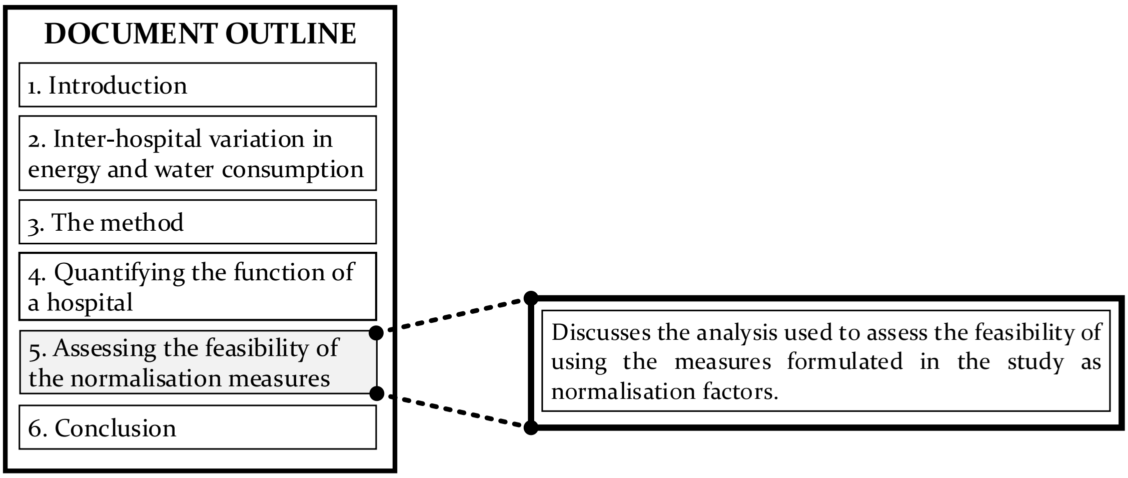


Figure 5.1: Thesis document outline: Chapter 5 contextualised

Section 5.1 provides an overview of each of the potential normalising factors and translates them into the statistical domain. This was achieved through the definition of the variables used to represent the respective normalising factors and the energy and water consumption of hospitals. Section 5.2 provides an overview of the dataset evaluated in the data analysis. The data analysis used to evaluate the relationships between the respective variables is discussed in Section 5.3. The findings of the data analysis were used to make recommendations on the feasibility of using complexity and specialisation as normalising factors. These findings and the recommendations made are discussed in Sections 5.4 and 5.5 for electricity consumption and water consumption respectively.

5.1 The variables investigated in the analysis

This section discusses the variables used to evaluate the feasibility of using the complexity and level of specialisation of a hospital's case mix as normalisation factors when comparing the energy and water consumption performance of hospitals. The set of potential normalisation factors being investigated consists of variables that represent the scale of the hospital facility, the hospital's output, and the diversity of the diagnostic mix of the hospital's patient population. The origin of these factors and the process that facilitated their selection was discussed in Chapter 2.

Five potential normalisation factors were evaluated for inclusion in the normalisation model. These normalisation factors were used as the independent variables in the data analysis. Subsection 5.1.1 introduces the respective normalising factors and discusses the variables that were used to represent these factors in the data analysis. Subsection 5.1.2 discusses the variables that were used to represent the concepts of energy and water consumption in the data analysis. These variables were the dependent variables in the data analysis.

5.1.1 The normalising factors

The normalising factors were classified into two categories: fixed factors and transient factors. The fixed factors represent characteristics of hospitals that are constant over extended periods of time (3 to 5 years). In this analysis these factors are represented by the scale of the hospital's facilities.

Two variables were used to capture the size of a hospital's facilities: its bed capacity and its building footprint. Bed capacity was specified in terms of the number of available inpatient beds at a hospital. The footprint of a hospital building was specified in terms of the total floor area occupied by the hospital building or campus. These variables are described in Table 5.1. They were used to test the extent to which the scale of a hospital's facilities is reflective of resource consumption at that facility.

Table 5.1: Description of the variables used to represent the size of a hospital in the data analysis

<table>
<tr><td>Variable ID:
BED</td><td>Variable units:
[bed]</td><td>Concept represented:
The scale of the hospital facility</td></tr>
<tr><td colspan="3">Aspect measured:
This variable is a measure of the number of beds in a hospital that are staffed and regularly maintained by healthcare practitioners within the hospital. These beds are immediately available for use by a patient admitted for treatment at a hospital.</td></tr>
<tr><td colspan="3">Scope and structure:
The BED variable represents the annual average number of beds available at a hospital. The scope of this variable includes all occupied and unoccupied beds in a hospital that are used for the acute care, rehabilitative care and long-term care of patients. However, the scope excludes temporarily available beds, beds associated with the long-term residential care of patients, and the beds in closed patient wards.</td></tr>
<tr><td>Variable ID:
TFA</td><td>Variable units:
$[\boldsymbol{m}^2]$</td><td>Concept represented:
The scale of the hospital facility</td></tr>
<tr><td colspan="3">Aspect measured:
This variable is a measure of the size of the total space located within the hospital building or campus. It tests the extent to which the total floor area of a hospital is reflective of resource consumption at that hospital.</td></tr>
<tr><td colspan="3">Scope and structure:
This is an aggregated measure of the floor area occupied by all the medical buildings on the hospital campus as well as all support services. This includes all diagnostic and emergency care spaces in a hospital, examination rooms, medical offices, laboratories, corridors and storage areas within the hospital.</td></tr>
</table>

The transient normalising factors represent the characteristics of hospitals that vary over extended periods of time and are dependent on the level of medical service provision at each hospital. In this study these factors are represented by the output of a hospital and the composition of the diagnostic mix of cases treated by the hospital. As introduced in Section 2.4.1, hospital output is specified using the patient day equivalent metric. The PDE variable described in Table 5.2 was used to represent this metric in the analysis.

Table 5.2: Description of the variables used to represent hospital output in the data analysis

Variable ID: **PDE**	*Variable units:* **[patient day]**	*Concept represented:* Hospital output
Aspect measured: Tests the extent to which the demand placed on the resources of a hospital by the treatment of a hospitalised patient is reflected in the recorded energy and water consumption data of hospitals.		
Scope and structure: Each PDE represents the hospitalisation of an inpatient for one day, or the corresponding outpatient, day patient, or emergency department equivalent as described by the ratios in equation (5.1).		

$$\text{PDE} = \text{inpatient day} + \frac{1}{2}\text{\# of day patients} + \frac{1}{3}\text{\# of outpatients} + \frac{1}{3}\text{emergency headcount} \quad (5.1)$$

The formulation of the second and third transient normalising factors was discussed in Chapter 4. These factors represent the relative complexity (CMPX) and level of specialisation (SPEC) associated with the case mix composition of a hospital's inpatient diagnostic caseload. Two sets of unitless measures were formulated in Chapter 4 to represent CMPX and SPEC respectively. These measures are described in Table 5.3 and Table 5.4 respectively.

These normalising factors were used as independent variables in the data analysis. From these variables a set of normalising factors whose combination best captures the variance in the energy and water consumption of the hospitals was selected. The selection of the normalising factors was done via a quantitative approach that compares the explanatory power provided by the different combinations of potential normalising factors and selects the set with the most statistically significant explanatory power.

Table 5.3: Description of the variable used to represent the complexity of a hospital's inpatient caseload in the data analysis

Variable ID: **CMPX**	*Variable units:* **N/A**	*Concept represented:* Case mix composition
Aspect measured: Tests the usefulness of the complexity of the structure of a hospital's inpatient diagnostic caseload in explaining the inter-hospital variations in energy and water consumption performance.		
Scope and structure: Complexity was defined in terms of the proportion of total provincial cases for each diagnostic grouping that were treated at a hospital. It evaluated whether the treatment of each diagnostic case type was concentrated in a select set of hospitals or distributed relatively evenly across the hospitals in the analysis. The scope of this variable is limited to the inpatient caseload of a hospital. The diagnostic structure of day patient, outpatient and emergency care activities that do not require hospitalisation are outside the scope of this variable.		

Table 5.4: Description of the variable used to represent the level of specialisation of a hospital's inpatient caseload in the data analysis

Variable ID: **SPEC**	*Variable units:* **N/A**	*Concept represented:* Case mix composition
Aspect measured: Tests the extent to which the breadth of the portfolio of diagnoses treated by a hospital is reflective of resource consumption at that hospital.		
Scope and structure: Specialisation was defined in terms of the distribution of a hospital's caseload across the range of diagnostic case types treated at a hospital. It evaluated whether treatment within the hospital was concentrated in a select set of diagnosis or distributed relatively evenly across an array of diagnosis. The scope of this variable is limited to the inpatient caseload of a hospital. The diagnostic structure of day patient, outpatient and emergency care activities that do not require hospitalisation are outside the scope of this variable.		

5.1.2 Resource consumption

The data analysis evaluated the energy and water consumption of hospitals as dependent variables. Energy and water consumption are specified on an input-oriented basis. Thus, consumption is specified in terms of the total quantity supplied to a hospital building or complex, and not the quantities used by the respective components of a hospital's systems. Therefore, all electric consumption or transmission losses, or water losses due to leaks, that occur within the hospital system are also included within the consumption figure specified by the respective AEC and AWC variables described in Table 5.5.

Table 5.5: Description of the variables used to represent the energy and water consumption of hospitals in the data analysis

Variable ID: **AEC**	*Variable units:* **[kWh/year]**	*Concept represented:* Energy consumption of a hospital
Aspect measured: Represents the total annual electricity consumption of a hospital within a given calendar year.		
Scope and structure: The scope of the AEC variable was limited to the grid-supplied electricity used by a hospital as stated on its electricity consumption bills. Electricity generated onsite or other forms of energy used such a diesel were excluded from the scope of the variable.		

Variable ID: **AWC**	*Variable units:* **[kL/year]**	*Concept represented:* Water consumption of a hospital
Aspect measured: Represents the total annual water consumption of a hospital within a given calendar year.		
Scope and structure: The scope of the AWC variable was limited to the total amount of municipally supplied water consumed at the hospital. Water supplied by onsite boreholes or greywater systems was excluded from the scope of the variable.		

5.2 The dataset

This section discusses the formulation of the dataset that was used in the data analysis to model the relationship between the normalising factors and resource consumption, and the techniques applied during the data cleaning process.

5.2.1 The initial dataset

The sample studied in the data analysis consisted of 18 district hospitals in the Western Cape province in South Africa. These hospitals were selected because of data availability and quality constraints. Secondary data pertaining to the resource consumption, building characteristics and patient ratios of the hospitals in the 2016 calendar year was provided by the WCDoH. The Directorate of Engineering and Technical Support Services at the WCDoH provided the data on the energy and water consumption measurements and the size of the 18 hospitals. This data was used to formulate the AEC, AWC, TFA and BED variables for the hospitals in the analysis.

The variables for the transient normalisation factors (CMPX, SPEC, and PDE) were formulated from the patient statistics data provided by the WCDoH's Directorate of Information Management. The formulation of CMPX and SPEC variables was discussed in Chapter 4. The PDE variable associated with each hospital in the analysis was calculated using the inpatient days, day patient annual headcount, outpatient annual headcount, and emergency patient annual headcount data for each hospital.

Table 5.6 shows the initial dataset consisting of data on the variables collected at the 18 hospitals studied in the analysis. Data cleaning was applied to this dataset to ensure that it was suitable for analysis. This process focused on two issues: standardising the dataset to account for the effect of the different units of the variables, and the different orders of magnitude of the respective variables, and testing for outliers within the dataset. The steps taken to detect and address these issues are discussed in Subsections 5.2.2 and 5.2.3 respectively.

Table 5.6: The initial dataset for statistical analysis

Hospital code	AEC [kWh/year]	AWC [kL/year]	CMPX	SPEC	BED [beds]	TFA [m^2]	PDE [patient-days/year]
ABH	385491	829.499	21.78	1.03	30	1822	12908
BWH	1071816	9399	48.01	1.62	57	6450	26000
CLD	1064937	17451	39.94	0.91	50	7358	21261
CLH	251552	1786.667	31.31	0.77	50	2627	16510
CRS	501175.2	11972.4	114.70	0.97	86	6779	43067
HER	848482	10108.996	56.56	0.59	71	8449	28597
KNY	1183417	21733.091	73.18	0.62	90	10634	39354
LAP	170781.333	5986.667	9.01	1.30	10	1979	5589
LBH	124021.2	4802.4	10.80	0.94	20	1591	6442
MON	408693.818	15598.909	23.73	0.83	40	3163	14150
PRH	285714.545	9118.8	16.60	1.40	29	2620	7269
RIV	1030725.333	10029.333	38.93	0.77	50	6251	14770
RKH	342327.273	4442.182	23.60	0.99	31	1979	20035
ROB	267824.64	10981.2	36.37	0.82	46	3057	24196
STB	859746.667	14908.364	59.30	1.28	85	6251	39373
SWE	598065	5426	29.55	0.86	51	3812	17211
UDH	359205.926	2202	8.35	3.73	13	1137	4695
VRE	710390	2861.455	73.29	1.15	75	4151	32844

5.2.2 Unit normal scaling

The variables in the dataset are specified in terms of different units and with respect to different scales. Thus, before applying any of the regression analysis methodologies onto the dataset, unit normal scaling was used to standardise the dataset. The aim of the standardisation was to transform the variables and normalise for the effects of their units of measurement and scales. This ensured that the variables were stated with respect to a common and comparable scale.

The unit normal scaling standardisation technique (see equation (5.2)) imposes a standard normal distribution on the data (Freudenberg 2003). It converts the variables to a common standardised scale with a mean of zero and a standard deviation of one (OECD & JRC 2008). The result of applying this normalisation scheme is that all the normalised variables for each alternative will have a common mean and a similar dispersion:

$$X_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j} \tag{5.2}$$

Where,

X_{ij} is the normalised variable for the *j-th* variable associated with the *i-th* hospital,

x_{ij} is the raw value of the *j-th* variable associated with the *i-th* hospital,

$\bar{x}_j$ is the mean of the set of x_j variables associated with the hospitals in the analysis,

s_j is the standard deviation of the set of x_j variables associated with the hospitals in the analysis.

Table 5.7 shows the standardised version of each of the variables in the dataset used in the data analysis. The standardised versions of the variables were used to assess the feasibility of using CMPX and SPEC as normalisation factors when comparing the energy and water consumption of hospitals. The range of the standardised versions of the variables in the dataset have similar magnitudes and are unitless.

The standardisation reduced any potential numerical instabilities that may be caused by the large differences in the scales of magnitude of the variables in the analysis. In Table 5.7, the scaling within each variable set is due to the standard deviation (s_j) of the set of values for each variable and not the range of the distribution of the variable ($x_{j_{max}} - x_{j_{min}}$) (Freudenberg 2003). This prevents the presence of extreme values from having an overly significant influence on the results of the analysis. However, outliers still have an influence on the results because the range ($x_{ij} - \bar{x}_j$) of the outliers rewards or punishes extreme values (Jacobs et al. 2004) but with this approach the effect of outliers is somewhat dampened because the scaling factor is the standard deviation.

Table 5.7: The standardised dataset used in the statistical analysis

Hospital code	AEC	AWC	CMPX	SPEC	BED	TFA	PDE
ABH	-0.563	-1.365	-0.651	-0.159	-0.772	-0.959	-0.652
BWH	1.410	0.090	0.300	0.678	0.319	0.730	0.431
CLD	1.390	1.457	0.008	-0.338	0.036	1.061	0.039
CLH	-0.948	-1.202	-0.305	-0.539	0.036	-0.666	-0.354
CRS	-0.230	0.527	2.719	-0.254	1.491	0.850	1.843
HER	0.768	0.211	0.611	-0.788	0.885	1.459	0.646
KNY	1.730	2.184	1.213	-0.741	1.652	2.257	1.535
LAP	-1.180	-0.489	-1.114	0.231	-1.581	-0.902	-1.258
LBH	-1.314	-0.690	-1.049	-0.296	-1.176	-1.044	-1.187
MON	-0.496	1.143	-0.580	-0.445	-0.368	-0.470	-0.549
PRH	-0.850	0.042	-0.839	0.368	-0.813	-0.668	-1.119
RIV	1.292	0.197	-0.029	-0.533	0.036	0.657	-0.498
RKH	-0.687	-0.751	-0.585	-0.222	-0.732	-0.902	-0.063
ROB	-0.901	0.359	-0.121	-0.464	-0.126	-0.509	0.282
STB	0.800	1.025	0.710	0.197	1.450	0.657	1.537
SWE	0.048	-0.584	-0.369	-0.404	0.076	-0.233	-0.296
UDH	-0.638	-1.132	-1.138	3.695	-1.459	-1.209	-1.332
VRE	0.371	-1.020	1.217	0.014	1.046	-0.109	0.997

5.2.3 Testing for outlier

The Multiples of IQR test was applied to identify potential outliers in the dataset. This method uses quartiles to define upper and lower limits for each variable in the dataset. These limits were then used to identify outliers. The upper and lower limits are based on the 25th quartile, 75th quartile and the interquartile range (IQR) of a dataset. The equations used to calculate the respective limits are shown in equation (5.3) (Chatterjee & Simonoff 2013).

$$\begin{aligned} Upper\ limit &= 75\%\ quartile + 1.5 * IQR \\ Lower\ limit &= 25\%\ quartile - 1.5 * IQR \end{aligned} \tag{5.3}$$

To calculate the limits, a constant multiple of the IQR of each variable was added or subtracted from the respective 75 percent and 25 percent quartile of each variable's dataset. The IQR represents the difference between the 75th quartile and the 25th quartile. It is a useful and robust method of quantifying the scatter in a dataset as it is insensitive to the effects of extreme datapoints (Chatterjee & Simonoff 2013).

Extreme datapoints are found in the first and fourth quartile and tend to distort the size of these quartiles making them disproportionally large.

Thus, by defining limits based on a multiple of the IQR, the Multiples of IQR tests defines limits based on the scatter of the datapoints. Data points that lay outside of the range defined by these limits are classified as outliers. Table 5.8 outlines the upper limits, lower limits, and the corresponding minimum and maximum values calculated for each variable in the analysis. The distribution of datapoints within these variables is visually represented by the box and whiskers plots shown in Table 5.9.

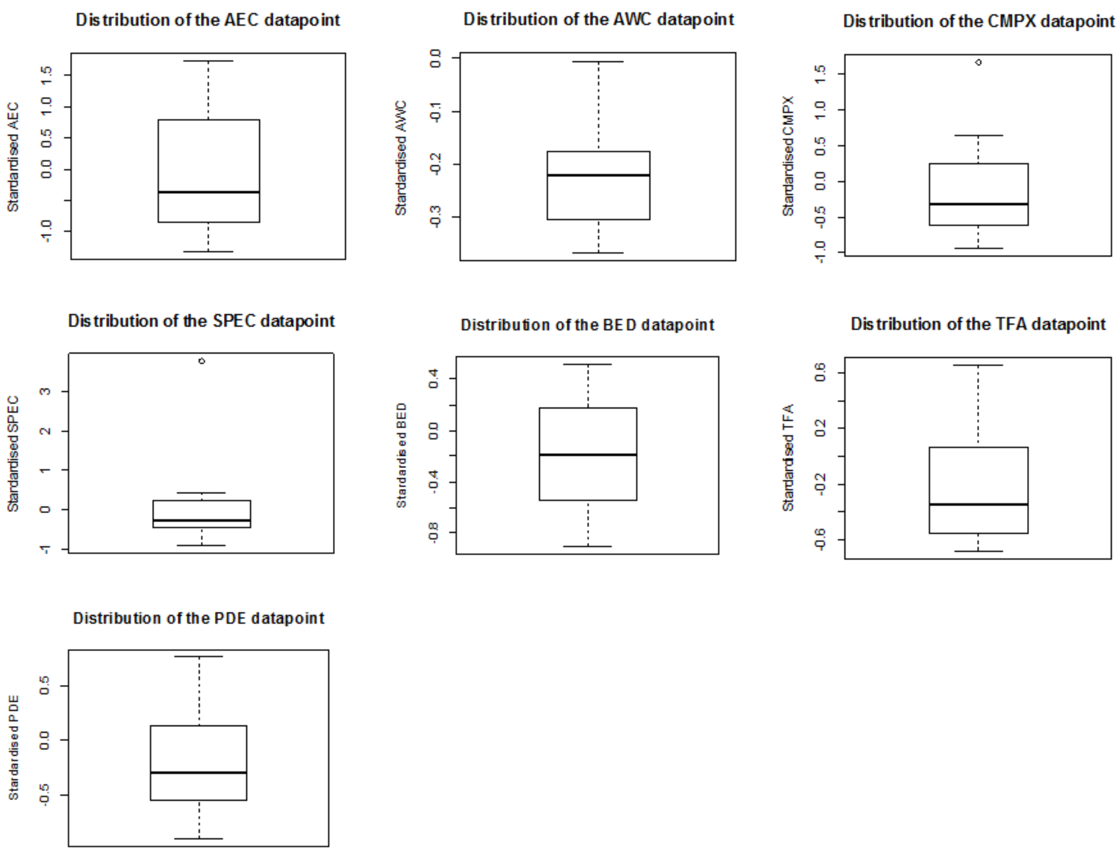
Table 5.8: Multiple of IQR test statistics

	AEC	AWC	CMPX	SPEC	BED	TFA	PDE
Upper limit	3.6014	0.062	1.6165	1.27	1.3243	1.0413	1.3485
Maximum	1.73	-0.006	1.674	3.773	0.519	0.657	0.756
Lower limit	-3.541	-0.5361	-2.0035	-1.4833	-1.677	-1.5208	-1.7635
Minimum	-1.314	-0.367	-0.945	-0.905	-0.902	-0.679	-0.888

Since the sample size used in the study is small it is susceptible to the effects of outliers. The presence of outliers in the data may reduce the correctness of the results obtained from the data analysis. Using the Multiples of IQR test, the dataset associated with two of the variables in the analysis was identified as containing outliers. These outliers correspond to the maximum values for CMPX and SPEC respectively. The outlier for the CMPX variable is associated with Ceres District Hospital (CRS), while, the outlier for the SPEC variable is associated with Uniondale District Hospital (UDH).

The data used in this analysis is secondary in nature. This limited the potential mitigation strategies for addressing the outliers as the author had no control over the process that produced the data. Thus, two options for addressing the outliers were identified: rejecting the outliers and removing them from the dataset or proceeding with the outliers and noting their effect on the observed results. As discussed in Section 4.5 these outliers are caused by the distribution of diagnostic cases at the respective hospital and are part of the internal mechanism that the analysis modelled. Thus, it was decided to keep these datapoints in the dataset.

Table 5.9: Box and whisker plots showing the distribution of the datapoints for each variable in the dataset



5.3 The quantitative analysis

5.3.1 The data analysis strategy

The aim of the data analysis discussed in this chapter was to develop regression models that characterise the relationship between the normalisation factors, and the energy and water consumption within district hospitals. These regression models were used to assess the significance of the explanatory power provided by the respective normalising factors in explaining the variance in the energy and water consumption behaviour of hospitals.

Figure 5.2 illustrates the process flow diagram of the data analysis strategy applied in this study. This analysis was conducted in RStudio version 3.4.2. An exhaustive set of multiple linear regression (MLR) models were generated for all the possible combinations of normalising factors using the standardised dataset. Energy and water consumption were analysed independently and MLR models were generated for each resource. The normalising factors were used as independent variables in these MLR analyses.

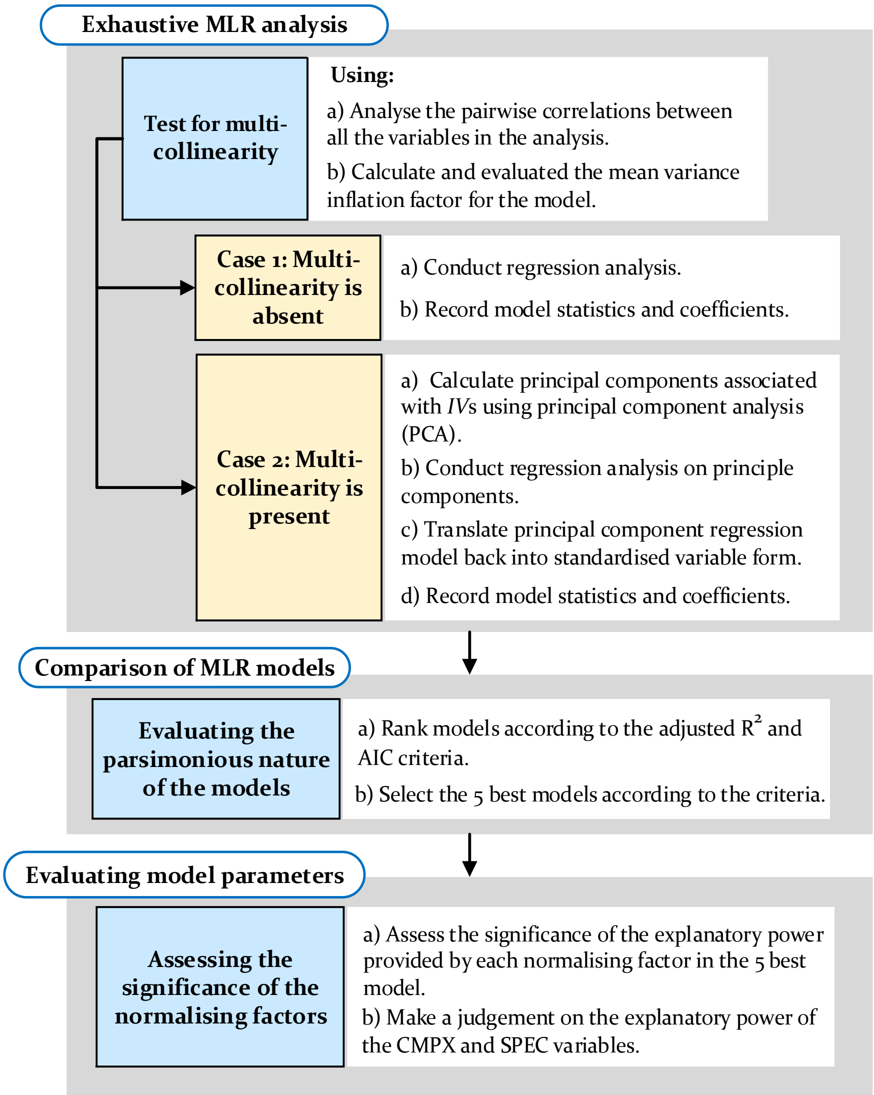


Figure 5.2: The data analysis strategy for the statistical analysis

The first step in the data analysis strategy was to test for multi-collinearity within the dataset used to develop each model. This occurs when the independent variables used in the regression analysis are correlated and may lead to the development of incorrect and misleading MLR models if not detected and addressed. The process used to detect and address multi-collinearity where it is present is discussed in Subsection 5.3.2.

After multi-collinearity was addressed, the MLR models were developed. The model coefficients and model statistics data associated with each of the MLR models for the respective combinations of independent variables were recorded and compared. These were used to identify and select the most parsimonious model, the model with the best balance between goodness of fit and model complexity. The selected model must also be consistent with what is known about the problem, thus making it intuitively practical.

As introduced in Chapter 3, parsimony aims to maintain a balance between the simplicity of a model (as described by the number of terms in the model) and the goodness of fit of the model (as described by the variance explained by the model). The AIC and R_a^2 criteria were used to assess and rank the goodness of fit of the respective MLR models and identify the most parsimonious models. This process is discussed in Sections 5.4 and 5.5.

The individual parameters associated with the terms in the best models were studied using the sample t-test to assess the significance of the explanatory power provided by each term in these models. This was used to make inferences on the significance of the explanatory power provided by each of the potential normalising factors identified in Section 2.4. This allowed us to evaluate the significance of the contribution of the CMPX and SPEC variables as normalisation measures and to make recommendations on the inclusion of SPEC and CMPX into the normalisation model.

5.3.2 Detecting multi-collinearity

This subsection discusses the detection of multi-collinearity within the dataset. Multi-collinearity occurs when two or more independent variables in an analysis are highly correlated. This occurs when two or more independent variables statistically control for the same underlying factor within an acceptable margin of error; or when there is a significant amount of statistical overlap between the underlying factors represented by the respective independent variables. Thus when studying the data, the predictive power of the analysis is not significantly increased by including the additional variables, because little additional variance is explained by the inclusion of additional variables (Monts & Blissett 1982).

Furthermore, from a normalisation point of view, multi-collinearity is associated with double counting. Since the qualitative significance of all the independent variables in the model has been determined, if two or more of the independent variables are collinear, then they potentially represent the same underlying attribute. Therefore, including both variables in the normalisation model accounts for the effect of the same underlying attribute two or more times (Jacobs et al. 2004).

Pairwise correlation analyses were conducted using RStudio version 3.4.2 to determine the degree of correlation between the variables in the analysis. These analyses evaluated the relationships between pairs of each of the variables in the dataset. A visual representation of the pairwise correlations between the variables in the respective AEC and AWC analyses is presented in Figure 5.3 and Figure 5.4 respectively.

In these figures, the top row shows the pairwise correlations of the normalising factors with the respective electricity and water consumption variables. The first column states the correlation coefficients associated with each pair of factors. These figures show that there is a significant correlation between the normalising factors and the respective AEC and AWC variables.

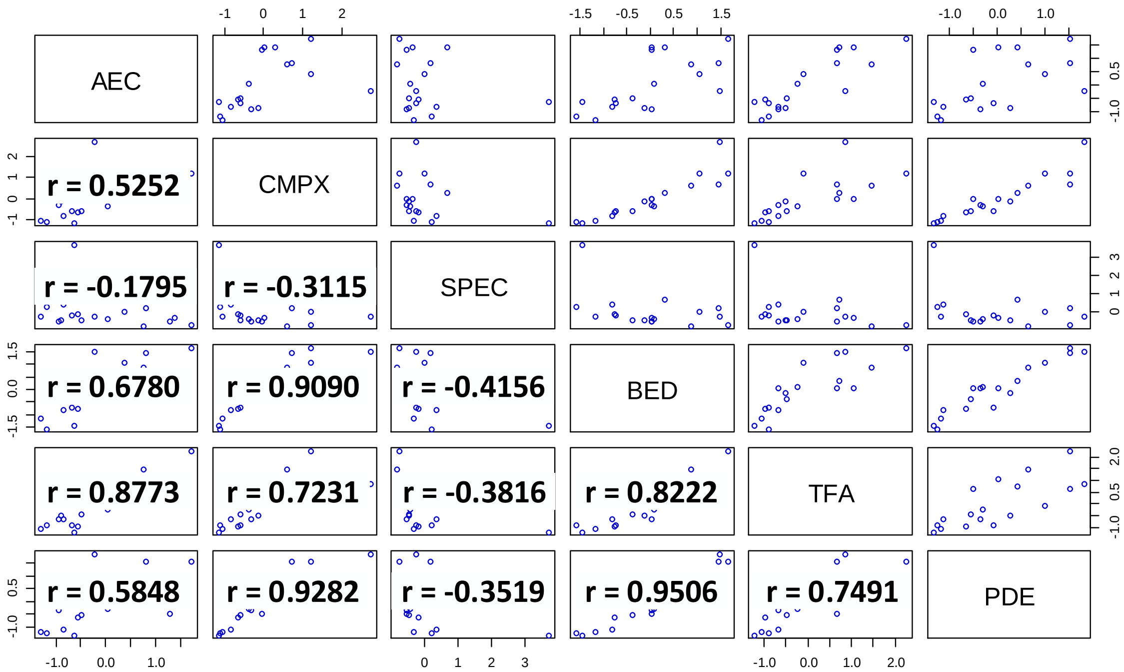


Figure 5.3: A visual representation of the pairwise correlations and the correlation coefficients for AEC analysis

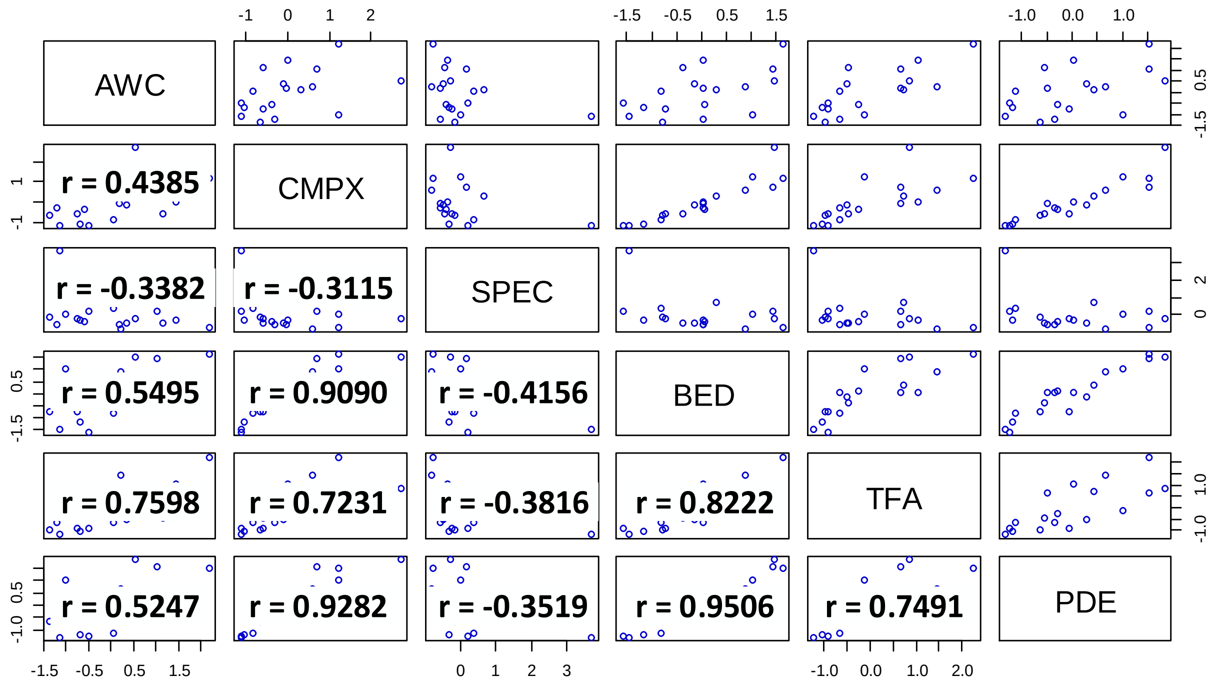


Figure 5.4: A visual representation of the pairwise correlations and the correlation coefficients for AWC analysis

A strong and positive correlation was found between four of the five normalising factors and both AEC (ranging from r= 0.5252 to r= 0.8773) and AWC (ranging from r= 0.4385 to r= 0.7598). This is intuitively consistent with the relationship expected between these normalising factors and electricity and water consumption: as CMPX, BED, TFA and PDE increase, so do AEC and AWC.

The fifth normalising factor (SPEC) is not significantly correlated to AEC, AWC or any of the other independent variables at a significance level of $\alpha = 0.05$. The critical value for the correlation coefficient at a significance level of $\alpha=0.05$ and 16 degrees of freedom is r= 0.468. Thus, the results of the correlation analysis of SPEC in relation to all the other variables in the study, show that it is not significantly correlated to any of the other variables as the correlation coefficients for each of these analyses are less than the critical value. This non-significant correlation has two implications. Firstly, SPEC does not covary with any of the other independent variables. However, it might covary with combinations of the other variables. Secondly, individually SPEC does not significantly contribute to explaining the variance in AEC or AWC.

The significant correlation found between the normalising factors, and AEC and AWC respectively, motivates the applicability of using linear regression analysis to study the relationships between the prospective normalisation factors and resource consumption. From the pairwise correlation analysis it can be seen that the normalising factors are also significantly correlated with each other, for example, for BED-PDE (r= 0.9506), and for CMPX-BED (r= 0.9090). This is indicative of the presence of multi-collinearity in the variable set. These high correlations may result in numerical instabilities within the model, or interpretive difficulties as the regression coefficients may intuitively differ from what is expected.

Some of these high correlations were expected. For example, BED and TFA are both representative of the size of a hospital, thus it is expected that they would be strongly correlated. This is due to the design of the data analysis strategy. The strategy exhaustively tests all the possible combinations of normalising factors to identify the most parsimonious combination. This model was used to assess the feasibility of including CMPX and SPEC in the normalisation model. Thus, combinations with multi-collinear variables are also tested.

These pairwise correlation comparisons are not conclusive for determining multi-collinearity as more complex correlations may exist within the dataset. For example, CMPX may be correlated to the sum of BED and PDE. Thus, the mean variance inflation factor (VIF) was calculated for the respective combinations of independent variables in the exhaustive MLR analysis. It was used to conclusively identify models with collinear independent variables.

The presence of two or more correlated independent variables in an MLR model inflates the regression coefficients of the correlated terms. The VIF as given by equation (5.4), measures the level of inflation in each regression coefficient associated with an independent variable in an MLR model (Chatterjee & Simonoff 2013).

$$VIF_i = \frac{1}{1 - R_i^2} \tag{5.4}$$

The VIF for the coefficient of the *i-th* independent variable (x_i) in an MLR model is calculated by regressing x_i against all the other independent variables in the model. In equation (5.4), R_i^2 is the coefficient of determination associated with this MLR model. It measures how well the variance in x_i is explained by the other independent variables in the original MLR model.

VIF_i is measured relative to the case where there is no correlation between x_i and the other independent variables in the model. This corresponds to the case where $R_i^2 = 0$ and $VIF_i = 1$. The other extreme value occurs when there is a perfect correlation between x_i and the other independent variables in the model, thus $R_i^2 = 1$ and $VIF_i \rightarrow \infty$. In this case, all the explanatory power provided by x_i can be obtained from the combination of the other regression variables. Thus, the larger the VIF_i the more severe the multi-collinearity.

In the data analysis VIF_i was calculated for all the independent variables in the respective MLR model. After this, a mean variance inflation factor was calculated and used to assess the level of multi-collinearity in each MLR model. Models with a mean variance inflation factor greater than 4 were categorised as having a significant amount of multi-collinearity (Chatterjee & Simonoff 2013). The sets of independent variables in 16 of the 31 MLR models for both AEC and AWC were found to have a significant amount of multi-collinearity. The mean VIF for the set of independent variables in the models are presented in Table 5.12 and Table 5.13 respectively.

Since these models were used to interpret and evaluate the relationships between the normalising factors, the high *VIF*s could not be ignored and mitigating measures were taken to address the multi-collinearity. These measures are discussed in Section 5.3.3.

5.3.3 **Addressing multi-collinearity**

Principal Component Analysis (PCA) was used to address multi-collinearity in the models that were identified as having collinear independent variables. The main premise behind this method is that for a dataset of n independent variables $(x_1, x_2, \cdots, x_n)$ that are highly correlated, most of the variance in this dataset is due to a smaller set of m uncorrelated independent variables $(PC_1, PC_2, \cdots, PC_m)$.

Principal component analysis evaluates and then transforms the set of n highly correlated independent variables into m uncorrelated variable groupings known as principle components (*PC*), where $m \leq n$ (Jacobs et al. 2004). The m principle components, as shown in equation (5.5), are linear combinations of the original set of independent variables that still contain the essence of the original dataset but are orthogonal, and thus uncorrelated.

$$
\begin{aligned}
PC_1 &= w_{11}x_1 + w_{21}x_1 + \cdots + w_{n1}x_1 \\
PC_2 &= w_{12}x_1 + w_{22}x_2 + \cdots + w_{n2}x_2 \\
&\vdots \\
PC_m &= w_{1m}x_m + w_{2m}x_m + \cdots + w_{nm}x_m
\end{aligned}
\tag{5.5}
$$

Thus, each of the principal components explains a different statistical dimension of the original dataset (OECD & JRC 2008). The principal components are arranged chronologically in order of decreasing explained variance. The weights (w_{ij}) are a measure of the statistical overlap between correlated variables and are used to correct the principle components for this overlap (OECD & JRC 2008).

These weights are the components of a rotation matrix which contains the parameters of the equation used to convert the standardised components of the independent variables into principal component form. An example of a rotation matrix is shown in Table 5.10 for the MLR model consisting of five independent variables. The rotation matrix changes with changes in the composition of independent variables in the model. Thus, different sets of PCs were calculated for each of the multi-collinear combinations of independent variables.

Table 5.10: Rotation matrix for a five-independent-variable model

	PC1	**PC2**	**PC3**	**PC4**	**PC5**
CMPX	0.4824	0.2183	-0.3621	0.7626	-0.0837
SPEC	-0.2573	0.9561	0.1231	-0.0566	-0.0371
BED	0.5050	0.0897	-0.0700	-0.4577	-0.7229
TFA	0.4494	0.0227	0.8703	0.1383	0.1448
PDE	0.4941	0.1723	-0.3024	-0.4320	0.6694

Table 5.11: The rotated model design matrix for AWC

	PC1	PC2	PC3	PC4	PC5
ABH	-0.6506	-0.1587	-0.7723	-0.9593	-0.6523
BWH	0.3005	0.6781	0.3188	0.7297	0.4308
CLD	0.0080	-0.3380	0.0359	1.0611	0.0388
CLH	-0.3051	-0.5385	0.0359	-0.6655	-0.3543
CRS	2.7190	-0.2535	1.4907	0.8498	1.8426
HER	0.6107	-0.7884	0.8846	1.4593	0.6456
KNY	1.2134	-0.7415	1.6524	2.2567	1.5354
LAP	-1.1140	0.2307	-1.5805	-0.9020	-1.2577
LBH	-1.0489	-0.2964	-1.1764	-1.0436	-1.1871
MON	-0.5801	-0.4448	-0.3682	-0.4699	-0.5495
PRH	-0.8387	0.3682	-0.8127	-0.6681	-1.1187
RIV	-0.0290	-0.5327	0.0359	0.6571	-0.4982
RKH	-0.5846	-0.2222	-0.7319	-0.9020	-0.0627
ROB	-0.1215	-0.4638	-0.1257	-0.5086	0.2815
STB	0.7100	0.1970	1.4503	0.6571	1.5369
SWE	-0.3688	-0.4039	0.0763	-0.2330	-0.2963
UDH	-1.1377	3.6946	-1.4593	-1.2093	-1.3316
VRE	1.2174	0.0139	1.0462	-0.1093	0.9969

Table 5.11 shows the rotated design matrix for the five-predictor-variable model. It details the contribution of the original independent variables associated with each hospital to the composition of the respective principal components. The benefit of this transformation is that the principal components are orthogonal and there is no multi-collinearity in this new dataset. Thus, it can be used to conduct a multiple linear regression analysis without potentially resulting in numerical instabilities or misleading results.

5.3.4 Exhaustive MLR analysis

An exhaustive set of 31 multiple linear regression analyses were conducted for each of the dependent variables: energy consumption and water consumption. These analyses generated regression models consisting of all the possible combinations of the normalising factors. The respective independent variable sets were classified into two categories: multi-collinear independent variable sets and non-multi-collinear independent variable sets. A different approach was applied to formulate multiple linear regression models for each type of variable set.

In the case where there is no multi-collinearity between the independent variables, the dataset used in the regression analysis consisted of the standardised versions of the original variables. This dataset was used to perform the regression analysis and to determine the model coefficients and model statistics of each model. In the case where multi-collinearity was present in the set of independent variables, the independent variables were converted into principal component form and the rotated design matrix for that variable set was regressed against the standardised version of the dependent variable.

Converting to principal component form corrected for the multi-collinearity in the independent variable set. The regression analysis was performed on the variable set in principal component form and the overall model statistics corresponding to the principal component regression were recorded. After this, the regression coefficients calculated using principal component regression (PCR) were transformed into their standardised original variable form and recorded.

These two types of model analysis procedures were used to perform the regression analysis. The data transformations and multiple linear regression analyses corresponding to these two cases were performed in RStudio version 3.4.2. Copies of the script files used for the analyses corresponding to the respective cases are presented in Appendix H. The model coefficients and statistics of the respective analyses are shown in Table 5.12 and Table 5.13 for AEC and AWC respectively. In these tables the models that were developed using principal component regression are denoted with “*Y*” in the PCR column.

Table 5.12: Model coefficients and statistics for electricity consumption MLR models

Model	Constant	CMPX	SPEC	BED	TFA	PDE	mean VIF	PCR	F-stat.	p-value	R.2_adj	AIC	SEE
1	0	0.5252							6.094	0.0252	0.2305	50.2444	0.8772
2	0		-0.1795						0.5328	0.4760	-0.0283	55.4633	1.0140
3	0			0.6780					13.61	0.0020	0.4259	44.9717	0.7577
4	0				0.8773				53.45	0.0000	0.7552	29.6283	0.4947
5	0					0.5848			8.315	0.0108	0.3008	48.5198	0.8362
6	0	0.5197	-0.0177				1.1074		2.86	0.0886	0.1796	52.2374	0.9058
7	0	-0.5249		1.1552			5.7583	Y	7.73	0.0049	0.4419	45.3025	0.7471
8	0	-0.2288			1.0430		2.0958		29.02	0.0000	0.7672	29.5620	0.4825
9	0	-0.1270				0.7027	7.2187	Y	3.937	0.0422	0.2568	50.4585	0.8621
10	0		0.1236	0.7294			1.2088		6.714	0.0083	0.4020	46.5456	0.7733
11	0		0.1817		0.9466		1.1705		29.6	0.0000	0.7709	29.2760	0.4787
12	0		0.0300			0.5953	1.1413		3.911	0.0430	0.2551	50.4983	0.8631
13	0			-0.1335	0.9870		3.0860		25.89	0.0000	0.7455	31.1713	0.5045
14	0			1.2734		-0.6261	10.4442	Y	7.417	0.0058	0.4302	45.6756	0.7549
15	0				1.0010	-0.1650	2.2788		26.84	0.0000	0.7524	30.6700	0.4976
16	0	-0.5902	0.1709	1.2855			4.5570	Y	5.283	0.0120	0.4305	46.4251	0.7547
17	0	-0.2160	0.1728		1.0990		1.8328		21.26	0.0000	0.7815	29.1835	0.4675
18	0	-0.1306	0.0322			0.7173	5.2765	Y	2.459	0.1057	0.2048	52.4336	0.8917
19	0	-0.3910		0.2460	0.9577		5.8355	Y	18.86	0.0000	0.7592	30.9321	0.4907

Continued on next page

Model	Constant	CMPX	SPEC	BED	TFA	PDE	mean VIF	PCR	F-stat.	p-value	R.2_adj	AIC	SEE
20	0	-0.3905		1.3812		-0.3662	10.8219	Y	5	0.0146	0.4138	46.9444	0.7656
21	0	-0.3321			1.0219	0.1275	5.8566	Y	18.28	0.0000	0.7531	31.3819	0.4969
22	0		0.1723	-0.0794	1.0080		2.5088		18.64	0.0000	0.7569	5.8566	0.4930
23	0		0.1602	1.4089		-0.6986	7.7552	Y	5.014	0.0144	0.4147	46.9179	0.7651
24	0		0.1710		1.0470	-0.1393	1.9514		19.42	0.0000	0.7647	30.5111	0.4850
25	0			0.1759	0.9568	-0.2993	9.5540	Y	16.91	0.0001	0.7373	32.4950	0.5125
26	0	-0.4678	0.2083	0.3860	0.9777		4.8761	Y	16.61	0.0001	0.7860	29.4732	0.4626
27	0	-0.4422	0.1807	1.5483		-0.4135	8.7454	Y	3.866	0.0278	0.4027	47.9469	0.7728
28	0	-0.3620	0.1807		1.0724	0.1810	4.7472	Y	15.23	0.0001	0.7700	30.7720	0.4796
29	0	-0.3723		0.2821	0.9538	-0.0525	10.0633	Y	13.15	0.0002	0.7409	32.9144	0.5090
30	0		0.1905	0.3195	0.9721	-0.3802	7.7101	Y	14.12	0.0001	0.7553	31.8826	0.4946
31	0	-0.4321	0.2105	0.4578	0.9703	-0.1023	8.5134	Y	12.33	0.0002	0.7692	31.3922	0.4804

Table 5.13: Model coefficients and statistics for water consumption MLR models

Model	Constant	CMPX	SPEC	BED	TFA	PDE	mean VIF	PCR	F-statistic	p-value	R.2_adj	AIC	SEE
1	0	0.4385							3.81	0.0687	0.1418	52.2084	0.9264
2	0		-0.3382						2.066	0.1699	0.0590	53.8667	0.9700
3	0			0.5495					6.921	0.0182	0.2583	49.5823	0.8612
4	0				0.7598				21.85	0.0003	0.5508	40.5554	0.6702
5	0					0.5247			6.079	0.0254	0.2300	50.2560	0.8775
6	0	0.3690	-0.2233				1.1074		2.334	0.1311	0.1356	53.1763	0.9297
7	0	-0.3512		0.8687			5.7583	Y	5.496	0.0323	0.2092	50.7376	0.8893
8	0	-0.2323			0.9278		2.0958		11.39	0.0010	0.5501	41.4242	0.6708
9	0	-0.3500				0.8496	7.2187	Y	3.098	0.0748	0.1980	51.8294	0.8956
10	0		-0.1327	0.4943			1.2088		3.474	0.0576	0.2254	51.2024	0.8801
11	0		-0.0565		0.7382		1.1705		10.36	0.0015	0.5240	42.4390	0.6899
12	0		-0.1752			0.4631	1.1413		3.249	0.0673	0.2092	51.5749	0.8893
13	0			-0.2319	0.9504		3.0860		11	0.0011	0.5406	41.7977	0.6778
14	0			0.5278		0.0228	10.4442	Y	3.245	0.0674	0.2089	51.5810	0.8894
15	0				0.8356	-0.1012	2.2788		10.43	0.0014	0.5260	42.3629	0.6885
16	0	-0.3100	-0.1079	0.7864			4.5570	Y	2.327	0.1190	0.1897	52.7713	0.9002
17	0	-0.2372	-0.0663		0.9060		1.8328		7.2	0.0037	0.5225	43.2535	0.6910
18	0	-0.3315	-0.1695			0.7728	5.2765	Y	2.17	0.1371	0.1712	53.1786	0.9104
19	0	-0.2206		-0.0178	0.9339		5.8355	Y	7.089	0.0039	0.5180	43.4225	0.6943

Continued on next page

Model	Constant	CMPX	SPEC	BED	TFA	PDE	mean VIF	PCR	F-statistic	p-value	R.2_adj	AIC	SEE
20	0	-0.4758		0.6591		0.3395	10.8219	Y	2.316	0.1201	0.1885	52.7981	0.9008
21	0	-0.5244			0.8688	0.3606	5.8566	Y	7.593	0.0030	0.5378	42.6668	0.6799
22	0		-0.0874	-0.2594	0.9397		2.5088		7.028	0.0041	0.5154	43.5167	0.6961
23	0		-0.1372	0.4118		0.0849	7.7552	Y	2.168	0.1374	0.1709	53.1847	0.9105
24	0		-0.0650		0.8181	-0.1110	1.9514		6.587	0.0053	0.4965	44.2085	0.7096
25	0			-0.6040	0.9867	0.3599	9.5540	Y	7.197	0.0037	0.5224	43.2580	0.6911
26	0	-0.1939	-0.0724	-0.0665	0.9270		4.8761	Y	5.025	0.0114	0.4864	45.2313	0.7167
27	0	-0.4425	-0.1167	0.5513		0.3701	8.7454	Y	1.693	0.2113	0.1403	54.5036	0.9272
28	0	-0.5159	-0.0512		0.8545	0.3454	4.7472	Y	5.338	0.0091	0.5051	44.5622	0.7035
29	0	-0.4570		-0.4736	0.9831	0.6628	10.0633	Y	5.631	0.0074	0.5214	43.9583	0.6918
30	0		-0.1067	-0.6844	0.9782	0.4052	7.7101	Y	5.21	0.0100	0.4976	44.8330	0.7088
31	0	-0.4323	-0.0867	-0.5460	0.9763	0.6833	8.5134	Y	4.267	0.0184	0.4900	45.6618	0.7141

5.3.5 Overall significance of the MLR models

The overall significance of each of the regression models described in Table 5.12 and Table 5.13 respectively were evaluated using the F-test. The F-test evaluates the overall statistical significance of the regression model. It was used to assess whether as a group, the independent variables significantly contributed to explaining the variance observed in the dependent variable (Chatterjee & Simonoff 2013).

The null hypothesis (H_0) corresponds to the case where none of the independent variables in the model accounted for a statistically significant amount of the variance in the dependent variable. Thus, all the regression coefficients for the independent variables in the model are zero. The alternative hypothesis (H_1) corresponds to the case where at least one of the independent variables accounted for a statistically significant amount of the variance in the dependent variable. Thus, at least one of the regression coefficients is not zero. This is given by:

$$\begin{aligned} &H_0: \beta_1 = \cdots = \beta_n = 0, \\ &H_1: at\ least\ one\ of\ \beta_i \neq 0, \qquad i = 1, 2, \dots, p. \end{aligned} \tag{5.6}$$

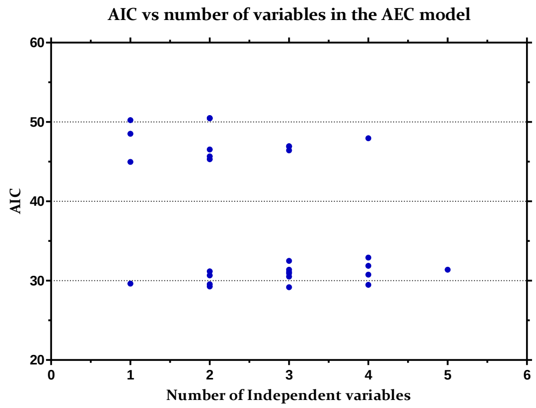
Each of the regression models was evaluated at the significance level $\alpha = 0.05$. The F-statistic and associated p-value of each model are presented in Table 5.12 and Table 5.13 for electricity and water consumption respectively. The p-values for 3 of the 31 AEC regression models in Table 5.12, and 11 of the 31 AWC regression models in Table 5.13 were larger than $\alpha = 0.05$.

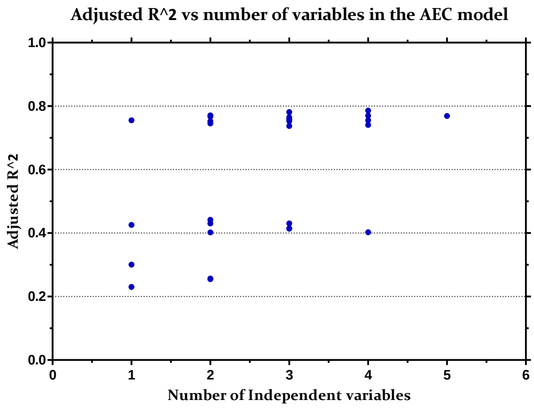
Thus, as a group these models do not significantly contribute to explaining the variance in the respective electricity (AEC) and water (AWC) consumption data of the hospitals in the analysis. These models are marked in red in the corresponding results tables. The null hypothesis was accepted for each of these models, and they were excluded from the model comparison process.

5.4 Quantitative findings: AEC models

The AIC and R_a^2 statistics corresponding to each of the energy consumption regression models in the analysis are provided in Table 5.12. These two criteria were used to evaluate the goodness of fit of each of the models. The most parsimonious model minimises the AIC statistic, while maximising the R_a^2 statistic. These statistics are visually represented by the plots in Table 5.14, where each model selection criterion statistic is plotted against the number of variables in that model.

Table 5.14: Plots of the model selection criterion statistics vs the number of independent variables in the model





The model-selection criteria statistics of the ten best regression models for the electricity consumption of the hospitals in the analysis are presented in Table 5.15. These models are sorted according to their AIC statistic values. According to the AIC criterion, the regression model with CMPX, SPEC and TFA as independent variables (C-S-T) is the best model for accounting for the variance in AEC. Taken as a set, the combination of these independent variables accounts for 78.15 percent of the variance in electricity consumption.

However, according to the R_a^2 statistic, the C-S-T model only explains 1.06 percent more of the variance in the electricity consumption variable than the best two-predictor-model (the S-T model). Furthermore, when compared to the best single-predictor-model (the T model), the C-S-T model only explains 2.63 percent more variance.

A comparison of the AIC and R_a^2 plots presented in Table 5.14 shows that there is not a significant difference between the AIC values of the best single-predictor-, two-predictor-, three-predictor-, and four-predictor-variable-models. This suggests that the additional variables in the two-predictor-, three-predictor-, and four-predictor-variable-models do not provide a significant amount of additional explanatory power. Thus, the most parsimonious model is the single-predictor-model, with TFA as the independent variable which explains 75.52 percent of the variance in the annual energy consumption dataset according to the R_a^2 measure.

Table 5.15: Ten best regression models for electricity consumption

Model	Number of Variables	R_a^2	AIC
C-S-T	3	0.7815	29.1835
S-T	2	0.7709	29.2760
C-S-B-T	4	0.7860	29.4732
C-T	2	0.7672	29.5620
T	1	0.7552	29.6283
S-T-P	3	0.7647	30.5111
T-P	2	0.7524	30.6700
C-S-T-P	4	0.7700	30.7720
C-B-T	3	0.7592	30.9321
S-B-T	3	0.7569	31.1003

Where: C=CMPX, S=SPEC, B=BED, T=TFA, and P=PDE.

All of the ten best electricity consumption regression models presented in Table 5.15 contain the TFA predictor variable. Due to the high amount of explanatory power attributed to the single-predictor TFA model, it was suspected that TFA alone accounts for most of the statistically significant variance in the other models listed in Table 5.15.

Since the presence of multi-collinearity between the predictor variables in each model was addressed before the regression analysis, a sample t-test was used to determine whether each independent variable in a regression model is a statistically significant predictor of the electricity consumption of the set of hospitals in the analysis. The t-test was used to evaluate whether the model coefficient associated with each predictor variable was significant. It tested the following hypothesis:

$$\begin{aligned} H_0&: \beta_i = 0, \qquad i = 1, 2, \dots, p \\ H_1&: \beta_i \neq 0, \qquad i = 1, 2, \dots, p. \end{aligned} \tag{5.7}$$

The p-value associated with every model coefficient for a sample t-test was used to determine the significance of the model coefficients. The p-value of each model coefficient was evaluated at the significance level $\alpha = 0.05$. If the p-value is greater than $\alpha = 0.05$, then from the sample t-test we cannot reject the hypothesis that the model coefficient associated with that value is equal to zero. If the model coefficient is zero, then it does not account for a statistically significant amount of unique variance in the electricity consumption of the hospitals in the analysis.

An evaluation of the significance of the individual parameters of the models showed that in all three of the two-predictor-variable models listed in Table 5.15, TFA was the only predictor variable that significantly contributed to explaining the variance in the electricity consumption of hospitals. Similar results were obtained for the four three-predictor-variable models listed in Table 5.15. Thus, it was concluded that TFA is the factor that most significantly explains the variance in the electricity consumption of hospitals. The results of these analyses are presented in Appendix I.

In the regression models that do not have TFA as an independent variable, BED is the factor that most significantly accounts for a unique amount of variance in AEC. However, there is a significant drop in the explained variance due to the exclusion of TFA from the normalisation model (see Table 5.12). The R_a^2 of the CMPX-BED model is 31.33 percent less than that of the TFA model alone. Also, an analysis of the individual parameters and confidence intervals of the 5 best models that do not include TFA, shows that in these models BED is the only factor that significantly contributes to explaining the variance in AEC.

The contribution of both CMPX and SPEC in the top ten models listed in Table 5.15 was found to be non-significant in all the models. Furthermore, the overall explained variance of these models was smaller or just marginally greater than that of the single-predictor TFA model. Thus, it was concluded that CMPX and SPEC do not provide any significant additional explanatory power when evaluating the variance in the energy consumption of the hospitals in the data sample.

5.5 Quantitative findings: AWC models

Table 5.16 illustrates the model selection criterion statistics for each of the significant water-consumption models plotted against the number of variables in that model. These plots are based on the results of the water-consumption regression analyses described in Table 5.13. As in the case of electricity consumption, the aim of the analysis is to identify the most parsimonious model for explaining the variance in the water-consumption of hospitals. Thus, the AIC and R_a^2 statistics corresponding to each regression model were used to analyse the goodness of fit of the respective models.

Table 5.16: Plots of the model selection criterion statistics vs the number of independent variables in the model

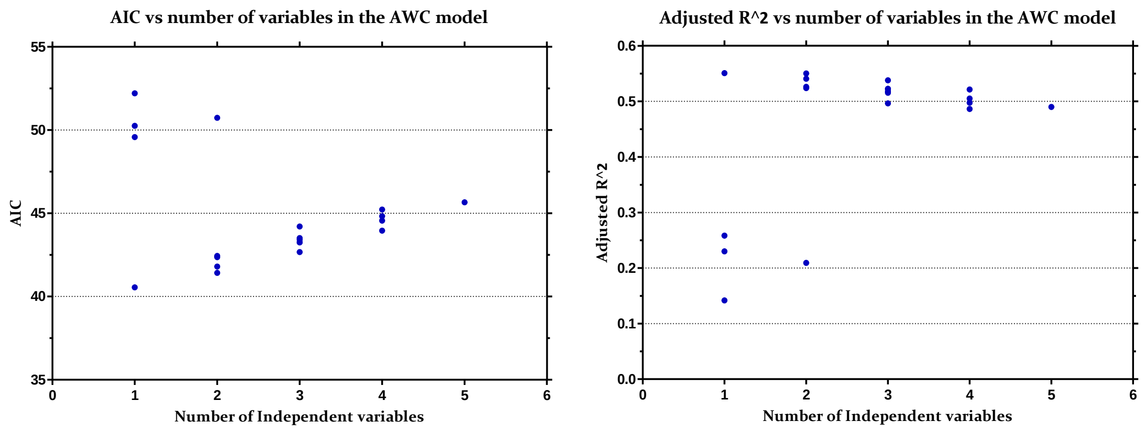


Table 5.17 lists the ten best models for describing the variance in the water-consumption data. The models are ranked with respect to their AIC statistics. An analysis of Table 5.16 in conjunction with the corresponding figures presented in Table 5.17 showed that a single-predictor model best explained the overall variance in the water-consumption dataset analysed. It was found that total floor area (TFA) has the lowest AIC statistic and the highest R_a^2 statistic of the models in the analysis. Thus, TFA accounted for the most significant amount of variance in the water consumption dataset ($R_a^2 = 0.5508$).

Furthermore, an analysis of Table 5.16 and Table 5.17 showed that increasing the amount of normalisation factors in the model decreases the amount of variance explained by the model. Thus, because of this decrease in explanatory power due to the inclusion of additional predictor variables in the regression model, the significance of these individual parameters in the corresponding regression models was not evaluated as they provided no additional explanatory power.

It was concluded that the total floor area (TFA) of hospitals is the most significant predictor of the water consumption of hospitals. Its contribution is so significant that the model with the most significant explanatory power that does not contain TFA as a predictor variable is the eighteenth most powerful explanatory model. This model consists of only the BED variable and has an R_a^2 = 0.2583.

As in the case of electricity, it was concluded that CMPX and SPEC do not significantly contribute to explaining the variance in the water-consumption of hospitals. Thus, changing the normalisation model used to compare the water consumption of hospitals to include terms that account for CMPX and/or SPEC would reduce the explanatory power of that normalisation model.

Table 5.17 Ten best AWC regression models

Model	Number of variables	R_a^2	AIC
T	1	0.5508	40.5554
C-T	2	0.5501	41.4242
B-T	2	0.5406	41.7977
T-P	2	0.5260	42.3629
S-T	2	0.5240	42.4390
C-T-P	3	0.5378	42.6668
C-S-T	3	0.5225	43.2535
B-T-P	3	0.5224	43.2580
C-B-T	3	0.5180	43.4225
S-B-T	3	0.5154	43.5167

Where: C=CMPX, S=SPEC, B=BED, T=TFA, and P=PDE.

5.6 Conclusion

The data analysis discussed in this chapter evaluated the explanatory power provided by the prospective normalising factors in accounting for the inter-hospital variations in the energy- and water-consumption of hospitals. This was used to rank the explanatory power provided by each normalising factor, and thus evaluate the feasibility of using complexity and specialisation as normalising factors. The analysis found that the size of the hospital as represented by its total floor area (TFA) explains the most statistically significant amount of variance in the energy and water consumption of the hospitals in the sample studied.

Furthermore, it was found that the explanatory power provided by the complexity (CMPX) and level of specialisation (SPEC) of a hospital's diagnostic caseload does not significantly contribute to explaining the variance in the energy and water consumption data. Thus, it was concluded that these parameters should not be included in the normalisation model, as this would increase the complexity of the normalisation model without contributing any additional statistically significant explanatory power.

Chapter 6 Conclusion

This chapter contextualises the findings of the research study and discusses the studies main conclusions. The implications of the statistical findings on the research problem that was investigated are discussed and the feasibility of the addition of the proposed normalisation factors to the current normalisation model are also outlined.

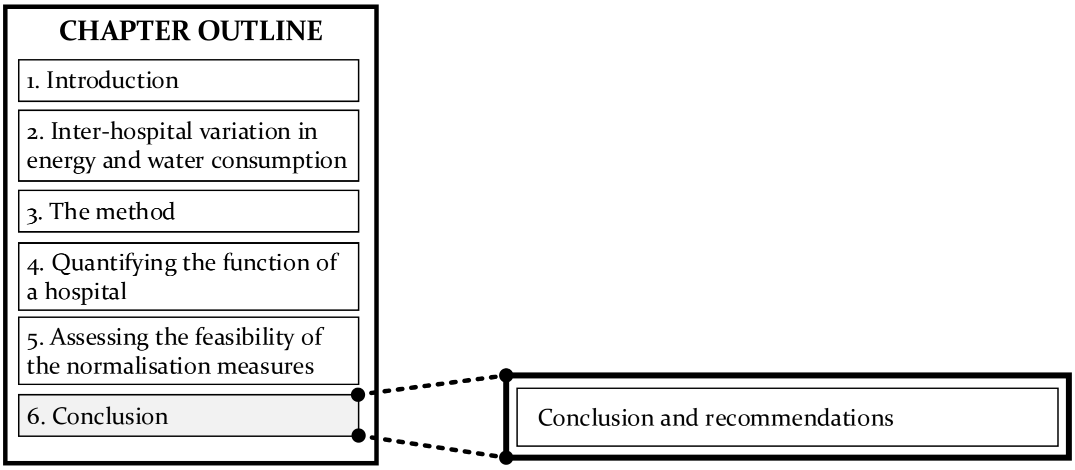


Figure 6.1: Thesis document outline: Chapter 6 contextualised

6.1 Project summary

This research study assessed the feasibility of including factors that are representative of a hospitals function[10] into the normalisation model. Normalisation models were developed for different combinations of a set of normalising factors representing hospital size and hospital function. The explanatory power provided by these models in explaining the variance observed in the energy and water consumption data of hospitals was compared and used to evaluate the feasibility of accounting for the function of a hospital in the normalisation model.

[10] The function of a hospital is represented by its level of medical service provision.

This was achieved through the following research objectives:

- Identifying factors to quantitatively represent the characteristics and function of a hospital as normalisation factors when evaluating the inter-hospital variations in resource consumption.
- Identifying and formulating methods for quantifying and comparing the explanatory power provided by each normalisation factor.
- Quantitatively assessing whether the normalisation factors that are representative of hospital function provided any significant and additional explanatory power to the normalisation model.

In Chapter 2, a literature analysis was conducted to identify the factors that affect the energy and water consumption of hospitals and establish a relationship between these factors and resource consumption. From this analysis five prospective normalisation factors that are representative of the size and function of a hospital were identified. Methods for quantifying these normalisation factors were also identified. Chapter 3 discussed these methods as well as the research design and methodology used to evaluate the feasibility of accounting for hospital function in the normalisation model.

The function of a hospital was quantified using its output (as represented by its PDE) and the complexity (CMPX) and level of specialisation (SPEC) of the hospital's caseload. In Chapter 4, the patient statistics data of a set of district hospitals in the Western Cape was collected and used to develop CMPX and SPEC measures for the caseloads of each hospital relative to the other hospitals in the analysis.

The patient statistics data was also used to quantify the output of a hospital using the PDE measure. Hospital size was quantified using the number of available inpatient beds (BED) and the total floor area (TFA) of a hospital. Data pertaining to the size of a hospital and its energy and water consumption was also collected and used to formulate the measures for hospital size (TFA and BED) and resource consumption respectively (AEC and AWC).

In Chapter 5, these measures were used in a set of MLR analyses to formulate MLR models that evaluated the explanatory power provided by all the possible combinations of normalising factors when assessing the inter-hospital variance in electricity and water consumption. These models were used to assess the significance of the explanatory power provided by the normalisation factors that were used to represent the function of a hospital. From these analyses' conclusions were drawn on the feasibility of accounting for hospital function in the normalisation model.

6.2 Limitations of study

The ICD 10 diagnostic coding system was recently rolled out to all the hospitals in the Western Cape, implementation started in 2014. Thus, the use of the system to record patient diagnosis is not implemented to the same degree at all the hospitals in the analysis. The resulting variations in data quality are reflected in the number of errors observed in the initial patient statistics dataset from the data recorded at the respective hospitals. This significantly limited the number of hospitals in the sample studied and resulted in a reduction in the size of the sample studied from 32 hospitals to 18.

However, the quality of patient statistics data recorded at the hospitals is likely to improve in the coming years as healthcare practitioners at the respective hospitals become more familiar with the ICD-10 MIT coding standard. This will potentially address this limitation as it will lead to the collection of more accurate patient statistics data that will result in a more accurate and larger data sample for future analyses.

Furthermore, the consumption data was based on the electricity and water metering measurements recorded by municipal officials at the facilities. The consumption measurement records for all the hospitals were not available for some of the calendar years initially considered. Consequently, this limited the years from which eligible annual data could be analysed. Hence, the dataset used in the study only consisted of data from the 2016 calendar year.

In addition, the analysis of data in this study could not be replicated using consumption data from earlier years, and its findings could not be validated by analysing this data. In the coming years, the quality of the data used in studies similar to this study is expected to improve in terms of accuracy and availability. This is likely to result from the installation of smart metering systems for both electricity and water consumption at the hospitals in the province. Hence, the replication of data analysis and the validation of the research findings by comparing consumption data from different years is likely to be feasible in the near future.

Another implication of the installation of smart metering systems is that data on the total electricity and total water consumption of hospitals will become available. The smart metering systems will record the total electricity consumption of hospitals (including the electricity that is generated onsite) and their total water consumption (including water from onsite boreholes and greywater systems). This data will provide a more representative characterisation of the consumption performance behaviour of the hospitals being evaluated.

6.3 Main conclusions

As discussed in Section 5.3, 62 MLR models were developed for the respective combinations of normalising factors. An analysis of these MLR models and their respective statistics, led to the conclusion that when comparing the inter-hospital variations in energy consumption (AEC) and water consumption (AWC), the size of a hospital is the normalising factor that provided the most statistically significant explanatory power.

In these analyses, two variables were used to represent the size of a hospital, total floor area (TFA) and number of beds (BED). Of these two variables, the TFA had the most statistically significant explanatory power. The single-predictor-TFA-model accounted for 75.52 percent of the variance in the electricity consumption (R_a^2=0.7552), and 55.08 percent of the variance in the water consumption (R_a^2 = 0.5508) of hospitals. This suggests that the single-predictor-TFA-model of hospital size accounts for the most statistically significant amount of variance in both the electricity and water consumption of a hospital.

Furthermore, for the multiple-predictor variable models, the models with TFA as one of the predictor variables accounted for approximately 30 percent more variance in AEC, and 25 percent more variance in AWC than the models without TFA. In addition, analyses of the significance of the individual parameters in the multiple-predictor models that contain TFA as a predictor variable were also conducted. From these analyses, it was concluded that for these models, TFA was the only variable in these models that significantly contributed to explaining the variance in AEC and AWC. The other normalising variables in the models did not account for a statistically significant amount of unique variance in AEC or AWC. Taken together, these results show that the inclusion of TFA in a normalisation model increases the robustness of that model.

The second variable used to represent the size of a hospital was BED. It accounted for the most statistically significant amount of variance in the multiple-predictor models that did not contain TFA as a predictor variable. BED and TFA both represent the size of a hospital in the normalisation models. It was thus concluded from the MLR analysis, that the size of a hospital was the most significant factor in explaining the variance in the energy and water consumption of hospitals. Of these two factors, the TFA of a hospital is the most significant normalising factor and it is more powerful than any combination of the other normalising factors.

Tests of the overall significance of the single-predictor-CMPX- and single-predictor-PDE-models, showed that both CMPX and PDE are statistically significant in accounting for the variance in the energy and water consumption performance of hospitals. However, the amount of individual variance accounted for by these factors is substantially less than that explained by models containing BED or TFA as normalising variables. Furthermore, another single-predictor model, SPEC, did not account for a statistically significant amount of variance in both AEC and AWC. This suggests that the variance that is explained by the single-predictor SPEC model could potentially be due to random chance.

Interestingly, the multiple-predictor-variable models with combinations of CMPX, SPEC and PDE all explained less variance than the models containing BED or TFA. From these models, it was concluded that the PDE, CMPX and SPEC variables as representative of the level of medical service provision at a hospital do not significantly contribute to explaining the variance in the energy and water consumption data of hospitals. In addition, the high correlation observed between the potential normalising variables (PDE, CMPX and SPEC) and the respective resource consumption variables (AEC and AWC) did not translate into a significant level of explained variance according to the results of the regression analyses.

This suggests that individually and in combination with each other, PDE, CMPX and SPEC, do not significantly represent any underlying unique drivers of energy and water consumption that are not captured in the TFA and BED variables. Thus, the variance captured by the PDE, CMPX and SPEC metrics used to represent the level of medical service provision at a hospital overlaps with that of the metrics used to represent the size and capacity of a hospital.

It was thus concluded that using the PDE, CMPX and SPEC measures to account for the function and output of a hospital in the normalisation model does not improve the comprehensiveness of the normalisation model. The addition of these measures to the model would complicate the normalisation model without providing any significant explanatory power or increasing the objectivity of hospital consumption performance comparisons.

Furthermore, individually, TFA is a much better predictor of energy and water consumption in hospitals than any combination of the normalising variables considered. Combining TFA with those variables only marginally increases the amount of explained variance in the consumption data. The trade-off between added complexity by increasing the number of terms in the model and the amount of explained variance does not justify the addition of those variables to the normalisation model. Thus, normalising for the TFA of a hospital alone will produce the most comprehensive comparisons of hospital consumption performance.

The study also found that of the two factors currently being used to normalise the electricity and water consumption of hospitals. The total floor area of a hospital is a significantly better predictor of inter-hospital variations in energy and water consumption performance. Thus, the robustness of the resource consumption performance comparisons that are normalised for the number of beds in a hospital could be improved by normalising them using the total floor area of the hospitals. The incorporation of this change would improve the current normalisation process.

6.4 Summary of contributions

The contributions of this study to academic literature and the body of knowledge are centred around the field of energy and water management in hospitals in a South African context. These are:

1. This study has conducted an extensive literature review of the factors affecting energy and water consumption. This review can be used as a point of departure for future studies into the energy or water consumption of hospitals.
2. The study evaluated the comprehensiveness of the measures currently used to normalise the energy and water consumption performance of hospitals in the Western Cape Province. The study has thus contributed to the body of knowledge by analysing and quantifying the effectiveness of the measures currently being used to normalise the energy and water consumption in hospitals in the Western Cape health system context.
3. The study identified and tested potential changes to the measures currently used to normalise energy and water consumption performance comparisons in hospitals in the Western Cape. It assessed whether the incorporation of these changes improved the performance evaluation process. It found that the current comparison measures are robust, and their comprehensiveness was not significantly improved by the addition of the proposed normalisation factors. The study has thus contributed to the body of knowledge by identifying factors that do not improve the robustness of the current normalisation model.
4. The study quantified the robustness of the current normalisation models. It identified that models that normalise for the total floor area of a hospital yield significantly more robust comparisons than those that normalise for the number of available beds in a hospital. Thus, the robustness of the comparisons that are currently used at a policy formulation or management level that normalised for the number of available beds in a hospital can be improved by normalising for the total floor area of a hospital instead.

5. The data availability and quality constraints encountered in the research process of this study may aid in highlighting the need to improve data recording procedures at hospitals.
6. The application of the information theory approach to electricity and water consumption used in this study, may inform future studies to a new technique of performance comparison that may be applicable and significant in another context.

6.5 Suggestions for further research

From the data analysis conducted and the results of this study, it is recommended that investigating the following topics may contribute to the understanding of normalisation in a resource consumption performance comparison context:

1. The inclusion of the average length-of-stay associated with each treated diagnostic case into the formulation of the CMPX and SPEC measures. This will introduce an element that captures the severity of the patient's diagnosis and the extent of the curative measures that were required to treat the diagnosis. Emphasis can be placed on length of stay as some wide spread cases may be associated with a longer stay in hospital and thus a higher demand on the resource consumption of the hospitals.
2. Develop complexity measures for all the cases treated by a hospital, not just the inpatient cases. This can be achieved by applying the ratios used in the development of the PDE metric to develop composite complexity and specialisation measures. These measures will have a more rounded outlook and be more representative of the caseloads of the respective hospitals in the analysis.
3. Apply replication test to check that the results are not just specific to this patient population but can be replicated on another population. It is recommended to study the variation in the case mix of hospitals over time and see if they were constant. This can be used to test the stability of the CMPX and SPEC measures and the formulated normalisation models over time.

Replication test can be used to test whether the findings of the MLR analysis are retained over time or whether they were unique to the 2016 calendar year.

4. Assess whether offsetting the BED metric of each hospital in the analysis by its corresponding bed occupancy rate improves the explanatory power of the BED metric. The bed occupancy rate is a measure of the average percentage of occupied beds during a given reporting period. Thus, offsetting the BED metric by its corresponding bed occupancy rate would account for both the capacity of hospitals and the utilisation of that capacity.
5. Future improvements in the quantity and quality of the available data will make it possible to address some of the constraints and limitations of this study. For example, the exclusive use of municipally recorded electricity and water consumption data thus excluding electricity that was generated onsite, thermal energy consumption, and water from onsite boreholes and greywater systems. This can be address by using the total energy consumption data provided by smart metering systems.

References

Abu Bakar, N.N. et al., 2015. Energy efficiency index as an indicator for measuring building energy performance: A review. *Renewable and Sustainable Energy Reviews*, 44, pp.1–11.

Alshayeb, M., Mohamed, H. & Chang, J.D., 2015. Energy Analysis of Health Center Facilities in Saudi Arabia : Influence of Building Orientation , Shading Devices , and Roof Solar Reflectance. *Procedia Engineering*, 118, pp.827–832. Available at: http://dx.doi.org/10.1016/j.proeng.2015.08.520.

Amunjela, A.S., de Kock, I.H. & Brent, A.C., 2017. Developing Normalised Metrics for Comparing the Energy Use Of Hospitals. In *26th International Association for Management of Technology Conference*. Vienna.

Araújo, C., Barros, C.P. & Wanke, P., 2013. Efficiency determinants and capacity issues in Brazilian for-profit hospitals. *Health Care Management Science*, 16(2), pp.126–38. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23912550.

Bagnasco, A. et al., 2015. Electrical consumption forecasting in hospital facilities : An application case. *Energy & Buildings*, 103, pp.261–270. Available at: http://dx.doi.org/10.1016/j.enbuild.2015.05.056.

Barer, M.L., 1982. Case mix adjustment in hospital cost analysis. Information theory revisited. *Journal of Health Economics*, 1(1), pp.53–80.

BIS, (Bio Intelligence Service) & Cranfield University, 2009. *Study on water performance of buildings*, Paris.

Böhme, T. et al., 2013. Methodology challenges associated with benchmarking healthcare supply chains. *Production Planning and Control*, 24(10–11), pp.1002–1014.

REFERENCES

Bryman, A. et al., 2014. *Reerch methodology: Business and Management Context methodology: Business and Management Context* 3rd ed., Cape Town: Oxford University Press Southern Africa (Pty) Ltd.

Buonomano, A. et al., 2014. Dynamic energy performance analysis : Case study for energy ef fi ciency retro fi ts of hospital buildings. *Energy*, 78, pp.555–572. Available at: http://dx.doi.org/10.1016/j.energy.2014.10.042.

Burman, E. et al., 2014. A comparative study of benchmarking approaches for non-domestic buildings : Part 2 – Bottom-up approach. *International Journal of Sustainable Built Environment*, 3(2), pp.247–261.

Caldera, M., Corgnati, S.P. & Filippi, M., 2008. Energy demand for space heating through a statistical approach: application to residential buildings. *Energy and Buildings*, 40(10), pp.1972–1983.

Catalina, T., Iordache, V. & Caracaleanu, B., 2013. Multiple regression model for fast prediction of the heating energy demand. *Energy & Buildings*, 57, pp.302–312.

Catalina, T., Virgone, J. & Blanco, E., 2008. Development and validation of regression models to predict monthly heating demand for residential buildings. *Energy and Buildings*, 40(10), pp.1825–1832.

Chatterjee, S. & Simonoff, J.S., 2013. *Regression Analysis Handbook of Regression Analysis*, New Jersey: John Wiley & Sons Inc.

Chung, M. & Park, H., 2015. Comparison of building energy demand for hotels , hospitals , and of fi ces in Korea. *Energy*, 92, pp.383–393. Available at: http://dx.doi.org/10.1016/j.energy.2015.04.016.

Chung, W., 2011. Review of building energy-use performance benchmarking methodologies. *Applied Energy*, 88(5), pp.1470–1479.

Chung, W., Hui, Y. V. & Lam, Y.M., 2006. APPLIED Benchmarking the energy efficiency of commercial buildings. *Applied Energy*, 83, pp.1–14.

Chung, W., Hui, Y. V. & Lam, Y.M., 2006. Benchmarking the energy efficiency of commercial buildings. *Applied Energy*, 83(1), pp.1–14.

REFERENCES

Collett, S. et al., 2016. Water usage in a multi-speciality hospital and its effective management Samuel Collett, Ilia Samarin, Ramkrishna Bhalchandra, Jeeva Soundaranayagam, Subrata Garai and Mammen Chandy. *Journal of The Academy of Clinical Microbiologists*, pp.135–137.

Cunninghame, A., 2015. *A Novices Guide to Planning Health Infrastructure: The Benchmarking Aprroach*, Cape Town.

D'Alessandro, D.D. et al., 2016. Water use and water saving in Italian hospitals. A preliminary investigation. *Annali dell'Istituto Superiore di Sanità*, 52(1), pp.56–62.

Deru, M. et al., 2011. *U.S. Department of Energy commercial reference building models of the national building stock*, Available at: http://digitalscholarship.unlv.edu/renew_pubs/44.

EHMI, (Ethiopian Hospital Management Initiative), 2017. *Hospital Performance Monitoring Improvement Manual*, Addis Ababa.

Evans, R.G. & Walker, H.D., 1972. Information theory and the analysis of hospital cost structure. *Canadian Journal of Economics*, 5(3), pp.398–418.

Faezipour, M., Ferreira, S. & Ph, D., 2013. Developing a Systems Thinking Perspective of Hospital Water Sustainability. In *SysCon 2013 - 7th Annual IEEE International Systems Conference*. Orlando: Institute of Electrical and Electronics Engineers.

de Fátima Castro, M. et al., 2015. Development of benchmarks for operating costs and resources consumption to be used in healthcare building sustainability assessment methods. *Sustainability (Switzerland)*, 7(10), pp.13222–13248.

Freudenberg, M., 2003. *Composite Indicators of Country Performance: A Critical Assessment*, Paris. Available at: http://www.oecd.org/sti/working-papers.

Fumo, N., Mago, P. & Luck, R., 2010. Methodology to estimate building energy consumption using EnergyPlus Benchmark Models. *Energy and Buildings*, 42(12), pp.2331–2337.

García-Sanz-Calcedo, J. et al., 2017. Analysis of the Average Annual Consumption of Water in the Hospitals of Extremadura (Spain). *Energies*, 10(479), pp.1–10.

REFERENCES

González, A.G., Salgado, D.R. & Mena, A., 2016. A Quantitative Analysis of Cold Water for Human Consumption in Hospitals in Spain. *Journal of Healthcare Engineering*, 2016.

Gray, A., Vawda, Y. & Jack, C., 2011. *South African Health Review 2011* A. Padarath & R. English, eds., Durban: Health Systems Trust. Available at: http://www.hst.org.za/publications/south-african-health-review-2011.

Gupta, S.K. et al., 2007. *Modern trends in planning and designing of Hospitals*, New Delhi: Jaypee Brothers, Medical Publishers.

Hofstee, E., 2006. *Constructing a good dissertation: a practical guide to finishing a Master's, MBA or PhD on schedule* 1st ed., Johannesburg: Exactica.

Hong, S., Paterson, G., et al., 2014. A comparative study of benchmarking approaches for non-domestic buildings : Part 1 – Top-down approach. *International Journal of Sustainable Built Environment*, 2(2), pp.119–130.

Hong, S., Burman, E., et al., 2014. A Comparative Study of Benchmarking Approaches for Non-domestic Buildings : Part 1 – Top-down Approach for non-domestic buildings : Part 1 – Top-down approach. *International Journal of Sustainable Built Environment*, 2(2), pp.119–130. Available at: http://dx.doi.org/10.1016/j.ijsbe.2014.04.001.

Hu, S.C., Chen, J.D. & Chuah, Y.K., 2004. Enery Cost and Consumption in a Large Acute Hospital. *International Journal on Architectural Science*, 5(1), pp.11–19.

Hygh, J.S. et al., 2012. Multivariate regression as an energy assessment tool in early building design. *Building and Environment*, 57, pp.165–175.

Jacobs, R., Smith, P. & Goddard, M., 2004. *Measuring performance: An examination of composite performance indicators*, York, United Kingdom.

Korolija, I. et al., 2011. Influence of building parameters and HVAC systems coupling on building energy performance. *Energy & Buildings*, 43, pp.1247–1253.

REFERENCES

Korolija, I. et al., 2013. UK office buildings archetypal model as methodological approach in development of regression models for predicting building energy consumption from heating and cooling demands. *Energy & Buildings*, 60, pp.152–162. Available at: http://dx.doi.org/10.1016/j.enbuild.2012.12.032.

Leipziger, D., 2013. *Comparing Building Energy Performance Measurement*, Washington DC.

Ma, H. et al., 2017. Analysis of typical public building energy consumption in northern China. *Energy and Buildings*, 136, pp.139–150.

Ma, H. et al., 2016. Public building energy consumption level and influencing factors in Tianjin. *Energy Procedia*, 88, pp.146–152. Available at: http://dx.doi.org/10.1016/j.egypro.2016.06.039.

Massyn, N. et al., 2017. *District Health Barometer 2016/17*, Available at: http://www.hst.org.za/publications/District Health Barometers/District Health Barometer 2016-2017.pdf%0Ahttp://www.hst.org.za.

Monts, J.K. & Blissett, M., 1982. Assessing Energy Efficiency and Energy Conservation Potential Among Commercial Buildings : a Statistical Approach. *Energy*, 7, pp.861–869.

Morgenstern, P. et al., 2016. Benchmarking acute hospitals: Composite electricity targets based on departmental consumption intensities? *Energy and Buildings*, 118, pp.277–290.

Mouton, J., 2005. *How to succeed in your masters and doctoral studies: A South African guide and research book* 1st ed., Pretoria: Van Schaik Publishers.

Murray, J., Pahl, O. & Burek, S., 2008. Evaluating the scope for energy-efficiency improvements in the public sector: Benchmarking NHSScotland's smaller health buildings. *Energy Policy*, 36(3), pp.1236–1242.

MWRA, (Massachusetts Water Resources Authority), Water Use Case Study: Norwood Hospital. *MWRA Online*. Available at: http://www.mwra.state.ma.us/04water/html/bullet1.htm [Accessed October 20, 2016].

REFERENCES

Nayar, P. & Ozcan, Y.A., 2008. Data envelopment analysis comparison of hospital efficiency and quality. *Journal of Medical Systems*, 32(3), pp.193–199.

OECD, (Organisation for Econmonic Co-operation and Development) & JRC, (Joint Reasearch Centre) European Commision, 2008. *Handbook on Constructing Composite Indicators: Methodology and User Guide*, Paris: OECD Publishing. Available at: http://www.oecd-ilibrary.org/economics/handbook-on-constructing-composite-indicators-methodology-and-user-guide_9789264043466-en.

Pacheco, R., Ordonez, J. & Martínez, G., 2012. Energy efficient design of building: A review. *Renewable and Sustainable Energy Reviews*, 16, pp.3559–3573.

Pizzol, M. et al., 2017. Normalisation and weighting in life cycle assessment: quo vadis? *International Journal of Life Cycle Assessment*, 22(6), pp.853–866.

Radwan, A.F. et al., 2016. Retrofitting of existing buildings to achieve better energy-efficiency in commercial building case study : Hospital in Egypt. *Alexandria Engineering Journal*, 55(4), pp.3061–3071. Available at: http://dx.doi.org/10.1016/j.aej.2016.08.005.

Rajagopalan, P. & Elkadi, H., 2014. Energy Performance of Medium-sized Healthcare Buildings in Victoria, Australia- A Case Study. *Journal of Healthcare Engineering*, 5(2), pp.247–260. Available at: http://www.hindawi.com/journals/jhe/2014/974863/.

Rohde, T. & Martinez, R., 2015. Equipment and Energy Usage in a Large Teaching Hospital in Norway. *Journal of Healthcare Engineering*, 6(3), pp.419–434.

SANDoH, (South African National Department of Health), 2002. *A District Hospital Service Package for South Africa: National norms and standards for district hospitals*, Pretoria.

SANDoH, (South African National Department of Health), 2012a. National Health Act, 2003. *Goverment Gazette*, (35101), pp.3–28.

REFERENCES

SANDoH, (South African National Department of Health), 2012b. *Technical User Guide compiled by the Ministerial ICD-10 Task Team to define standards and guidelines for ICD-10 coding implementation*,

SANDoH, (South African National Department of Health), 2014. *The South African ICD-10 Morbidity Coding Standards and Guidelines*,

Sherman, H.D. & Zhu, J., 2006. Improving Service Performance using Data Envelopment Analysis (DEA). In *Service Productivity Management*. New York: Springer Science+Business Media, LLC.

Singer, B.C. et al., 2009. *Hospital Energy Benchmarking Guidance - Version 1 . 0*, Berkeley CA.

Szklo, A.S., Soares, J.B. & Tolmasquim, M.T., 2004. Energy consumption indicators and CHP technical potential in the Brazilian hospital sector. *Energy Conversion and Management*, 45(13–14), pp.2075–2091.

Taleb, H.M., 2016. Enhancing the skin performance of hospital buildings in the UAE. *Journal of Building Engineering*, 7, pp.300–311. Available at: http://dx.doi.org/10.1016/j.jobe.2016.07.006.

Thopil, G.A. & Pouris, A., 2016. A 20 year forecast of water usage in electricity generation for South Africa amidst water scarce conditions. *Renewable and Sustainable Energy Reviews*, 62, pp.1106–1121.

VDoH, (Victorian Government Department of Health), 2009. *Guidelines for water reuse and recycling in Victorian health care facilities - Non-drinking applications*, Melbourne.

Velasquez, M. & Hester, P.T., 2013. An Analysis of Multi-Criteria Decision Making Methods. *International Journal of Operations Research*, 10(2), pp.56–66.

Verlicchi, P. et al., 2010. Hospital effluents as a source of emerging pollutants : An overview of micropollutants and sustainable treatment options. *Journal of Hydrology*, 389(3–4), pp.416–428.

REFERENCES

WCDEADP, (Western Cape Department of Environmental Affairs and Development Planning), 2008. *A Guide to Energy Management in Public Buildings*, Cape Town. Available at: http://www.cityenergy.org.za/uploads/resource_156.pdf.

WCDoH, (Western Cape Department of Health), 2009. *L1/L2/L3 Acute hospital packages of care*, Cape Town.

Weimar, B.D. & Browning, A., 2010. Reducing Water Costs in Building HVAC Systems. *Facility Engineering Journal*, (June), pp.24–26.

Wong, L.T. & Mui, K.W., 2008. Epistemic water consumption benchmarks for residential buildings. *Building and Environment*, 43, pp.1031–1035.

Zhu, Y., 2006. Applying computer-based simulation to energy auditing: A case study. *Energy and Buildings*, 38, pp.421–428.

Appendix A Literature review procedure and results

A.1 Factors that affect the energy use of hospitals

The reviewed publications were obtained by tailoring the search parameters of the online database for academic journals; Scopus. The list of reviewed publications was expanded to include the relevant publications obtained in the references of the publications found in the Scopus search. Table 3.1 shows the parameters used to define the search and the number of publications each search found.

Table A.1: Search parameters used in the energy use literature analysis

Category	Limited to	Publications	
		First search	Reviewed
Article title, Abstract, Keywords	("energy use") OR ("energy consumption") AND ("hospital" OR "healthcare facilit*")	296	25
Year of publication	2000 – 2017		
Subject area	("engineering" OR "energy" OR "environmental science")		

In all, 296 Publications were found during the first Scopus search. The focus of these publications and their relevance to the objective of the review (finding the factors that affect energy use in hospitals) were used as the inclusion-exclusion criteria. The keywords, abstracts and the body of the publications were analysed in relation to the inclusion-exclusion criteria. This yielded 25 publications for inclusion in the literature analysis.

The 25 publications were grouped into 5 categories according to their focus, namely: building design, end use and internal load, energy consumption, energy efficiency retrofits and reference building models. These categories speak to the focus of the publications and how they relate to the energy consumption within the hospital. Most of the publications reviewed (16 of the 25) fall into the energy consumption category (see Figure A.1). Due to the size of this category, the publications within it were subdivided according to their secondary area of focus into five smaller subcategories: benchmarking models, energy-demand models, energy-use analysis, forecasting and overview of hospital energy consumption.

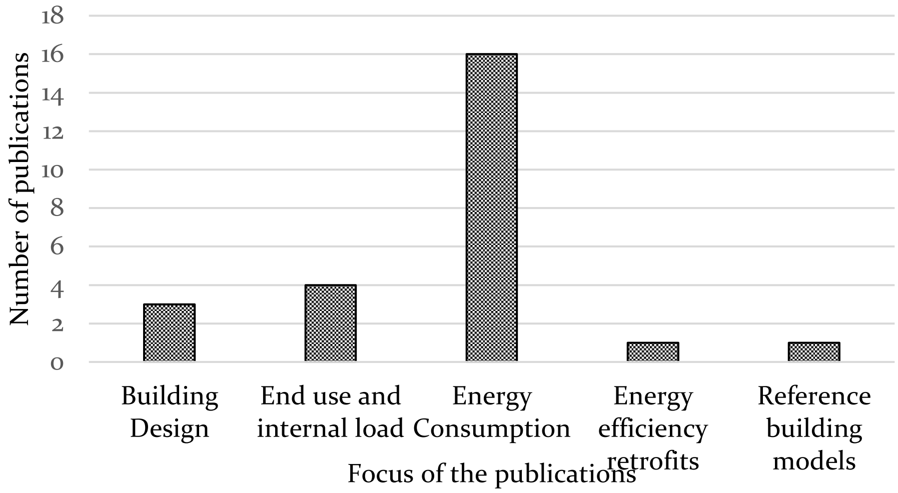


Figure A.1: Reviewed publications

Figure A.2 presents the distribution of the publications in the 'energy-consumption' category with respect to their secondary area of focus. The energy-consumption category dominated the publications and the general state of consumption within hospitals will thus significantly influence the types of factors found in the analysis. This is advantageous as it does not allow one set of factors to significantly influence the factors that are identified.

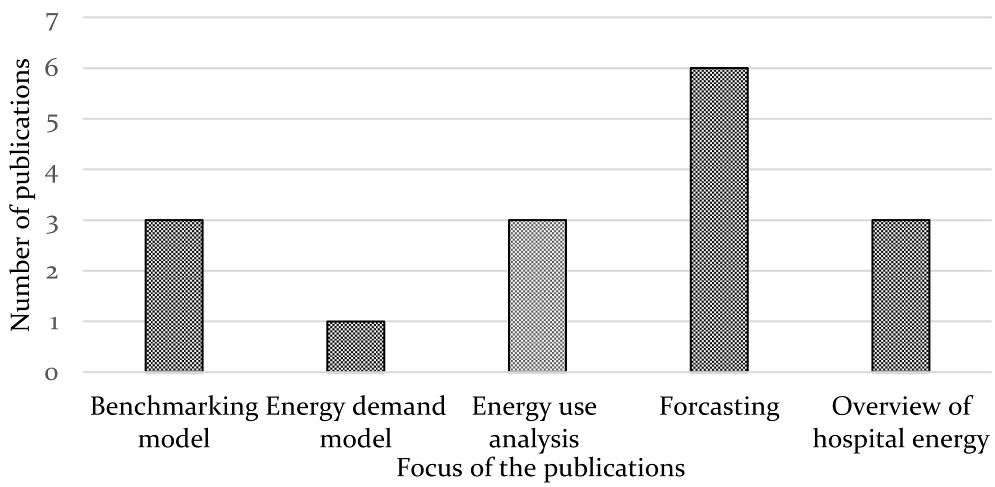


Figure A.2: Focus of the publications in the energy- consumption category

APPENDIX A

Table A.2: Factors affecting energy consumption as identified in literature

		Factors	Martini et al. (2005)	Szklo et al. (2004)	Zhu (2005)	Caldera et al. (2008)	Murray et al. (2008)	Fumo et al. (2009)	Singer et al. (2009)	Deru et al. (2011)	Catalina et al. (2012)	Kolokotsa et al. (2012)	Korolija et al. (2011)	Korolija et al. (2013)	Pacheco et al. (2012)	Zhao et al. (2012)	Rajagopalan & Elkadi (2014)	Buonomano et al. (2014)	Alshayeb et al. (2015)	Bagnasco et al. (2015)	de Fátima Castro et al.	Chung & Park (2015)	Rodhe & Martinez (2015)	Ma et al. (2016)	Radwan et al. (2016)	Taleb (2016)	Ma et al. (2017)
Hospital characteristics and construction	Building characteristic	Year of construction				E	E			E		E					E				E			E			E
		Location					E	E	E																		E
		Size (floor area, volume)		E	E	E	E		E	E							E				E						E
		Geometric layout			E									E													
		Thermal zoning						E				E	E					E									
	Building fabric	Envelope thermal properties	E		E	E		E		E	E			E	E	E	E							E		E	E
		Glazed surfaces				E				E	E		E	E	E										E		
		Air-tightness								E				E										E			E
		Level of insulation										E		E													
	Building Form	Aspect ratio						E		E																	
		Orientation						E	E						E				E						E		
		Window to wall ratio						E					E				E								E		E
		Shading												E	E				E								E
		Compactness ratio											E		E		E										
Weather and climate factors		Seasonal factors		E				E												E		E					
		Temperature				E					E					E	E			E							
		Climate						E					E									E		E			
		Humidity															E			E							
		Solar irradiance									E						E							E			

APPENDIX A

		Factors	Martini et al. (2005)	Szklo et al. (2004)	Zhu (2005)	Caldera et al. (2008)	Murray et al. (2008)	Fumo et al. (2009)	Singer et al. (2009)	Deru et al. (2011)	Catalina et al. (2012)	Kolokotsa et al. (2012)	Korolija et al. (2011)	Korolija et al. (2013)	Pacheco et al. (2012)	Zhao et al. (2012)	Rajagopalan & Elkadi (2014)	Buonomano et al. (2014)	Alshayeb et al. (2015)	Bagnasco et al. (2015)	de Fátima Castro et al. (2015)	Chung & Park (2015)	Rodhe & Martinez (2015)	Ma et al. (2016)	Radwan et al. (2016)	Taleb (2016)	Ma et al. (2017)
Function of hospital	Clinical services	Complexity of service offered		E					E			E	E					E						E			
		Capacity (number of beds)		E					E	E											E						
		Clinical specialties		E			E		E		E	E	E					E			E		E	E			
		Catchment area																			E						
	Equipment and lighting	Equipment type								E													E				
		Prevalence	E																				E				E
		Use patterns	E		E			E								E	E						E				E
		Level of maintenance on equip.										E															
	End uses	System type & configuration			E		E	E	E	E			E				E								E		E
		Equipment power rating																					E				
		Operational parameters								E			E				E										
		Efficiency of equipment		E				E		E			E														E
	Building Use	Type of spaces			E					E							E	E							E		
		Occupancy rates & density	E							E	E						E								E		
		Heat gains: equip. & lighting									E			E									E		E		
		Heat gains from occupants												E											E		
		Operating schedule			E			E		E			E				E					E					
		Occupant behaviour			E											E											

A.2 Factors that affect the water use of hospitals

The second part of the literature analysis focused on the identification of the factors affecting the water consumption of hospitals. As in Section 2.1, the online database for academic journals; Scopus, was used to select the reviewed publications. Water consumption in hospitals is an under-researched field. García-Sanz-Calcedo et al. (2017) encountered a similar problem in their study of water consumption in Spanish hospitals. Table A.3 shows the parameter used to tailor the literature search and the publications found during the search.

The initial Scopus search generated a limited amount of relevant published literature. Of the nine publications included in the study, two came from the Scopus search. The rest of the publications were excluded on the grounds of their relevance to the focus of the literature analysis. Two publications were excluded because they were written in Japanese and Mandarin respectively. The literature analysis was expanded by studying the reference lists of the publications found in the Scopus search and including relevant publications. Figure A.3 shows the type of literature used in this review: because of the limited amount of publications both academic and grey literature was reviewed.

Table A.3: Search parameters used in the water use literature analysis

Category	Limited to	Publications	
		First search	Reviewed
Article title, Abstract, Keywords	("water use") OR ("water consumption") AND ("hospital" OR "healthcare facilit*")	58	9
Year of publication	2000 - 2017		
Subject area	("engineering" OR "energy" OR "environmental science")		

Figure A.3 and Figure A.4 show the area of focus of each publication in the review. More than half of the publications focused on water-consumption analysis and provided a good overview of water consumption within hospitals. Two studies modelled water use within hospital buildings, focusing on benchmarking consumption and forecasting consumption respectively.

APPENDIX A

BIS & Cranfield University (2009) studied the water consumption of residential and non-residential, public and commercial buildings. In their publication, building water consumption is divided into two categories which represent the type of water use in all types of buildings: human water needs, and productive water uses. Human water needs are defined as the water used for sanitary and leisure activities. Productive water use is defined as the water that is applied to fulfilling the building's intended purpose. In the case of hospitals, this purpose is providing an environment that facilitates the improvement of the health of patients.

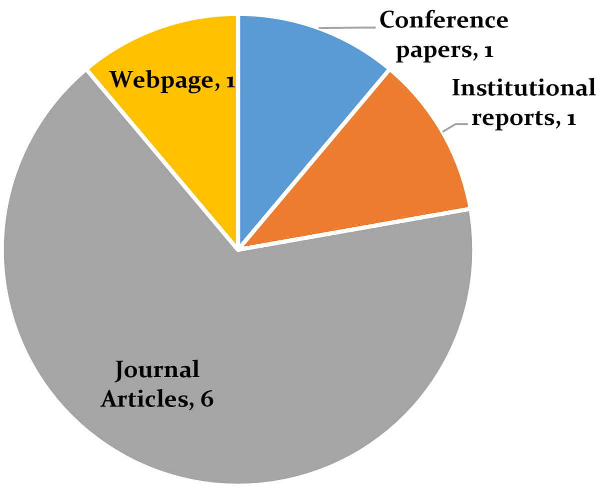


Figure A.3: Type of literature used in review

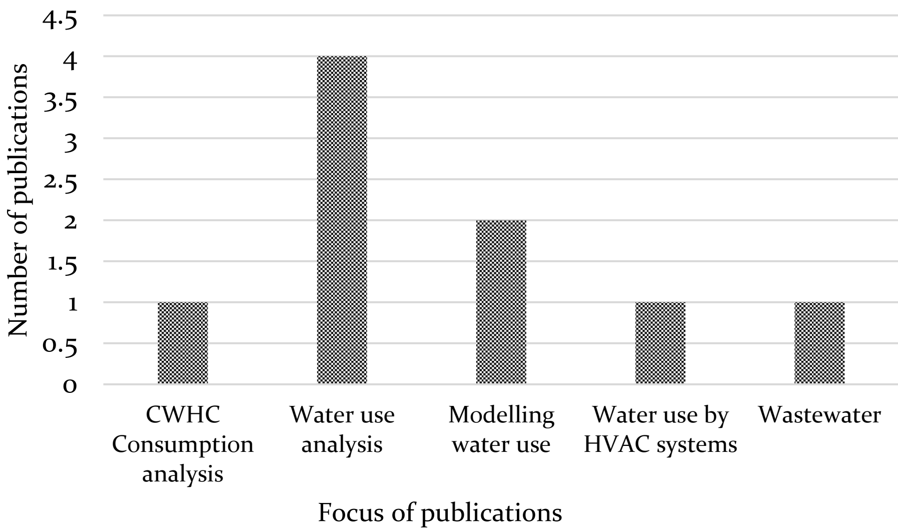


Figure A.4: Area of focus of the publications

APPENDIX A

Table A.4: Factors affecting water consumption as identified in literature

	Factor	(MWRA n.d.)	Wong & Mui (2008)	BIS & Cranfield University (2009)	Verlicchi et al. (2010)	Weimar & Browning (2010)	D'Alessandro et al. (2016)	Faezipour & Ferreira (2016)	González et al. (2016)	García-Sanz-Calcedo et al. (2017)
Hospital characteristics and Context	Year of construction	W			W				W	W
	Location				W				W	
	Size		W	W				W	W	W
	Culture		W		W					
	Economy		W						W	
	Climate		W	W	W					
	Seasons				W					
	Temperature								W	
Management policies & practices	Management policies and awareness	W			W		W	W	W	W
	Maintenance practices	W								
	Incentives for implementing water sustainability measures							W		
Clinical service	Capacity (number of beds)	W		W	W		W	W	W	W
	Type and complexity of service offered	W			W	W			W	W
	Complementary services provided by the hospital									W
	Mix of areas (space area ratios)		W							
Building Use	Occupancy	W	W						W	W
	Amount of green areas in the hospital								W	W
	Average flow rates of water taps		W	W						
	Type of HVAC cooling system	W				W	W		W	

Appendix B Statistical performance analysis approaches

The approaches discussed in this Appendix are quantitative in nature and focus on determining a baseline or reference to which the alternatives are compared, and their performance is evaluated, thus benchmarking the performance of hospitals against said reference.

B.1 Multivariate Linear Regression (MVLR)

Multivariate linear regression (MVLR) determines a reference consumption performance index for a set of buildings and then assesses the performance of each building in that set relative to the reference index. A multivariate linear regression equation is used to generate the reference index, where the resource consumption is the dependent variable, and the building attributes that need to be accounted for are the independent variables. The approach generates a benchmark by analysing a dataset that consists of the resource consumption measurements of a set of hospitals and their corresponding attributes.

Regression is performed to fit a line or plane that best represents the relationship between the dependent variable and the independent variables for all the data points in the dataset. The difference between each data point and the regression line or plane represents the error associated with the performance of each dependent variable for that independent variable (Chung 2011). The regression model yields an equation that can be used to predict the energy consumption of the building as a function of its key parameters. The estimated multiple regression equation is of the following form:

$$\hat{y}(x_1, x_2, x_3, \dots, x_n) = \beta_0 \pm \sum_{i=1}^{n} \beta_i x_i + \varepsilon \tag{0.1}$$

Where,

$\hat{y}_j(x_i)$	represents the estimated energy or water consumption for the *j-th* hospital;
x_i	represents the attributes of the respective hospital;
β_0	represents the intercept; and
β_i	represents the estimates of the coefficients of the regression equation.

Each coefficient (β_i) represents the estimated change in y that corresponds to a unit change in x_i when all other $x_i's$ are held constant. Thus the coefficients of the regression equation (β_i) are proportional to the estimated dependent variable's ($\hat{y}_j(x_i)$) sensitivity to changes in an independent variable (x_i) (Hygh et al. 2012).

The regression model generates a MVLR equation for a dataset consisting of the resource consumption level and the attributes of each alternative. This equation can be used to predict the energy consumption of the building as a function of its key parameters. The equation is used to estimate a reference performance index for each alternative by substituting its parameters into the regression equation. Thus, in order to attribute a score to the performance of a hospital (y_j), the actual resource consumption is normalised by its estimated performance ($\hat{y}_j(x_i)$).

$$H_j = \frac{y_j}{\hat{y}_j(x_i)} \tag{0.2}$$

MVLR is a sophisticated, yet straightforward and inexpensive approach that can be carried out using most spreadsheet software packages (Chung 2011). However, it requires a large set of data to generate an accurate and robust predictive model (Catalina et al. 2008). Furthermore, the residuals that capture the inefficiency in the consumption performance of the hospitals are calculated relative to the fitted average function. This fitted average does not correspond to an efficiency frontier and does not quantitatively say anything about the consumption efficiency of the hospital with respect to the most efficient hospital within the sample being studied (Chung 2011).

The fitted average captures inefficiencies in the performance of the hospitals as well as statistical noise due to unexplained factors and data errors. MVLR groups the effects of these errors and unexplained factors within the metric that measures the relative efficiency levels for each alternative in the analysis. Monts & Blissett's (1982) energy utilisation index (EUI) approach is based on multivariate linear regression.

The regression-based approach described by Monts & Blissett (1982) has the same theoretical basis as various benchmarking approaches for non-domestic buildings found in literature. Furthermore, the US Environmental Protection Agency's Energy Star scheme is based on this method (Hong, Burman, et al. 2014), and benchmarks for commercial buildings in Hong Kong were developed based on this method (Chung, Hui & Y. M. Lam 2006).

B.2 Data Envelopment Analysis (DEA)

This approach uses data envelopment analysis (DEA) to formulate an efficiency frontier. This frontier is used to benchmark the performance of each alternative in the study (OECD & JRC 2008). DEA can be used to compare and benchmark the 'performance' of different alternatives with respect to a set of attributes and identify the best performing alternative(s) or develop an artificial best performing alternative against which the other alternatives are benchmarked (Sherman & Zhu 2006). This best performing alternative(s) forms the efficiency frontier. The distance between the frontier and each alternative represents the inefficiency of that alternative.

$$A_j = \frac{D_j}{D_j^*} = \frac{\sum_{i=1}^{n} w_{ij} a_{ij}}{\sum_{i=1}^{n} w_{ij} a_{ij}^*} \tag{0.3}$$

Where,

- A_j the normalised and weighted resource (energy or water) consumption performance score awarded to the *j-th* hospital;
- a_{ij} The performance of the *j-th* hospital on the *i-th* attribute;
- a_{ij}^* The optimal performance of the *j-th* hospital on the *i-th* attribute, represented by its projection from the origin onto the frontier; and
- w_{ij} The weighting factor associated with a_{ij}; and $w_{ij} \in [0,1]$.

The efficiency of an alternative can be defined as the ratio of its current level of performance (D_j) over its optimal operating level (D_j^*) (Chung 2011). The DEA approach assumes that the variance in the performance of alternatives is because of differences in the degree to which the respective attributes affect the resource consumption at the respective alternative. Thus, there exists an optimal version (D_j^*)

of each alternative (D_j) that operates at the optimal operating level for that alternative given its attributes.

The set of these optimal alternatives forms the efficiency frontier. Figure B.1 illustrates this for a simple six alternative, two attribute example, where D_j represents the alternatives and D_j^* represents the projection of that alternative on the efficiency frontier. The analysis outputs a single composite score on the interval [0,1] that represents the performance of the alternative. The alternatives on the efficiency frontier (D_2, D_4, and D_5) are awarded the maximum score ($A_j = 1$). All non-efficiency-frontier alternatives (D_1, D_3, and D_6) are awarded a score $A_j < 1$. Since the approach computes and analyses the efficiency of an alternative relative to the frontier, the frontier provides a means of identifying and determining potential areas of improvement and quantifying the improvements needed to increase the efficiency of an inefficient alternative (Chung 2011).

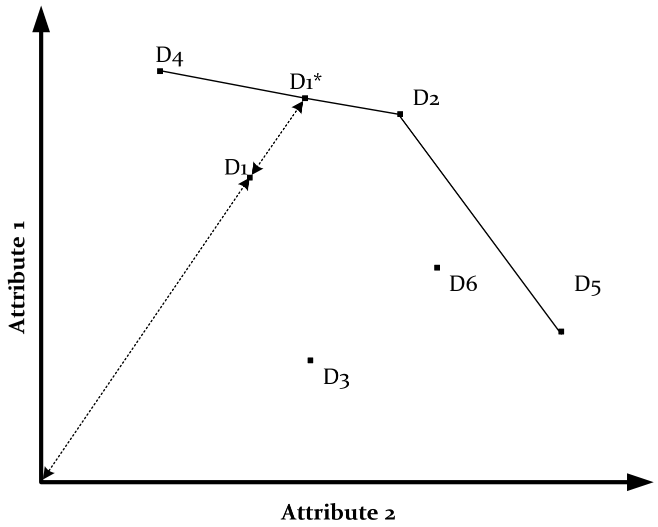


Figure B.1: Illustrative example of an efficiency frontier

One of the limitations of DEA is that the choice of the frontier is sensitive to outliers. Outliers (i.e. due to data errors) can cause large inefficiency values. There are two types of outliers; outliers that overestimate consumption, thus significantly increasing the inefficiency of one alternative, and outliers that underestimate consumption, thus significantly increasing the inefficiency of all the alternatives.

Another limitation is that adding attributes never decreases the individual efficiency score of the alternatives in the analysis. This may be misleading, especially if the analysis consists of a large number of attributes and a small alternative sample size; every alternative can be located on the frontier (Chung 2011). Thus there can be multiple optimal solutions, no unique solution and thus no frontier against which to benchmark the alternatives (Pizzol et al. 2017).

Other limitations of the DEA approach include its requirement for precise data which is not always possible in a real-world context (Velasquez & Hester 2013). Furthermore, low inefficiency scores may be caused by factors that are not included in the model. Thus the method may not account for the effect of some of the attributes on the resource consumption because the underlying attributes could not be modelled and were not testable (Chung 2011).

Appendix C ICD-10 MIT chapter descriptions

Table C.1 details the macro-level clinical diagnosis and cause of morbidity groupings defined in the ICD-10 MIT table. The caseloads of each clinical speciality provided at the hospitals in the analysis were defined by classifying the cases treated under each speciality into these groupings

Table C.1: ICD-10 MIT chapter descriptions used to classify hospitals case load

Chapter	Description
CHAPTER I	Certain infectious and parasitic diseases (A00-B99)
CHAPTER II	Neoplasms (C00-D48)
CHAPTER III	Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism (D50-D89)
CHAPTER IV	Endocrine, nutritional and metabolic diseases (E00-E90)
CHAPTER V	Mental and behavioural disorders (F00-F99)
CHAPTER VI	Diseases of the nervous system (G00-G99)
CHAPTER VII	Diseases of the eye and adnexa (H00-H59)
CHAPTER VIII	Diseases of the ear and mastoid process (H60-H95)
CHAPTER IX	Diseases of the circulatory system (I00-I99)
CHAPTER X	Diseases of the respiratory system (J00-J99)
CHAPTER XI	Diseases of the digestive system (K00-K93)
CHAPTER XII	Diseases of the skin and subcutaneous tissue (L00-L99)
CHAPTER XIII	Diseases of the musculoskeletal system and connective tissue (M00-M99)
CHAPTER XIV	Diseases of the genitourinary system (N00-N99)
CHAPTER XV	Pregnancy, childbirth and the puerperium (O00-O99)
CHAPTER XVI	Certain conditions originating in the perinatal period (P00-P96)
CHAPTER XVII	Congenital malformations, deformations and chromosomal abnormalities (Q00-Q99)
CHAPTER XVIII	Symptoms, signs and abnormal clinical and laboratory findings not elsewhere classified (R00-R99)
CHAPTER XIX	Injury, poisoning and certain other consequences of external causes (S00-T98)
CHAPTER XX	External causes of morbidity and mortality (V01-Y98)
CHAPTER XXI	Factors influencing health status and contact with health services (Z00-Z99)
CHAPTER XXII	Codes for special purposes (U00-U99)
CHAPTER M	Morphology of neoplasms (M000-M999)

Appendix D Data collection

D.1 Annexure A: application for health data

Western Cape Government
Health

Annexure A

APPLICATION FOR ACCESS TO HEALTH DATA

The following application form is to be completed by all person/persons/organisations/groups who wish to access health-related data from Western Cape Department of Health and is to be completed in accordance with the Departments' *Guidelines on requests for access to patient data and patient information systems from the Department of Health*. Please note that application for use of data does not guarantee that the data request will be approved. If the intended purpose for data access is altered or extended in anyway, a new agreement must be entered into.

Applicant details: *(Refers to the detail of the person requesting the change.)*

Name:	Abimelek Shikongo	**Surname:**	Amunjela
Designation / Rank:	Masters Student	**Date:**	27/09/2017
Organisation:	Department of Industrial Engineering, Stellenbosch University		
Email:	16208803@sun.ac.za	**Tel/Cell:**	+27 785 706 461

Please supply the contact detail of the person to whom the processed application must be returned.

ASA

Details of Data Request: (please append any additional information where necessary)

Type of Data Requested : (please tick appropriate option)	Aggregated data X	Non-identified individualised data	Identified individualised data

Please provide a short description of the data requested. Please attach a list/attach a list of the variables required.

- Monthly electricity bills for each facility,
- Monthly water bills for each facility,
- Monthly energy audit data for each facility,
- Monthly water audit data for each facility,
- Climate data (temperature and humidity) at every facility,
- Occupancy pattern data (number of occupants and period of occupancy),
- Facility characteristic data (age, total floor area, number of beds etc),
- Building use data (departments present at different facilities, types of equipment used at facilities)

Time period the data should cover:	Start date: 01/01/2014	End date: 31/12/2016
Frequency of Access: (please tick appropriate option)	Once-off X	Periodically

ASA

If periodically, please specify time frames for access:

Is the data to be used for research purposes?	Yes	No

Please provide a brief motivation for this request, highlighting the purpose for which the data will be used

The data will be used to provide a quantitative base for the development of a framework for quantify the actual energy and water consumption performance potential of hospitals. The framework can be used to assess the energy and water performance of hospitals and to compare hospitals energy and water performance accurately. This is useful at a both the local and the policy making levels.

Do you have a security protocol for handling the data (attach detail if necessary)?	Yes	No

APPENDIX D

Outcome of Application: (To be completed by the Designated Health Authority)

Name:		**Surname:**	
Designation / Rank:		**Signed:**	
Application Approved:	Yes / No	**Date:**	

TERMS OF AGREEMENT FOR DATA ACCESS

The Western Cape Department of Health is committed to ensuring availability of data that supports the provision of health care and other essential services to authorised Users. This agreement aims to ensure the authorisation, maintenance of confidentiality and appropriate use of the data provided to Users.

This agreement is between:

The Western Cape Government: Department of Health, hereafter "the Department"

AND

Abimelek Shikongo Amunjela, hereafter "the User"

1. Application for use of data must be made through the channels identified in the "*Guidelines on requests for access to patient data and patient information systems*" document.
2. This agreement sets forth the terms and conditions to which the Department will disclose certain confidential health information in the form of a Data Set(s).
3. The User agrees that the Department is the owner of the Data Set(s).
4. Permitted Uses and Disclosures

4.1. Except as otherwise specified herein, the User may make all uses and disclosures of the [Insert name of Data Set(s)] necessary to conduct the Master of Science in Engineering Management thesis titled: "The normalisation of resource efficiency measures in healthcare facilities: the case of energy and water" for the period starting 20/10/2017 and ending 27/09/2018. ASA

4.2. The User will receive the Data Set(s) once off, from the designated Department official. ASA

4.3. In addition to the User, the individuals, or classes of individuals, who are permitted to use or receive the Data Set(s) for purposes of the Identified Project include: Abimelek Shikongo Amunjela. ASA

5. User Responsibilities

5.1. The User will not use or disclose the Data Set(s) for any purpose other than permitted by this Agreement pertaining to [insert project name/report name] for which written approval was granted.

5.2. The User agrees that the Data Set(S) provided will not be released to any third party that is not included by the provisions of the agreement between the primary parties, without the written permission of the Department. A third party will be required to complete an agreement as well.

2

5.3. The User agrees that the Department will be provided with an opportunity to comment and give feedback prior to the finalisation of any report/publication derived from the Data Set(s) according to the following conditions:

5.3.1. The data will be used to compile Master of Science in Engineering Management thesis titled: "The normalisation of resource efficiency measures in healthcare facilities: the case of energy and water" for presented in partial fulfilment of the requirements for the degree of Master of Science in Engineering Management at Stellenbosch University. ASA

5.3.2. The report will be sent to the Department for perusal prior to finalisation. The latter should respond or react on the report issued within 31 working days, if this period lapses it would be interpreted as a confirmation that the Department acknowledges the presentation and interpretation of data as correct and factual in the report.

5.4. The User will ensure that the Department is acknowledged in any output resulting from the use of the data including.

5.5. The User will communicate any data quality issues identified to the Department to improve the dataset.

5.6. The User agrees that any use of the Data Set(s) or reliance by the User on any of the Data Set(s) is at the User's own risk and that Department shall not be held liable for any loss or damage howsoever arising as a result of such use.

5.7. The User agrees that he/she will make no statement nor permit others to make statements indicating or suggesting that interpretations/views drawn from the findings are those of the Department.

5.8. The User agrees that he/she will maintain confidentiality in accordance with item 6. Below.

6. Data Security and Confidentiality

6.1. All Data Set(s) from the Western Cape Department of Health are to be treated as confidential and used in accordance with the following security standards:

6.1.1. Database storage: At a minimum the database must have user-level security, may not be housed on laptops or external media unless these are encrypted. Ideally the data should be stored on a central server with restricted access and not be stored on portable computer equipment like memory sticks, external hard drives and laptops.

6.1.2. The Data Sets(s) must be password protected and that such passwords are not to be shared with anyone other than the principle user.

6.1.3. Data may not be linked to personally identifiable records from any other source unless prior approval has been granted.

6.1.4. File storage: At a minimum files will be stored with AES encryption e.g. 7-zip, and 15 character passwords which include numbers, special characters and letters.

6.1.5. Passwords and files may not be provided together but using two different methods of communication e.g. data zipped and e-mailed while password is SMS'ed to User.

6.1.6. When the timeframe for the agreed utilisation of the data expires (see item 4.1. above) the data must be destroyed in all its forms.

APPENDIX D

7. In making information available, the Department of Health reserves the right to set conditions in which its staff (including academic staff in joint provincial posts) should be invited to participate in any research undertaken that uses the data they have generated with a view to co-authorship of the final report/s.

8. The User accepts that this data is routinely collected as part of service delivery and therefore the data quality may not be of the highest quality.

9. Failure to adhere to the written agreement can and may be sanctioned

Signatories

Abimelek Shikongo Amunjela		**27/09/2017** ASA
User's Name (Print)	Signature	Date
Department of Health (Designated authority)	Signature	Date

D.2 Ethical clearance

UNIVERSITEIT·STELLENBOSCH·UNIVERSITY
jou kennisvennoot • your knowledge partner

Approval Notice
New Application

30-Sep-2016
Amunjela, Abimelek AS

Proposal #: SU-HSD-002329
Title: An investigation into the normalisation of resource efficiency measures in healthcare facilities: The case of energy and water

Dear Mr Abimelek Amunjela,

Your **New Application** received on **09-Aug-2016**, was reviewed
Please note the following information about your approved research proposal:

Proposal Approval Period: **29-Sep-2016 -28-Sep-2019**

Please take note of the general Investigator Responsibilities attached to this letter. You may commence with your research after complying fully with these guidelines.

Please remember to use your **proposal number** (**SU-HSD-002329**) on any documents or correspondence with the REC concerning your research proposal.

Please note that the REC has the prerogative and authority to ask further questions, seek additional information, require further modifications, or monitor the conduct of your research and the consent process.

Also note that a progress report should be submitted to the Committee before the approval period has expired if a continuation is required. The Committee will then consider the continuation of the project for a further year (if necessary).

This committee abides by the ethical norms and principles for research, established by the Declaration of Helsinki and the Guidelines for Ethical Research: Principles Structures and Processes 2004 (Department of Health). Annually a number of projects may be selected randomly for an external audit.

National Health Research Ethics Committee (NHREC) registration number REC-050411-032.

We wish you the best as you conduct your research.

If you have any questions or need further help, please contact the REC office at .

Included Documents:

DESC Report
REC: Humanities New Application

Sincerely,

Clarissa Graham
REC Coordinator
Research Ethics Committee: Human Research (Humanities)

Appendix E Data collection

E.1 Hospitals in the initial dataset

Table E.1: List of district hospitals in the study's initial dataset

Hospital Name	Hospital Code
Ladismith (Alan Blyth) Hospital	ABH
Otto Du Plessis Hospital	BRE
Beaufort West Hospital	BWH
Caledon Hospital	CLD
Clanwilliam Hospital	CLH
Ceres Hospital	CRS
Eerste River Hospital	ERH
False Bay Hospital	FBH
GF Jooste Hospital	GFJ
Hermanus Hospital	HER
Helderberg Hospital	HHH
Karl Bremer Hospital	KBH
Khayelitsha Hospital	KHA
Knysna Hospital	KNY
LAPA Munnik Hospital	LAP
Laingsburg Hospital	LBH
Murraysburg Hospital	MBH
Mossel Bay Hospital	MBY
Montagu Hospital	MON
Mitchells Plain Hospital	MPH
Oudtshoorn Hospital	OUD
Prince Albert Hospital	PRH
Riversdale Hospital	RIV
Radie Kotze Hospital	RKH
Robertson Hospital	ROB
Stellenbosch Hospital	STB
Swartland Hospital	SWA
Swellendam Hospital	SWE
Uniondale Hospital	UDH
Victoria Hospital	VHW
Vredendal Hospital	VRE
Wesfleur Hospital	WFH

E.2 Hospitals in the final dataset

Table E.2: List of district hospitals in the study's final dataset

Hospital Name	Hospital Code
Ladismith (Alan Blyth) Hospital	ABH
Beaufort West Hospital	BWH
Caledon Hospital	CLD
Clanwilliam Hospital	CLH
Ceres Hospital	CRS
Hermanus Hospital	HER
Knysna Hospital	KNY
LAPA Munnik Hospital	LAP
Laingsburg Hospital	LBH
Montagu Hospital	MON
Prince Albert Hospital	PRH
Riversdale Hospital	RIV
Radie Kotze Hospital	RKH
Robertson Hospital	ROB
Stellenbosch Hospital	STB
Swellendam Hospital	SWE
Uniondale Hospital	UDH
Vredendal Hospital	VRE

Appendix F MATLAB script file for information theory analysis

The following MATLAB script was used to calculate the complexity and specialisation measures for each of the hospitals in the analysis. The script is divided into five sections in accordance with the data analysis procedure outlined in Figure 4.10. The script file imports the final dataset into MATLAB from an MS Excel file, applies the information theory approach on the dataset and then outputs complexity and specialisation measures to an MS Excel file.

```
clear all;
close all;
clc;

%% Phase 1: Importing the dataset matrix
D = xlsread('C:\Documents\MATLAB\Thesis\Calc', 'Sheet1', 'B2:JP20');
col1= D(:,1);
N = length(col1);
C_i = sum(D,2);
C_j = sum(D,1);
C = sum(C_i,1);

%% Phase 2: Front matter
Q = D ./ C_j;
test2 = sum(Q,1);
P = D ./ C_i;
test1 = sum(P,2);
Q_j = C_j ./ C;
P_i = C_i ./ C;

%% Phase 3: COMPLEXITY
% Expected information gain
A1 = Q .* log(N*Q);
A1(isnan(A1))=0;
EIG = sum(A1,1);
% Standardisation to average of 1
EIG_a = EIG ./ (sum((EIG .* Q_j), 2));
test3 = sum((EIG_a .* Q_j), 2);
%Complexity index
COMP = sum((EIG_a .* P_i), 2);

%% Phase 4: SPECIALISATION
% Expected information gain
A2 = P .* log(P ./ Q_j);
A2(isnan(A2))=0;
IG = sum(A2,2);
```

APPENDIX F

```
% Standardisation to average of 1
IG_a = IG ./ (sum((IG .* P_i), 1));
test4 = sum((IG_a .* P_i), 1);
%Specialisation index
SPEC = sum((IG_a .* Q_j), 2);

%% Phase 5: Back matter
[~, txt, ~] = xlsread('C:\Documents\MATLAB\Thesis\Calc', Sheet1',
'A2:A19');
col_header= {'Hospital', 'Complexity','Specialisation'};
xlswrite('XS.xlsx', col_header, 'Metric','A1');
xlswrite('XS.xlsx', txt, 'Metric','A2');
xlswrite('XS.xlsx', COMP, 'Metric','B2');
xlswrite('XS.xlsx', SPEC, 'Metric','C2');
xlswrite('XS.xlsx', EIG, 'EIG-CMPX','A1:HU1');
xlswrite('XS.xlsx', EIG_a, 'EIG_a-CMPX','A1:HU1');
xlswrite('XS.xlsx', IG, 'EIG-SPEC','A1:A18');
xlswrite('XS.xlsx', IG_a, 'EIG_a-SPEC','A1:A18');
xlswrite('XS.xlsx', C_i, 'C_i','A1:A18');

xlswrite('XS.xlsx', C_j, 'C_j','A1:HU1');
```

Appendix G Correlation analysis

G.1 Statistics associated with CMPX and SPEC measures

Table G.1 details the statistics associated with the set of complexity and specialisation measures for the 19 hospitals in the analysis. These statistics describe the distribution of the measures associated with the caseloads of the respective hospitals. This distribution is illustrated in the scatter diagram presented in Figure 4.14 and was discussed in Section 4.5.

Table G.1: Statistics describing the distribution of CMPX and SPEC measures

Statistic	CMPX	SPEC
Minimum	8.353	0.5916
25% Percentile	20.49	0.8065
Median	33.84	0.9505
75% Percentile	57.25	1.286
Maximum	114.7	3.725
Mean	39.72	1.143
Std. Deviation	27.57	0.699
Std. Error	6.499	0.1648

G.2 Correlation analysis results for Chapter 4

G.2.1 Standardised EIG vs number of cases treated

A correlation analysis was conducted to evaluate the relationship between: the standardised EIG for complexity associated with each diagnostic case type grouping vs the number of hospitals that treated that diagnostic case type grouping. This analysis was discussed in Section 4.5.

Table G.2: Results of the correlation analysis: standardised EIG vs number of cases treated

Statistic	Value
Pearson r	-0.9226
95% confidence interval	-0.9398 to -0.9006
P value (two-tailed)	< 0.0001
R square	0.8511

G.2.2 Diagnostically different cases treated vs complexity

Table G.3 and Figure G.1 detail the results of a correlation analysis comparing the complexity measure associated with the caseload of each hospital to the number of diagnostically different cases treated at each hospital. This analysis was discussed in Section 4.5.

Table G.3: Results of the correlation analysis: diagnostically different cases treated vs complexity

Statistic	Value
Pearson r	0.8904
95% confidence interval	0.7324 to 0.9574
P value (two-tailed)	< 0.0001
Is the correlation significant? (α=0.05)	Yes
R square	0.7928

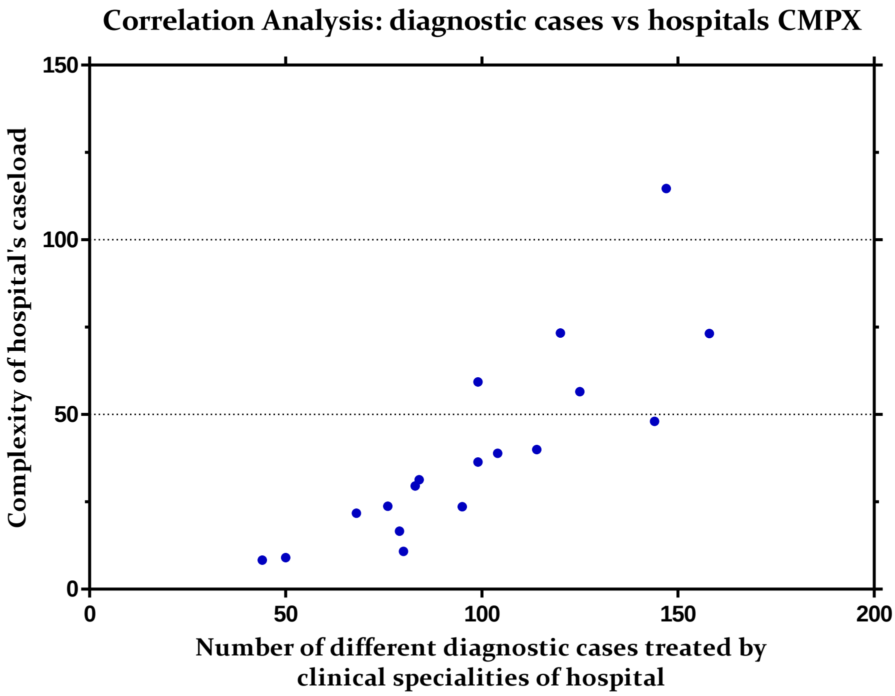


Figure G.1: Scatter diagram: diagnostically different cases treated vs complexity

G.2.3 Diagnostically different cases treated vs specialisation

Table G.4**Error! Reference source not found.** and Figure G.2**Error! Reference source not found.** detail the results of a correlation analysis comparing the specialisation measure associated with the caseload of each hospital to the number of diagnostically different cases treated at each hospital. This analysis was discussed in Section 4.5.2.

Table G.4: Results of the correlation analysis: diagnostically different cases treated vs specialisation

Statistic	**Value**
Pearson r	-0.4402
95% confidence interval	-0.7525 to 0.03371
P value (two-tailed)	0.0675
Is the correlation significant? (α=0.05)	No.
R square	0.1938

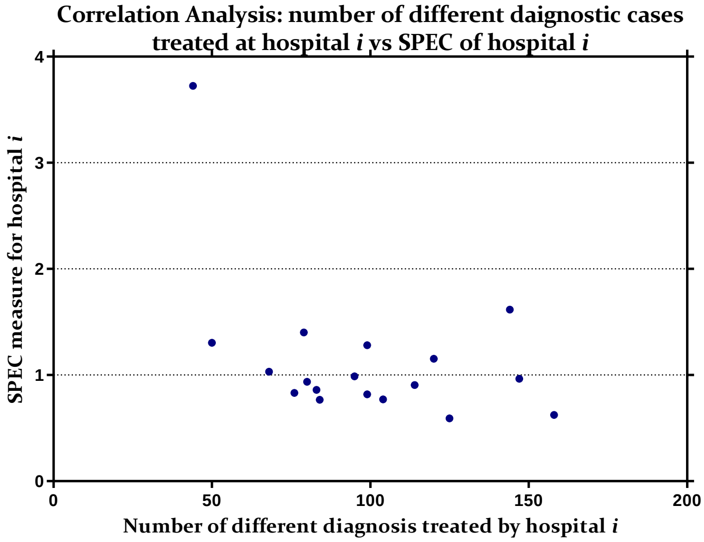


Figure G.2: Scatter diagram: diagnostically different cases treated vs specialisation

G.2.4 Number of cases treated vs specialisation

Table G.5**Error! Reference source not found.** and Figure G.3**Error! Reference source not found.** detail the results of a correlation analysis comparing the specialisation measure associated with the caseload of each hospital to the number of cases treated at each hospital. This analysis was discussed in Section 4.5.2.

Table G.5: Results of the correlation analysis: number of cases treated vs specialisation

Statistic	Value
Pearson r	-0.3115
95% confidence interval	-0.6796 to 0.1820
P value (two-tailed)	0.2084
Is the correlation significant? (alpha=0.05)	No.
R square	0.09701

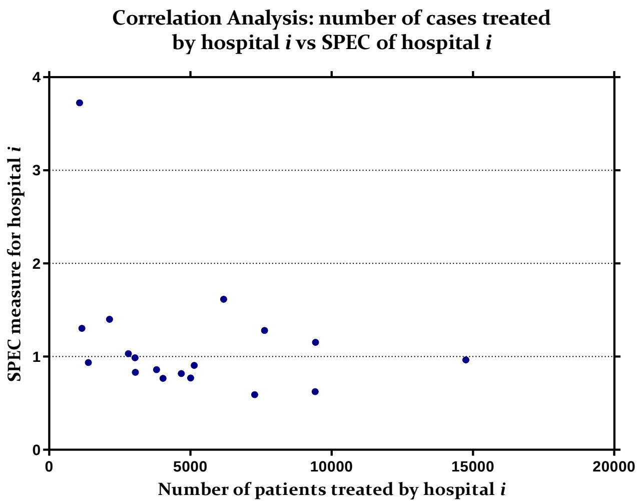


Figure G.3: Scatter diagram: number of cases treated vs specialisation

Appendix H RStudio script

H.1 MLR without principle component analysis

```
# Phase 1: Importing variables and installing libraries
#---------------------------------------------------------------------#
# Importing data for water consumption analysis
library(readxl)
elec <- read_excel("elec4.xlsx")
library(e1071)
library(outliers)
#---------------------------------------------------------------------#
# Phase 2: Standardise the variables: unit normal scaling
#---------------------------------------------------------------------#
#standardise each variable (z-score: subtract mean, divide by sd)
elec2= data.frame(scale(elec))
#---------------------------------------------------------------------#
# Phase 3: Detecting Outliers
#---------------------------------------------------------------------#
# Creates QQ-Plots
qqout = qqnorm(elec2$AEC, main = 'QQ-Plot of AEC', ylab = 'Sample
Quantiles',
               col= "blue3")
qqline(elec2$AEC)

qqout = qqnorm(elec2$TFA, main = 'QQ-Plot of TFA', ylab = 'Sample
Quantiles',
               col= "blue3")
qqline(elec2$TFA)
#--------------------------------------------------------------------#
# Phase 3: Correlation analysis
#--------------------------------------------------------------------#
#calculates the correlation matrix for standardised dataset
mycorr2 =cor(elec2)
mycorr2

#Scatter plot of the respective combinations of variables in the dataset
plot(elec2, main ="Pairwise correlation analysis of stardardised
variables",
     col= "blue3")
#--------------------------------------------------------------------#
# Phase 4: Multiple Regression analysis
#--------------------------------------------------------------------#
#Regression via a linear model
MLR = lm(AEC~.,data=elec2)
summary(MLR)

# AIC - Akaike's Information Criterion using the built-in AIC function
AIC.pcr_a = AIC(MLR, k=2)
AIC.pcr_a
#-------------------------------------------------------------------#
```

H.2 MLR with principle component analysis

```
# Phase 1: Importing variables and installing libraries
#--------------------------------------------------------------------#
# Importing data for water consumption analysis
library(readxl)
library(car)
library(carData)
elec <- read_excel("elec31.xlsx")
#-------------------------------------------------------------------#
# Phase 2: Standardise the variables: unit normal scaling
#-------------------------------------------------------------------#
#standardise each variable (z-score: subtract mean, divide by sd)
elec2= data.frame(scale(elec))
#-------------------------------------------------------------------#
# Phase 3: Correlation analysis
#-------------------------------------------------------------------#
#calculates the correlation matrix for standardised dataset
mycorr2 =cor(elec2)
# Scatter plot of the respective combinations of variables in the dataset
plot(elec2, main ="Pairwise correlation analysis of standardised
variables",
     col= "blue3")
#--------------------------------------------------------------------#
# Phase 4: Multiple Regression analysis
#--------------------------------------------------------------------#
#Regression via a linear model
MLR = lm(AEC~.,data=elec2)
summary(MLR)
#--------------------------------------------------------------------#
# Phase 5: Multicollinearity
#--------------------------------------------------------------------#
#Variance inflation factors
vf<- vif(MLR)
#mean Variance inflation factor
mean.vf<- mean(vif(MLR))
mean.vf
#--------------------------------------------------------------------#
# Phase 6: Principal component analysis
#--------------------------------------------------------------------#
#Remove the first column (elec2) from the data matrix
elec2x= elec2[,-1]

#Peform Principal Component Analysis
elec2x.pca <- prcomp(elec2x, center = TRUE, scale = TRUE)

#Rotation matrix:
print(elec2x.pca$rotation)

#shows all the data that assoicated with each hospital used to generate
PC
elec2x.pca$x[,]

#Variance explianed by each principal component
eigen(cor(elec2x))$values
#--------------------------------------------------------------------#
```

APPENDIX H

```
# Phase 7: Principal component Regression
#-------------------------------------------------------------------#
#Creates a dataframe with AEC in col1 and PC in other col's
elec2.pca=cbind(elec2[,1], data.frame(elec2x.pca$x))
colnames(elec2.pca)[1] <- "AEC"

#Correlation analysis of AEC another principal component
mycorr2.pca= cor(elec2.pca)[,1]

# Computes PCR model
elec2.pcr <- lm(AEC~., data= elec2.pca)
summary(elec2.pcr)
#--------------------------------------------------------------------#
# Phase 9: PCR Model statistics
#--------------------------------------------------------------------#
# Calculating AIC using the built-in AIC function
AIC.pcr = AIC(elec2.pcr, k=2)
AIC.pcr
#--------------------------------------------------------------------#
# Phase 10: PCR model coefficients
#---------------------------------------------------------------------#
# Converts PC into the original variables
model.coef = elec2.pcr$coefficients[-1]
betas2 = elec2x.pca$rotation %*% model.coef
betas2
#---------------------------------------------------------------------#
```

Appendix I Test of individual parameters

I.1 Two-predictor AEC models

I.1.1 CMPX-TFA model

Table I.1: Results of the t-test and confidence intervals for CMPX-TFA regression model

	Estimate	t-stat	p-value	Lower CI (2.5%)	Upper CI (97.5%)
Constant	1.24E-16	0	1	-0.2424	0.2424
CMPX	-0.2288	-1.351	0.197	-0.5899	0.1323
TFA	1.0427	6.155	1.84E-05	0.6817	1.4038

I.1.2 SPEC-TFA model

Table I.2: Results of the t-test and confidence intervals for SPEC-TFA regression model

	Estimate	t-stat	p-value	Lower CI (2.5%)	Upper CI (97.5%)
Constant	1.57E-16	0	1	-0.2405	0.2405
SPEC	0.1817	1.447	0.168	-0.0860	0.4494
TFA	0.9466	7.537	1.78E-06	0.6789	1.2143

I.1.3 TFA-PDE model

Table I.3: Results of the t-test and confidence intervals for TFA-PDE regression model

	Estimate	t-value	p-value	Lower CI (2.5%)	Upper CI (97.5%)
Constant	1.44E-16	0	1	-0.2500	0.2500
TFA	1.0010	5.494	0.0001	0.6126	1.3891
PDE	-0.1650	-0.906	0.379	-0.5532	0.2233

I.2 Three-predictor AEC models

I.2.1 CMPX-SPEC-TFA model

Table I.4: Results of the t-test and confidence intervals for CMPX-SPEC-TFA regression model

	Estimate	t-value	p-value	Lower CI (2.5%)	Upper CI (97.5%)
Constant	1.64E-16	0	1	-0.2363	0.2363
CMPX	-0.2160	-1.314	0.21	-0.5686	0.1366
SPEC	0.1728	1.406	0.181	-0.0907	0.4363
TFA	1.0994	6.505	1.39E-05	0.7369	1.4619

I.2.2 SPEC-TFA-PDE model

Table I.5: Results of the t-test and confidence intervals for SPEC-TFA-PDE regression model

	Estimate	t-value	p-value	Lower CI (2.5%)	Upper CI (97.5%)
Constant	1.79E-16	0	1	-0.2452	0.2452
SPEC	0.1710	1.336	0.203	-0.1036	0.4456
TFA	1.0470	5.787	4.71E-05	0.6589	1.4348
PDE	-0.1393	-0.78	0.449	-0.5224	0.2438

I.2.3 CMPX-BED-TFA model

Table I.6: Results of the t-test and confidence intervals for CMPX-BED-TFA regression model

	Estimate	t-value	p-value	Lower CI (2.5%)	Upper CI (97.5%)
Constant	1.64E-16	0	1	-0.2363	0.2363
CMPX	-0.2160	-1.314	0.21	-0.5686	0.1366
BED	0.1728	1.406	0.181	-0.0907	0.4363
TFA	1.0990	6.505	1.39E-05	0.7369	1.4619

I.2.4 SPEC-BED-TFA model

Table I.7: Results of the t-test and confidence intervals for SPEC-BED-TFA regression model

	Estimate	t-value	p-value	Lower CI (2.5%)	Upper CI (97.5%)
Constant	1.47E-16	0	1	-0.2492	0.2492
SPEC	0.1723	1.3070	0.2124	-0.1105	0.4551
BED	-0.0794	-0.3710	0.7165	-0.5386	0.3799
TFA	1.0080	4.7860	0.0003	0.5564	1.4602