A Probabilistic Estimation of the Capacity of Solar PV SSEGs Installed on a LV Feeder Network

Lewis Sakwa Waswa

Thesis presented in partial fulfilment of the requirements for the degree of Master of Engineering in Electrical Engineering in the Faculty of Engineering at Stellenbosch University

Supervisor: Dr Bernard Bekker

Co-supervisor: Dr Justice Chihota

March 2020

Plagiarism declaration

Plagiarism is the use of ideas, material and other intellectual property of another's work and to present it as my own.

I agree that plagiarism is a punishable offence because it constitutes theft.

I also understand that direct translations are plagiarism.

Accordingly, all quotations and contributions from any source whatsoever (including the internet) have been cited fully. I understand that the reproduction of text without quotation marks (even when the source is cited) is plagiarism.

By submitting this thesis/dissertation 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.

Name: Waswa Lewis Sakwa ....

Date: March 2020

Abstract

Increased solar photovoltaic (PV) installation on to the grid has led to increased technical challenges in electricity network operations. These challenges mainly stem from the design structure of the grid, which only allows unidirectional power flow. This results in several challenges including violation of voltage limits, tripping of network protection systems and distribution line overloads among other issues. These challenges are mainly restricted to the distribution networks, as most solar PV small-scale embedded generators (PV SSEGs) are connected to the distribution networks, whose conditions are, in most cases, not remotely monitored. This results in increased challenges experienced by the networks in terms of network planning, distribution network operations, maintenance, regulation and grid control.

To manage these challenges, the distribution operator needs to estimate the total capacity of solar PV installed on the distribution network, in addition to how much of that capacity is embedded in the network's net demand, which is important in determining the condition of the network at any particular time. Several methods have been used to estimate the capacity of solar PV SSEGs installed in an area. Most studies apply remote sensing and computer vision algorithms to count the number of solar PV panels found in an area. Analysis of these studies indicate that the results obtained cannot be used in determining the condition of the network as they only determine the capacity of solar PV in an area. Secondly, disaggregation studies have largely been used to quantify the installed solar PV capacity embedded in the net demand of a feeder or network. These methods assume a multi-variable approach which requires multiple inputs that are not readily available.

This study introduces a novel probabilistic method that applies Monte Carlo methods to quantify the solar PV SSEGs embedded in the net demand of a low voltage feeder. Historical demand, net demand and the solar PV output is used to determine the solar PV capacity embedded in the net demand of a feeder. The accuracy of the method is tested using simulated net demand and actual measured net demand metered from households connected on carefully selected feeders.

Results demonstrate that the method performs well where the historical demand and the net metered demand are obtained from similar customer classes. Therefore, it is concluded that it is possible to estimate the capacity of solar PV SSEGs embedded in the net demand obtained from a feeder by analysing and comparing the net demand of that feeder and the historical demand of a similar customer class feeder.

Opsomming

Verhoogde sonkrag-fotovoltaïese (FV) installasies op die netwerk het gelei tot verhoogde tegniese uitdagings in elektrisiteitsnetwerkbedrywighede. Hierdie uitdagings spruit hoofsaaklik uit die ontwerpstruktuur van die netwerk, wat slegs eenrigtingskrag moontlik maak. Dit lei tot verskeie uitdagings, insluitend die oortreding van spanningsbeperkings, die uitklop van netwerkbeskermingstelsels en oorlading van verspreidingslyne, onder andere. Hierdie uitdagings is hoofsaaklik beperk tot die verspreidingsnetwerke, aangesien die meeste kleinskaalse ingeboude kragopwekkers op die sonkrag (FV KSIK's) aan die verspreidingsnetwerke gekoppel is, waarvan die toestande in die meeste gevalle nie op afstand gemonitor word nie. Dit lei tot verhoogde uitdagings wat die netwerke ervaar ten opsigte van netwerkbeplanning, verspreidingsnetwerkbedrywighede, instandhouding, regulering en netwerkbeheer.

Om hierdie uitdagings te hanteer, moet die verspreidingsoperateur die totale kapasiteit van sonkrag-FV wat op die verspreidingsnetwerk geïnstalleer is, skat, benewens hoeveel van die kapasiteit ingebed is in die netto aanvraag van die netwerk, wat belangrik is om die toestand van die netwerk te bepaal op enige spesifieke tyd. Verskeie metodes is gebruik om die kapasiteit te bereken vir PV-sonkrag-KSIK's vir sonkrag in 'n gebied. Die meeste studies gebruik algoritmes vir afstandwaarneming en rekenaarvisie om die aantal sonkrag-FV-panele in 'n gebied te tel. Analise van hierdie studies dui daarop dat die resultate wat verkry is nie gebruik kan word om die toestand van die netwerk te bepaal nie, aangesien dit slegs die kapasiteit van sonkrag-FV in 'n gebied bepaal. Tweedens is verdeeldheidstudies grootliks gebruik om die geïnstalleerde sonkrag-FV-kapasiteit wat in die netto vraag van 'n voerder of netwerk ingebed is, te kwantifiseer. Hierdie metodes veronderstel 'n multi-veranderlike benadering wat veelvuldige insette benodig wat nie geredelik beskikbaar is nie.

Hierdie studie stel 'n nuwe waarskynlikheidsmetode bekend wat die Monte Carlo-metodes toepas om die KSIK-sonkrag-PV-sonkrag te bepaal wat ingebed is in die netto aanvraag van 'n laespanning-voerder. Historiese aanvraag, netto aanvraag en die sonkrag-FV-uitset word gebruik om die sonkrag-FV-kapasiteit wat in die netto aanvraag van 'n voerder ingebed is, te bepaal. Die akkuraatheid van die metode word getoets met behulp van gesimuleerde netto aanvraag en werklike gemete netto aanvraag gemeet van huishoudings wat op noukeurig geselekteerde voerkrale gekoppel is. Resultate demonstreer dat die metode goed presteer waar die historiese aanvraag en die netto gemeet aanvraag van soortgelyke kliënteklasse verkry word. Daarom word die gevolgtrekking gemaak dat dit moontlik is om die kapasiteit te bepaal van FV-sonkrag wat ingebed is in die netto aanvraag wat van 'n voerder verkry word, deur die netto aanvraag van die voerder te analiseer en te vergelyk met die historiese aanvraag van 'n soortgelyke klantklas-voerder.

Acknowledgements

All glory be unto the Lord God, Creator of the Universe and the men who live in it, for enabling me to finish this work. Secondly, to the men He created. This work would not have been completed in its form without the support of my supervisor Dr Bernard Bekker. Your guidance and patience are appreciated. To Dr Justice Chihota who came in at the right time, I thank God for your important support and feedback on this work.

I am equally indebted to the larger team at the Centre for Renewable and Sustainable Energy Studies, the Stellenbosch Municipality Electricity Department team as well as the Media Lab for your support and brotherhood.

To the able leaders and my benefactors at the Mandela Rhodes Foundation, your support, guidance, encouragement and character-building moments are forever remembered. This work would not have been without your support. I am deeply honoured to be part of the Foundation.

This work is dedicated to those whom we love and are yet to see, Niel Rapando. And to my mother, Jane. This is also for those we loved, deeply, but who went before us in our absence; Daniel Rapando, my grandfather and brother and to you Phanice Mango, my grandmother. This is for your tender souls, wherever they are. I am, because you lived among us.

And to the new-found family of my African sisters and brothers. May we never stop learning and unlearning.

Table of contents

Table of figures

Page

List of tables

List of abbreviations

AMEU	Association of Municipal Electricity Utilities
ADD	After Diversity Demand
ADMD	After Diversity Maximum Demand
DGs	Distributed Generations
DSM	Demand-side Management
ED	Economic Dispatch
EE	Energy Efficiency
FITs	Feed in Tariffs
GHI	Global Horizontal Irradiation
GIS	Geographical Information Systems
IEA	International Energy Agency
LSM	Living Standard Measure
PDF	Probability Distribution Function
PV	Photovoltaics
SSEG	Small Scale Embedded Generators
STL	Seasonal Trend decomposition by Loess
SWH	Solar Water Heaters
VREs	Variable Renewable Energy sources

1 Introduction

This chapter introduces the research problem, which is how to estimate the capacity of solar photovoltaic (PV) small-scale embedded generators (SSEGs) embedded in the low voltage network feeder demand. In this chapter, a brief review on the global status of solar PV, its growth patterns, and the impacts arising from increased solar PV penetration is provided. Subsequently the project motivation, description, formulation of hypothesis and objectives are established.

1.1 Background

Studies show that there is an increasing adoption of solar PV globally and this is likely to continue as the cost of solar PV reduces (Muaafa *et al.*, 2017; Sandiford *et al.*, 2015; Zhou, 2018)

The rise in solar PV installations can be attributed to several factors. These are discussed in both local and global context in literature. Globally, these factors have been categorized into economic factors, environmental factors as well as regulatory factors. Lopes *et al.* (2007) attributes the increasing uptake of solar PV to the increasing global awareness of greenhouse gas emissions. The authors argue that climate change awareness has significantly contributed to the changing policy framework leading to many economies adopting green sources of energy. Additionally, the authors point out that the depletion of fossil fuels plays a key role in the increased adoption of variable renewable energy systems. According to Sionshansi (2016), global policy frameworks are being changed in the wake of the natural catastrophes as in the case of the Japan Fukushima incident and the prevalent hurricanes in the USA towards ensuring greater system resilience.

A study by the International Energy Agency (IEA) (2017), shows that the adoption of Distributed Generators (DGs) has the potential to reduce investment in distribution networks by 30%. This report further points to the declining costs of PV as one of the factors leading to increased solar PV adoption.

A study by Sewchurran *et al*. (2016) links factors such as load shedding, rising electricity tariffs and delays in construction of new power stations as to the increased uptake of solar PV SSEGs in South Africa.

Increased solar PV adoption has led to an increased penetration of solar PV systems on to the grid, which has in turn increased the technical challenges affecting the network. To manage the impact of increased penetration of solar PV on the grid, utility companies and distribution network operators must regulate the installation of solar PV on their networks. Vrettos *et al.* (2019) points out that increased penetration of solar PV SSEGs reduces the monitoring efficiency of the utilities as the power production of these generators are neither controlled nor monitored by the utilities.

1.2 Challenges of increased solar PV penetration

Experience has shown that increasing penetration of solar PV results in many technical challenges that impact the performance of existing networks. These challenges are attributed to the fact that the power system was traditionally designed for unidirectional flow of power from centrally located generation sources to the loads.

The integration of solar PV on the network exposes the network to reverse power flow. This may lead to voltage variations along the line, line overloads and may require power system overhaul to accommodate the arising challenges. (Sewchurran *et al.*, 2016), (Jenkins *et al.,* 2015). Additionally, an increase in both grid-tied and off-grid solar PV installations may lead to reduced energy demand from the utility, resulting in revenue losses (Korsten, 2016).

Solar energy resource is characterized by variability and intermittency. Variability and intermittency cause technical challenges such as rapid voltage changes. This is captured in many literature texts, among them Jenkins *et al*. (2015) and Liu *et al*. (2016). From these literature sources, network planning is identified as a key challenge associated with system variability and intermittency. Solar irradiance intensity exhibits spatial-temporal variation, rising steadily from morning and peaking at midday before reducing to zero. In high solar PV SSEG penetration scenarios, temporal variation in the intensity of the solar irradiance impacts the daily net demand profile (California Independent System Operator, 2012) (Roselund, 2018). Temporal variation in the irradiance may result in higher net demand in the morning and evenings and reduced net demand at midday. This may result in a net demand profile which is duck-shaped as seen in Figure 1. This phenomenon has been reviewed in several literature texts (California Independent System Operator, 2012)(Lew & Miller, 2017).

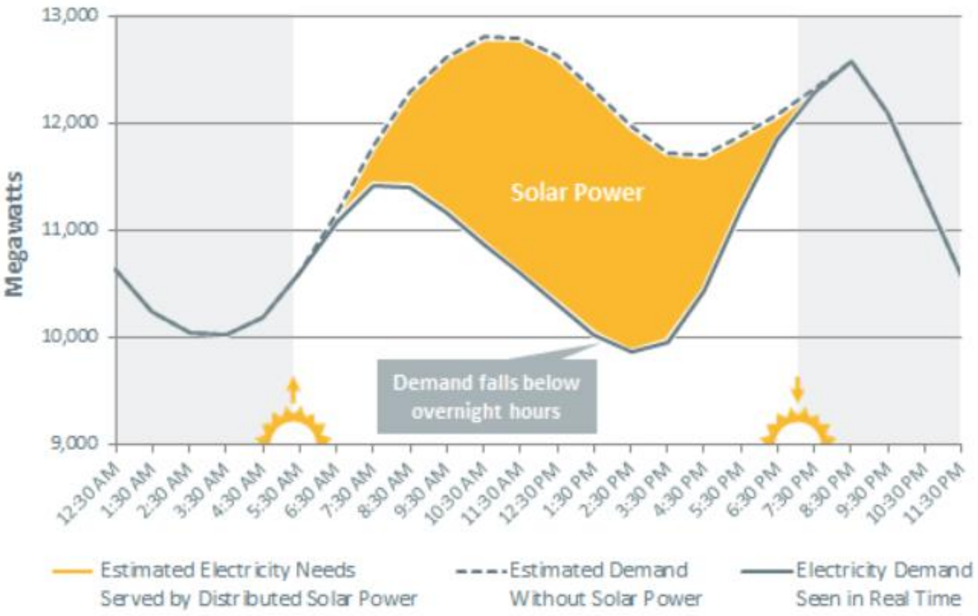


Figure 1: The 'duck-curve' representing the net demand profile in New England, USA. Adapted from Roselund (2018)

A study by Howlader *et al.* (2018) attempts to address the challenge posed by the 'duck–curve' through thermal unit commitment. A similar approach is explored by Or *et al.* (2017). Lastly, an MIT Energy Initiative report (2016) identifies the challenges that are attributed to increasing solar PV and suggests solutions to these.

Apart from the temporal variation, solar PV output can be affected by changes in weather in terms of cloud movement. According to Mills *et al*. (2009) a rapid change in cloud cover can significantly alter the generation of solar PV by about 60% of its peak. Solar output is thus intermittent in nature and depended on other factors that are not easily predictable. The effects of intermittency range from power quality, voltage regulation, load following and unit commitment. These challenges are identified in Sayef *et al*., (2012), and Rowe *et al*. (2016). Other system-wide issues attributed to intermittency are challenges in frequency control (Lew & Miller, 2017).

Rowe *et al.* (2016) captures the impact of intermittency on different time scales ranging from seconds to days. Table 1 shows the potential impacts of intermittency on a grid. Sionshansi (2016) provides a detailed context into intermittency and its effects on the system.

Table 1: Potential impacts of intermittency on a grid (Rowe *et al*., 2016)

Timescale of intermittency	Potential power system impact
Seconds	Power quality (e.g., voltage flicker)
Minutes	Regulation Reserves
Minutes to hours	Load Following
Hours to days	Unit Commitment

Effective mitigation strategies have been used in meeting the challenges resulting from the high penetration of solar PV. For large scale PV systems connected to the high and medium voltage networks, strategies used include: generation curtailment and economic dispatch of the solar PV (Lew & Miller, 2017), optimal thermal unit commitment (Howlader *et al.*, 2018), and improving the solar PV forecasting (Kaur, Nonnenmacher & Coimbra, 2016). Additionally, solar PV connected on both high voltage (HV) and medium voltage (MV) networks are constantly monitored by the utility operators. This improves the fault detection and response time in resolving the detected challenges for the MV and HV installations. This level of monitoring is typically not the case for distribution networks (Vrettos *et al.*, 2019)

1.3 Global growth of solar PV SSEGs

Shaker *et al.,* (2015) estimated that by 2013, the global generation of solar PV SSEG stood at 23 GW. In 2017, a report by the International Energy Agency (IEA) indicated that global solar PV capacity had reached 398 GW. 40% of this capacity was from solar PV distributed generations (International Energy Agency, 2019). Another report from the International Renewable Energy Agency, indicated that as of 2019, the total solar PV installed capacity was at 480 GW (Bellini, 2019).

According to the Australian solar PV Institute (2018), there was about 1.84 million new solar PV installations in Australia in 2018. According to another study, the average national Australian PV penetration stood at 18.7% (Jaeger-Valdau, 2017). However, some states registered about 30% penetration individually. Zhou (2018) noted that about 1.3 GWp of rooftop PV was installed in Australia in the small scale sector in 2018. According to the Solar Naturally report (2018), 3.5 million rooftop solar PVs were installed in Australian homes with 20% of these selling back to the grid.

A report by the UK Department of Business indicated that as of 2018, about 19.9% of the total installed PV in the UK came from installations that were less than 4kWp (Clark, 2018).

As of 2018, 44 GWp of the total installed generation capacity in Germany was solar (Fraunhofer, 2018). A report by the German Development Agency, (2016) estimates that 74% of the solar installations in Germany were roof mounted with 70% of these being less than 10kWp. This translates to about 50% of the installations being less than 10kWp.

1.3.1 Solar PV SSEGs in South Africa

In alignment with the global trends explored in the previous section, South Africa has also seen an increase in recent PV installations. In 2016 a report by PQRS (PQRS, 2016) found that 417 MWp of solar PV SSEGs was installed in South Africa

A more recent report obtained from the Association of Renewable Energy Practitioners (2019), suggests that the installed solar PV SSEGs capacity at the end of 2019 will be more than 1 GWp. This is more than 100% growth in less than 3 years. Analysis of the 2016 data reveals that 89% of the solar PV SSEGs in South Africa lie in the commercial sector and are predominantly roof-mounted. Further, most of these installations are connected on low voltage networks that are operated by municipalities.

A study conducted in 2018 by the Western Cape Department of Agriculture (DoA) (Department of Agriculture Western Cape, 2018), suggests that there were 4248 rooftop PV panels installed in the Stellenbosch municipality at that time. From this it was estimated that the municipality had an installed solar PV capacity of about 752kWp, of which 545kWp is from residential solar PV. Comparing the statistics in this study to the PV systems formally registered with the Stellenbosch Municipality for the year 2019 (Dyusha, 2019) reveals significant discrepancies in that a large number of installations could not be accounted for. Consequently, it can be concluded that some of the systems that were captured during the DoA study were illegal installations or unregistered expansions of registered existing sites. Figure 2 shows the satellite imagery of the Stellenbosch area that was mapped during the DoA study.

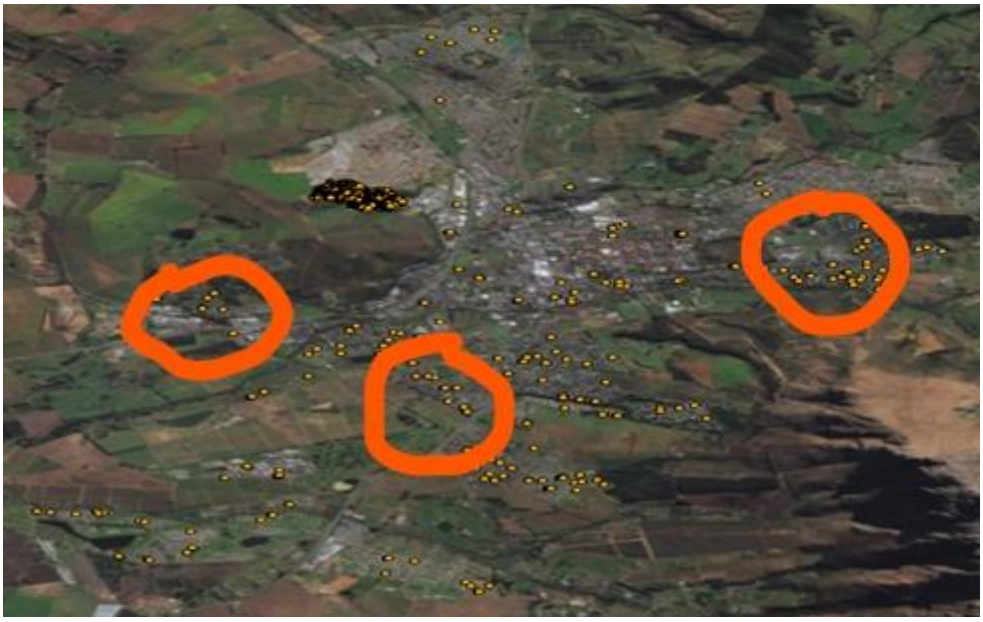


Figure 2: Satellite imagery of the drone flyover area around Stellenbosch showing PV Clusters (Department of Agriculture Western Cape, 2018).

1.3.2 Impact of high penetration of solar PV SSEGs on distribution networks

The impact of solar PV SSEGs on the distribution networks is likely to be more severe as compared to the impact experienced by high voltage and medium voltage connected networks, on which large scale solar PV installations are connected.

The challenges resulting from increased penetration of solar PV affect network planning and performance in terms of voltage variations (voltage-drop, voltage-rise and flicker), frequency control, and phase unbalance among others (Kara *et al*., 2018) (Shaker *et al*., 2016). The impact of high solar PV penetration on the distribution networks is likely to present more issues to the operator due to several factors: firstly, most standards do not enforce the requirement for remote monitoring and control of inverters. Secondly, most SSEGs are distributed in nature and as such, the conditions of each of the solar PV connection affects the performance of the feeder differently, depending on where it is connected. Finally, the location and capacity of future SSEGs are unknown and cannot be easily determined as they depend on the customer's attitude, investment preference, and affordability.

Furthermore, due to the distributed nature of solar PV SSEGs, there is an increased likelihood of illegal installations being connected to the grid. This negatively impacts the performance of the grid. Unauthorized expansion of the existing installations increases the technical challenges associated with increasing penetration of solar PV. Additionally, it is difficult to ascertain whether or not the approved standard set of inverters and the connection requirements have been adhered to in these illegal installations. This in turn increases safety concerns.

Increasing the penetration of solar PV in general and solar PV SSEGs on the distribution networks in particular, may result in technical challenges. Therefore, it is important for the distribution operators to regulate the installation of solar PV SSEGs on their networks. This is because the performance of the networks depends on the capacity and installation configurations of the solar PV SSEGs. It is clearly important for the operator to know the capacity of installed solar PV SSEGs. However, this is difficult as some clients do not follow the utilities' regulations for registrations while some, once registered, expand the installed capacity without notice. At the same time, the operator may notice changes in the electricity revenue or the demand profile, but cannot immediately determine the capacity of SSEGs embedded in the network. In such cases, techniques for estimating the installed SSEG capacity become very important for planners.

1.4 Project motivation

By reviewing the possible impacts of solar PV SSEGs on a network, it is evident that there are challenges associated with high penetration scenarios of solar PV SSEGs on the distribution networks. Therefore, it is important to know the capacity of the installed solar PV on a network as well as the embedded capacity in the net demand of a network or feeder. The following points summarize why the knowledge of the total installed capacity is important to the distribution operators.

i) Increase in the penetration of solar PV SSEGs on the network leads to the increased likelihood of technical challenges which affect the performance of the network.

ii) Network planning requires knowledge of current capacity of distributed generations installed. This is significant for network reinforcement and upgrades with the aim of maintaining the performance of networks within regulated limits of voltage and unbalance among other issues.

iii) Since there is no remote monitoring of the status of distribution networks, solar PV SSEGs estimates can be used as indicators of the conditions of a particular feeder, based on regulated penetration limits. With such information, early mitigation of faults, possible feeder overloads as well as voltage regulation can be implemented.

Given these reasons, it is important to estimate the capacity of solar PV SSEGs embedded in the net demand of a network. Two main approaches have been proposed to estimate the capacity of solar PV installed in an area. These are broadly categorized in Geographical Information Systems (GIS) based studies (Arjun *et al*., 2017; Bradbury *et al*., 2016; Bradbury *et al*., 2015; Devarajan *et al.*, 2018) and disaggregation studies.

GIS based studies use remote sensing to capture imagery through computer vision algorithms. While these approaches have shown a high performance, they cannot solely provide information relating to how much of detected capacity is embedded in the net demand.

Unlike the GIS methods, the reviewed disaggregation methods are capable of determining the capacity of installed solar PV embedded in the system demand. To this extend they have been applied to testing if they can disaggregate accurately solar PV contribution in the demand. However, these studies as captured in Vrettos *et al. (*2019) and Kara *et al*. (2018), do not consider the uncertainty that is inherent in both the PV output as well as the load. Additionally, these methods are complex as the approaches use multiple variables and data which is not readily available in some instances. Other disaggregation approaches identified in Shaker *et al*. (2016)(2015), use data reduction techniques which may lead to the loss of demand characterization. Lastly, the approaches applied in disaggregation studies are modelled with the aim of separating the solar generation from the total generation in scenarios where the capacity installed is known.

From literature review, the two main methods of estimating the present-day PV SSEG capacity of an area do not address two key gaps. These gaps include:

i) The need for a method that uses commonly available network data and encompasses simplicity in its implementation.

ii) The need to consider the probabilistic nature of both the demand and solar PV output in the models.

This research is focused on evaluating the residential solar PV SSEGs penetration and estimating their contribution to the distribution network load. This is primarily for the following reasons.

i) Residential PV systems are mainly connected to the distribution network.

ii) Residential PV systems growth and installation are not controlled by the distribution operators and increasing penetration levels of these solar PV systems on the distribution network aggravates the system technical challenges.

iii) For most LV networks, there is no requirement for remote monitoring and as such the conditions of the network are not visible. This poses challenges in the distribution network operation.

1.5 Project description and hypothesis

This research proposes a method of estimating the capacity of solar PV SSEGs embedded on a low voltage feeder network based on net demand. The underlying concept is that, by studying the differences between historical demand data prior to the installation of solar PV SSEGs and presently measured net demand data, while allowing for uncertainties in both the demand and solar irradiance, accurate estimates of the installed solar PV SSEGs capacities on the residential feeder can be obtained.

Accordingly, measured net demand data can be compared with net demand profiles simulated from historical load data offset with different PV capacity assumptions and optimised for accuracy. Therefore, the following sets of data and parameters are required as inputs:

i) Historic Domestic Demand data

ii) Measured Net Demand data from selected aggregators like kiosks or mini-substations

iii) Solar irradiance data

In South Africa, historical load data for residential customers exists in two forms; raw collected data for selected communities and generalized statistical classifications based on Living Standard Measure (LSM) class. The historic demand data proposed for use in this thesis is in the form of the Domestic Load Research data obtained from the DLR report (University of Cape Town, 2018).

In this study, the net demand data is in two forms; by collection of measurements from carefully selected residential feeders in Stellenbosch and through simulation using historic demand data and solar irradiance data (obtained from SAURAN and Helio-Clim databases).

The accuracy and performance of the proposed method is validated using four case studies. The objective of these case studies is to evaluate the accuracy and performance of the proposed method under different conditions and using different forms of the historic demand.

In summary, the research hypothesis which this thesis aims to test is the following:

“The total capacity of solar PV SSEGs embedded in the aggregated net demand of a residential LV feeder can be estimated by comparing the net demand profile with historical demand from a similar customer class using a stochastic technique.”

1.6 Research objectives

The following are the objectives of this study.

i) Create a statistical method that can be used to estimate the capacity of solar PV SSEGs embedded in the net demand of a distribution feeder.

ii) Test whether the use of general statistical models for clustered customers, such as the LSM classes, in the proposed method is justifiable.

iii) To determine the accuracy of estimates derived using the proposed method.

iv) To establish the limitations of the study and recommend possible improvements.

1.7 Research questions

This research answers the following questions.

i) What existing methods have been used in estimating the capacity of solar PV SSEGs?

ii) What are the gaps identified in these estimation methods?

iii) How can net demand measurements and historical load data be used to compute the total embedded capacity?

iv) Can LSM classifications and the related models be used in accurately estimating the capacity of solar PV SSEGs currently installed or does the approach require feeder-specific data?

v) How can the effects of other factors such as energy efficiency (EE), solar water heaters (SWHs) and tariff changes be accounted for?

vi) What are the limitations of the proposed method?

1.8 Scope and limitations of the research

The proposed method is centred on estimating the capacity of residential solar PV that is embedded in the network demand and as such the focus of the study is on residential solar PV SSEGs as opposed to commercial PV SSEGs or any other category. Further, this research does not account for the effects of battery energy storage systems (BESS) on the load profile for the considered households.

1.9 Document overview

The rest of the document is divided into different chapters as follows:

- Chapter 2 looks at the estimation methods reviewed from literature, where and how they have been used and their shortfalls as far as the estimation of the installed solar PV SSEGs on the network are concerned.
- Chapter 3 provides a high-level review of the proposed method, defines the terms used in this study, defines the processes and the essence of those processes and the statistical tests that are carried out in this study.
- Chapter 4 is focussed on the development of the estimation model. It defines the specific data used, the resolution of the data used, the pre-processing of the data and the statistical testing that is carried out in validating the proposed model.
- Chapter 5 provides the results and findings of the validation procedure.
- Chapter 6 concludes the dissertation with a summary of the findings, conclusions from the findings, summary of contributions and recommendations for future research.

The appendix section contains a brief introduction to the probability concepts used in this study, a background review of the data collection process and characterization of the areas that are used in this study.

2 Review of current estimation methods

This section reviews the available literature on the methods used in estimating the capacity of solar PV in an area. It is subdivided into three sections which first broadly examines the methodologies that have been used in assessing PV uptake in an area. Several methods are reviewed under this sub-section. The second sub-section of this chapter reviews the methodologies used in forecasting net demand as a way of assessing the impact of solar PV on the system demand. Finally, the last subsection reviews the methods that have addressed the objective of this study. It discusses the two main approaches that have been used in estimating the capacity of installed solar PV in an area or on a network. The identified methods in this sub-section include Geographical Information Systems methods as well as disaggregation methods.

2.1 Review of methods applied in determining the PV SSEGs uptake

Various methods have been used in literature in estimating the solar PV uptake. These methods range from the use of GIS models to diffusion models. They are essential in assessing the potential PV holding capacity of an area. Methods that have been used here include socio-demographic reviews. These methods have exploited demographic indicators for sites in determining the likelihood of solar PV uptake.

This is captured in Vivian *et al*. (2016). Socio-demographic methods also use GIS software such as ArcMap to locate and place different sites in different classes. In the review of other uptake estimation models, roof space assessment has been broadly used.

Reinecke *et al.* (2013) investigates the adoption of rooftop solar PV in Riversdale municipality using GIS.

Bergamasco *et al*.(2011), uses solar radiation data, availability of rooftop space while employing a rooftop coefficient among other factors to assess the potential for several provinces in Italy. The rooftops are classified in residential, commercial and even industrial. GIS tools are extensively used in the mapping of irradiance maps and the areas of study. A similar approach is adopted by Ali *et al.* (2018). Data driven models are also used in the determination of potential uptake of solar PV. This is captured in the studies by Zhao *et al.* (2017).

Diffusion models have been used in several studies (Graziano & Gillingham, 2015; Margelou, 2015; Snape, 2016; Wang *et al.*, 2017; Zhao *et al*., 2017) to assess the uptake of solar PV in areas. The end goal of most of these studies is to predict how many installations an area would hold at a particular time, for future planning purposes. Baas' generalized model for diffusion is employed by Wang *et al*. (2017) to estimate the uptake of solar PV for commercial purposes in California. A similar approach that uses learning curves is used by Masini & Frankl (2003) to forecast the diffusion of solar PV in Southern Europe.

In a study assessing the long term solar PV diffusion in Switzerland, (Margelou, 2015), the author investigates the reasons for adoption of solar PV by residential areas through classification of several demographic and socio-economic categories. Other methods used in determining the uptake of solar PV in an area include spatial pattern diffusion models as used by Zhao *et al*. (2017) in the analysis of the diffusion of PV in Pudong region of Shanghai, China. Graziano & Gillingham (2015), investigates the adoption of residential solar PV using a similar approach. Another study by Snape (2016), investigates the spatial and temporal characteristics of PV adoption in the UK. This study points out that the use of S-curves and other prediction models are not sufficient to accurately show the adoption pattern for PV while considering policy analysis.

Zhang *et al.* (2014) investigates the solar PV uptake using agent based models to predict the adoption agent-based model. Several factors are included in Zhang *et al*. (2014), which include occupancy of the buildings, the type of building, and the roof space of the building and the value of the property among many other issues.

2.2 Review of studies focussed on PV output and net demand forecasting

Net demand refers to the total system load, less the amount of load catered for by the intermittent variable renewable sources. In a case where the intermittent energy source considered is solar PV, the net demand refers to the total system load, less the demand met by solar PV.

Net demand estimation studies have also been conducted in various studies these include but are not limited to Wang *et al*.,(2018), Kaur *et al*.,(2016) and Zhang *et al*., (2018). Zhang, *et al*. (2018) reviews the concept of net demand focusing on the need to decompose the net demand into various subsections to include the actual load, the PV output and the residual load by exploring demand data.

Further, Wang *et al*. (2018), addresses the question of net demand forecasting for high penetration scenarios of distributed solar PV systems. The measured load data is broken down into its constituents which include solar PV output, the residual and the actual demand.

A study by Kaur *et al*.(2016), focuses on net demand forecasting for a high penetration scenario. This study gives prominence to the trading balance between the utilities and the micro-grids. Kaur *et al*. (2016) aims to provide a model that can improve the forecasts of the solar PV output, with the intention of improving the net demand forecasts for electricity markets.

Lastly, a study by Kuppannagari *et al*., (2017), investigates how to carry out optimal net demand balancing. Through this study, optimal net demand estimation methods are assessed with the aim of mitigating the disparity between supply and demand in integrated grids.

2.3 Review of the methods used in estimating the solar PV SSEGs capacity in an area or on networks

2.3.1 Geographical information systems (GIS) oriented methods

Several studies have used GIS to estimate the amount of solar PV that is installed in an area (Arjun *et al.*, 2017; Bradbury *et al.*, 2016, 2015; Kara *et al.*, 2018; Malof *et al*., 2015). A large proportion of these studies use remote sensing technologies to enumerate the amount solar PV in an area. Malof *et al.* (2015) points out how satellite imagery can be used to automatically detect and enumerate how many solar panels exist in an area, using computer vision algorithms to detect solar PV SSEGs installed on the rooftops.

A similar approach using remote sensing is applied by Bradbury *et al.,* (2016)(2015) to assess the quantity of solar PV installed in an area. The results obtained showed that the technique by Bradbury *et al*.,(2015) performs well in imagery detection. These methods use a PV identification algorithm to extract solar PV in a region using a technique called maximally stable extremal regions (MSER). A similar method is implemented in the study carried out by the DoA, Western Cape, in the mapping of solar PV resources. This was done using drones to determine how much solar SSEGs are installed in Stellenbosch(Department of Agriculture Western Cape, 2018).

According to Pierro *et al.* (2017), it is possible to estimate the solar PV generation of an area through spatial clustering of the solar PV found in a region by using remote sensing techniques to assess the capacity of solar PV output in an area. This is done by factoring in weather prediction data. This method has been used and extended to forecast the PV output generation.

2.3.2 Shortfalls of the GIS oriented methods

Firstly, the use of satellite imagery has qualitative limitations as far as the estimation of the amount of solar PV installed on a feeder is concerned. This is because while it is possible to estimate the quantities of solar PV in an area, it may not be possible to estimate how much of the solar PV SSEGs feeds back into the grid. As such, these methods cannot provide the qualitative feedback required in assessing the conditions of the network.

Secondly, through physical inspection of some of the sites identified in the DoA study (2018), it was realized that some of the objects identified and categorized as solar PV panels were not. This is likely to be a source of estimation error thereby exaggerating the capacity of solar PV in an area. To enhance accuracy, there is need to improve or change the computer vision algorithm that was applied in that study.

Lastly, it is not possible to assess the impact of the solar PV capacity determined from the imagery obtained, without the knowledge of their connection status with respect to the grid. Review of the demand data collected during the study shows that some of the solar PV connected on the rooftops are actually off-grid and this cannot be captured through remote sensing.

2.3.3 Energy disaggregation methods

According to Ebrahim & Mohammed (2018), energy disaggregation is defined as a computational approach through which an individual energy output or contribution from a source can be retrieved from a measured set of aggregated energy outputs, energy usage or aggregated energy measurements. Disaggregation studies strive to decompose aggregated measurements into the individual source measurements.

Energy disaggregation is an important concept as far as the unmasking of the ‘unseen’ demand is concerned. Several studies have used this method to detect the presence of hidden PV installations. Shaker *et al*. (2015), Vrettos *et al*. (2019) and Kara *et al*. (2018) (2016) have used this method to estimate the invisible solar PV in the system demand.

Solar PV SSEGs, like most behind the meter generations, are connected to LV networks and may not be directly monitored and controlled by the utility operator in real-time. In essence, these connections are not visible to the utility operator. This makes it difficult to monitor and mitigate the challenges that may arise during their operation. Disaggregation models can be used to monitor the solar PV plant power output in real time and thus provide useful information for the efficient operation and response to technical challenges on the network (Vrettos *et al*., 2019). Several methodologies have been adopted in the use of energy disaggregation. Below is a review of approaches that have been encountered in literature.

2.3.3.1 Wavelet based variability model

The wavelet based variability model is used by Zhu *et al*.(2016) to predict the solar PV output of PV power plants. It combines artificial neural networks and wavelet decomposition methods to predict the PV generation. The wavelet decomposition method is used to separate the PV output from noise.

Lave *et al*.,(2013) used the wavelet based variability model to improve the solar PV output forecasts by reducing the effect of variabilities. In doing this, this method erases the inherent variabilities that characterize the solar irradiance and provide accurate forecasts of the solar output. This method does not provide a pathway of estimating the solar PV capacity embedded in the demand. However, it provides a basis through which accurate measurements of solar PV output can be made.

2.3.3.2 Machine learning data driven disaggregation models

A study by Shaker *et al.* (2015) proposes a method that estimates the contribution of different solar sites given their location, their quantity, the day factor among other issues. This is a learning model that classifies geographical regions based on their active power generation through the use of either hybrid-k class or principal component analysis. This is then used to create a generation aggregate for a whole region. The findings note that their algorithm’s performance decreased with increasing spatial coverage due to the nature of Global Horizontal Irradiation (GHI) used.

Sossan *et al.* (2018) proposes a model using disaggregation while employing machine learning methods to solve the problem. In his model, irradiance is used to model power output and as such investigates the patterns of solar PV output in relation to the aggregated power flow measurements. This method cannot sufficiently estimate the amount of solar PV embedded in a system as it can only work in data aggregates that have a high solar penetration. Secondly, it may be difficult to capture the randomness that is inherent in the demand.

2.3.3.3 Time series models

Kloibhofer *et al*. (2017) uses Seasonal Trend decomposition by Loess (STL) to disaggregate demand. This study makes use of anonymous smart meter measurements of about 40 houses which have PV systems installed on them and monitors the aggregated individual household loads. This method is adopted for forecasting purposes.

A similar method investigates how to decompose power demand into uncertain features and through that, the ability to separate an aggregate signal into the source signals. This is captured and used in a study by Imanishi *et al* .(2017). The study uses seasonal decomposition where possible transient changes in a power time series can be identified and separated.

2.3.3.4 Regression and contextually supervised source separation model

Kara *et al*. (2018) proposes a method that can be used to disaggregate solar generation from the feeder level measurements. Kara *et al*. (2018a) makes use of the relationship between the load active power and consumption and reactive power to give a real time disaggregation. The model minimizes prediction error of linear regression models and the results indicate good performance of the model for both distributed and centralized PV systems. A similar method is adopted by Vrettos *et al*., (2019).

2.3.3.5 Support vector machines

Mohan *et al*. (2014) introduce the concept of solar disaggregation in order to separate the power production of solar panels, from the home’s net consumption. Mohan *et al*. (2014) used support vector machines to predict the solar output. Additionally, Mohan *et al*. (2014) uses a comparison metric between the solar output and the net demand to smooth the output from the possible current inrushes which would misrepresent the solar PV. Essentially, the method is anchored on accurate estimation of the solar PV.

2.3.4 Shortfalls of the present disaggregation methods

Firstly, the application of the disaggregation methods studied in this section reveal that their application is limited by their complexity. Most of the methods use multi-objective approaches which interweaves several methods in order to estimate the embedded capacity.

Secondly, the discussed disaggregation methods do not address the probabilistic nature of both demand and solar PV output. In this regard, they do not provide a means to addressing the uncertainties from these two sets of data.

Lastly, to use these methods, more individual solar PV output and demand data is required from smart meters. Such data is not readily available in our case.

2.4 Conclusion

From the above review of methods in literature, it can be concluded that disaggregation provides the possible pathway to separate aggregated demand according to their sources or use. This is shown in the texts that have been reviewed.

It has been demonstrated from literature that disaggregation methods take form through the use of many other models. Among the models adopted and reviewed are machine learning, time series decomposition and multilinear regression. However, it has also been demonstrated that these methods perform effectively when there is a lot of data obtained from sites with PV as well as individual demand data.

2.5 Existing research gap

From this literature review, it can be seen that there are several methods which have been previously used to estimate the capacity of solar PV installations. However, several gaps exist in these methods. The existing gaps which the method proposed in this thesis address include:

i) Most reviewed methods are highly complex and require various input parameters to estimate the solar PV capacity embedded in net demand.

ii) Most of the reviewed methods do not address how to deal with the problem of stochasticity in the demand as well as the solar PV output. There is need to develop a model that captures both the stochasticity in the solar PV output as well as the demand.

Table 2 below summarizes some of the reviewed estimation methods, their complexity as well as the existing shortfalls.

Table 2: A summary of the reviewed Solar PV estimation methods

Title	Year of Publication	Approach adopted	Specific method used	Functionality	Shortfalls
Estimating PV Power from Aggregated power Measurements (Kara *et al.*)	2019	Contextually Supervised Source Separation	Regression model	Determines the amount of solar PV embedded in the demand	Multi-variant inputs required
Estimating Behind the meter solar Generation with existing measurement infrastructure (Kara *et al.*)	2016	Contextually Supervised Source Separation	Similar to the above but with lesser regression factors	Determines the amount of solar PV embedded in the demand	Multi-variant inputs required
Disaggregating solar generation from feeder level measurements (Kara *et al.*)	2018	Multiple linear regression analysis Compares the reactive power and the active power though regression analysis. Uses Aggregated demand data	MLR and Contextually Supervised Source Separation	Determines the embedded capacity of solar PV in the demand Does not use historical demand data	Multi-variant inputs required No accounting for the stochasticity of both PV and demand
Unsupervised Disaggregation of Photovoltaic Production from Aggregated Power Flow Measurements of Heterogeneous Prosumers (Sossan *et al.*)	2018	Uses several algorithms to disaggregate PV generation from demand.	Regression and spectral density analysis	Determines the embedded capacity of solar PV in the demand Does not use historical demand data	No accounting for the stochastic nature of the demand in the aggregates. Relatively complex method
A novel method for decomposing electricity feeder load into elementary profiles from customer information (Gerossier *et al.)*	2017	Statistical Blind Source Model & Augmented Lagrangian method	Statistical Blind Source Model & Augmented Lagrangian method	Carries out signal decomposition. It is used for source signal separation	Not tested for solar PV
Automated Rooftop Solar PV Detection and Power Estimation through Remote Sensing (Devarajan *et al.*)	2016	Geographical Information systems	Remote sensing and satellite imagery	Determines the quantity of solar PV panels installed in an area	Does not estimate the amount of solar PV output in the demand. Can be used for further investigation but would not solve the problem
Solar Power Estimation Through Remote Sensing (Devarajan *et al.*)	2018	Geographical Information systems	Remote sensing and satellite imagery	Determines the quantity of solar PV panels installed in an area	Does not estimate the amount of solar PV output in the demand. Can be used for further investigation but would not solve the problem
sUncover: Estimating the Hidden Behind-the-meter Solar Rooftop and Battery Capacities in Grids (Padullaparthi *et al.)*	2019	Exploits the patterns in the Battery Energy Storage Systems(BESS) and solar PV output	Energy Balance Equation	Determines the quantity of solar PV capacity in an area/with a client	Relies on the presence of the BESS system. Cannot estimate without this parameter
Automatic Solar Photovoltaic Panel Detection in Satellite Imagery (Malof *et al.)*	2015	Geographical Information systems	Remote sensing and satellite imagery	Determines the quantity of solar PV panels installed in an area	Does not estimate the amount of solar PV output in the demand. Can be used for further investigation but would not solve the problem
Data-driven upscaling methods for regional photovoltaic power estimation and forecast using satellite	2017	Geographical Information systems	Remote sensing and satellite imagery; spatial clustering and neural networks	Determines the solar PV generation for a region	Provides estimates based on imagery and hence cannot be used to determine the capacity embedded

and numerical weather prediction data (Pierro *et al.)*					
A Data-Driven Approach for Estimating the Power Generation of Invisible Solar Sites (Shaker *et al*)	2015	Maps the PV output of an area to the longitude and latitude using clustering methods and reduction techniques.	Kalman Filters, Linear Regression	Determines the embedded solar PV in the demand of a large area	
Estimating Power Generation of Invisible Solar Sites Using Publicly Available Data (Shaker *et al*.)	2016	Uses clustering modes and Principal Component analysis in the selection of areas and uses fuzzy model to associate PV output with an area.	Fuzzy model	Determines the embedded solar PV in the demand of a large area.	Accounts for the spatial uncertainty in the PV output for different sites. There is no representation of stochasticity in the demand
Solar Energy Disaggregation using Whole-House Consumption Signals (Mohan *et al)*	2014	Smoothing functions applied on the net demand and focusses mainly on the solar PV prediction. SVM used in the solar PV Prediction.	Regression model used. Support Vector Machines used in solar prediction	Determines the embedded solar PV	No representation of the probabilistic nature of solar PV or in demand.
Comparing and improving residential demand forecast by disaggregation of load and PV generation (Kloibhofer *et al.)*	2017	Time series analysis of the individual household profiles. Explores the property of stationarity in the load profile.	Time series analysis	Determines the value of the embedded solar PV	By assuming stationarity in the demand profile, the method eliminates the need for changing aggregated demand; hence it does not address the question of stochasticity. Assumes that the error between the total demand and the net demand is only solar PV; Does not provide a method that accounts for other demand influencing factors.

3 Overview of the proposed methodology

This chapter outlines the proposed methodology used in this study. It defines the terms used and explains the purpose of the statistical tests carried out in this research.

3.1 Brief introduction of the proposed method

In this study, we propose a method that compares and analyses historical demand and net measured demand of a feeder with the aim of estimating the capacity of solar PV embedded in the net demand. This is done by factoring in the impact of solar irradiance. The capacity of embedded solar PV is established by determining the point of the minimum difference between the aggregated historical and aggregated net demand. The capacity of solar PV embedded in the historical demand is taken to be zero or of known value. The capacity of solar PV embedded in the measured net demand is unknown.

The proposed method is based on the Monte-Carlo approach. This is premised on the fact that Monte-Carlo methods are anchored in probability and they are efficient in the representation of random processes (Glasserman, 2003). Additionally, Monte Carlo methods are capable of handling large computational problems in less time. The Monte Carlo approach was selected due to the inherent stochasticity and the large size of data that is used in the analysis.

To determine the capacity of solar PV embedded in the feeder demand, we apply a direct search optimization technique. Generally, direct search is a classical analysis technique applied on unconstrained optimization problems which do not require the use of derivatives (Kolda *et al*., 2003; Lewis *et al.*, 2000). Such is the case with the proposed approach. Analysis using these methods has shown that they are reliable.

We use least error direct search approach in this study to develop our objective function. To determine the capacity x of PV embedded in the measured demand, we search for a minimum point in the difference between the historical demand and the present measured demand. Below is the objective function adopted for this study.

$$\min\left[\, L_{unknown}^{Net\,Meas.} - \left(L_{known.}^{hist} - x.PV\right)\right]$$

Where

$L_{unknown}^{Net\,Meas.} = Current\ Measured\ Net\ demand\ with\ unknown\ PV\ penetration\ (Amps)$

$L_{known.}^{hist} = Historical\ demand\ with\ known\ PV\ (Amps)$

$$x = variable\ representing\ installed\ PV\ SSEG\ on\ feeder\ (kWp)$$

$$PV = PV\ capacity\ factor\ for\ measured\ location\ (\frac{Amps}{kWp})$$

The objective function is applied under the assumption that for a customer classified under the same LSM class, the historical demand is greater than the current measured demand. In addition, by applying this objective function, further investigation is conducted to determine whether historical demand of one feeder can be used together with net demand from a different feeder, to estimate the solar PV embedded in the feeder demand considering the LSM class.

3.2 Processes and terms used in modelling

3.2.1 Stochastic expansion

This is a process from which a larger pool of individual customer demand profiles are generated from a smaller pool using a random 'sampling with replacement' Monte-Carlo based approach. The larger pool of demand profiles provides a source of similar profiles from which aggregated demand profiles for that community can be generated. The optimal sample size for the larger pool of profiles is determined by testing for convergence and the preservation of different selected statistical parameters.

3.2.2 Stochastic aggregation

This is a process through which a specific number of individual customer demand profiles are randomly sampled and summed to create an aggregated demand for that particular feeder or community. This process is carried out by the application of Monte-Carlo process on the expanded demand profiles.

For a given community with a specific number of customers connected on a feeder, we use stochastic aggregation to determine different aggregated demand profile scenarios for that community. Aggregated feeder demand varies from time to time. We therefore determine the optimal representative number of samples for which the statistical properties of the aggregated demand are preserved for a selected community. We then use the representative samples in the testing and validation of the method.

3.3 The essence of stochastic expansion

The stochastic expansion process produces a larger pool of demand profiles used in the process of stochastic aggregation. By randomly sampling and adding a number of individual load profiles, aggregated demand for a group of customers can be determined.

3.4 The essence of stochastic aggregation

This process is utilised to generate possible aggregated demand profile scenarios for a particular feeder or community. It is implemented using the Monte-Carlo approach. In this study, aggregated demand data represents the total demand for a clustered community connected on a selected feeder.

3.5 Statistical tests carried out

3.5.1 Intra-hour statistical coherence testing

It is a process of evaluating the probability distributions of the original historical demand and the stochastically *expanded* historical demand. This process is carried out by comparing the probability distribution of the original historical demand, to that of the *expanded* historical demand. The purpose of this test is to check for the optimal sample size of *expanded* historical demand at which intra-hour statistical coherence is maintained for two sets of demand data.

3.5.2 Temporal coherence testing

This is a process of evaluating the optimal sample size for which the demand profiles converge. Demand profiles for different tested sample sizes are plotted. These are load profiles for the mean, demand profile for the 5th percentile and the demand profile for the 95th percentile. The different statistical parameters are tested in order to evaluate the robustness of the proposed method. Monte Carlo methods are used in this process. The purpose of this test is to determine the optimal sample size for which the temporal statistical characteristics of the demand data are preserved.

3.5.3 Coefficient of determination (R-square)

This is a test of evaluating the goodness of fit for the estimated net demand compared to the measured net demand, towards identifying the minimum error as shown in Figure 3. The purpose of the test is to measure the R-square value between the estimated net demand and the measured / simulated net demand at the particular estimated capacity.

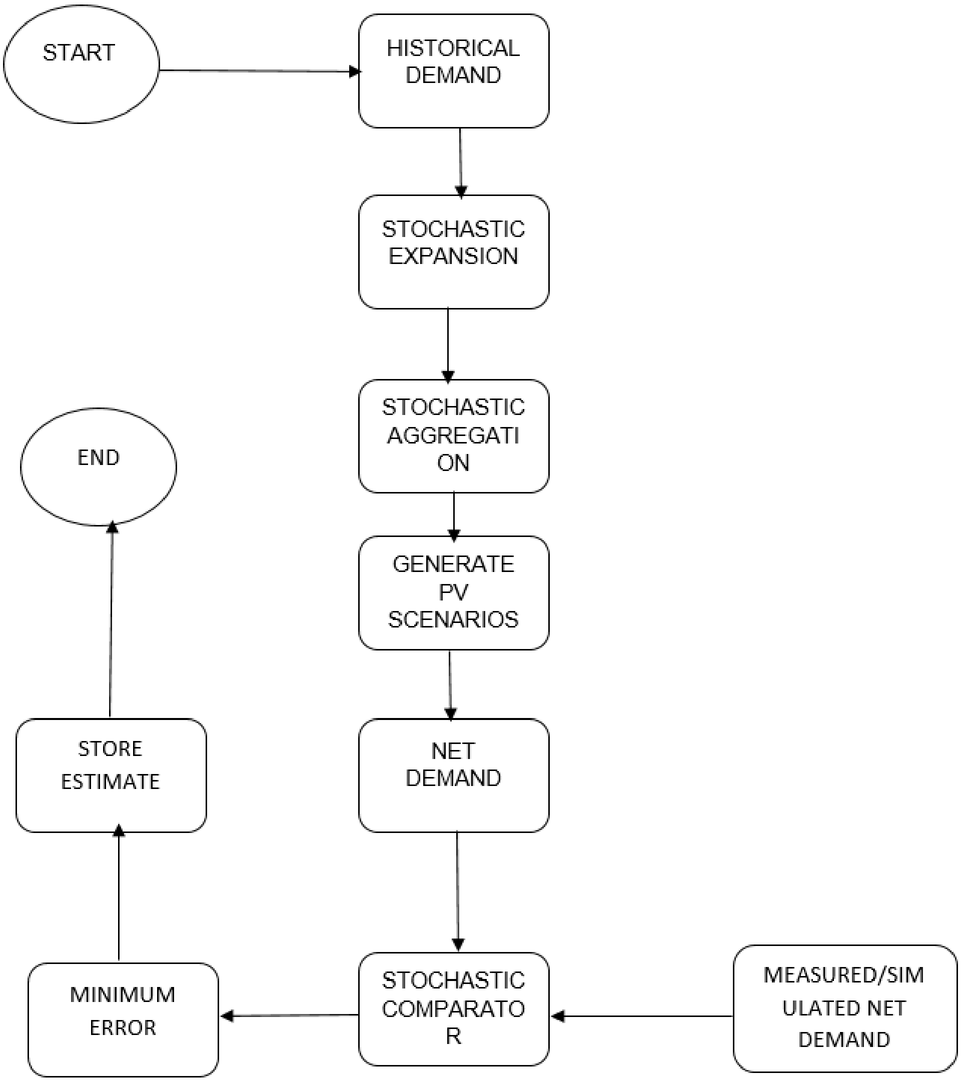


Figure 3: High level illustration of the proposed methodology

4 Development of probabilistic estimation framework

This chapter describes the development of the stochastic estimation method. It describes the sets of data used in the research, the modelling procedure and the statistical tests carried out. Furthermore, the chapter presents the testing and validation of four case studies adopted in the research. The purpose of this section is to describe in detail the procedure that was used in the modelling process. It covers the pre-processing of the historic and net demand data as well as the pre-processing of the solar irradiance data to yield the PV capacity factor. The section on data collection from the kiosks and its cleaning and preparation is included in Appendix A.

4.1 Data used in modelling

The primary data required for the development of a probabilistic estimation model is historical demand data prior to the installation of PV and current measured net demand of a feeder. Additionally, irradiance data obtained for the geographical area where the feeder is located, is also required in order to model the anticipated solar PV output. For uniformity, the temporal resolution of the three sets of data: the historical demand; the measured net demand and the solar PV output must the same. Data used in this modelling process has an hourly resolution. Additionally, hourly resolution is used for faster data processing.

To validate the hypothesis, net demand is simulated from historical demand by factoring in the solar PV output. Holding other variables constant, net demand is obtained by differencing the solar PV output, calculated for a given kWp capacity, from the total aggregated historical demand. The proposed model is then used to test how much solar PV is embedded in the net demand. Additionally, this method is tested using measured net demand obtained through the measurement of demand from selected feeders in Stellenbosch. The results obtained are compared to the estimates from Stellenbosch municipality.

4.2 Data collection process

4.2.1 Historic demand

Historical demand data used in this study was obtained from the DLR report (University of Cape Town, 2018). The amount of the solar PV embedded in the historical demand is assumed to be zero considering the period within which the data used was collected and from the estimations obtained from PQRS South Africa's residential PV data (Ballack, 2017). For purposes of extending this model and using other sets of demand data, historical demand may be used to refer to demand collected from individual customers with a known total solar PV installation.

4.2.1.1 Selection of the historic demand

Data used in the validation of the proposed method was carefully selected from several sources. First, Moreletta Park and Welgemoed demand data was selected as representative samples for upper income quantile of customers. We used demand data from West Ridge and Gasese to represent the possible demand profiles of middle income quantiles and low income quantiles respectively. Details of these selected areas is as below:

i. Site A: Moreletta Park 2002

The DLR report (University of Cape Town, 2018) characterizes Moreletta Park as an upper customer class area with most customers having a 60A supply. This area has large houses of an average size of 380m^2. The average income of households in this area was R29000 per month as of 2002. It had full penetration of household appliances. The figure below shows a sample residence in Moreletta.

Figure 4: A residential area in Moreletta park (University of Cape Town, 2018)

The following table 3 shows the demand characteristics of data obtained from Moreletta Park.

Table 3: Demand parameters for Moreletta park (University of Cape Town, 2018)

Peak #	Date	Total I (A)	Number of customers	Average I (A)	Std dev (A)
1.	24/01/2002 06:30	969.76	62	15.64	11.95
2.	06/02/2002 06:30	937.59	65	14.42	9.98
3.	30/01/2002 06:45	933.49	57	16.38	11.99
4.	13/08/2002 19:15	930.99	53	17.57	9.87
5.	10/04/2002 06:30	929.16	64	14.52	10.44

ii. Site B: Welgemoed, Cape town

According to the DLR report, (University of Cape Town, 2018), the average income for Welgemoed was approximately R29000 per household in 2002. The area is also classified in the upper income quantile with a significant home ownership of about 95%. The average size of the houses is 330 square meters with an 80A supply. The following table shows selected demand parameters for the year 2002.

Table 4: Demand parameters for Welgemoed (University of Cape Town, 2018)

Peak #	Date	Total I (A)	Number of customers	Average I (A)	Std dev (A)
1.	09/09/2002 20:50	973.73	63	15.46	11.62
2.	10/09/2002	955.95	62	15.42	12.74
3.	26/07/2002 20:45	925.84	59	15.69	11.78
4.	24/07/2002 19:10	909.38	58	15.68	11.52
5.	13/08/2002 20:50	904.73	54	16.75	13.38

The following figure shows a sample residential area in Welgemoed.

Figure 5: A residential area in Welgemoed (University of Cape Town, 2018)

iii. West Ridge, Cape Town

Estimations from the DLR report indicate an average household income of R5200 per month. The area is well serviced with roads and has access to water. There is a full penetration of major electrical appliances. Table 5 shows the demand statistics obtained.

Table 5: Demand parameters for West Ridge (University of Cape Town, 2018)

Peak #	Date	Total I (A)	Number of customers	Average I (A)	Std dev (A)
1.	04/08/2002 11:55	641.53	64	10.2	7.31
2.	28/05/2002 18:45	635.91	67	9.49	7.80
3.	02/08/2002 18:40	620.69	67	9.26	7.72
4.	12/08/2002 19:40	609.09	67	9.09	6.69
5.	08/07/2002 19:05	607.61	66	9.21	7.32

The figure below shows a sample of the houses found in West Ridge, as indicated in the DLR report.

Figure 6: A residential area in West Ridge (University of Cape Town, 2018)

iv. Site 3 – Gasese, Northern Cape

The average household income of this selected area was estimated to be about R660 per month. Most of the dwellings are fitted with a 60A breaker. The houses have an average built space of 50 square meters. There is approximately 47% penetration of major electrical appliances. The table below shows the statistical demand parameters obtained from Gasese.

Table 6: Demand parameters for Gasese (University of Cape Town, 2018)

Peak #	Date	Total I (A)	Number of customers	Average I (A)	Std dev (A)
1	03/09/2002 06:30	184.32	74	2.49	3.35
2	05/08/2002 19:05	178.69	66	2.71	3.89
3	22/08/2002 06:45	173.96	76	2.29	3.52
4	03/08/2002 06:45	170.97	74	2.31	3.22
5.	21/08/2002 06:45	169.68	76	2.23	3.71

The figure below shows a sample residential building in Gasese, as stated in the DLR report (University of Cape Town, 2018).

Figure 7: A residential area in Gasese (University of Cape Town, 2018)

The selection of these four areas provide data to test whether the proposed method is subject to LSM classifications.

4.2.2 Measurement of the 2019 demand data

Background information on the survey conducted by the Department of Agriculture, which informed the selection of the areas that were metered, is provided in appendix A. However, the selection was conducted by picking areas categorized as wealthy through physical inspection and using energy consumption data.

A study conducted by Korsten *et al.* (2017) clustered different areas in Stellenbosch according to their average consumption. The consumption data combined by physical inspection of the area to be metered is indicative of the status of the different areas in Stellenbosch. Additionally, selection of the wealthier areas of Stellenbosch was premised on the Stellenbosch Municipality data (Dyusha, 2019), indicating that most of the solar installations were located in these areas. The following areas were selected for metering.

i. Dalsig

ii. Brandwacht

iii. Karindal

iv. Paradyskloof.

Kiosks to be metered were selected based on the mapping information provided in the DoA study (Department of Agriculture Western Cape, 2018). The table below indicates the average consumption details as reported by Korsten *et al*. (2017).

Table 7: Average and maximum consumption for different areas of Stellenbosch (Korsten *et al.*, 2017)

Suburb	Average kWh per household per month	Maximum Consumption per household per month
Uniepark and Karindal	2178	18830
Dalsig and Brandwacht	2011	18620
Die Boord and Paradyskloof	1735	26698
Onder Papegaiiberg	1323	10677
Welgevonden	518	3117
Idas Valley	1136	9308
Cloetesville	813	9326
Kayamandi	578	3877

4.2.3 Irradiance data

Irradiance data is first used to model the solar PV output factor, which is the output in amps for every 1 kWp installed per GHI value (Amps/kWp/kW/m^2). These are hourly values for the whole year that represent the output of 1 kWp of installed PV for each corresponding value of the GHI. The solar PV output factor is used to calculate the solar PV capacity factor (Amps/kWp).

However, one-year data cannot sufficiently represent the uncertainties in the solar irradiation for an areas over a long period of time. To do this, 14 years of GHI data was additionally compiled from Helio-Clim database (SodaPro, 2019). A sampling procedure was applied to this data to create a larger pool of irradiance values. A statistically acceptable sample of over 30 irradiance values per interval were created. A total of 42 entries per hour were made by systematically creating two additional irradiance values, sampled 24 hours before and after. This process was carried out in order to express different possible irradiance values that may occur in an area. This process was carried out to capture the uncertainty in the intra-interval irradiance values, which, in this case represents irradiance variability. A total of 42 samples per interval were generated through this exercise and used to calculate the capacity factor.

4.2.4 Living standard measure classes

LSM is a metric that is used to classify different people according to the socio-demographic indicators which represent the quality of life that they lead (Kironji, 2010). In the DLR project, it was demonstrated that different areas could be clustered into different LSM classes based on the electricity consumption patterns. Areas in the same LSM class exhibited similar probability distributions for loads. LSM classes are used in this investigation of the proposed methodology in cases where there are no historical demand data from the feeder. This study uses data drawn from a possible LSM class 10-high, LSM class 5- high and LSM class 1.

Accordingly, the model investigates whether it is possible to estimate the capacity installed in area A, using the historical demand from a different area B with the same LSM classification; where areas A and B are communities classified under the same LSM class, in which A has the net demand but lacks historical demand whereas both datasets are available for B. This investigation is carried out with the LSM class as a varying factor. This is done using both the simulated demand as well as the measured net demand. These sites are described in the Appendix B.

4.3 Specific data used

For this research, the following data inputs are used in modelling.

i. October to November 2002 Historic Demand for Moreletta Park in Pretoria, Gauteng and Welgemoed in Cape Town, Western Cape

ii. May 2002 Historic Demand for Moreletta Park, Welgemoed, Gasese in the Northern Cape and West Ridge in the Western Cape.

iii. May 2019 net demand for selected areas in Stellenbosch

The respective historical data was selected by checking for the months with least or no missing values in order to reduce the errors in estimation.

4.4 Data pre-processing

The historical demand data used in this study is provided in the form of DLR load data values obtained from the DLR study report. These are individual customer load current measurements available at a 5-minute temporal resolution. The demand data is available in two files, each representing half a year. These were combined to form one dataset, prior to pre-processing. For faster data processing and conformity, the data was averaged to hourly demand values and cleaned before undergoing the statistical expansion and aggregation processes, which produce the final aggregated demand data. Figure 8 shows the flow chart of the processes which the historical demand data undergo.

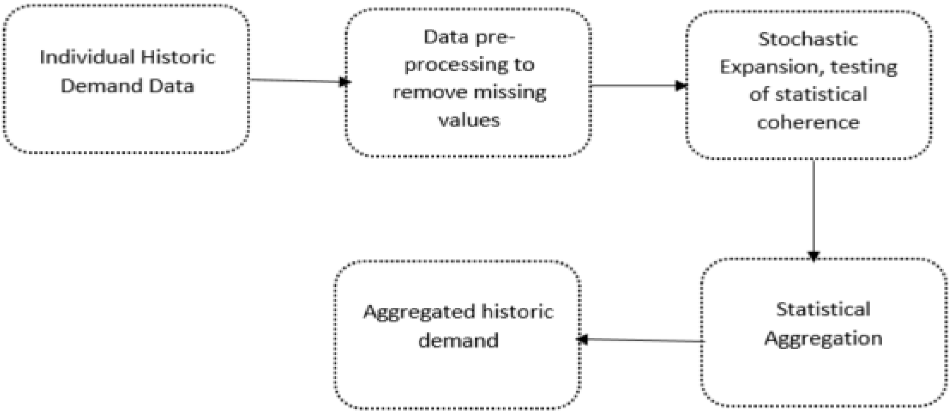

Individual Historic Demand Data
Data pre-processing to remove missing values
Stochastic Expansion, testing of statistical coherence
Statistical Aggregation
Aggregated historic demand

Figure 8: Procedure of preparation and cleaning of historic DLR data

4.4.1 Pre-processing of historical demand data

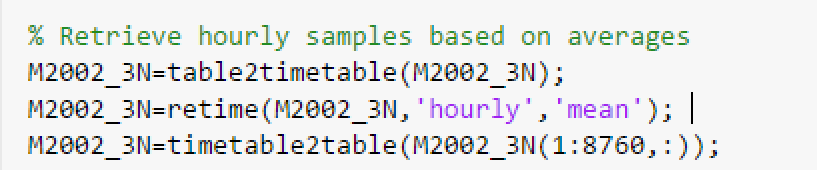
The following procedure is undertaken in the pre-processing exercise:

i) The value -999 was used to represent missing values in the historic demand data. These values were removed and then replaced with NaN values. Thereafter, the missing values were replaced by the closest values from the next interval. It is assumed that there is a close correlation between demand within one hour and the next.

ii) The data is selected such that there is minimal missing values in the selected profiles to minimize the chances of error in the process. This was done by carefully selecting the months with little or no missing values for use in the validation.

iii) The 5 minutes demand data was hourly resolved for faster data processing and for conformity with the solar irradiance data which had an 1-hour resolution.

Figure 9 shows a snippet of the code as used in the hourly sampling from the 5 minutes data.



```
% Retrieve hourly samples based on averages
M2002_3N=table2timetable(M2002_3N);
M2002_3N=retime(M2002_3N,'hourly','mean');
M2002_3N=timetable2table(M2002_3N(1:8760,:));
```

Figure 9: Code section used to average the 5 minutes to 1 hour

4.5 Stochastic expansion of individual historic demand data

'Stochastic expansion' is a process that creates a larger and optimal sample of individual demand data profile from a smaller sample. The purpose of stochastic expansion is to create a larger sample of individual customers' profiles which are then aggregated to form possible aggregated profiles. The DLR data used had 42 customers demand profiles, resolved at an hourly interval for conformity purposes. The stochastic aggregation process was implemented using Monte Carlo simulation. The following procedure outlines the process involved.

i. Initialize a random process

ii. From n customer demand profiles, randomly select a profile. Repeat this process k times, recording and selecting each selected profile for every kth time.

iii. Determine the appropriate number of runs by inspecting for the convergence of selected central tendencies. At this point of convergence, k=N. K represents the number of Monte Carlo runs at which convergence is achieved. N is the optimal sample size of individual customer profiles for which convergence is achieved.

iv. Store the number N and use for aggregation.

To test for the preservation of statistical independence of the original demand data and its statistical representative sample, a convergence test was carried out.

4.5.1 Convergence testing

To determine the optimal sample size from the stochastic expansion process, a convergence test was carried out. To do this, we examined by inspection, the demand profiles for three central tendencies. The mean, the 95th percentile and the 5th percentile historical demand profiles are examined for each installed PV SSEG capacity under consideration. The mean time series is reviewed since it closely follows the After Diversity Maximum Demand (ADMD) in a given interval which is useful in the classification of LSM clusters. Additionally, the time series plot of the mean often shows the general variation of the load over a period.

However, the mean on its own cannot fully capture the diversity of the load in a given interval as well as the temporal correlation between different load segments. For this reason, the 95th and the 5th percentile profiles are used to represent possible maximum and minimum time series plots that captures most demand values.

In general, uncertainty analysis in Monte-Carlo method requires an adequate number of trials. To do this, we require a large number of samples to represent the stochasticity of the inputs. The purpose of evaluating the convergence test is to determine the optimal sample size at which statistical parameters of the data are preserved and do not exhibit significant variations.

For this study, samples of 100, 300, 1000, 3000, 5000 and 10000 individual customers are tested. From these generated samples, the mean, the 5th percentile values and the 95th percentile values were plotted. The optimal sample size of the pool is validated by evaluating the number of samples for which there is convergence of the mean demand, convergence of the 5th percentile demand profiles as well as the convergence of the 95th percentile demand profiles. The evaluation process was carried out by inspection of the profiles obtained.

It was found that 5000 samples provide a best convergence for these statistical parameters. Figures 10 and 11 below are a reasonable representative of the whole series obtained and as such the order of convergence exhibited in the figures persists throughout the series. The illustrated profile used four days to achieve better resolution.

It can be seen that the profiles merge into each other as the samples start to increase from 5000. Figure 10 shows the convergence of the mean and the 95th percentile load profiles at 5000 samples. The 5000 data samples thus represent a stochastic sample of the load. It can be seen that the 42 customers can be represented stochastically by a larger sample of 5000 data points in each interval.

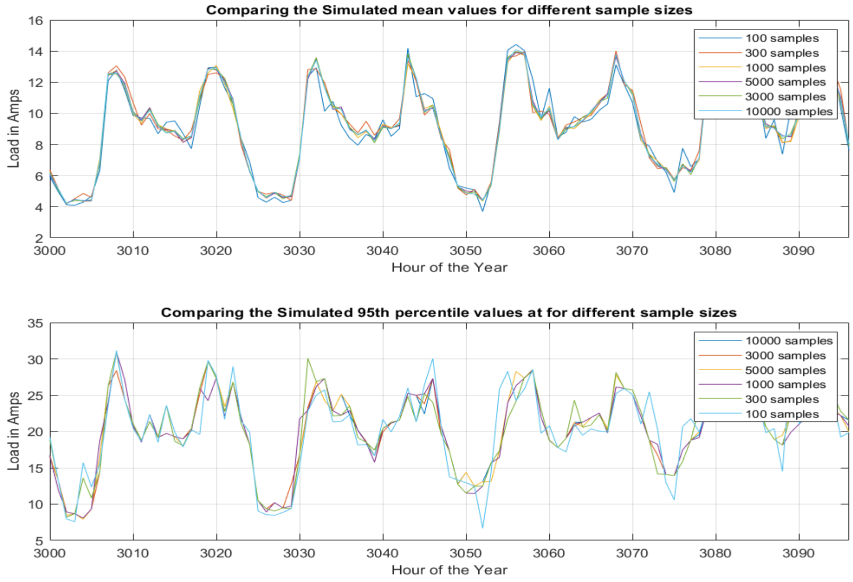


Figure 10: Optimal sample size testing

The following figure shows the means and the 95th percentile values for sample sizes 5000 and 10,000.

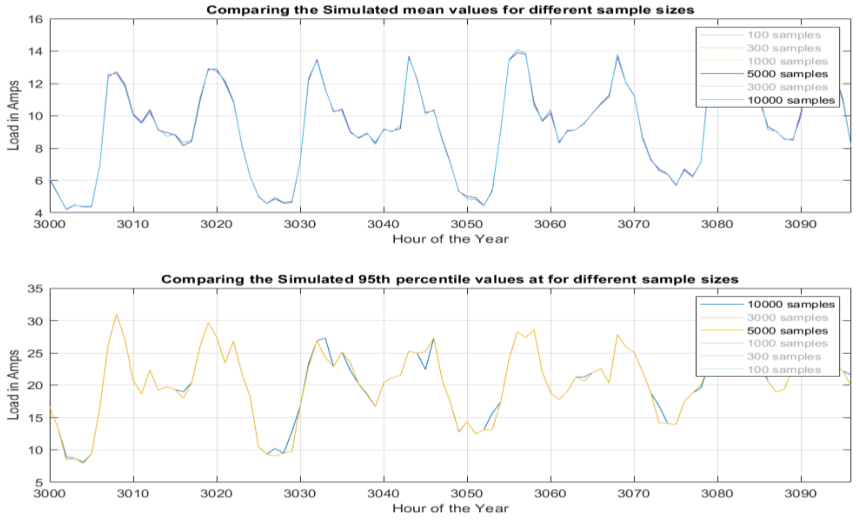


Figure 11: Optimal sample size testing for the 10000 and the 5000 samples

The validity of the optimal sample size is further illustrated by assessing the mean error distributions. These errors are obtained by differencing the mean of the *expanded* demand data from that of actual demand data and plotting it to determine the spread of errors from the mean. It is established that by using at least 5000 samples, a tolerance of 0.2 amperes can be achieved. It was determined that using sample sizes with lesser samples than 5000 resulted in as large deviations as 1.0 amperes. For different sample sizes tested, this is illustrated in the figures below.

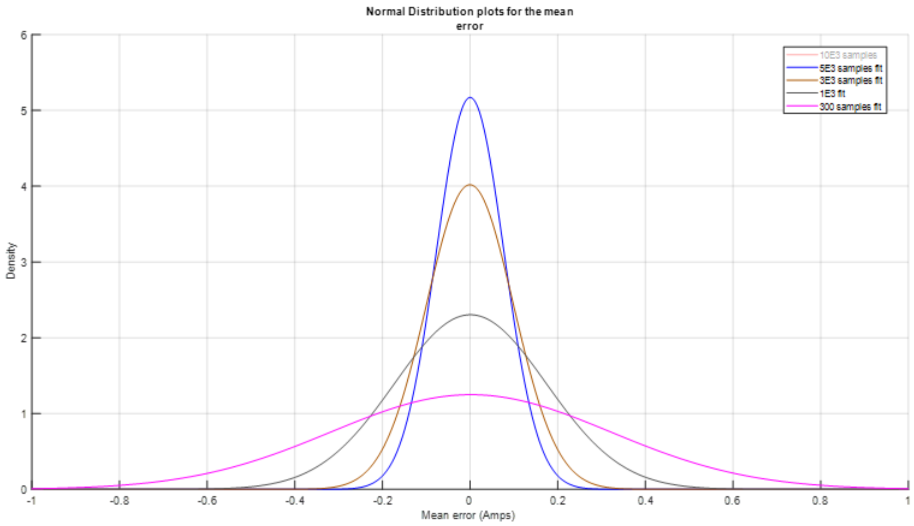


Figure 12: Mean error distribution plots

To further determine the distribution of these errors for different sample sizes, we plotted a cumulative distribution function for the mean errors. The figure below illustrated the results obtained.

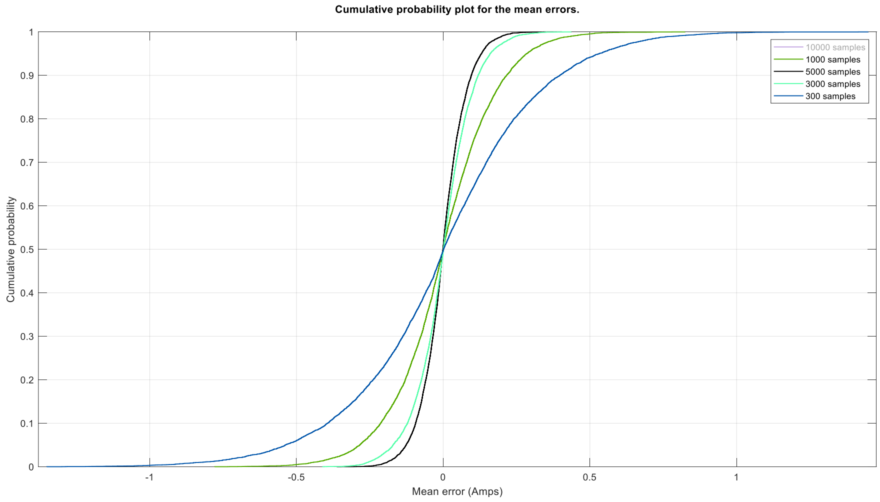


Figure 13: Cumulative distribution plot for the mean errors

From the figure 13 below, it can be seen that the 5000 sample distribution plot exhibits a smaller deviation as compared to the rest of the plots. Further, it can be deduced that, for 5000 samples, 90% of the entries have a mean error which is less than 0.2 amperes. Additionally, the maximum deviation from the mean using 5000 samples is approximately 0.25

amperes. It is expected that mean error converges to zero as the optimal size is arrived at. From this analysis, 5000 samples were deduced to be the size for which convergence begins.

4.5.2 Statistical coherence for the individual load measurements

A valid pool of load profiles from the Monte Carlo simulation will demonstrate comparable intra-interval variability, which is a measure of the diversity of the grouped customer loads, and inter-interval variability, which measures the correlation between loads in different intervals and the possible ADD profile of the load. The conformance of the simulated samples to these characteristics can be referred to as *statistical coherence* and the term is used consistently throughout this chapter, and some parts of the thesis. To carry out this, we investigate both the intra-interval as well as the inter-interval (temporal) coherence.

4.5.2.1 Assessing intra- interval coherence

To check for the intra-interval coherence of the sampled data sets, several random hours are selected, and their non-parametric distributions plotted. For each hour of the actual historic demand data with 42 customers, and the stochastic representative sample of 5000 samples, distribution plots are generated and compared for consistency. Figure 12 shows the distribution plots obtained for the actual demand and the *expanded* optimal demand data.

There is no significant deviation in the distribution fits between the actual demand distribution and the *expanded* demand distribution. The distributions show a very small variation in the density of the load currents. It can also be inferred from the distributions that both the mean and the spread of the optimal *expanded* demand and the actual original demand are consistent with each other. This points to a likely preservation of intra-interval demand diversity for the actual and the *expanded* sample.

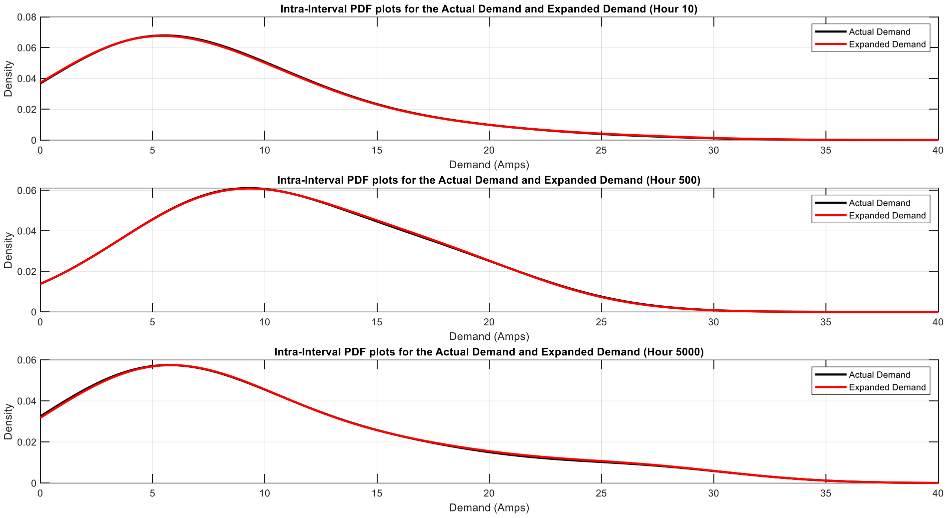


Figure 14: Distribution fits for selected hours to show the intra-interval preservation of demand diversity from the stochastic expansion process.

4.5.2.2 Assessing temporal coherence of the data

Temporal coherence refers to the inter-interval variability, which measures the correlation between loads in different intervals and the average demand profile of the load. Temporal coherence was determined through a comparison of the average profiles of both the *expanded demand* sample and the *actual demand* sample. A similar evaluation was carried out on the 95th as well as the 5th percentile values of the two sets of data. The load profiles were plotted and compared, and convergence was observed.

The variation in the load profiles of the mean, the 95th percentile and the 5th percentile for the two sets of data was investigated. This was done for two randomly selected months; February and June. Figures 13 and 14 show the results obtained.

From the results, it is evident that the mean time series plot of the actual demand data and that of the *expanded demand*, have a better convergence than that of the 5th and the 95th percentile time series, for the obtained optimal sample size. This may be due to the fact that, in calculating the average, the law of large numbers dictates that the expectation is always close to the theoretical mean (Reddy, 2001), for a large set of random variables. However, the law of large numbers is not applicable to outlying values such as the 95th and the 5th percentile values. They therefore exhibit greater variation when compared to that of the average. A similar pattern is observed in the average load profiles for the two months which have been evaluated.

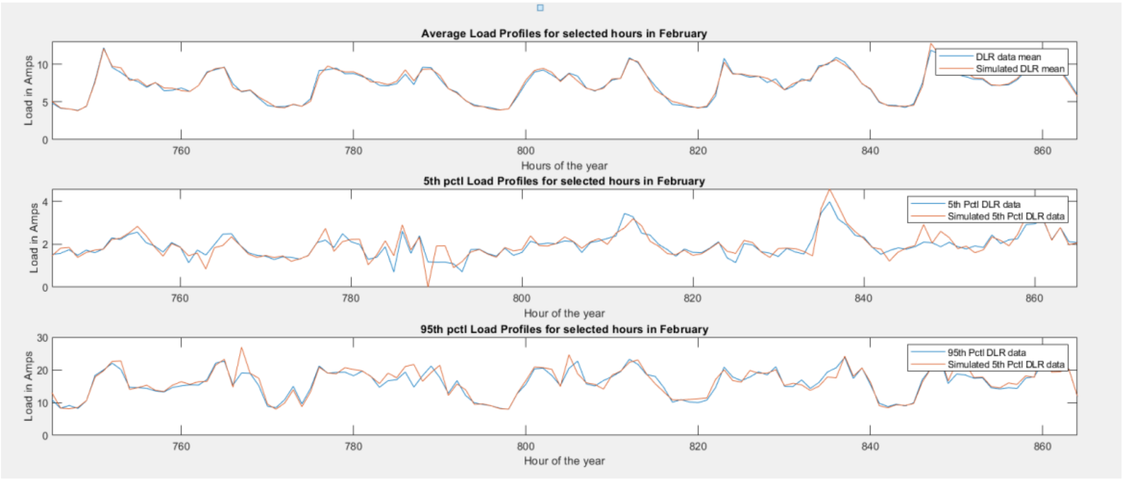


Figure 15: Comparison of the mean, 5th and 95th percentiles and simulated data for the selected

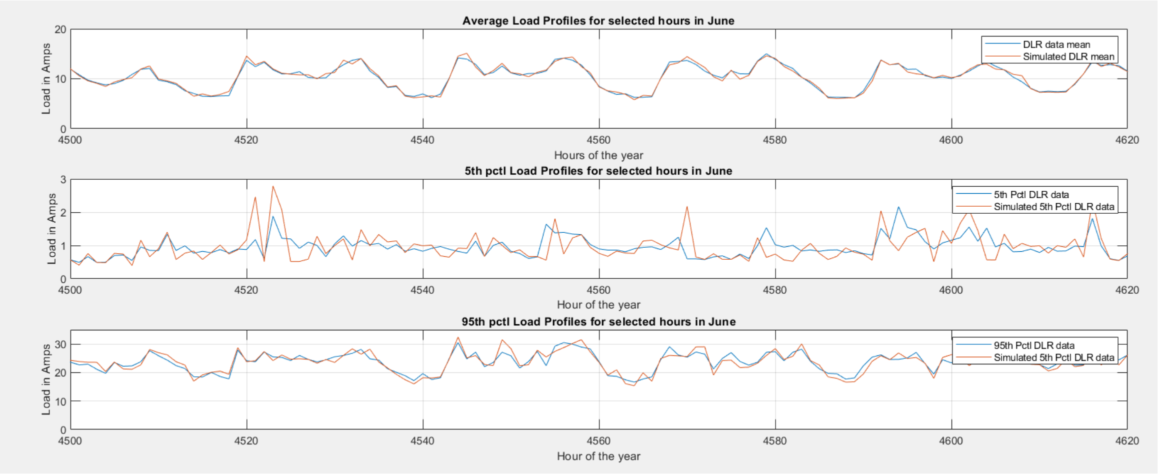


Figure 16: Demand profile for the selected hours for the mean, 5th percentile and the 95th percentile for the actual historical demand data, expanded historical data for June

4.5.2.3 Determining the correlation coefficient

To establish whether the expanded demand data does not significantly deviate from the actual demand data, a correlation plot comparing the two sets of demand data is plotted. To do this, the relationship between demands in a 24 hour period was examined. A curve fitting exercise was carried out to determine the standard error as well as the correlation coefficient. Figure 15 shows the results.

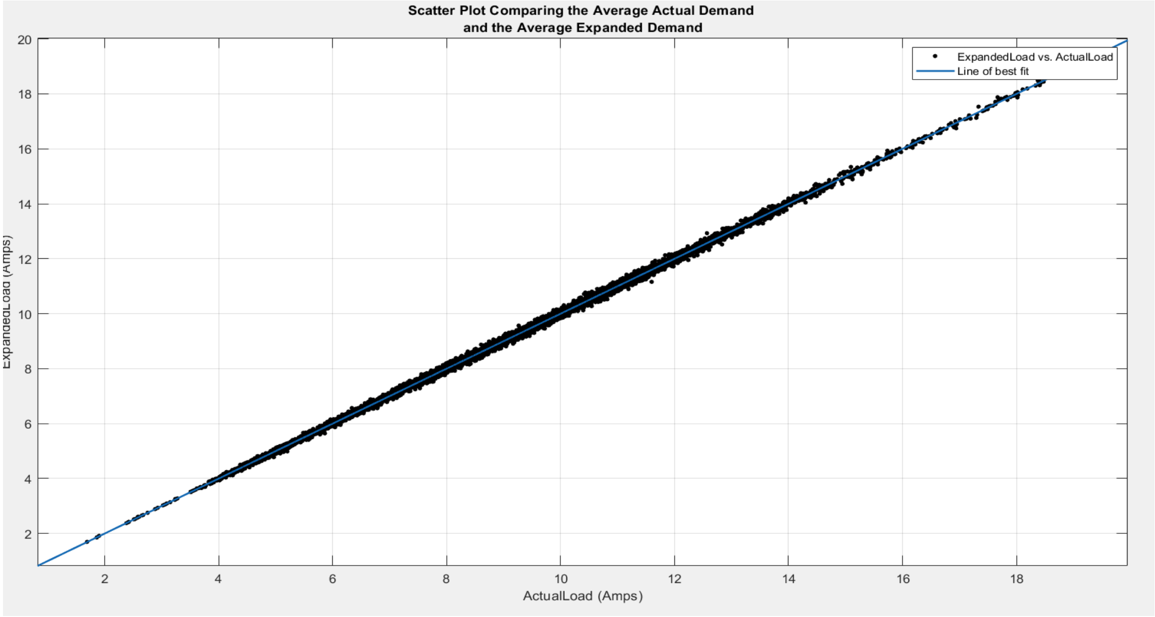


Figure 17: Correlation plot of the expanded DLR data vs the original set of data

The correlation coefficient was determined to be 0.9993. The root mean square error (RMSE) was determined to be 0.0778. These two values are of significant in terms of the temporal variance of the expanded demand from the actual measured demand. The high correlation value of 0.99 indicates that there exist negligible variations between the actual and the expanded demand. The value of the RMSE is a likely indicator that the temporal characteristics of the actual demand are preserved, and the expanded demand can thus be used in the aggregation process.

4.5.3 Conclusion

The purpose of the three tests described above is to ascertain whether the *expanded optimal demand* preserves the *statistical parameters* of the actual demand. These test results are therefore a validation of the stochastic expansion process. Additionally, they validate the further use of the *expanded demand* data in the next process of stochastic aggregation.

From the validation processes, it is concluded that the statistical expansion process preserves both temporal and intra- hour characteristics of the historical demand data and the expanded demand data can be used in the stochastic aggregation process.

4.6 Stochastic aggregation of demand data

Stochastic aggregation describes the process of creating aggregated demand profiles from the expanded sets of individual demand profiles. The purpose of stochastic aggregation is to create sets of aggregated demand profiles representing possible aggregated feeder profiles that could be recorded at a particular time. This process is carried out for a feeder with a known number of customers. For each period of time, *n* customers are selected randomly from the *expanded historic demand* and summated to generate a scenario of aggregated demand profiles. The process is carried out using Monte Carlo process.

4.6.1 Determining the optimal sample size of aggregated demand

As earlier indicated, each customer load has a large number of possible states, due to the fact that the load is stochastic. As such, the *aggregated load* represents '*combinations of load scenarios'* of the customer loads. The extent of the possible scenarios, which depends on the total number of customers and the variability in the load, can be indicated using mathematical calculations based on permutations and combinations. It can be demonstrated that the total number of simulations easily reaches the scale of millions. It is impossible to simulate every scenario, hence a Monte Carlo Simulation with reduced trials and tested for convergence is a viable approach where analytical methods are avoided. The purpose of this exercise is to determine the optimal sample size for which the intra-interval variance and the inter-interval correlation of the *aggregated demand* is preserved. The following procedure is followed.

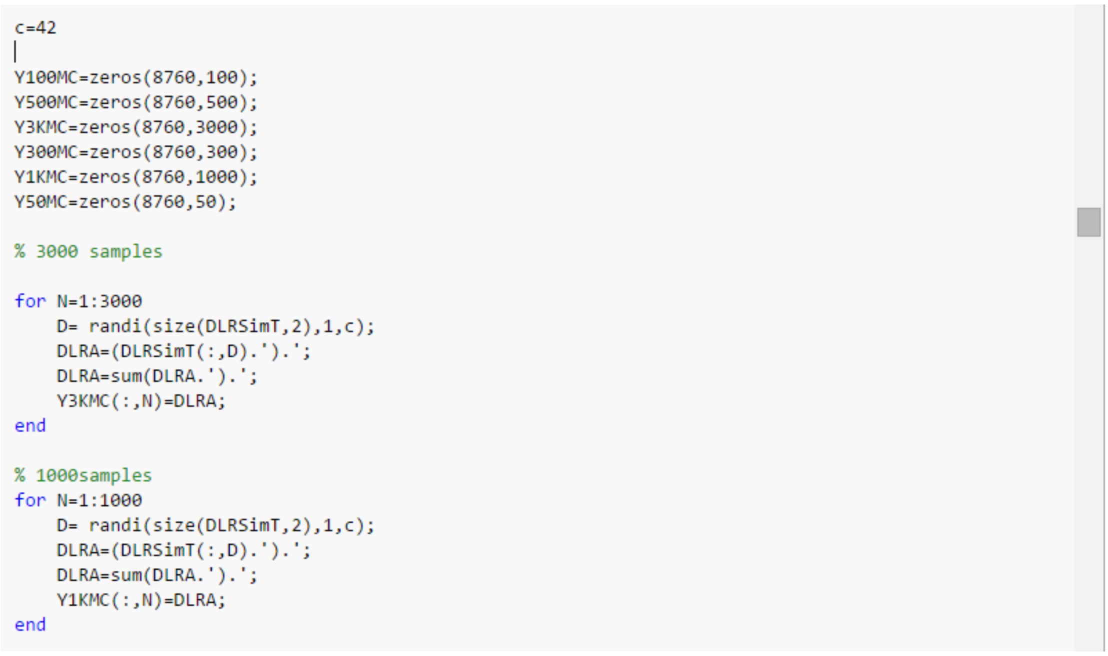
i) From the selected sample size of individual customer demand data, randomly select *n* load vectors from the characteristic pool of load vectors and sum them to achieve a scenario of the aggregated load vector.

ii) To get *m* sets of aggregated data, implement *m* Monte Carlo simulations.

iii) To test the adequacy of *m* MCS trials, different sample sizes per interval are simulated and tests are carried out to test whether there is preservation of: diversity of the demand; inter-interval correlation of loads; and the ADMD profile of the load

iv) Plot the mean, the 5th percentile and the 95th percentile values for each set of data for different *m* time series.

By running a Monte Carlo simulation, it was observed that at 1000 samples, represented by 1000 Monte Carlo runs convergence of the three statistical parameters begin. As such, *m* was selected to be 1000. Figure 18 shows the code snippet used in the implementation and Figures 19 and 20 shows the results obtained from the process.



```
c=42

Y100MC=zeros(8760,100);
Y500MC=zeros(8760,500);
Y3KMC=zeros(8760,3000);
Y300MC=zeros(8760,300);
Y1KMC=zeros(8760,1000);
Y50MC=zeros(8760,50);

% 3000 samples

for N=1:3000
    D= randi(size(DLRSimT,2),1,c);
    DLRA=(DLRSimT(:,D).').';
    DLRA=sum(DLRA.').';
    Y3KMC(:,N)=DLRA;
end

% 1000samples
for N=1:1000
    D= randi(size(DLRSimT,2),1,c);
    DLRA=(DLRSimT(:,D).').';
    DLRA=sum(DLRA.').';
    Y1KMC(:,N)=DLRA;
end
```

Figure 18: Monte-Carlo simulations for 3000 and 1000 runs

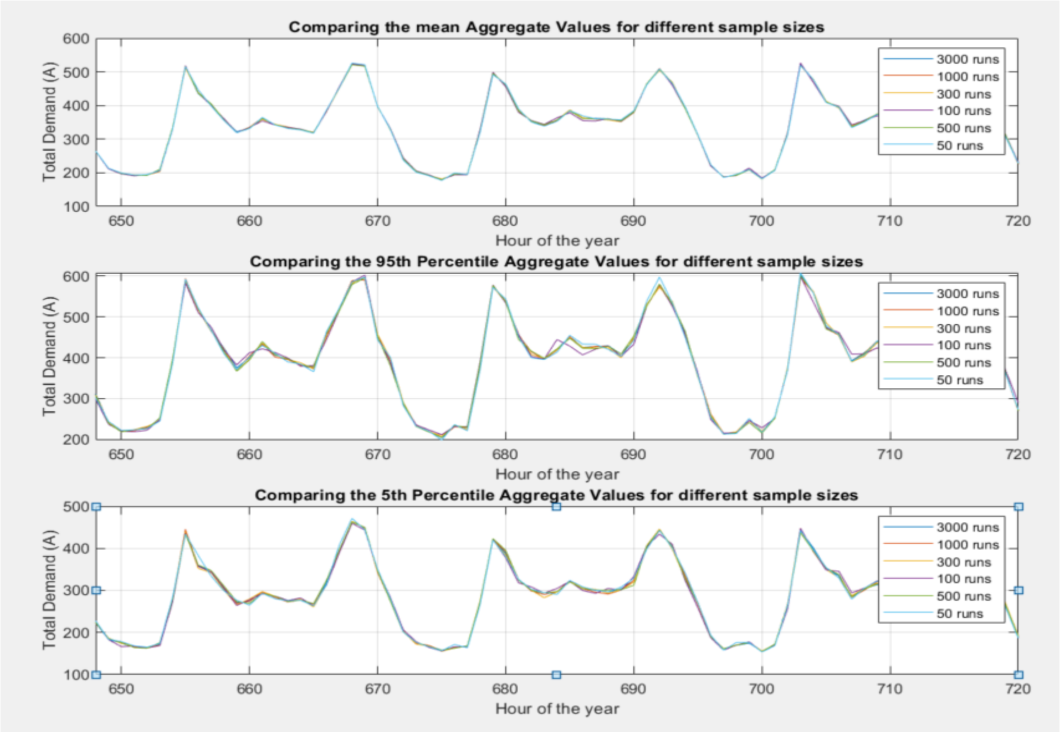


Figure 19: Plots showing the convergence of statistical parameters with different runs

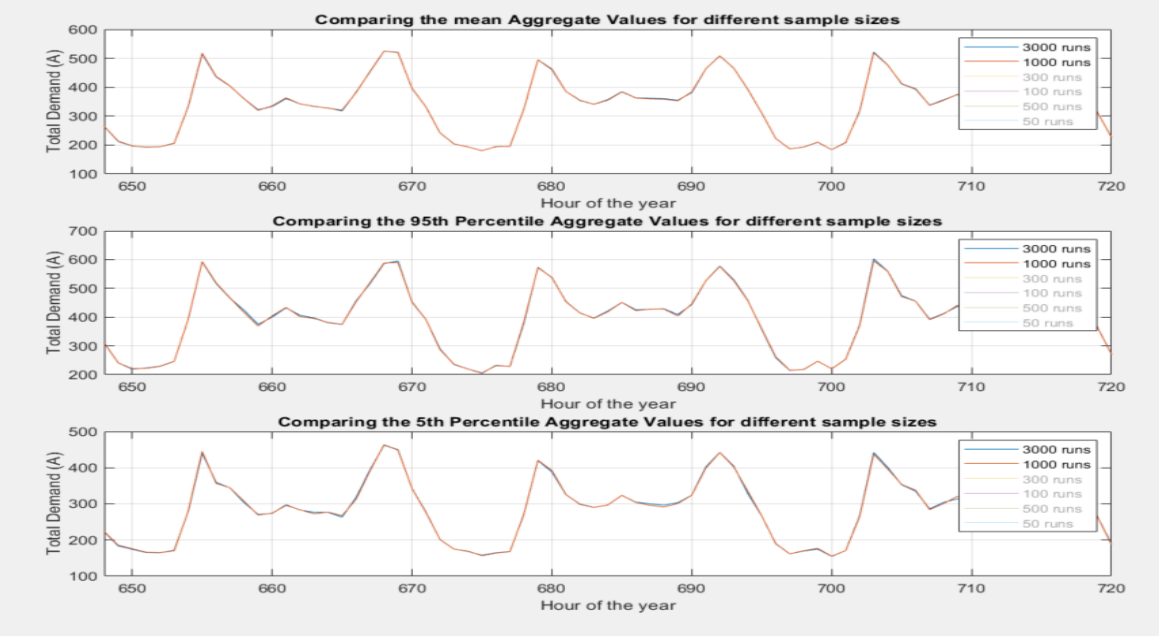


Figure 20: Convergence of the means and percentiles for 3000 and 1000 runs

4.7 Summary of the demand data processing procedure

The flow chart below shows the summary of the demand data processing as captured in the procedure outlined before.

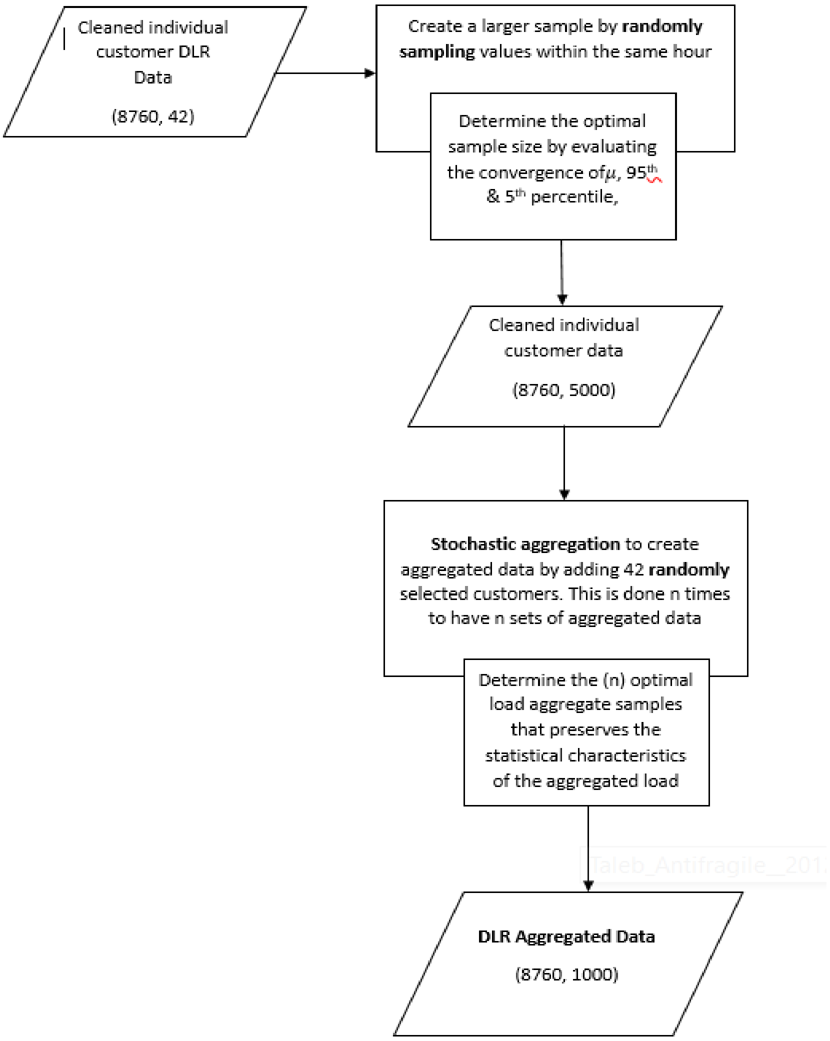
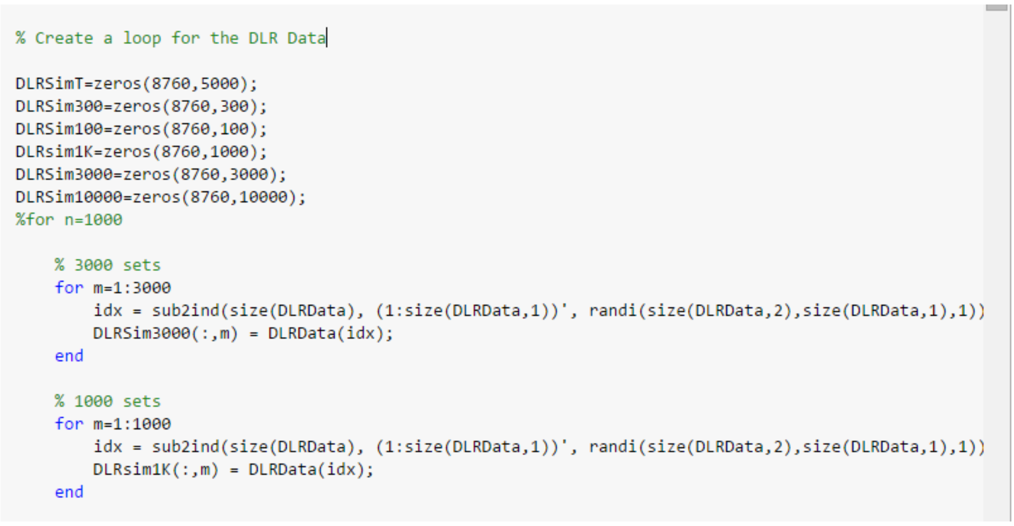

Cleaned individual customer DLR Data
(8760, 42)
Create a larger sample by **randomly sampling** values within the same hour
Determine the optimal sample size by evaluating the convergence of μ, 95th & 5th percentile,
Cleaned individual customer data
(8760, 5000)
Stochastic aggregation to create aggregated data by adding 42 **randomly** selected customers. This is done n times to have n sets of aggregated data
Determine the (n) optimal load aggregate samples that preserves the statistical characteristics of the aggregated load
DLR Aggregated Data
(8760, 1000)

Figure 21: Flow chart diagram showing how the aggregated demand data was created



```
% Create a loop for the DLR Data

DLRSimT=zeros(8760,5000);
DLRSim300=zeros(8760,300);
DLRSim100=zeros(8760,100);
DLRsim1K=zeros(8760,1000);
DLRSim3000=zeros(8760,3000);
DLRSim10000=zeros(8760,10000);
%for n=1000

    % 3000 sets
    for m=1:3000
        idx = sub2ind(size(DLRData), (1:size(DLRData,1))', randi(size(DLRData,2),size(DLRData,1),1))
        DLRSim3000(:,m) = DLRData(idx);
    end

    % 1000 sets
    for m=1:1000
        idx = sub2ind(size(DLRData), (1:size(DLRData,1))', randi(size(DLRData,2),size(DLRData,1),1))
        DLRsim1K(:,m) = DLRData(idx);
    end
```

Figure 22: Code section used to create expanded demand samples

4.8 Modelling the PV output factor

The purpose of this exercise is to generate the yearly solar PV output in Amps for every 1kWp of installed PV based on the corresponding values of GHI.

Generating a PV output factor was done using PVsyst software. PVsyst is a software that can be used in both sizing, studying and analysis and design of PV systems (PVsyst, 2019). It provides an interface through which one can specify the desired power output of a PV system, select a PV module and inverter from its resource database and finally model several types of PV systems and configurations. In our case study, the irradiance data used in generating these values was obtained from South African Universities Radiometric Network (SAURAN) data based at the Centre for Renewable and Sustainable Energy Studies (CRSES) at Stellenbosch University (Brooks, 2015).

The following conditions are assumed in the simulation process that was used by PVsyst.

i) Effects of shading are ignored.

ii) The plane tilt is set to 29 degrees based on standard angle used in PVsyst simulation.

iii) The Azimuth angle is set to 0 degrees based on the assumption that all panels are north facing.

iv) All solar panels are of Polycrystalline type, as this is the most commonly used globally.

The PV output factor is used in determining the PV capacity factor. The following flowchart describes the method used to determine the PV output factors.

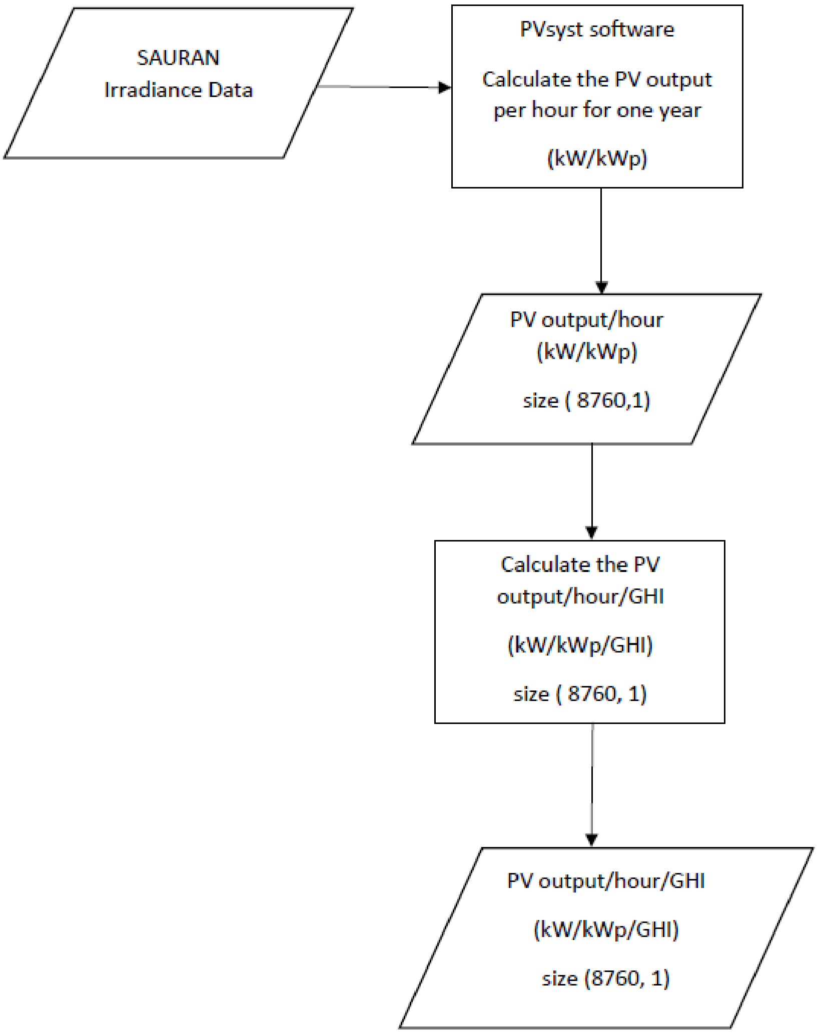


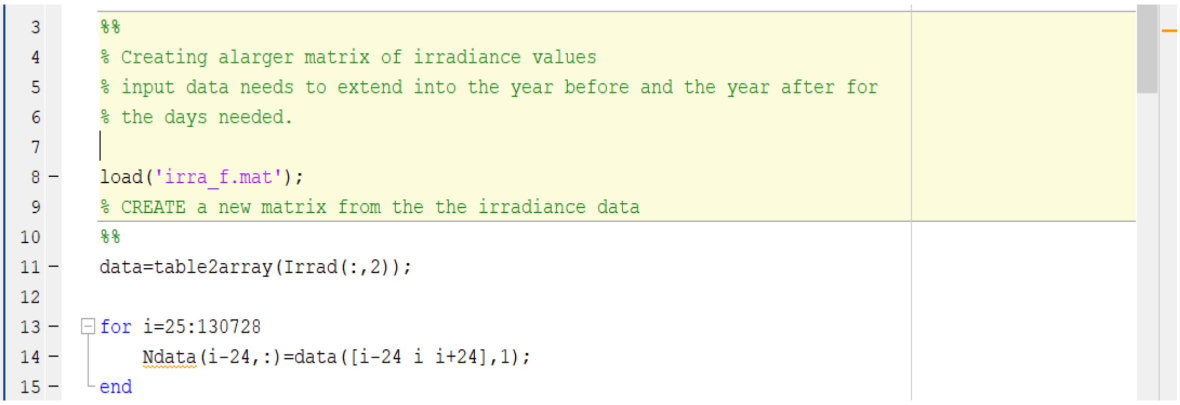
Figure 23: Flow chart diagram showing the methodology for deriving PV output per GHI ratio

4.8.1 Modelling the uncertainty in solar irradiation

In this section we expound on how we represented and modelled the uncertainty that characterizes solar irradiation. It is known that the amount solar irradiation reaching the solar panel determines the solar PV generation.

To do this, we obtained 14 years of GHI data from Helio-Clim database which was hourly resolved. Thus, for every hour of the years, 14 different GHI values were available for modelling. The different values for each hour were used to represent different possible irradiation scenarios. However, 14 scenarios cannot adequately represent the uncertainties that characterize the irradiation. In accordance to the law of large numbers, adequate sample size is required to fully represent the statistical parameters of the population from which the sample is derived (Boyce & Well, 1990). Sample sizes are therefore related to the distribution of the population. In their study, Guwaeder & Ramakuma,(2018) determined that solar irradiation takes either the Weibull or the normal distribution. Based on this study, we use a normal distribution to determine the minimum sample size that we can generate.

For normal distribution, a minimum of 30 data points is required (Reddy, 2001). Based on the assumption that solar irradiance follows a normal distribution, the dataset was expanded by systematic sampling and creation of 2 more sets of data for every hour of the year, by sampling the values of a day before and a day after. This yielded 42 irradiance values for every time of the year which was more than 30 irradiance values for each interval of time. Figure 24 shows the snippet of the code used in the sampling process.



```
%%
% Creating alarger matrix of irradiance values
% input data needs to extend into the year before and the year after for
% the days needed.

load('irra_f.mat');
% CREATE a new matrix from the the irradiance data
%%
data=table2array(Irrad(:,2));

for i=25:130728
    Ndata(i-24,:)=data([i-24 i i+24],1);
end
```

Figure 24: Creation of new values by picking of a day before and a day after

4.8.2 Modelling of the PV capacity factor table

Figure 25 shows a flow chart representing the process used to generate the PV capacity factor It shows the modification that was carried out on the GHI data from the Helio-Clim database, the modelling to create a larger sample that can be used for statistical distribution and the eventual steps of calculating the PV capacity factor.

The PV capacity factor shows the variation of the PV output in each hour of the year for 1kWp of installed solar PV. It thus captures the intra-hour variations that may be experienced in PV output due to intermittency. The GHI data undergoes stochastic expansion in order to increase the sample size.

$$PV\ Capacity\ Factor\ \left(\frac{kW}{kWp}\right)(8760{,}42) = PVoutput\ Factor\ (8760{,}1)(\frac{kw}{kWp}/GHI) \times GHI(8760{,}42)\ldots \text{(iv)}$$

For an installed capacity of *G* kWp, the PV capacity factor is used to calculate the solar PV output. For conformity purposes, The PV capacity factor is converted to Amps per kWp

4.8.3 Modelling solar PV output for installed capacity G

For a scalar *G,* representing the installed capacity on a feeder network, the PV output is generated by the following equation:

$$PV\ output = G \times PV\ Capacity\ factor$$

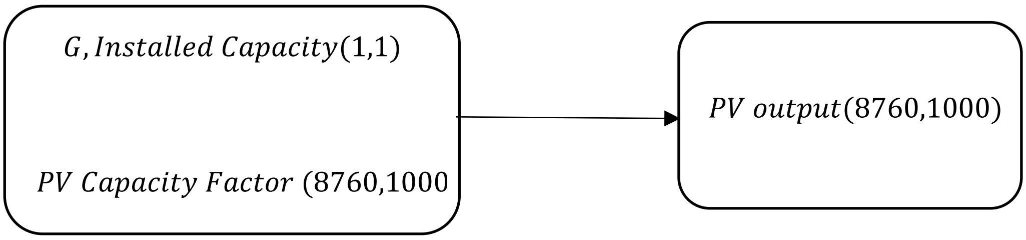

$G, Installed\ Capacity(1{,}1)$
$PV\ Capacity\ Factor\ (8760{,}1000$
$PV\ output(8760{,}1000)$

Figure 25 shows the procedure used in determining the PV capacity factor.

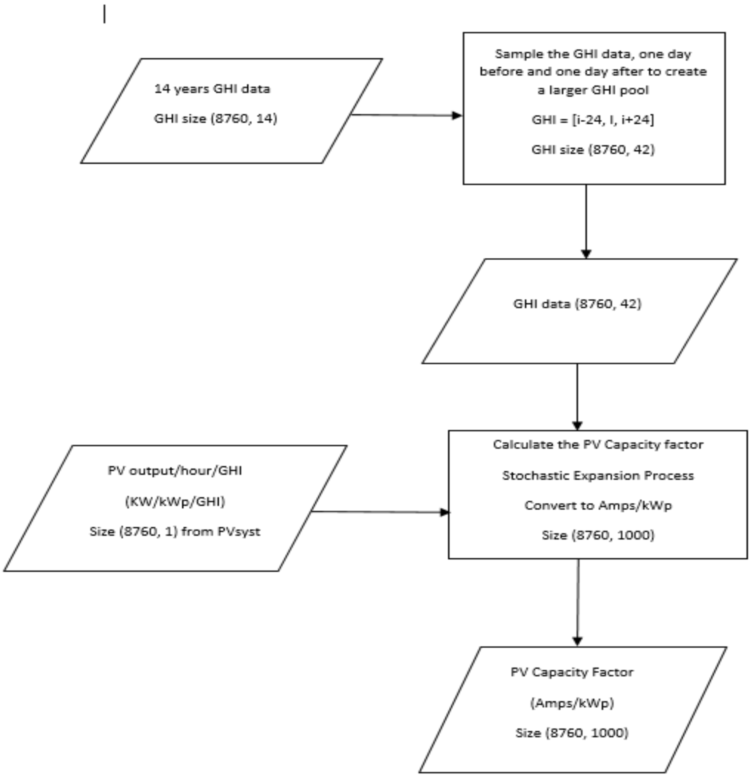


Figure 25: Flow chart diagram showing the generation of the PV capacity factor

4.9 Modelling using measured net demand

The objective of this section is to explain the procedure used in the estimation process by means of measured net demand.

4.9.1 Accounting for the effect of other factors on demand profile

To determine the changes in the demand profile that can be attributed to solar PV with fair accuracy, it is important to consider the effect of other energy factors on the demand profile and account for them. Demand profile changes can be attributed to a range of factors. These factors include but are not limited to SWH adoption, changes in tariffs, changes in household sizes, improved energy efficiency, changes in consumer electricity consumption patterns, modified household patterns due to other range of factors as well as change in income. Other differences may result from the incompatibility between the surveyed households. Accounting for these changes require an in depth study and modelling of the changes in these factors.

The measured net demand data was collected in 2019 from households in Stellenbosch from selected feeders. For the purpose of validation, knowledge of the installed PV capacity was required. As such, the metered households were separated into those with solar PV and those without. These two sets of data were used to determine the solar PV capacity embedded in the measured demand. The measured demand data is divided into households based on whether the identified and metered household had solar PV. The identification of the households for each category is based on the following:

i) The DoA study (Department of Agriculture Western Cape, 2018) was used to estimate the number of solar panels on rooftops within the study area. The DoA used Google Earth and physical inspection to identify households with PV panels.

ii) Stellenbosch Municipality information on the distribution network points of measurement as well as the solar PV installation details.

iii) Through the analysis of the metered net demand data, households with solar PV could also be identified based on their demand patterns. In some instances, substantial reverse current levels were identified indicating presence of solar PV. This analysis was used to cross-check and validate the DoA study report data.

Accordingly, the differentiated households were firstly used in the estimation of the capacity of solar PV SSEGs installed in the metered area. These results were used in testing case studies involving the use of measured net demand in the follow up cases. These results were compared to the results obtained by reviewing the report and the satellite imagery obtained from the DoA report.

By testing for the results obtained using different historical demand, there was a need to factor in other ‘energy factors’ such as solar water heating, change in customer behaviour and energy efficiency which impacts are also embedded in the net demand. This is because the net demand and the historical demand are years apart and as such the net demand has been affected by these energy factors which need to be accounted for in the modelling process.

A total of 27 households connected on selected feeders were metered for the month of May 2019. The demand current was measured for these households over this period. The number of rooftop PV on each of the metered houses were determined using the GIS survey report from the DoA. For the metered households, 108 panels were counted. According to the assumption made in the report, this amounts to approximately 27kWp installed capacity. Of the 27 households, data was reviewed to assess for quality. From the 27 households, 18 were selected and differentiated into those with solar PV SSEGs and those without. These were later used in the follow up modelling cases.

The two sets of households were taken through stochastic sampling to increase the sample size from 9 to 126. The sample size was arrived at using the combinatorial formula below.

$$C(n, r) = n!/(r!\,(n-r)!)$$

For 9 readings for each set of households, the maximum combinations there can be are 126. Stochastic aggregation was done and the demand aggregated. The set of data without solar PV was used to generate the capacity of solar PV embedded in the net demand. The capacity of solar PV obtained was then compared with what was obtained from the DoA and the discrepancies were discussed.

To validate the results obtained in this case, as well as the method, demand data for households without solar PV was replaced with historical demand data from other areas. The historical demand data used was from 2002. This means that there need to account for other variables that may affect the demand.

4.9.2 Constraints with the measured net demand

The number of customers with usable data reduced due to several operational constraints which included metering equipment breakdown and unreliable lack of data on some of the meters used. The data used reduced from 35 to 18 customers.

4.9.3 Accounting for other variables in the historical demand.

To account for the changes in demand from the 2002 to 2019, an error adjustment ratio was introduced. To do this, the study analysed and compared the average demand profile of the 2019 passive net demand and the 2002 historic demand for an equal number of customers.

The first objective was to adapt the model to account for changes in demand that are a result of changes in energy efficiency, solar water heating, changes in consumer behaviour as well as changes in tariffs. The second objective was to use this adapted model to estimate how much solar PV is embedded in the measured net demand.

To estimate the variation in the demand through the years (2002 to 2019), we calculated an error adjustment ratio using 2002 and 2019 demand data. The following procedure is followed in modelling the error adjustment factor.

i. From the measured demand, identify the number of households without solar PV installations on them. This was done using the 2019 demand data from Stellenbosch.

ii. Aggregate the demand values to obtain an aggregated demand profile for the selected number of customers. Determine the average daily load profile from this set of demand.

iii. From the historical individual demand, randomly select an equal number of customers as that in (i) above.

iv. Aggregate the historical demand to obtain the aggregated demand profile and determine the average daily profile as in (ii) above.

v. Obtain $Adj.Ratio = \frac{Aggregated\ Measured\ \ Demand_{Daily\ Avg}}{Aggregated\ Historical\ \ Demand_{Daily\ Avg}}$

There are several assumptions made in the derivation and application of this adjustment ratio.

i) That feeders classified under the same LSM class have similar demand patterns.

ii) The feeder demand in 2002 is more than the feeder demand in 2019 for an equal number of customers, in most cases.

The derived error adjustment ratio is an approximate scaling factor used to adjust the historical demand in order to account for all the factors that are not solar PV. These factors include but are not limited to solar water heating, changes in energy efficiency, changes in tariffs and other demand-changing variables. Figure 26 compares the average daily demand profiles for the year 2019, 2002 and the adjusted average daily demand profile for 2002.

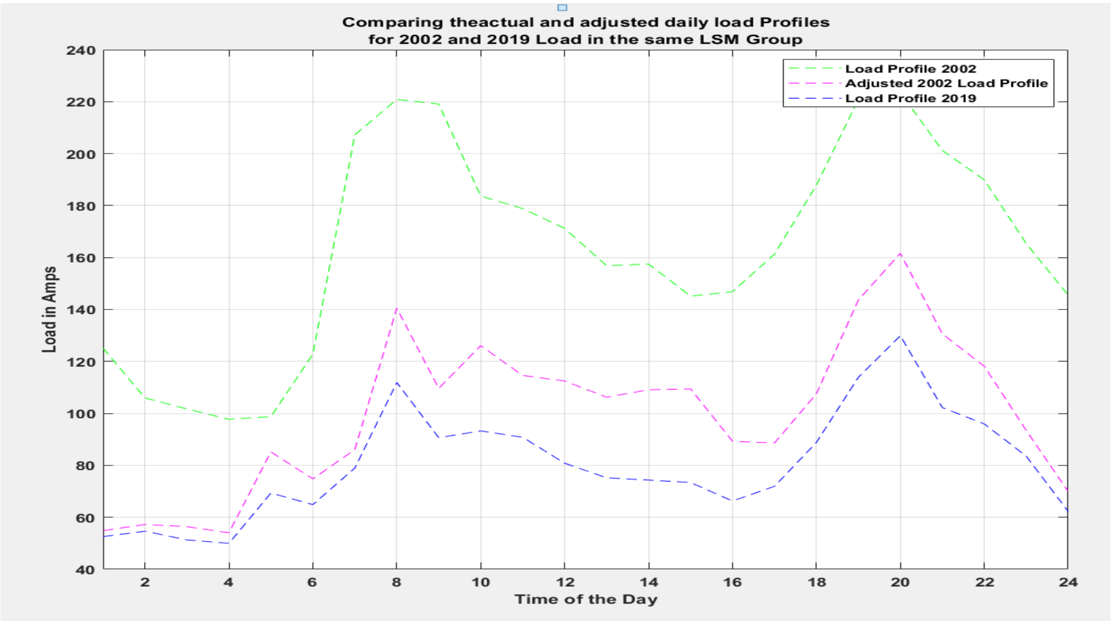


Figure 26: Plot of the actual and the adjusted 2002, and 2019 daily average demand after scaling for a feeder with 18 customers.

The figure below shows a representative segment of the adjusted aggregated demand profile for 2002, the actual aggregated demand of 2002 as well as that of 2019.

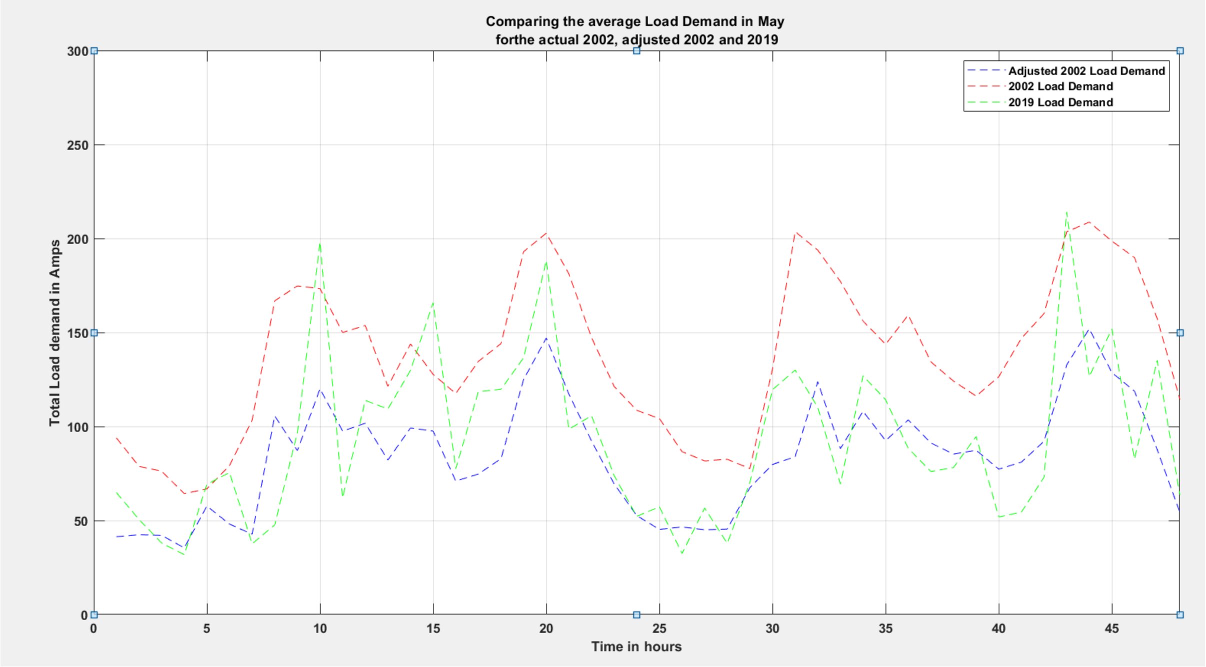


Figure 27: Comparing the measured 2019 demand, the actual and the adjusted 2002 demand for an 18 customer feeder

From Figure 27, it can be deduced that application of the error adjustment ratio scales the historical demand accordingly. The adjusted historical demand profile is closer to the 2019 demand as compared to the unadjusted demand and this caters for the impact of other factors other than solar PV SSEGs.

4.9.4 Error analysis and testing

This section evaluates the standard error of estimations. This evaluation was done by analysing and comparing the mean of the historical demand, or demand data with known solar PV capacity, and the aggregated net demand for both measured and simulated time series. The purpose of this test is to determine whether the use of standard error, coefficient of correlation and the determination of the standard error as a percentage of the peak demand, can be used as indicators of estimation accuracy. Furthermore, the mean and standard deviation of different datasets were evaluated to establish the relationship between the performances of the method when used with historical demands that have differing statistical properties.

Figure 28 shows the modified methodology flow chart that accounts for other variables that impact the demand over the years.

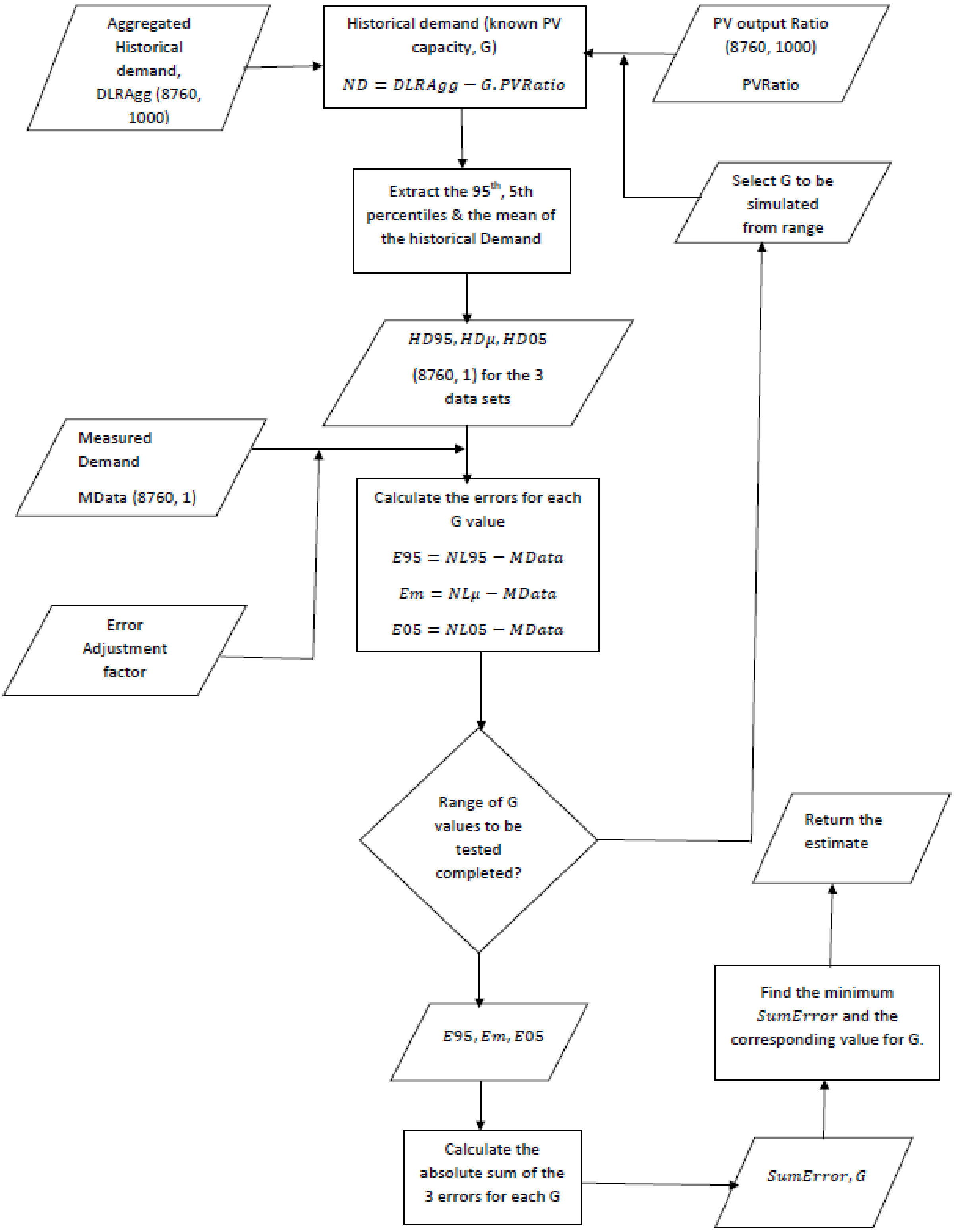

Aggregated Historical demand, DLRAgg (8760, 1000)
Historical demand (known PV capacity, G)
$ND = DLRAgg - G.PVRatio$
PV output Ratio (8760, 1000)
PVRatio
Extract the 95th, 5th percentiles & the mean of the historical Demand
Select G to be simulated from range
$HD95, HD\mu, HD05$
(8760, 1) for the 3 data sets
Measured Demand
MData (8760, 1)
Calculate the errors for each G value
$E95 = NL95 - MData$
$Em = NL\mu - MData$
$E05 = NL05 - MData$
Error Adjustment factor
Range of G values to be tested completed?
Return the estimate
$E95, Em, E05$
Find the minimum $SumError$ and the corresponding value for G.
Calculate the absolute sum of the 3 errors for each G
$SumError, G$

Figure 28: Flow chart of the methodology which includes the error adjustment ratio

4.9.5 Summary of the general procedure

In summary, the following procedure was used in the modelling and testing of the case studies in this research.

i) Aggregated historic demand sets are generated using a stochastic aggregation process described in previous sections.

ii) To account for the solar water heating, changes in tariffs, energy efficiency and others factors, an adjustment factor is applied.

iii) Using the PV capacity factor, possible values of the net demand for an area are created using the equation below, where *G* is a scenario of installed capacity estimate.

$$Net\ Load = Aggregated\ Historical\ Demand - G\ \times\ PV\ Capacity\ factor$$

For each value of *G*, an array of the net demand is generated and the mean, the 95th and the 5th percentile demand values are calculated.

iv) The error is calculated by finding the difference between mean, 95th and 5th percentile values of the historical demand and the simulated aggregated net demand.

$$E\mu = Agg.Historic\ (\mu) - Agg.Net\ Demand$$

$$E95 = Agg.Historic(95) - Agg.Net\ Demand$$

$$E05 = Agg.Historic(05) - Agg.Net\ Demand$$

Where E95, E05 and Eμ are the 95th percentile error, 5th percentile error and the mean error.

v) The absolute error is calculated by summing the errors obtained from vi.

vi) The value of G is found such that the sum of errors is minimum.

4.10 Conclusion

This chapter has detailed the structure and the procedures involved in developing and validating this model. Furthermore, a case by case analysis of steps involved in constructing a model from an ideal scenario, to that which accounts for other factors such as energy efficiency and change in tariffs is provided.

5 Results, findings and analysis

This chapter tests the developed approach to PV estimation and discusses the results obtained. The tests, which are defined in four case studies, investigate the accuracy of the approach and the scope of its application. We test the application of the proposed method using historical demand data, from which we obtain the respective simulated demand data with PV in case study 1. We also test the application of this method using actual measured demand data differentiated in data with PV and data without PV in case study 3. In the absence of historical demand data for a particular feeder, we investigate whether it is possible to use demand data obtained from a different feeder in an area classified under the same LSM. This is captured in case study 2, where we use simulated demand data from one area and historical demand data from a different feeder within the same LSM cluster. This is further tested in case study 4 where we estimate the installed capacity in demand measured from Stellenbosch using historical demand drawn from areas with both similar and dissimilar characteristics. The criteria used in the selection of these areas is captured in section 4.1.

5.1 Testing and validation of the methodology

The purpose of this section is to test and validate the proposed methodology. Four case studies were conducted. The performance of the methodology was tested in each of these cases. The validation of the methodology was carried out using both simulated and measured net demand. Simulated net demand was calculated from the historical demand by factoring in the solar PV output.

Measured net demand was obtained from metering households on selected feeders in Stellenbosch Municipality. The selection was based on the presence of rooftop PV and the likelihood of the area being classified as high income residential.

5.2 Test case studies

In the first case study the method was tested using historical demand obtained from Moreletta Park, Pretoria. It was validated using simulated net demand derived from the same historical demand.

In the second case study, in order to investigate the performance of this method in the case where one of the data sets was missing, we used historical demand obtained from Welgemoed. Welgemoed site was selected for the second part of the investigation since it is in the same LSM group as Moreletta. Further, based on the socio-demographic indicators from the DLR report, the two areas are likely to have similar demand characteristics. This case is validated using simulated net demand from different area.

The third case study used metered net demand obtained from Stellenbosch to determine the capacity of solar PV embedded in the measured Stellenbosch demand. The method adopted in the third case study identified metered customers without rooftop solar PV as “historic demand” and compared them to customers with PV as “net demand”. The results obtained was compared to the capacity calculated using the satellite imagery obtained from the DoA (Department of Agriculture Western Cape, 2018)) study. May 2019 measured demand data was used in this case.

The last case investigated the performance of the proposed method in the absence of historical demand. Measured net demand is obtained from Stellenbosch and validated using historical demand obtained from different areas which are classified under the same LSM. The condition of implementing this is that the area from which historical demand is drawn must have compatible characteristics as that which measured demand is obtained. The selection of the area should consider wealth indicator, the electricity consumption patterns, and the average household income among other issues. However, two more comparisons were also made to ascertain the performance of the method where incompatible areas are used.

To this effect, this this case used historical demand from Moreletta 2002, Welgemoed 2002, West Ridge 2002 and Gasese 2002 to test how much PV was embedded in the measured Stellenbosch demand. This case investigates the performance and constraints of using this method while considering the impact of energy efficiency, solar water heating, and change in tariffs among other energy factors.

5.2.1 Case study 1: Testing with simulated net demand

The objective of this case study is to test the performance of the developed methodology using ‘clean’ net demand data, which is without the impacts of other energy factors apart from the known capacity of embedded PV. This case, based on a ‘perfect scenario’, will allow a fair and easy assessment of the performance of the methodology. Since such data is impossible practically, simulated net demand data obtained using the historical demand and the solar PV output, tested for different kWp capacities, is used. Details on how this data is generated was discussed in Chapter 4.

In this case, it was assumed that both the historical demand data and net demand data for the tested feeder are available. In this case, historical demand data from Moreletta Park was used in the evaluation. The demand data of 2002 was used.

5.2.2 Case study 2: Simulated net demand and historical demand

This case investigates the performance of the method using in a scenario where net demand for a test area is available while the historical demand is not. The case study determines whether the performance of the proposed method is subject to LSM class from which historical demand is derived. The validation of this case was done using simulated net demand. In the first part of this case study, data from Moreletta Park and Welgemoed (both in LSM 10 - high) was used. Moreletta Park data provided the historical demand while Welgemoed provided the net demand. Two months' data was used in the study. The months were selected based on the fidelity of data. As such the months with least missing values were selected and in this case October and November were used.

Case study 2A: Case 2 with similar community

Case study 2A investigates the method's performance in a scenario where historical demand is sourced from areas that are in the same LSM class. Historical demand data spanning over 2 months is used in order to maintain the fidelity of the data. Historical demand from Moreletta Park is used to validate the amount of solar PV embedded in the simulated net demand from Welgemoed.

Case study 2B: Case 2 with different community

This case has a similar approach to the case study 2A. However, this case seeks to determine whether the LSM class of historical demand used in the case study 2 affects the performance of the method. This case study uses historical demand from the Moreletta Park to test the performance of the method using simulated net demand data for West Ridge and Gasese. Moreletta Park, West Ridge and Gasese are all categorized in different LSM classes. May 2002 data is used in this process and it considers a feeder with 42 customers connected on it.

5.2.3 Case study 3: Using actual measured demand

The purpose of case study 3 is to investigate the performance of the method using measured net demand data and compare the same to the actual estimation obtained from the DoA study. Historical demand data used in cases 1 and 2 is replaced by measured Stellenbosch demand data verified to have zero solar PV capacity installed. The simulated net demand is replaced with measured Stellenbosch net demand from households which have solar PV installed on them. Through the analysis and comparison of the measured net demand, with PV and without PV, we estimate the capacity of the solar PV embedded in the measured net demand.
A total of 18 customer demand data was used and differentiated into those without PV and those with PV in accordance to the process described earlier. We substituted historical demand as used in case study 2 with measured household demand without PV obtained from carefully selected feeders in Stellenbosch.

The two sets of households were taken through stochastic expansion and aggregation to obtain aggregated demand profiles. The model was applied to these two sets of data to determine how much PV was embedded in measured households with PV. The capacity of solar PV obtained was then compared with what was obtained from the DoA.

Further, this case is used to provide a basis for evaluating the performance of the model in case studies 4A to 4B

The results obtained in this case study is also used to determine the capacity of solar PV embedded in the measured net demand while comparing this to the possible total PV output expected from the total PV count recorded from the same feeders, using the DoA study.

5.2.4 Case study 4: Factoring in other demand variables

The objective of this case is to assess the performance of the model with practical demand data, which may have several ‘energy factors’ such as energy efficiency, solar water heating and impacts of demand-side management (DSM) approaches, for example tariffs. The error adjustment ratio obtained above is used to adjust the historical demand and cater for these changes that take place through the years.

5.2.5 Case Study 4A: Adjusted historical demand, measured net demand, same LSM

Historical demand data used was obtained for Moreletta Park and Welgemoed, to which the error adjustment ration was applied. The net demand data used in this study was obtained from Stellenbosch. May 2002 historical demand data was used in both cases since the available net demand data was taken in May. Both the net demand and the historical demand used have one-hour temporal resolutions.

5.2.6 Case study 4B: Adjusted historical demand, measured net demand, different LSM

Historical demand used was obtained for West ridge and Gasese. As indicated, Gasese, West ridge as well as Stellenbosch are found in different LSM clusters. The net demand data used in this study was obtained from Stellenbosch. May 2002 demand data was used in both cases, since the available net demand data was taken in May. Both the net demand and the historical demand used have one hour temporal resolutions. Unlike in the case study 4 A, the net demand and the historical demand were taken from areas with different socio-demographic indicators and their demand parameters are likely to be different.

5.3 Results for case study 1

Case study 1 investigated the performance of the proposed method using historical demand data taken from Moreletta Park in 2002. The case was validated using simulated demand data that was derived from the historical demand. As indicated earlier, one year's data was used for testing. Table 8 below shows the results obtained for various tested solar PV capacities in the net demand. The actual installed capacity is the embedded capacity in the simulated net demand while the estimated capacity is capacity obtained by analysing the simulated net demand using the proposed method.

Table 8: Estimated PV capacity for case study 1 using Moreletta demand data

Actual installed Capacity (kWp)	25	75	125	275	300
Estimated Capacity (kWp)	25	75	125	275	300

It can be deduced from the Table 8 that using simulated net demand obtained from historical demand of an area provides accurate estimations. There is no variation between the actual embedded capacity and the estimated embedded capacity. This is probably because, in using historical demand and simulated net demand from the same area and time, the only variation in the demand data is the solar PV. Figure 29 shows the error plot obtained for the 25kWp installed capacity using a feeder with 42 customers.

The proposed method works on the principle of error reduction. It can be seen that the minimum error for a tested embedded capacity of 25kW returns an estimate of 25kW. This shows that using simulated historical demand and simulated net demand from the area would return accurate estimates of the embedded capacity.

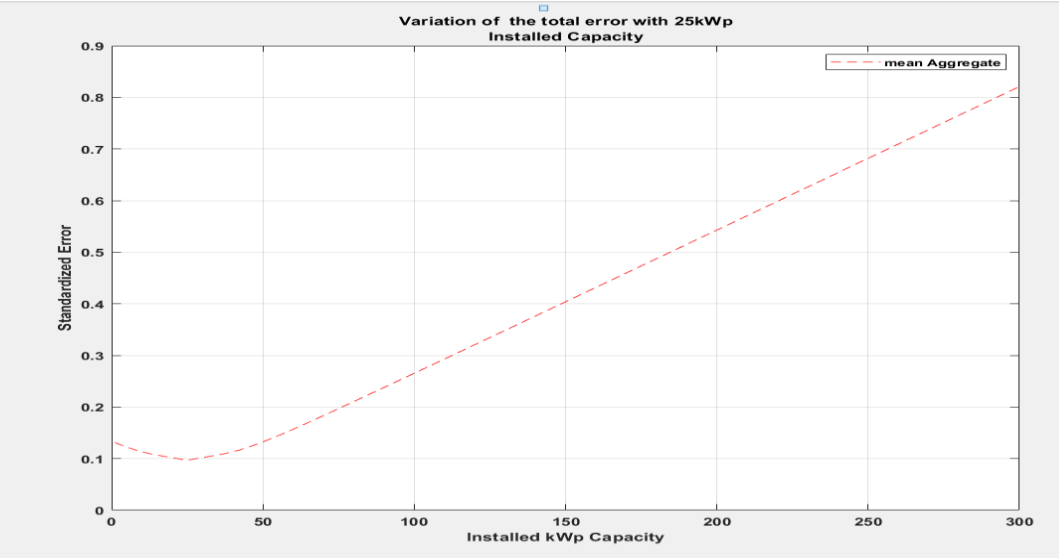


Figure 29: Plot of error estimation for the scenario with 25kWp installed PV.

The results obtained demonstrate that the method used in this study – i.e. the use of stochastic expansion and stochastic aggregation, is an effective estimation method in the simulated situation. This is because in using simulated net demand, the estimated embedded solar PV capacity and the actual capacity in the net demand obtains a 100% match. Case study 1 sets precedence for further investigations on the performance of the proposed method under different circumstances.

5.4 Results for case study 2A: Same LSM class

Table 9 compares the estimated solar PV capacities for different actual installed capacities in simulated net demand. This table shows estimations of the amount of solar PV embedded in the simulated net demand from Welgemoed using historical demand obtained from Moreletta Park.

Table 9: Estimated solar PV capacities for the same LSM class

Actual installed Capacity (kWp)	25	75	125	275	300
Estimated Capacity (kWp	27	78	127	277	302

From Table 9 above, the estimated capacity and the actual PV SSEG capacity in the simulated net demand deviate by a very small margin. The consistency of this error is a probable indicator of a sensitivity error in the estimation process. These results show that it may be possible to estimate the amount of solar PV embedded in the net demand obtained from an area using historical demand of another area classified under the same LSM class.

5.5 Results for case study 2B: Different LSM class

Historical demand data was taken from the Welgemoed data (which belongs to LSM class 10 - high). The simulated net demand used was obtained from two different areas drawn from different LSM class; these areas were West Ridge of LSM class 5 - high and Gasese of LSM class 1. Table 10 shows the results obtained from the simulation at different capacities.

Table 10: Variation of estimated solar PV capacities for different LSM Class

Actual embedded Capacity	Estimates using West Ridge (LSM 5 - High) (kWp)	Estimate using Gasese (LSM 1) (kWp)
250	239	180
200	190	132
150	134	80
100	87	33
10	1	1

From the results obtained in this study, it is noted that in the absence of historical demand data of one feeder, it is possible to use historical demand from another a feeder given that the two feeders are from the same LSM class. Results obtained using historical demand from a different feeder however show significant discrepancies.

Using historical demand obtained from Welgemoed (similar LSM class with simulated net demand), it is demonstrated that we can estimate with a fair degree of accuracy the amount of solar PV embedded in the net demand obtained from another feeder. It is also determined that for areas in the same LSM class, the results obtained have minimal errors. However, the error in estimation widens as the gap between the LSM classes widens. Different LSM clusters have different ADD profiles and the test areas which fall under the same LSM class may have similar ADD and ADMD values. This is a probable explanation for the wide error obtained in the analysis and comparison of those areas which fall under different LSM class.

For scenarios where the historical demand was drawn from an area in the same LSM class, the absolute error of estimation was about 2kWp. For the other areas with historical demand drawn from different LSM classes than that of the net demand, the absolute error of estimation is between 10kWp and 70kWp for West Ridge and Gasese. From these results, it can be concluded that the variation in the estimation error is probably tied to the difference in the LSM class.

These results show that it is possible to estimate the amount of solar PV SSEGs embedded in net demand, using historical demand data obtained from another feeder, given that the both the net demand data being tested and the historical demand data being used, are drawn from areas or feeders with the same LSM class. However, using data from areas classified under different LSM classes yielded results with huge discrepancies.

5.6 Results for case study 3

In case study 3, we test the application of our method using demand data obtained from Stellenbosch. The objective of this study is determine the capacity of solar PV embedded in the measured demand obtained from Stellenbosch and compare it with the estimates obtained from the DoA study. As indicated, it forms the basis of evaluating the performance of this method in the subsequent case studies.

Results obtained from this case study form a basis of analysis of the performance of this method using historical demand data. Figure 30 shows the error plot obtained. The minimum error is recorded at 14kW. Consequently, it is deduced that about 14kW of solar PV is embedded in the tested net demand. This error plot is obtained for a feeder with 18 customers.

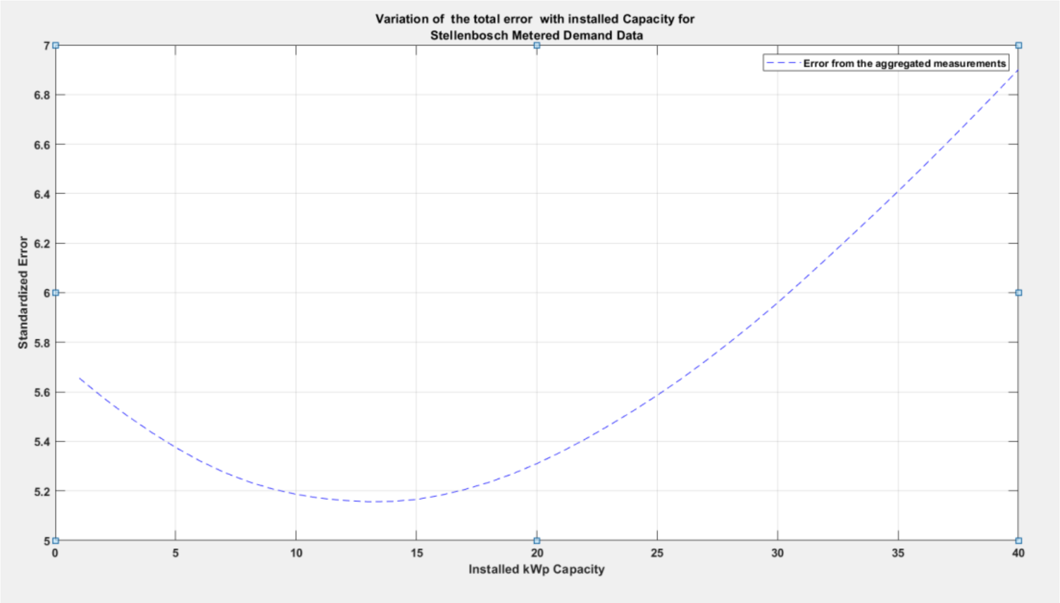


Figure 30: Error plot for the measured Stellenbosch demand data considering a feeder with 18 customers.

However, from the analysis of the DoA data of 2018 for the metered area, it was estimated that about 27kWp of solar PV could have been represented in the net demand data used. This estimation was arrived at using the aerial images and the calculations captured during the fly-over. Of the 27kWp of installed solar rooftop PV that was in the area under investigation, the proposed method reveals that only about 14kWp of the identified installations are embedded in the demand measurements.

The difference between the estimates obtained from the DoA study and those obtained using these methods may be attributed to the following reasons:

I. Possible discrepancies between the households that were metered and the households that were identified to be having installed rooftop PV from the GIS identification and mapping process.

II. Overestimation of the numbers of installed solar PV by the DoA. By physical inspection of the low-lying roofs in some areas, it was noticed that solar pool heater pipes were identified as solar PV.

In order to assess the performance as well as ascertain the results obtained using the practical net demand from Stellenbosch, further tests were carried out using historical demand drawn from areas in the same LSM class as that of Stellenbosch. Additional tests were carried out to assess the performance of the method using historical demand data from different LSM classes as with case study 2. These tests are carried out in the case study 4.

5.7 Result for case study 4A

Case study 4 uses both historical demand and measured net demand to estimate the embedded solar PV capacity in the net demand. The Stellenbosch historic demand data without solar PV in case study 3 is replaced with historic demand data from Moreletta Park. The net demand data from households with solar PV as used in case study 3, is replaced by the undifferentiated (the whole set) of measured net demand data obtained from Stellenbosch.

The objective of this test was to determine the performance of the method in the absence of historical demand data of the feeder where practical net demand is available. Moreletta Park and the selected households are found in the same LSM class (LSM 10 – high).

Apart from checking the performance of the method using actual measured net demand data, the method is used to ascertain whether the results obtained in case study 3 is a way of determining the total installed capacity impacting the demand. It is thus used to validate the estimates obtained in case study 3. Case study 4 makes use of the two sets of measured aggregate demand load data from May 2002 and May 2019. The feeders are assumed to have 18 customers.

To do this, the demand of 2002 was adjusted to cater for other ‘energy factors’. The Figure 31 below shows the demand time series for several hours comparing the adjusted demand and the actual demand for Moreletta Park, compared to the measured demand data in Stellenbosch. It is evident from Figure 31 that the adjusted demand is scaled to reduce the existing error present in the historical demand.

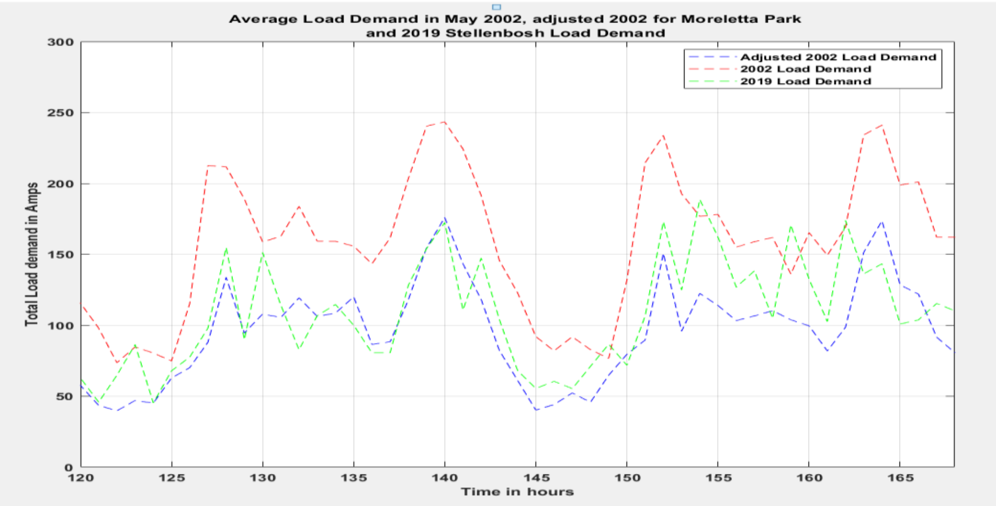


Figure 31: Demand variation for the May 2002 Moreletta Park load after adjustment for energy efficiency, solar water heating and other factors compared to the May 2019 load for 18 customers.

Correlation analysis was carried out by comparing the net demand and the mean historical demand, as well as the adjusted mean historical demand. This exercise evaluates whether adjustment improves the linear relationship between the measured net demand and the mean historical demand by checking on the R-Square as well as the standard error of the historical demand before and after adjustment. Table 11 below shows the results obtained from this exercise.

Table 11: Comparing correlations with May 2019 demand for the actual and adjusted May 2002 demand.

	R-Square	Adj. R-Square	RMSE (A)
Actual 2002 historical demand	0.3658	0.3649	31.39
Adjusted 2002 historical demand	0.5219	0.5212	27.25

From Table 11 it is evident that there is a remarkable change in the coefficient of determination, which indicates that by introducing the error adjustment, the adjusted demand follows the net demand more closely than before. Additionally, the standard error in the estimation is lowered to about 27.25 from 31.39. This change in the error may be attributed to the adjustment ratio used.

The adjusted aggregated historical demand data for 2002 was then used to determine how much solar PV SSEGs is embedded in the measured net demand of 2019 obtained through the metering of selected areas in Stellenbosch.

The method was firstly used to evaluate the capacity of solar PV embedded in demand before the adjustment of the historical demand data and after the application of the adjustment ratio. Figure 32 below shows the error plot obtained by evaluating the minimum error points using the adjusted and the unadjusted historical demand data. It is observed that when using the unadjusted historical demand, the estimated value of embedded solar PV is about 45kWp, while when using the adjusted value, the embedded capacity is about 15kWp. By comparing the estimates obtained from the error plot, it may be deduced that adjusting the historical demand to factor in the unknown energy factors improves the accuracy of estimation.

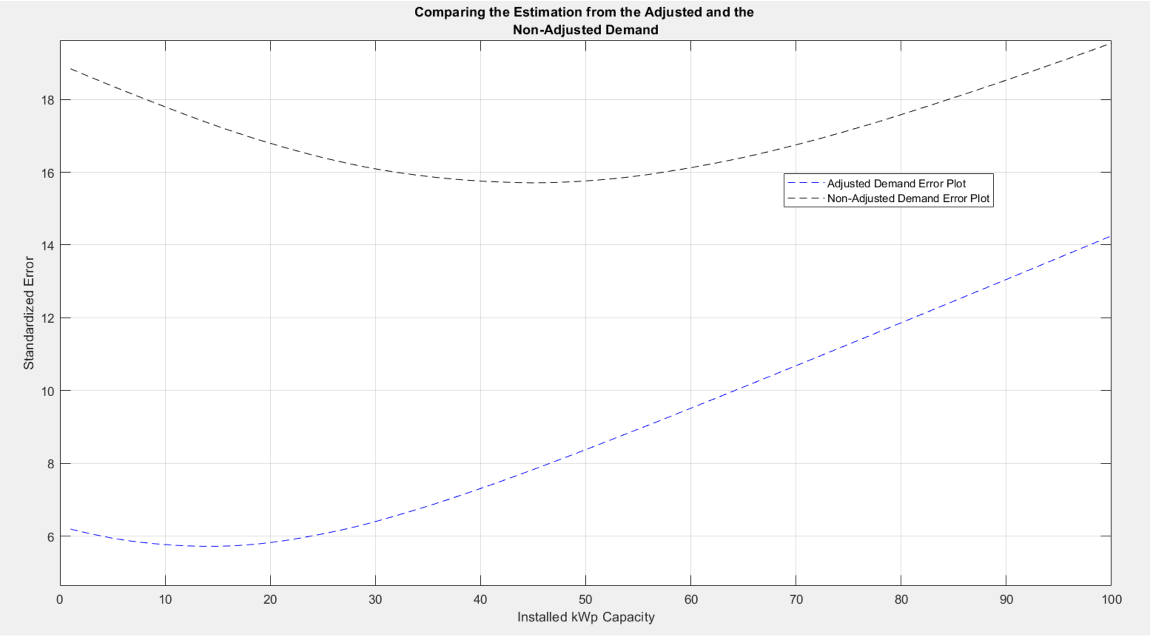


Figure 32: Comparing the estimated demand for the adjusted and the non-adjusted demand data where the validating data has 18 customers and 14kWp.

Figure 33 below shows the error plot obtained using the adjusted demand data alone to estimate the embedded solar PV Capacity in the case study 4.

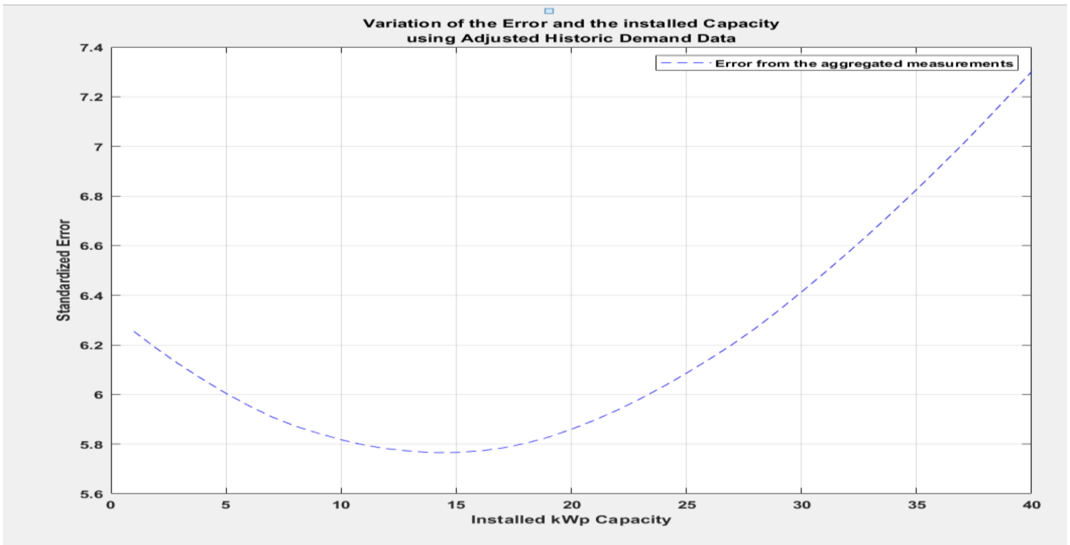


Figure 33: Error plot for installed capacity using adjusted historic for Moreletta Park using for an 18 customer feeder.

The plot indicates that an estimate of approximately 15kWp of solar PV is embedded in the Stellenbosch 2019 net demand. This is close to the estimation of 14kWp obtained in case study 3 using two sets of demand data obtained from Stellenbosch. From this, it can be concluded that it may be possible to carry out the estimation using historical demand data and net demand with acceptable accuracy levels.

To test the consistency of this method, historical demand data for Moreletta Park was replaced by historical demand from Welgemoed, which is in the same LSM class. Both Welgemoed and the metered Stellenbosch households can be classified under the same LSM cluster. Results obtained here are compared to those obtained in case studies 3 and to those obtained from Moreletta Park test and their accuracy compared. Further, the net demand data used in this case study is obtained from Stellenbosch.

Figure 34 shows the error curve obtained in the Welgemoed case study. From the plot, 13kWp of installed PV are embedded in the measured net demand, measured from the 18 households in Stellenbosch.

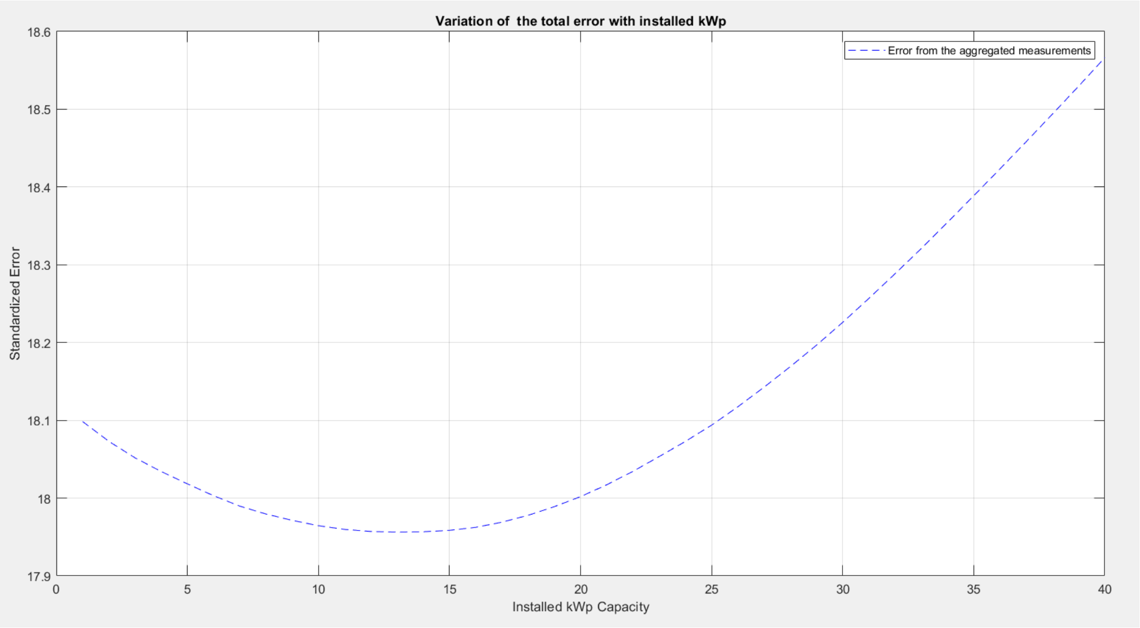


Figure 34: Case study 4 error plot showing the embedded solar PV using Welgemoed May 2002 historical demand and Stellenbosch May 2019 – tested for 18 customer feeder

By comparing the results obtained, it is demonstrated that there is consistency in estimates. As such it can be concluded from these results, based on the minimal variation in the estimates that the amount of solar PV embedded in the metered net demand from Stellenbosch was in the range of 13 to 15kWp. It can be further inferred that, given these results, it is possible to use adjusted historical demand obtained from an area in a different LSM class in this estimation process.

5.8 Results for case study 4B

The objective of this case is to determine the performance of the proposed method when measured net demand is used with historical demand data from different LSM classes. Historical demand data for this case was sourced from Gasese and West Ridge. Demographics from Gasese show that this area is a low-income area and there is hardly any proliferation of electrical equipment. West Ridge on the other hand is a middle-income area. Accordingly, West Ridge has a moderate proliferation of electric equipment. Other demographic patterns used in the selection of these two areas are shown in the appendix section. These two areas exhibit very different demographic patterns from the areas selected in Stellenbosch, from which the net demand was metered.

Estimations of the installed solar PV in this section shows that the amount of solar PV embedded in the measured net demand was 1kWp. This is an indicator that it is not possible to use historical demand from different LSM communities in the estimation process. Further, drawing from the results obtained using simulated demand, it was observed that the sensitivity error widens when comparing LSM class 10 and LSM class 1. This error is up to 70kWp.The error reduces as the gap between the LSM classes from which demand data sets are drawn from narrows. As such, it is not possible to provide an estimation in cases where the amount of solar PV embedded in the net demand where the net demand and the historical demand used in the estimation process are not drawn from the same LSM class.

5.9 Conclusions

Table 12 summarizes the results obtained from the estimation process using measured demand. Case 3 is the ideal case, where the validating demand and the net demand are derived from the same area. Case 4A shows results obtained using historical demand drawn from areas found in LSM 10 while case 4B shows the estimation using results obtained using historical demand from LSM 5 and LSM 1 respectively.

Table 12: Estimation of installed solar PV using net demand data from Stellenbosch and validated by different LSM classes

Case	Stellenbosch (Case 3)	Moreletta (Case 4A)	Welgemoed (Case 4A)	West Ridge (LSM 5 Case 4B)	Gasese (LSM 1 Case 4B)
LSM Class	LSM10 High	LSM 10 High	LSM10 High	LSM 5	LSM1
Estimation (kWp)	14	15	13	1	1

Figure 35 compares predicted net demand and measured net demand for the four tested cases. The first two subplots show the results obtained in case study 4A (same LSM class as the net demand). The subsequent subplots show the results obtained in case 4B where demand data was derived from Gasese and West Ridge areas (LSM 1 and LSM 5).

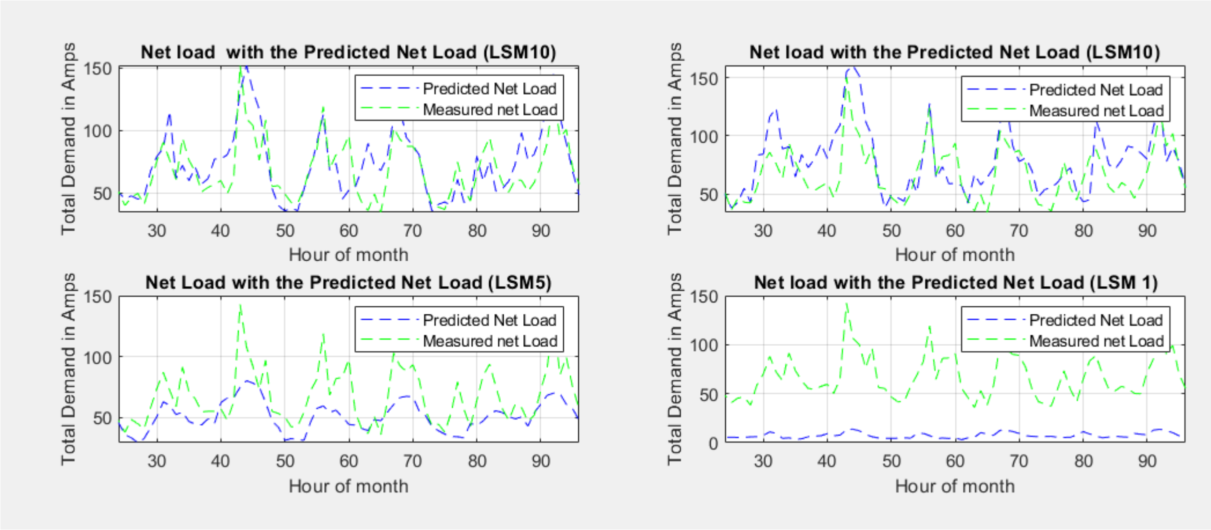


Figure 35: Comparing the measured net demand with predicted net demand

These results show a close relationship in patterns of average demand profiles of areas the same LSM cluster. It therefore follows that the accuracy of estimations using historical demand data is related to the LSM class from which the historical demand data is drawn. Communities found in the same LSM class are likely to produce accurate results using this method.

5.10 Estimation error analysis.

This section carries out analysis of results obtained from case studies 4A and 4B in order to determine the accuracy of installed solar PV predictions. To do this, the metered net demand and the predicted average net demand were compared to determine their correlation coefficient. The correlation coefficient was used to establish the precision of the methodology adopted. Additionally, the standard error of estimation was evaluated to determine the accuracy of the estimation. The standard deviation and the mean for both the predicted and measured net demand were evaluated. Further, the standard error was expressed as percentage of the peak demand in order to represent the relative error of estimation in each case study. Finally, conclusions were drawn based on the results.

5.10.1 Error analysis for similar communities

Firstly, a time series plot covering several hours was created for each set of net demand data in their respective areas. Figure 36 below shows the results obtained from the estimated mean net demand from Moreletta and the net demand obtained from Stellenbosch.

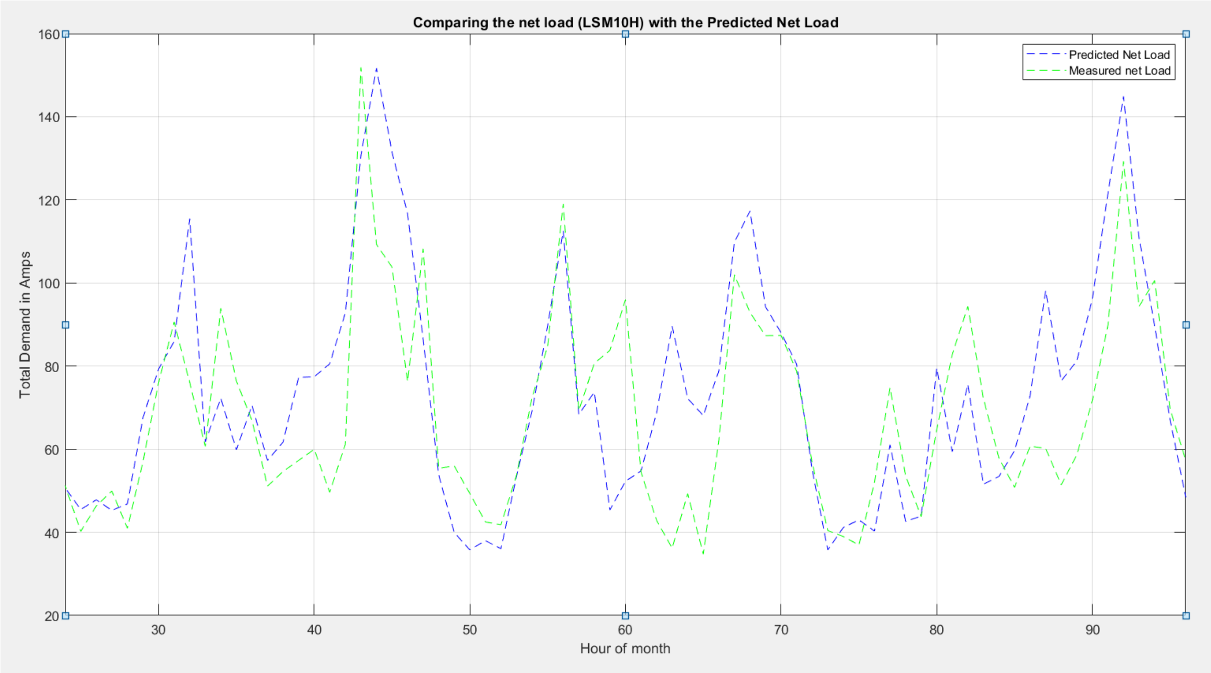


Figure 36: Segment comparing the net demand and the predicted net demand for Moreletta and Stellenbosch respectively in the month of May

From figure 36, it can be seen that the magnitude of the measured net demand and that of the predicted net demand exhibit small variations. This may be attributed to the fact that these two areas are classified under the same LSM class and they have close demand patterns. Accordingly, their net demand profiles are likely to be similar.

Figure 37 shows a scatter plot of the predicted net demand and the measured net demand as an indicator of the method's precision and accuracy. The predicted net demand is determined as a function of the measured net demand. The standard deviation and the mean of the two sets of net demand were calculated and compared. Furthermore, the coefficient of determination was calculated in order to determine if there exists a good linear relationship between the measured net demand and the predicted net demand.

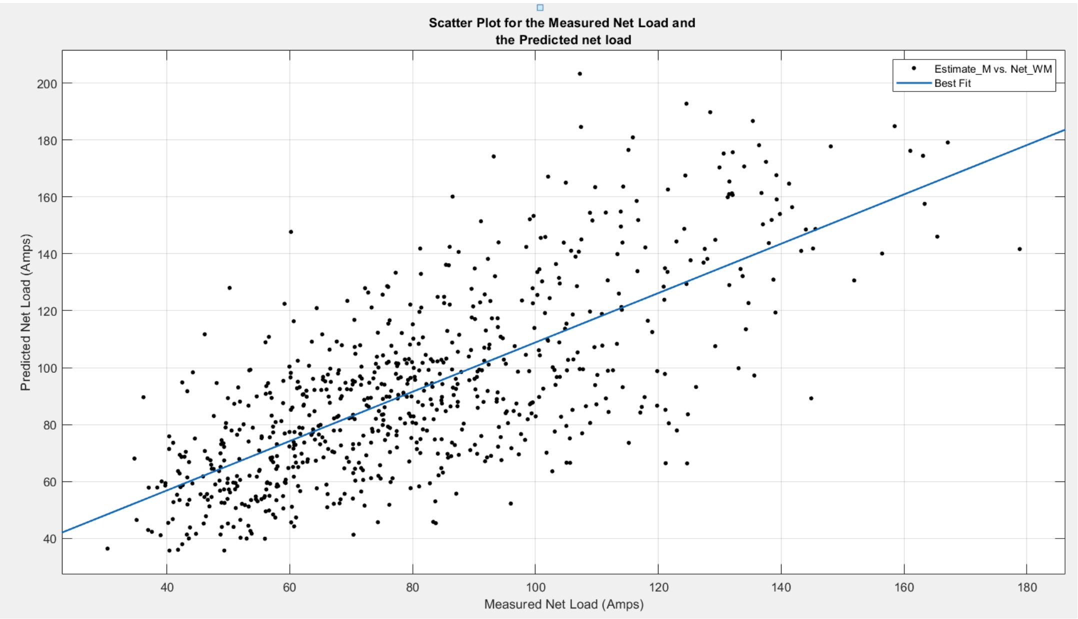


Figure 37: Scatter plot for the predicted net demand and the measured net demand (Moreletta Park versus Stellenbosch)

Table 13 shows the calculated coefficient of determination, the RMSE, the standard deviation and the means between both sets of demand data.

Table 13: Results for case study 4 with Moreletta's demand

Peak Net Demand	R-square	RMSE	% of Peak Demand	Estimated kWp	Mean (Measured net)	Mean (Predicted net)	Std(m-net)	Std (p-net)
178.84A	0.5196	22.21 A	12.42	15 kWp	80.47A	91.96A	26.65A	32.026A

From the Table 13 above, it can be deduced that there is an error in the predicted values. The mean of the predicted demand varies by 14% from the measured net demand, while the variation in the standard deviation is about 20%. This is probably due to the randomness of the process. Equally, this can be attributed to the sample size. The number of customers used in this study was restricted by the number of customers metered in the Stellenbosch. It is expected that the error of prediction becomes less as the number of customers increase. Additionally, it is expected that the larger the sample of customers, the closer the prediction is to the average demand of the feeder and hence the lower the observed error. The standard estimation error is determined to be 22.21A. This is about 12.42% of the peak demand of the metered net demand.

A second time series plot was created, as seen in Figure 38, to show the variation between the predicted net demand obtained using Welgemoed historical demand (LSM 10 - high) and that of the measured net demand obtained by measuring Stellenbosch. The plots demonstrate a likely relationship between the predicted net demand and the measured net demand. The demand has little variation, and this is probably because the two data sets are drawn from the same LSM class.

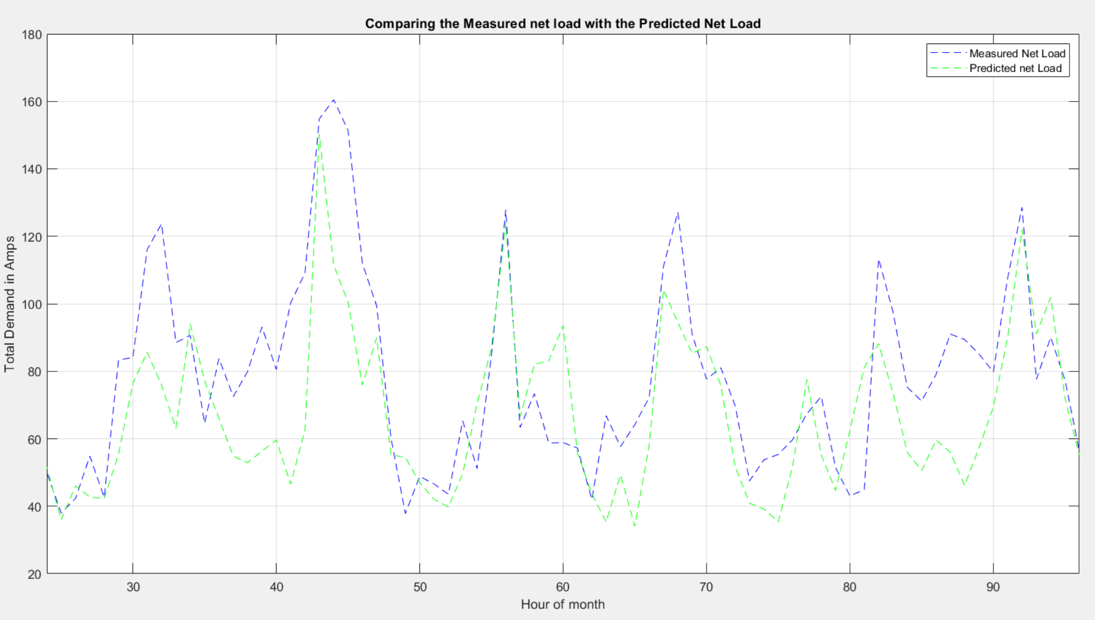


Figure 38: Comparing the actual mean net demand from Stellenbosch and the estimated net demand from Welgemoed, tested on an 18 customer feeder.

To determine the precision and accuracy of estimation in this case, a similar approach was taken where a scatter plot was created and a goodness of fit line drawn, as seen in Figure 39 below. The root mean square error (RMSE) was determined. Additionally, the standard deviation and the mean for the predicted as well as the measured net demand was evaluated. The purpose of this evaluation is to see how close the predicted value is from the actual value using statistical parameters.

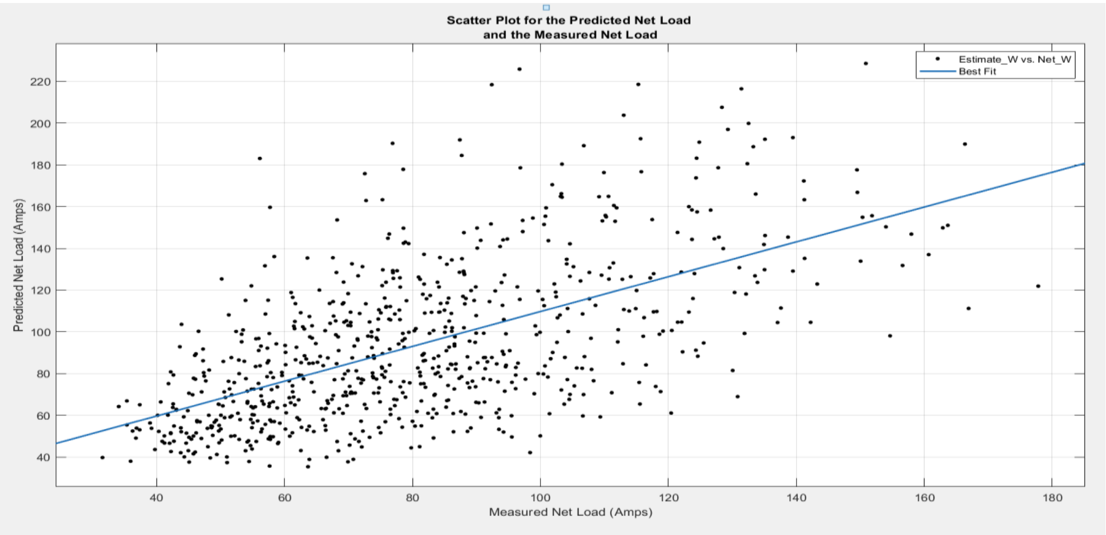


Figure 39: Scatter plot for the predicted net demand and the measured net demand (Welgemoed, Stellenbosch)

Table 14 shows the tested parameters and the results obtained from this case study.

Table 14: Results for case study 4 with Welgemoed demand

Measured Peak Demand[A]	R-square	RMSE [A]	% of Peak Demand	Estimated [kWp]	Mean (measured - net)	Mean (predicted -net)	Std (measured net)	Std (predicted net)
177.82	0.3626	29.67	16.7	12	80.39A	93.31A	26.85A	37.16A

From the results shown in the Table 14 above, the error in the predicted mean was calculated to be 16 % while the variation in the standard deviation was determined to about 38%. Further, the standard error as was calculated and found to be about 16.7%. The R-square value indicates that there is lack of linear relationship between the two sets of demand data. This is probably due to the stochastic approach used in this study.

Finally, to analyse the results obtained, net demand was obtained for different for the different tested areas for the each capacity of PV obtained. To investigate whether the variations in the distribution of the demand, PDF plots were created to compare the measured net demand and the predicted net demand as seen in Figure 40.

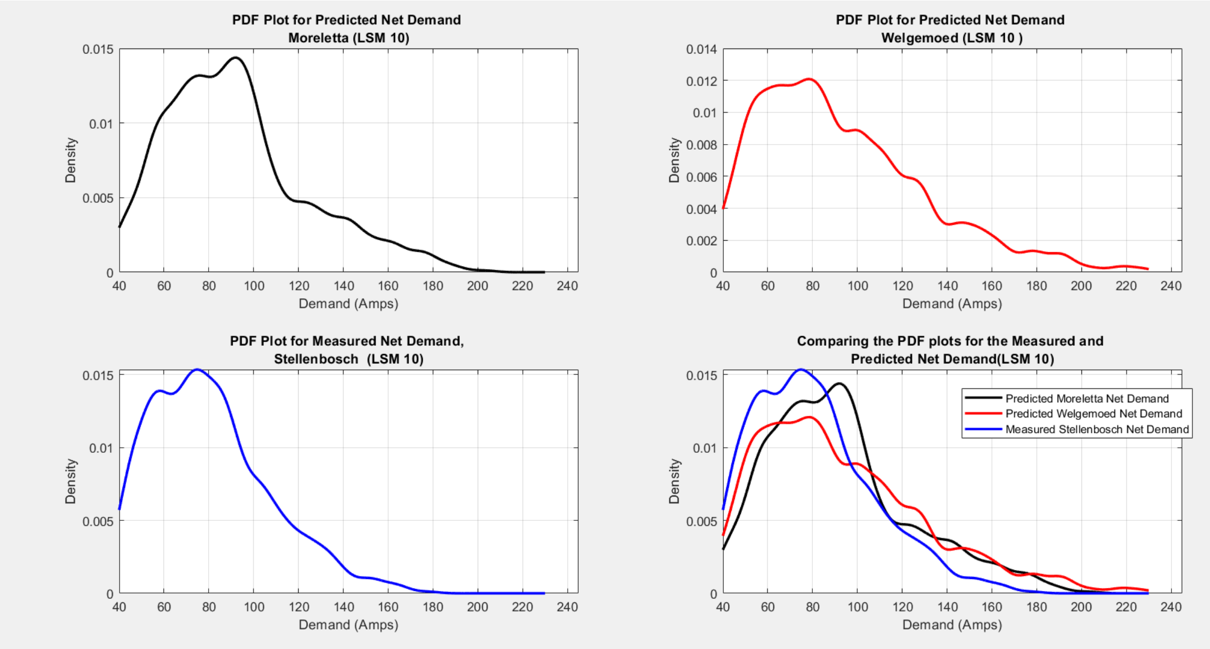


Figure 40: Probability distribution plots for the predicted demand and the measured net demand under the same LSM class

Based on the PDF plots in the Figure 40, it can be seen that the distributions exhibit a similar pattern with minor differences observed in terms of probability densities and the measure of spread. The discrepancies exhibited in the individual PDF plots are fairly tolerable and this is a likely pointer to good performance of the model using historical demand data drawn from a feeder with similar characteristics.

It is observed that the Welgemoed distribution has a larger spread than Moreletta Park. Consequently, it has a higher measure of standard deviation and the RMSE value as observed in the table. It is observed that the mean values of the predicted net demand are very close for both cases. This partly explains the differences in the estimation of embedded solar PV capacity using historical demand drawn from different areas in the same LSM class. It is noted that similar LSM classifications may show similarities in statistical parameters such as the mean, the peak demand and the standard deviation.

5.10.2 Error analysis for different communities

The last part of the error analysis evaluated the same parameters as those in the above exercises. The purpose was to analyse the errors using a different sets of demand data obtained from the case study 4B. Similarly, standard deviation and the mean of the predicted and measured net demand were calculated. The RMSE was determined and analysis carried out to determine the existing discrepancies between the measured net demand and the predicted net demand. The net demand from Stellenbosch in 2019 is compared with the estimated net demand from West Ridge and Gasese, which can be seen in Figure 41.

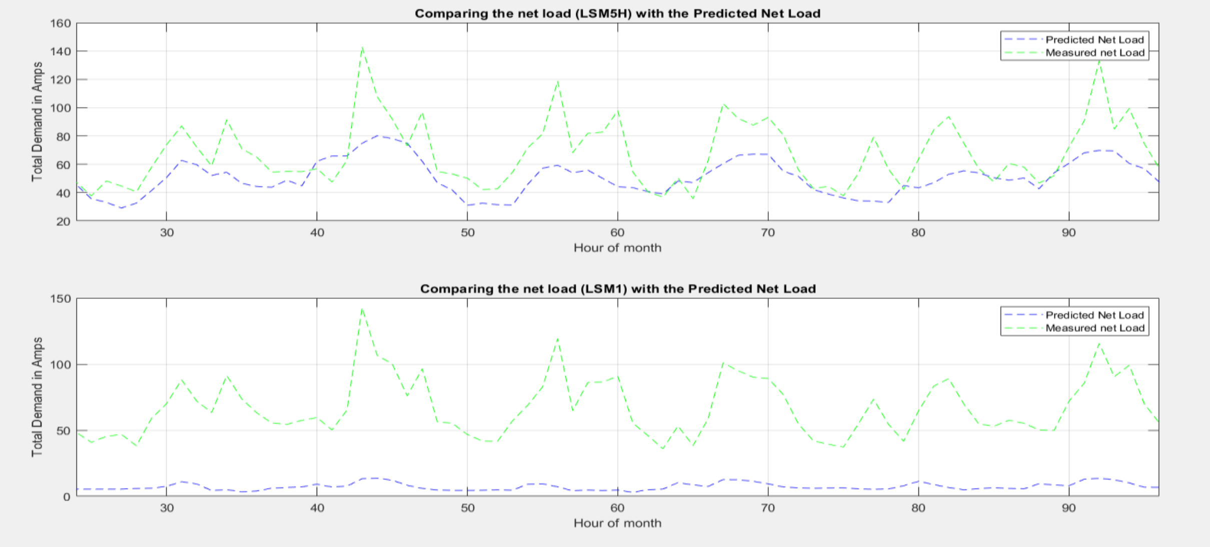


Figure 41: Comparing the actual mean demand and the estimated demand from West Ridge (top) and Gasese (bottom)

From the subplots in the Figure 41, it is evident that there exist large variations between the measured net demand and the predicted net demand in case study 4B. This is probably because the areas from which the historical demand was derived, used in generating the predicted demand, are not in the same LSM class as the area within which the net demand is metered. As such the closer the LSM class of the areas, the better the estimation. Additionally, it is evident that the closer the demand of two areas is, the more likely that they can be used to produce more accurate results.

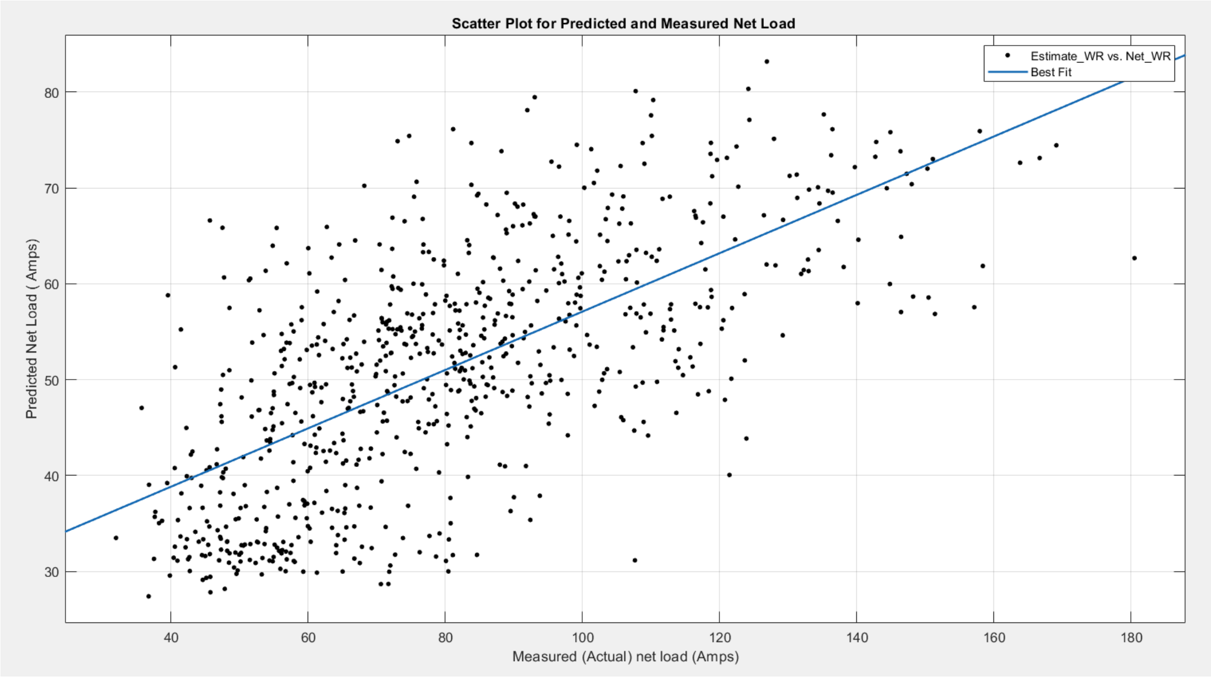


Figure 42: Scatter plot for Stellenbosch (measured net demand) and West Ridge (predicted net demand)

Table 15 compares various statistical estimates obtained through the analysis of net demand for both West Ridge and Gasese. The R-square is obtained by comparing the net demand from Stellenbosch and the predicted net demand for each of the cases.

Table 15: Test Results for case study 4 with West ridge & Gasese demand

Area	Peak Demand	R-square	RMSE	% of Peak Demand	Estimated kWp	Std (Measured net)	Std (Predicted net)	Mean (Measured net)	Mean (Predicted net)
West R	180.49	0.4328	9.36	5.19	1 kWp	26.82A	12.42A	80.49A	51,17A
Gasese	171.53	0.1832	2.7183	1.60	1 kWp	26.00A	3.01A	80.33A	6.20A

From Table 15 above, it can be observed that the standard deviation of the predicted net demand from West Ridge and Gasese are smaller than the measured net demand by a considerable margin. There is a wide deviation between the mean demand values obtained from West Ridge and the net demand measured from Stellenbosch. This deviation widens as the difference between the LSM classes widen, with results showing an increasing margin between the standard deviations of the net demand data and the predicted net demand data.

This is a likely indicator that the accuracy of prediction reduces drastically as the difference between the LSM class of the historical demand data and the measured net demand data widens. This may be the explanation as to why the estimated capacity of the embedded solar PV in the net demand using historical demand in case 4B returns a negligible answer. The Figure 43 shows the PDF plot for the predicted net demand as well as the measured net demand.

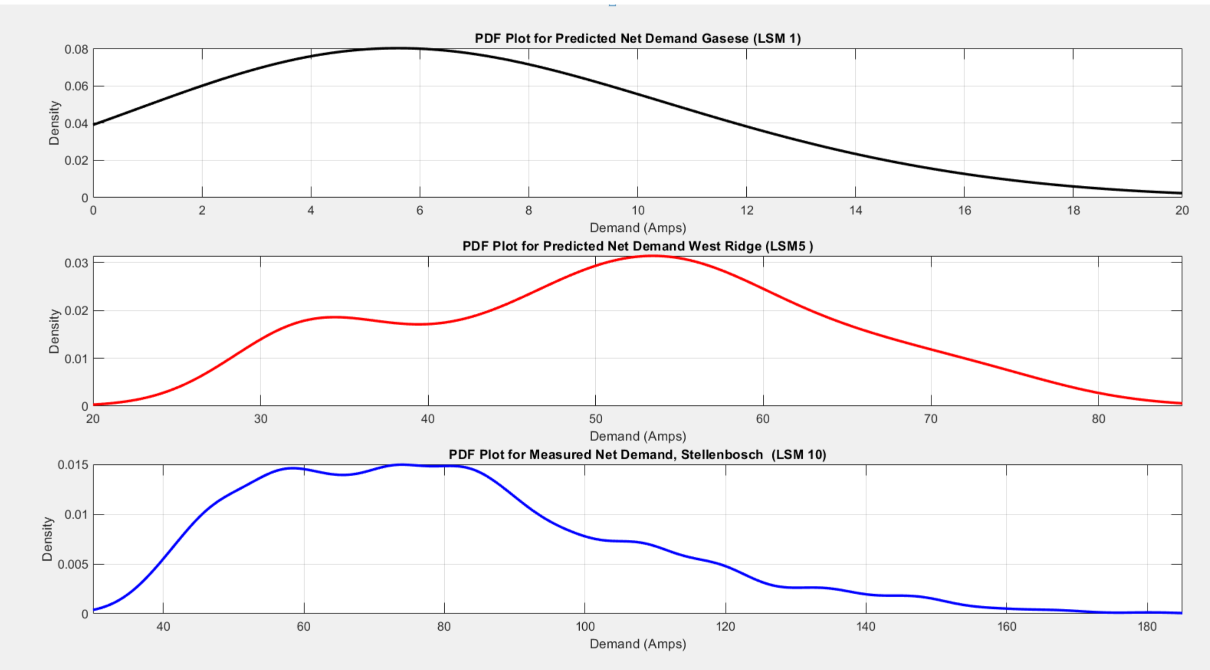


Figure 43: PDF plots comparing the predicted net demand for Gasese and West Ridge vs the actual net demand measured in Stellenbosch

From the above plots, it is observed that there is a large discrepancy in shape of the PDF plots of the three sets of demand. The metered net demand has a large range of values with a mean current demand of about 80 Amps. The predicted net demand from LSM class 5 has a mean demand of about 55 Amps with a range of 80 Amps. The predicted net demand from LSM class 1 has a mean of about 6 Amps and a range of about 20 Amps.

To emphasize the existing differences in the predicted net demand obtained from the different LSM classes, histogram fits were plotted to compare the measured net demand and the predicted net demand as seen in Figure 44 below.

It is obvious that there is a significant variation in the variance, the mean and the density of the different predicted net demand and that of the measured demand. This may partly explain the discrepancies in the results obtained when using demand data drawn from different LSM classes.

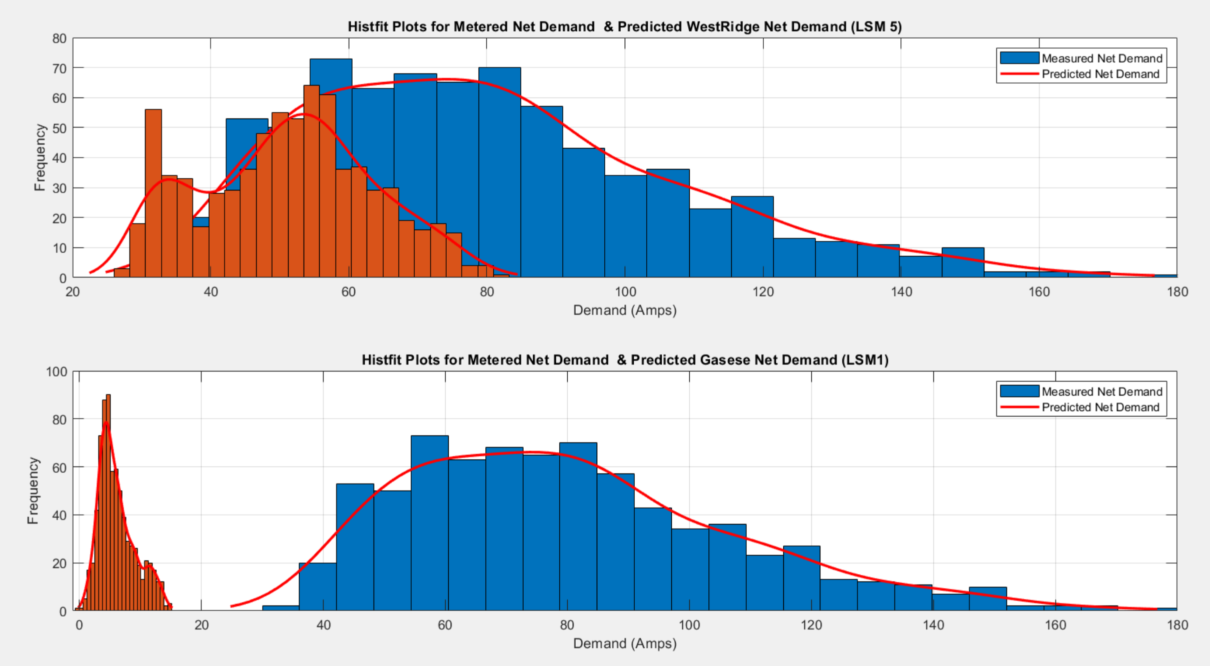


Figure 44: Distribution plots showing the distribution of the predicted net demand estimates for load under different LSM classes with the measured net demand

5.10.3 Conclusions from the error analysis

It may be concluded that, in absence of historical demand from a feeder, it is possible to estimate the capacity of embedded solar PV in the net demand using historical demand obtained from another area which falls under the same LSM class as the area of the feeder. This is because areas classified under the same LSM class exhibit similar demand characteristics in terms of the average demand, have relatively close standard deviations and they have a relatively small range in terms of their maximum and minimum values.

5.10.4 Constraints for case study 4

The main constraint of case study 4 is related to the choice of the historical demand used the estimation process. The case study shows that there is need to choose historical demand from an appropriate LSM class in order to accurately estimate the embedded solar PV

6 Conclusions

The objective of this research was to develop a method for estimating the capacity of installed solar PV SSEGs on a LV feeder by using aggregated historical demand, the measured net demand and solar irradiance data. The method was implemented using a Monte Carlo probabilistic approach and was tested for accuracy under various conditions. This chapter provides the summary of the relevant findings made in this research based on the hypothesis, research objectives as well as the research questions.

Summary of the findings

The objective of this study was to test the hypothesis which was stated as follows:

"The total capacity of solar PV SSEGs embedded in the aggregated net demand of a residential LV feeder can be estimated by comparing this net demand profile with historical demand from a similar customer class using a stochastic technique."

The research questions posed at the beginning of the study were answered through literature review and also through the development and testing of the proposed method. By answering the research questions, the findings and conclusions made by this research are discussed herein.

- ***What are the existing methods and their limitations?***

Chapter 2 reviewed the existing literature on the estimation of installed PV capacity on distribution networks. The review established that several methods have been proposed to estimate the amount and capacity of installed PV SSEGs on a network. Two main approaches to estimation were identified. These are GIS based methods and disaggregation methods.

Most reported GIS methods are focussed on determining the total number of solar PV panels installed in an area through remote sensing and computer vision algorithms. For this specific application, the accuracy of GIS based methods is acceptable. However, the use of the methods for the estimation of the capacity of solar PV embedded in the demand is limited. This is because it is very difficult to determine, from satellite imagery, the capacity of PV systems, let alone the contribution of each solar PV system to the demand. Furthermore, one cannot determine whether the identified panels are operated off-grid or grid tied and the utility company and grid operator are more concerned with the total capacity of grid-tied systems. Furthermore, GIS based methods do not provide any information on the demand of the area under study. Without information on this demand, the potential impacts of the installed PV SSEGs cannot be accurately determined. GIS based methods are therefore not able to provide sufficient details for an accurate estimation of installed PV capacity. Nonetheless, this method is important in applications where the number of solar PV systems in an area is relevant.

Secondly, most of the disaggregation models that have been reported use techniques such as linear regression, time series decomposition, machine learning methods among other methods to estimate the amount of solar PV embedded in the net demand. For this specific application, disaggregation methods are acceptable. Different methodologies used in the disaggregation process have different limitations, including complexity, requiring data that is difficult to obtain, and not accounting for demand and solar PV uncertainties. In some of the methodologies, decomposition models and machine learning models use data reduction techniques which alter the characteristics of the data used. This is a limitation and may affect the accuracy of the estimation results.

- ***Can net demand measurements and historical demand be used to compute the total capacity of embedded solar PV SSEGs?***

It was demonstrated that by using a feeder’s historical demand and its net demand (present) measurements, coupled with GHI solar irradiation data which informs the potential solar PV output per installed unit capacity, it is possible to accurately estimate the capacity of solar PV SSEGs embedded in the net demand of a residential LV feeder. To do this, the concept of stochastic expansion of demand, where more load profiles are generated using a Monte Carlo probabilistic approach, was developed, tested and validated. To represent the diversity of the loads on a feeder, stochastic aggregation processes were carried out and the generated load profiles were compared with the actual profile for accuracy.

Additionally, to estimate the capacity of solar PV SSEGs embedded in measured net demand data using historical demand data, it was demonstrated that adjustments needed to be made to the historical demand in order to reflect the changes attributed to factors such as energy efficiency, solar water heaters and customer behaviour changes.

Furthermore, the relationship between the predicted net demand and the actual measured net demand was investigated. It was established that variations in statistical properties of the predicted and actual measured net demand was minimal for the areas classified under the same LSM class. The maximum error in the mean values was estimated to be about 16% while the standard deviation was recorded as 38%.

From the testing and validation process, it can be concluded that is possible to use historical demand of an area to determine the capacity of solar PV embedded in the net demand of the area. Similarly, it was demonstrated that it is possible to estimate the embedded capacity of solar PV SSEGs in the net demand of one feeder using adjusted historical demand of another feeder, if the two feeders are in areas classified under the same LSM class.

- ***Can LSM classifications and the related models be used as historical data in the estimation of the capacity of solar PV SSEGs? Does the approach require feeder-specific data?***

The investigations conducted in Chapter 5 demonstrated that in cases where only net demand data is available, historical demand data obtained from another feeder, in an area classified under the same LSM, can be used to estimate the amount of solar PV SSEGs embedded in the net demand, with good accuracy. This therefore means that the selection of the alternative historical demand data must be based on a feeder that has customers with demand characteristics similar to those from which the net demand is obtained. Where the load characteristics differ significantly, as is expected with customers in different LSM classes, the margin of error is significant.

- ***How can the effects of other factors such as energy efficiency (EE), solar water heaters (SWHs) and tariff changes be accounted for?***

Comparing the average demand profile plots for 2002 and 2019, for demand without solar PV, for customers in the same LSM class, it was observed that 2002 in general has a higher demand than 2019. For this reason there was a need to adjust the 2002 demand to cater for unknown energy factors such as energy efficiency, solar water heating, changes in customer behaviour and preference changes. To consider these elements, a scaling factor is introduced to approximate the impact of these factors and was subsequently used to adjust the historical demand data. The scaling factor was found by determining the ratio of the average daily demand of 2019, measured without the impacts of solar PV and the average of 2002, measured at a time where PV is zero. This ration was then replicated to the respective hours of the year to create an adjusted set of demand data. By applying the scaling factor, it was possible to estimate the embedded PV capacity with minimal error. It was further demonstrated that the use of unadjusted demand consistently overestimates the embedded capacity. The methodology proposed therefore provides a procedure through which the impacts of other unknown energy factors can be assessed.

Based the answers found to the research questions stated in Chapter 1 of this thesis, it can be concluded that the thesis' hypothesis has been validated, within the constraints of the following limitations of the method specifically, and the study in general:

- There is need to have prior knowledge of the classification of the areas under assessment which is not always possible. Otherwise, it is difficult to determine which feeder data will lead to acceptable accuracy without detailed load research.

- The method used in the modelling the impact of *other energy factors* is not very comprehensive as it does not take into account the variation in demand as far as the difference in consumption on a weekday and that of a weekend is concerned. This method needs to be refined in order to reflect all these variants.

- The validation as presented in case study 3 requires precise knowledge of the actual installed solar PV SSEGs capacity within the net demand, which in the case of this study was not readily available.

Summary of contributions

The following are the key contributions of this research.

i) A statistical approach was developed to estimate the amount of solar PV SSEGs installed on a residential LV network by analysing and comparing measured aggregated demand data to the historic demand data of that feeder or from another area in the same LSM class.

ii) The method proposes a new approach on how to estimate the solar PV SSEGs while accommodating other energy factors such as changes in tariffs, changes in energy use patterns, energy efficiency among other factors. While this method offers an approximation, it is noted that this approach has not been reported elsewhere in the estimation methodologies.

iii) The proposed method allows for the development of other applications such as the analysis of demand changes on a feeder over the years using aggregated data.

iv) The developed method can be used to give an indication of the capacity of illegal connections and unapproved extensions.

v) Given that in most cases, LV networks are not monitored and are configured to only reflect the aggregated net demand, it is not possible for the utility operators to estimate the amount of solar PV generation present in the demand as both is metered as net demand. This method may be applied in assessing the amount of solar PV generation embedded in the net demand.

Future research

Further research can be conducted to investigate the impact of other factors such as energy efficiency, changes in energy tariffs, consumer preference among other factors that affect the demand profile. Such an investigation would determine how these factors vary among different LSM classifications with time. In this regard, it would be important to test the performance of the proposed method using a different error adjustment approach.

It would be an informative to conduct additional research geared at evaluating the performance of this method using demand data at a different resolution. This study used hourly resolved demand data to test and validate the method.

7 References

Ali, I., Shafiullah, G.M. & Urmee, T. 2018. A preliminary feasibility of roof-mounted solar PV systems in the Maldives. *Renewable and Sustainable Energy Reviews*. 83(May 2017):18–32.

Arjun, D., Rody, K., Cassidee, K. & Aaron, N. 2017. Automated Rooftop Solar PV Detection and Power Estimation through Remote Sen. *Duke Energy Initiative*. [Online], Available: https://bassconnections.duke.edu/sites/bassconnections.duke.edu/files/documents/automated-rooftop-solar-pv-detection.pdf [2019, September 11].

Australian PV Institute. 2018. *Australian PV market*. [Online], Available: http://pv-map.apvi.org.au/analyses.

Ballack, C. 2017. *PQRS Shared Data*. Shared with the author by the pubilsher on 09/04/2018. Available upon request from the publisher.

Ballack, C. 2019. *South Africa Solar PV Update 2019 – AREP*. [Online], Available: https://arepenergy.co.za/south-africa-solar-pv-update-2019/ [2019, July 04].

Baron, M. 2007. *Probability and Statistics for Computer Scientists*. New York: Chapman & Hall.

Bellini, E. 2019. *Global cumulative PV capacity tops 480 GW, IRENA says – pv magazine International*. [Online], Available: https://www.pv-magazine.com/2019/04/02/global-cumulative-pv-capacity-tops-480-gw-irena-says/ [2019, September 13].

Bergamasco, L. & Asinari, P. 2011. Scalable methodology for the photovoltaic solar energy potential assessment based on available roof surface area: Application to Piedmont Region (Italy). *Solar Energy*. 85(5):1041–1055.

Boyce, S.J. & Well, A. 1990. Understanding the Effects of Sample Size on the Variability of the Mean. *Organizational Behaviour and Human Decision Processes*. 312(47):289–312.

Bradbury, K., Pratt, L.C., Nicholas, T.J., Pratt, J.M. & Nicholas, R.N. 2015. Solar Power Estimation Through Remote Sensing. *Duke Energy Initiative*. 2007.

Bradbury, K., Pratt, L.C., Nicholas, T.J. & Nicholas, R.N. 2016. Automated Rooftop Solar PV Detection and Power Estimation through Remote Sensing. *Duke Energy Initiative*. 31(6):2016.

Brooks, M.. 2015. “SAURAN: A new resource for solar radiometric data in southern Africa”. *Journal of Energy in southern Africa*. 26:2–10.

California Independent System Operator. 2012. What the duck curve tells us about managing a green grid. *California ISO, Shaping a Renewed Future*. Fact Sheet:1–4.

Clark, H. 2018. *Solar Photovoltaics Deployment in the UK, March 2018*. London. [Online], Available: https://www.gov.uk/government/statistics/monthly-small-scale-renewable-deployment%09%09%09%09%09%09%09%09%09%09%0A.

Department of Agriculture Western Cape. 2018. *The Mapping of Agricultural Commodities and Infrastructure in the Western Cape*. Cape town: Available at Department of Agriculture, Elsenburg on request.

Devarajan, A., Manor, S., Perkins, J., Saboo, R. & Zhang, W. 2018. Solar Power Estimation Through Remote Sensing. *Duke University Energy Initiative*. [Online], Available: https://bigdata.duke.edu/sites/bigdata.duke.edu/files/images/SolarPoster_sm.pdf [2019, September 11].

Dyusha, V. 2019. *Stellenbosch Municipality Solar PV customers. Internal Data*. Stellenbosch: Accessed with permission from Stellellenbosch Municipality Electricity Department.

Ebrahim, A. & Mohammed, O. 2018. Pre-processing of Energy Demand Disaggregation Based Data Mining Techniques for Household Load Demand Forecasting. *Inventions-MDPI*. 1–16.

German Development Agency. 2016. The German solar rooftop experience – Applicability in the Indian context Indo-German Development Cooperation.

Glasserman, P. 2003. Monte Carlo Methods in Financial Engineering (Stochastic Modelling and Applied Probability). 616.

Graziano, M. & Gillingham, K. 2015. Spatial patterns of solar photovoltaic system adoption. *Journal of economic geography*. 15(4):815–839. [Online], Available: http://www.econis.eu/PPNSET?PPN=835395405.

Guwaeder, A. & Ramakumar, R. 2018. Statistical Analysis of PV Insolation Data Statistical Analysis of PV Insolation Data. (February).

Howlader, H.O.R., Sediqi, M.M., Ibrahimi, A.M. & Senjyu, T. 2018. Optimal Thermal Unit Commitment for Solving Duck Curve Problem by Introducing CSP, PSH and Demand Response. *IEEE Access*. 3536(c):1–1.

IEA. 2017. World Energy Outlook 2017. *International Energy Agency*.

Imanishi, T., Yoshida, M., Wijekoon, J. & Nishi, H. 2017. Time-series decomposition of power demand data to extract uncertain features. *IEEE International Symposium on Industrial Electronics*. 1535–1540.

International Energy Agency. 2019. *Solar*. [Online], Available: https://www.iea.org/topics/renewables/solar/ [2019, September 13].

Jaeger-Valdau, A. 2017. *PV Status Report 2017*. Ispra, Italy.

Jenkins, N., Long, C. & Wu, J. 2015. An Overview of the Smart Grid in Great Britain. *Engineering*. 1(4):413–421.

Kara, E.C., Tabone, M., Roberts, C., Kiliccote, S. & Stewart, E.M. 2016. Estimating Behind-the-meter Solar Generation with Existing Measurement Infrastructure. In New York, New York, USA: ACM Press *Proceedings of the 3rd ACM International Conference on Systems for Energy-Efficient Built Environments - BuildSys '16*. 259–260.

Kara, E.C., Roberts, C.M., Tabone, M., Alvarez, L., Callaway, D.S. & Stewart, E.M. 2018. Disaggregating solar generation from feeder-level measurements. *Sustainable Energy,* *Grids and Networks*. 13:112–121.

Kaur, A., Nonnenmacher, L. & Coimbra, C.F.M. 2016. Net load forecasting for high renewable energy penetration grids. *Energy*. 114:1073–1084.

Keith Bowen. 2017. Eskom Electricity Demand Data. *Eskom Demand Data, 2012-2017*.

Kironji, E. 2010. Measuring Quality of Life in South Africa: A Household-Based Development Index Approach. University of Pretoria. [Online], Available: https://repository.up.ac.za/bitstream/handle/2263/25060/Complete.pdf?sequence=7&isAllowed=y [2019, September 02].

Kloibhofer, S., Stifter, M., Leimgruber, F. & Rao, B.-V. 2017. Comparing and improving residential demand forecast by disaggregation of load and PV generation. *CIRED - Open Access Proceedings Journal*. 2017(1):1638–1641.

Kolda, T.G., Lewis, R.M. & Torczon, V. 2003. Optimization by direct search: New perspectives on some classical and modern methods. *SIAM Review*. 45(3):385–482.

Korsten, N. 2016. An investigation into the financial impact of residential Rooftop PV on Stellenbosch Municipality. Stellenbosch University.

Korsten, N., Brent, A.C., Sebitosi, A.B. & Kritzinger, K. 2017. The impact of residential rooftop solar PV on municipal finances: An analysis of Stellenbosch. *Journal of Energy in Southern Africa*. 28(2):29–39.

Kuppannagari, S.R., Kannan, R. & Prasanna, V.K. 2017. Optimal Net-Load Balancing in Smart Grids with High PV Penetration.

Lave, M., Kleissl, J. & Stein, J.S. 2013. A Wavelet-Based Variability Model (WVM) for Solar PV Power Plants. *IEEE Transactions on Sustainable Energy*. 4(2):501–509.

Lew, D. & Miller, N. 2017. Reaching new solar heights: integrating high penetrations of PV into the power system. *IET Renewable Power Generation*. 11(1):20–26.

Lewis, R.M., Torczon, V., Trosset, M.W. & William, C. 2000. Direct Search Methods : Then and Now. *Science*. 124(1–2):191–207.

Liu, T., Tan, X., Sun, B., Wu, Y., Guan, X. & Tsang, D.H.K. 2016. Energy management of cooperative microgrids with P2P energy sharing in distribution networks. *2015 IEEE International Conference on Smart Grid Communications, SmartGridComm 2015*. 410–415.

Lopes, J.A.P., Hatziargyriou, N., Mutale, J., Djapic, P. & Jenkins, N. 2007. Integrating distributed generation into electric power systems: A review of drivers, challenges and opportunities. *Electric Power Systems Research*. 77(9):1189–1203.

Malof, J.M., Rui Hou, Collins, L.M., Bradbury, K. & Newell, R. 2015. Automatic solar photovoltaic panel detection in satellite imagery. In IEEE *2015 International Conference on Renewable Energy Research and Applications (ICRERA)*. 1428–1431.

Margelou, S. 2015. Assessment of long term solar PV diffusion in Switzerland.

Masini, A. & Frankl, P. 2003. Forecasting the diffusion of photovoltaic systems in southern Europe. *Technological Forecasting and Social Change*. 70(1):39–65.

Mcmillan, R.W. & Kohlberg, I. 2018. A simple method for combining probability distribution functions relevant to radar and communications systems. *2017 IEEE International Conference on Microwaves, Antennas, Communications and Electronic Systems, COMCAS 2017*. 2017-Novem(4):1–5.

Mills, A., Ahlstrom, M., Brower, M., Ellis, A., George, R., Hoff, T., Kroposki, B., Lenox, C., et al. 2009. Understanding Variability and Uncertainty of Photovoltaics for Integration with the Electric Power System. *Electricity Journal*.

Mohan, R., Cheng, T., Gupta, A., Garud, V. & He, Y. 2014. Solar Energy Disaggregation using Whole-House Consumption Signals. *NILM Workshop 2014*. 1–4.

Muaafa, M., Adjali, I., Bean, P., Fuentes, R., Kimbrough, S.O. & Murphy, F.H. 2017. Can adoption of rooftop solar panels trigger a utility death spiral? A tale of two U.S. cities. *Energy Research and Social Science*. 34(May):154–162.

Ochi, M. 1976. *Applied Probability and Stochastic Processes in Engineering and Physical Sciences*. first ed. Florida: John Wiley & Sons.

Or, H., Howlader, R., Furukakoi, M., Matayoshi, H. & Senjyu, T. 2017. Duck Curve Problem Solving Strategies with Thermal Unit Commitment by Introducing Pumped Storage Hydroelectricity & Renewable Energy. (December):502–506.

Papoulis, A. & Pillai, S.-U. 1991. *Probabilities, Random Variables, and Stochastic Processes*.

Pengelly, J. 2002. Monte Carlo Methods.

Pérez Arriaga, I. & Knittel et al, C. 2016. *Utility of the Future. An MIT Energy Initiative response*. [Online], Available: energy.mit.edu/uof.

Pierro, M., De Felice, M., Maggioni, E., Moser, D., Perotto, A., Spada, F. & Cornaro, C. 2017. Data-driven upscaling methods for regional photovoltaic power estimation and forecast using satellite and numerical weather prediction data. *Solar Energy*. 158:1026–1038.

PQRS. 2016. *Demistifying the total installed PV capacity for South Africa*. [Online], Available: http://pqrs.co.za/data/demystifying-the-total-installed-pv-capacity-for-south-africa-nov-2016/.

PVsyst. 2019. *General description of the PVsyst software*. [Online], Available: https://www.pvsyst.com/help/general_descr.htm [2019, October 21].

Reddy, T.A. 2001. *Applied Data Analysis and Modeling for Energy Engineers and Scientists*. 1st Editio ed. Arizona: Springer.

Reinecke, O., Leonard, C., Kritzinger, K., Bekker, D.B., Niekerk, P.J.L. van & Thilo, J. 2013. Unlocking the Rooftop PV Market in South Africa. *Center for Renewable and Sustainable Energy Studies*. 27(March).

Roselund, C. 2018. *The duck curve comes to New England – pv magazine USA*. [Online], Available: https://pv-magazine-usa.com/2018/05/08/the-duck-curve-comes-to-new-england/ [2019, September 13].

Rowe, D., Sayeef, S. & Platt, G. 2016. Intermittency: It's the Short-Term That Matters. *Future of Utilities Utilities of the Future*. (January, 1):129–150.

Sandiford, M., Forcey, T., Pears, A. & McConnell, D. 2015. Five Years of Declining Annual

Consumption of Grid-Supplied Electricity in Eastern Australia: Causes and Consequences. *Electricity Journal*. 28(7):96–117.

Sayeef, S., Heslop, S., Cornforth, D., Moore, T., Percy, S., Ward, J.K., Berry, A. & Rowe, D. 2012. *Solar intermittency: Australia's clean energy challenge Characterising the effect of high penetration solar intermittency on Australian electricity networks*. [Online], Available: https://publications.csiro.au/rpr/download?pid=csiro:EP121914&dsid=DS1 [2019, April 04].

Sewchurran, S., Kalichuran, J. & Maphumulo, S. 2016. Drivers and application of small scale DG on municipal distribution networks. *65th AMEU Convention*. 65–75. [Online], Available: http://www.ee.co.za/wp-content/uploads/2016/11/AMEU-2016-pg-65-75.pdf.

Shaker, H., Zareipour, H. & Wood, D. 2015. A Data-Driven Approach for Estimating the Power Generation of Invisible Solar Sites. *IEEE Transactions on Smart Grid*. PP(99):1–11.

Shaker, H., Zareipour, H. & Wood, D. 2016. Estimating Power Generation of Invisible Solar Sites Using Publicly Available Data. *IEEE Transactions on Smart Grid*. 7(5):2456–2465.

Siddal, J. 1983. Monte Carlo Simulation 8.1. In First ed. New York: CRC Press *Probabilistic Engineering Design*. 1–16.

Sionshansi, F. 2016. *Future of Utilities – Utilities of the Future How Technological Innovations in distributed Energy Resources will Reshape the Electric Power Sector*. First Edit ed. London: Elsevier.

Snape, J.R. 2016. Spatial and temporal characteristics of PV adoption in the UK and their implications for the smart grid. *Energies*. 9(3):1–18.

SodaPro. 2019. *SoDa products - www.soda-pro.com*. [Online], Available: http://www.soda-pro.com/soda-products;jsessionid=743422AD4DFF279E377F5016E49583A9 [2019, April 21].

Solar Naturally. 2018. *Solar Power In Australia Statistics*. [Online], Available: https://www.solarnaturally.com.au/solar-power-in-australia-statistics/.

Sossan, F., Nespoli, L., Medici, V. & Paolone, M. 2018. Unsupervised Disaggregation of PhotoVoltaic Production from Composite Power Flow Measurements of Heterogeneous Prosumers. *IEEE Transactions on Industrial Informatics*. 14(9):1–10.

Terejanu, G.A. 2009. *Tutorial on Monte Carlo Techniques*. New York.

University of Cape Town. 2018. *Domestic Load Research Project*. [Online], Available: http://energydata.uct.ac.za/th/organization/about/dlr [2018, October 17].

Vivian, S., Bader, A., Norah, A., Yazeed, A. & Mansour, A. 2016. A Predictive Model to ForeCast Customer Adoption of Rooftop Solar. *International Symposium 0n computational and business Intelligence*.

Vrettos, E., Kara, E.C., Stewart, E.M. & Roberts, C. 2019. Estimating PV power from aggregate power measurements within the distribution grid. *J. Renewable Sustainable Energy*. 11(2):23707.

Wang, W., Yu, N. & Johnson, R. 2017. A model for commercial adoption of photovoltaic systems in California. *Journal of Renewable and Sustainable Energy*. 9(2):1–15.

Wang, Y., Zhang, N., Chen, Q., Kirschen, D.S., Li, P. & Xia, Q. 2018. Data-Driven Probabilistic Net Load Forecasting With High Penetration of Behind-the-Meter PV. *IEEE Transactions on Power Systems*. 33(3):3255–3264.

Waswa, L.S. & Bekker, B. 2018. Impact of PV Small Scale Embedded Generation on South Africa's System Demand Profile. In Durban: SASEC *South African Solar Energy Conference*. [Online], Available: https://www.sasec.org.za/full_papers/90.pdf [2019, April 23].

Yan, L. 2018. [Online], Available: http://web.stanford.edu/class/archive/cs/cs109/cs109.1188/lectureHandouts/LN12_indep_rv.pdf.

Zhang, H., Vorobeychik, Y., Letchford, J. & Lakkaraju, K. 2014. Predicting Rooftop Solar Adoption Using Agent-Based Modeling. *AAAI Fall Symposium*. 44–51.

Zhao, T., Zhou, Z., Zhang, Y., Ling, P. & Tian, Y. 2017. Spatio - temporal analysis and forecasting of distributed PV systems diffusion : A case study of Shanghai using a data - driven approach. *IEEE Access*. 5:5135–5148.

Zhou, N. 2018. Australia's solar power boom could almost double capacity in a year. *The Guardian*. 11 February. [Online], Available: https://www.theguardian.com/australia-news/2018/feb/11/australias-solar-power-boom-could-almost-double-capacity-in-a-year-analysts-say.

Zhu, H., Li, X., Sun, Q., Nie, L., Yao, J. & Zhao, G. 2016. A power prediction method for photovoltaic power plant based on wavelet decomposition and artificial neural networks. *Energies*. 9(1):1–15.

Appendix A Background study information

The background data used at the beginning of this study was retrieved from the Eskom national demand data (Keith Bowen, 2017). This was used in the preliminary study of the impact of the current distributed solar PV generations on the South Africa's demand profile (Waswa & Bekker, 2018). The results indicated that on such a high-level analysis of demand, the impact of distributed solar PV generations could not be seen through deterministic analysis. The study narrowed down to estimating the installed capacity on LV distribution networks as this is where most distributed generations are usually connected.

A.1 Measurement of consumer data

Two sets of data were collected and used in the modelling. The first source of data was the livewire repository from the municipality. Sets of data were obtained and they included data for consumption patterns for particular substations as well as some anonymized customers with prepaid meters; some among them were PV installers. However, data on a substation level was still on a high level of aggregation and the solar contribution would be lost in the noise of the consumption. It was decided that measurement be undertaken by the municipality guided by the distribution of PV SSEGs map obtained from the flyover study by the Department of Agriculture.

A.2 The Flyover study: A background

A section of Stellenbosch Municipality was selected based on the study that was conducted in Stellenbosch by the Western Cape Department of Agriculture. In this study, drones are used to fly over the Stellenbosch marking the arable land as well as identifying the households which have installed solar water heaters, swimming pool pumps and solar panels on their roofs.

In the study, it was estimated that there was about 545kWp installed capacity of solar rooftop PV in Stellenbosch, a figure that is almost 5 times what the municipal records indicated as of 2018. The flyover study formed a greater motivation for this study based on the implication that there was rampant installation of PV in the municipality.

The picture below shows the fly over area on which the study was conducted.

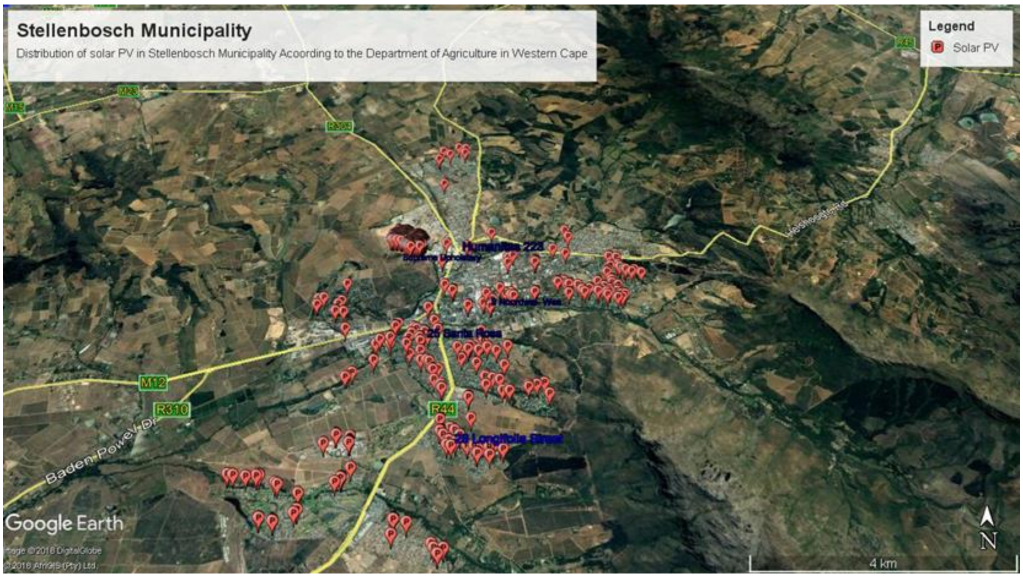

Stellenbosch Municipality
Distribution of solar PV in Stellenbosch Municipality According to the Department of Agriculture in Western Cape
Legend
Solar PV
M12
R310
R44
Baden Powell Dr
Humanitas 223
25 Santa Rosa
28 Longifolia Street
Google Earth
4 km
N

Figure 45: Recreated images for the drone surveillance by DoA incorporating substations.

A.3 Selection of the areas to meter

To select the areas to be metered, the data obtained from the department of Agriculture was used to locate areas which have a high PV potential. This, coupled with data from the municipality were used to recreate a map which showing the points of measurements. This included substations, mini-substations and eventually kiosks.

A GIS map was used to evaluate the possible areas that could be metered in Stellenbosch. Selection of areas to be metered was done based on:

i) High density of solar PV SSEGs in the area

ii) The classification of the area as a high income or wealthy area

iii) The classification of the area as a residential area.

A careful inspection reveals that high-income areas have a higher density of rooftop PV installed.

The figures 46 to 48 show the recreated maps that were used in the study. These maps were created by combining all the shape files obtained from GIS department of Stellenbosch municipality. This was done using QGIS software.

The combined map showing the location of solar PV, the closest mini substations as well as the kiosks. This is used to locate the areas for mapping.

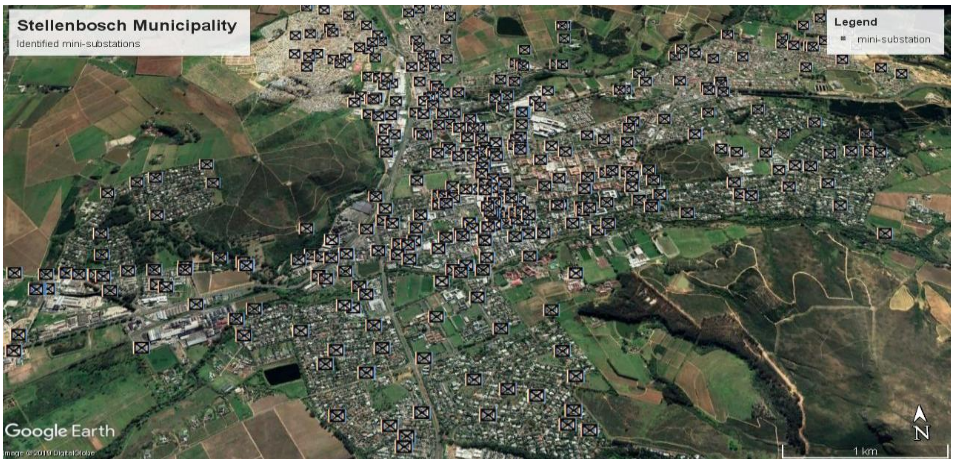

Stellenbosch Municipality
Identified mini-substations
Legend
▪ mini-substation
Google Earth
Image © 2019 DigitalGlobe
N
1 km

Figure 46: Recreated image showing the distribution of mini substations in Stellenbosch Municipality.

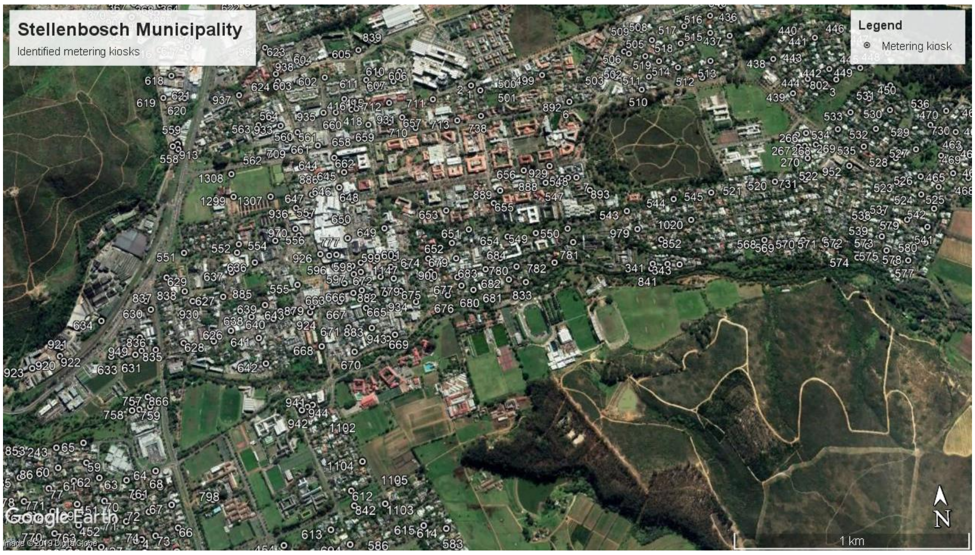


Figure 47: Recreated image showing the distribution of metering kiosks in Stellenbosch Municipality.

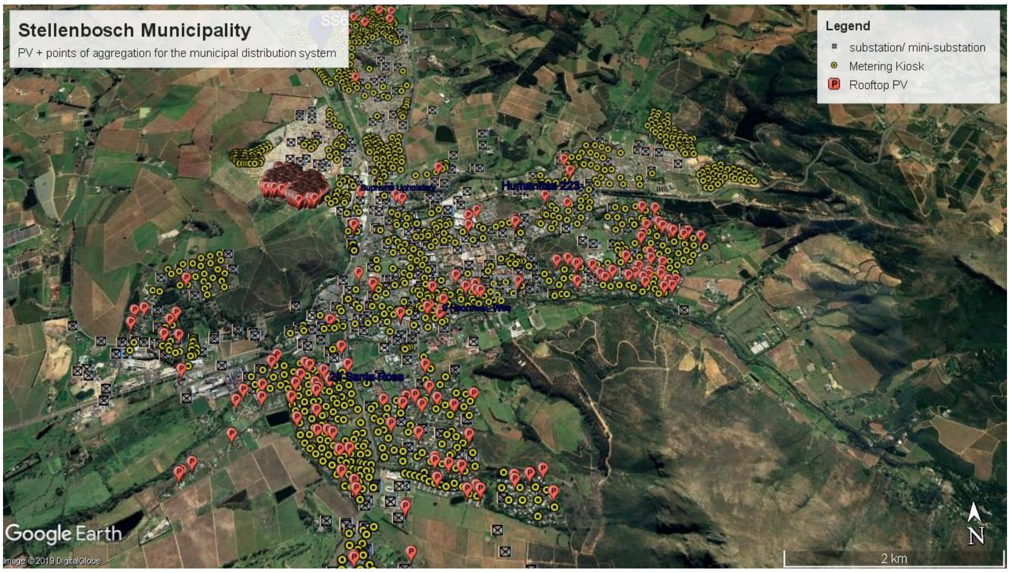

Stellenbosch Municipality
PV + points of aggregation for the municipal distribution system
Legend
substation/ mini-substation
Metering Kiosk
Rooftop PV
Google Earth
Image © 2019 DigitalGlobe
2 km
N

Figure 48: Combined map showing PV concentration and the points of Measurement.

Discarding the informal settlements, selected kiosks in several areas were to be measured. However, measurement of data was undertaken selected kiosks in three of these areas due to technical and resource constraints. The areas include:

i) Jonkershoek

ii) Paradyskloof

iii) Dalsig

Figure 49 below shows a sample spatial distribution of solar PV SSEGs, the mini-substations and the kiosks as captured on QGIS for Dalsig.

A.4 Selection of the kiosks

Kiosks to be metered were selected by through a physical survey of the area aided by street view option of Google Earth. The addresses of the kiosks closest to the houses with solar PV installations were identified for metering.

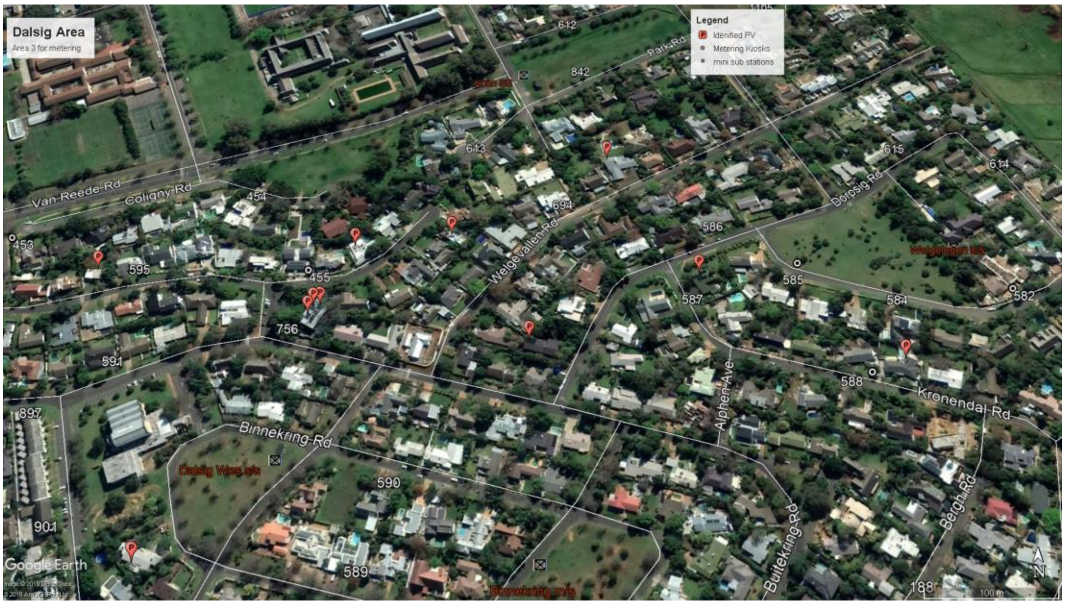


Figure 49: Areal View of Dalsig Area as captured in the DoA study (Department of Agriculture Western Cape, 2018). *Image reproduced courtesy of DoA, Western Cape.*

A.5 Metering

First, the first metering exercise was to be undertaken together with the municipality staff and using the municipality metering equipment. However, several challenges came up and the exercise had to be stopped. These constraints included the following:

i) Lack of adequate meters to measure about 30 selected customers that had been selected. The municipality uses the SL 7000 electricity meters series. This meant that they had to put at our disposal at least ten meters.

ii) There was an irregular distribution of the customers in the kiosks. This is in so far as the kiosks were concerned, there was a possibility of finding only one of the identified customers in one kiosks. This meant an increase in the number of meters used in the measurement.

iii) One meter could only measure three customers. Given the size of the meters, it was not possible to fit two meters in one kiosk and thus this would reduce the sample size.

iv) There was a challenge in so far as the retrieval of the measurements was concerned. This meter required disconnection and eventual removal of the SD card for the data to be read.

v) The installation of one of these meters required a lot of work and switching off of connected customers due to the need to connect the large current transformers.

vi) Some of the kiosks were either too small, too old or lacked a lock, which compromised the security of these expensive meters.

The following photos indicate how the set up for the metering exercise was done. Using this set up, 3 kiosks were measured with one kiosk being measured for individual customer data while the remaining two measuring the aggregated demand data for the two other kiosks.

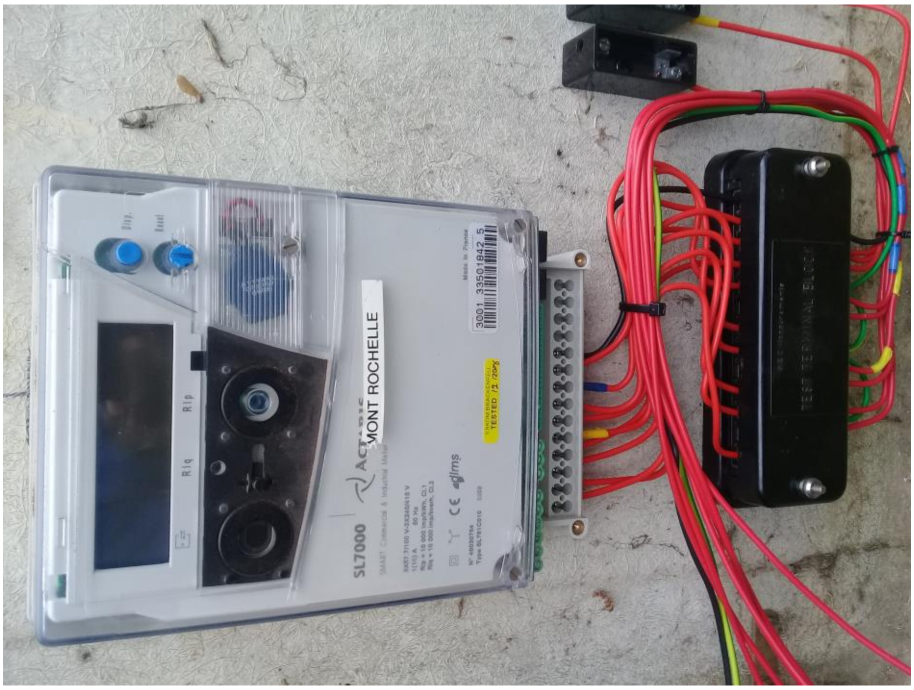


Figure 50: Meter set-up with the SL7000 meters

A greater challenge of short supply cables that could not sustain the weight of the current transformer also arose. The photos below indicate the customer supply side with short cables on which current transformers could not be put on.

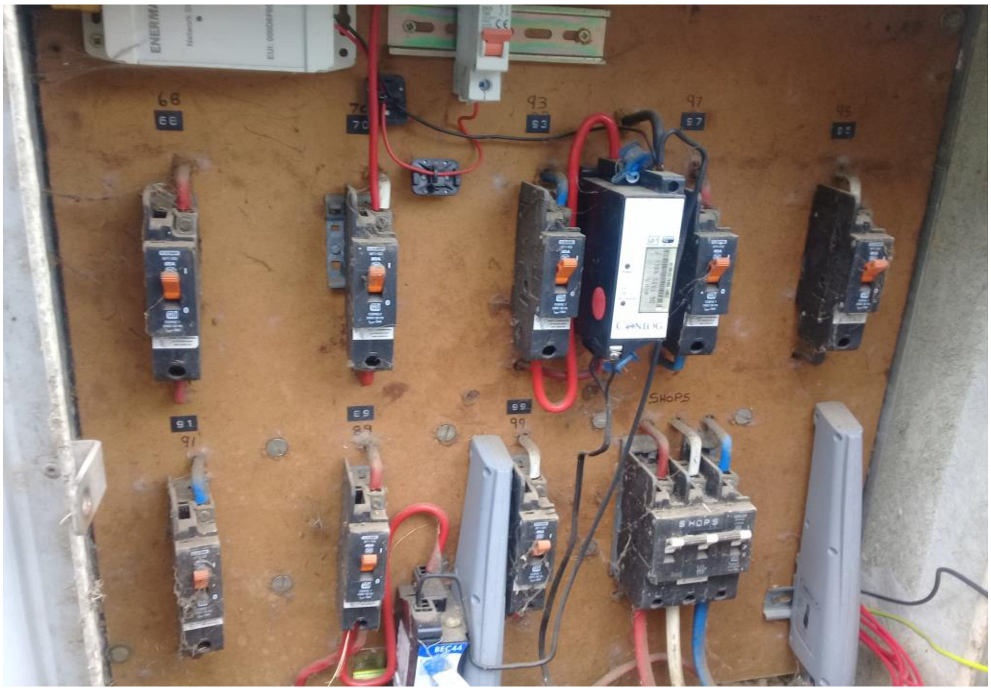


Figure 51: The customer supply side showing the lack of space to accommodate the current transformers.

Some kiosks were too small to accommodate the meters with the SL7000 size. Below is an indicator of the same.

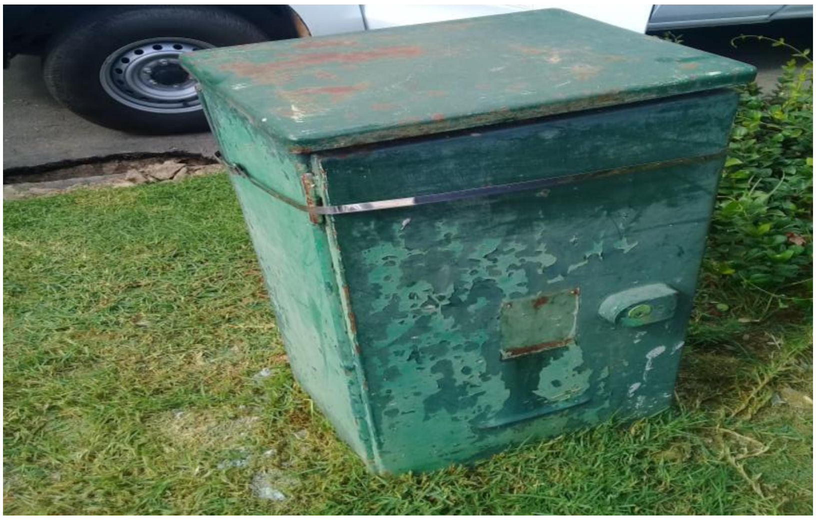


Figure 52: Old kiosks without enough space to accommodate the meters.

A.6 Metering with IoTaWatt Wi-Fi meter

To collect an equally good sample size, at least a sample size of 30 customers would be needed. It was decided therefore to procure smart meters small enough to fit in most of the kiosks to collect the individual customer data. Raspberry pi based open energy monitor meters were procured for the same exercise. The meter has a 0.2% accuracy on the values metered. Additionally, the recorded values are in watts. It can record both positive and negative values and therefore was useful in solar PV metering exercise. Figure 53 shows the collection after the set up ready for installation.

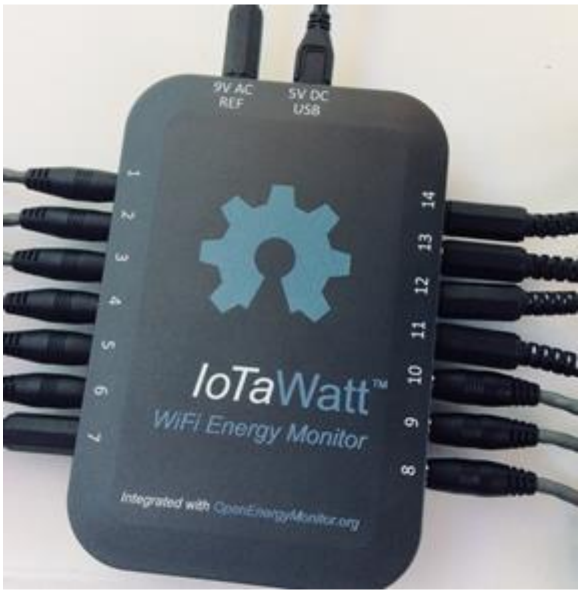


Figure 53: Sample IoTaWatt Energy monitor meter installed in the kiosks

The Wi-Fi energy monitor was chosen for various reasons:

i) It has a convenient size and thus can fit in most of the kiosks including the old ones without a problem.

ii) Additionally, it can log 14 inputs at once and as such can capture all the customer’s consumption within the kiosk.

iii) The current transformers used by the meter are equally small.

iv) Data retrieval did not require the removal of the meter as it was the case in the first meter, SL7000 series

About 35 individual consumptions were measured and the demand data retrieved for analysis.

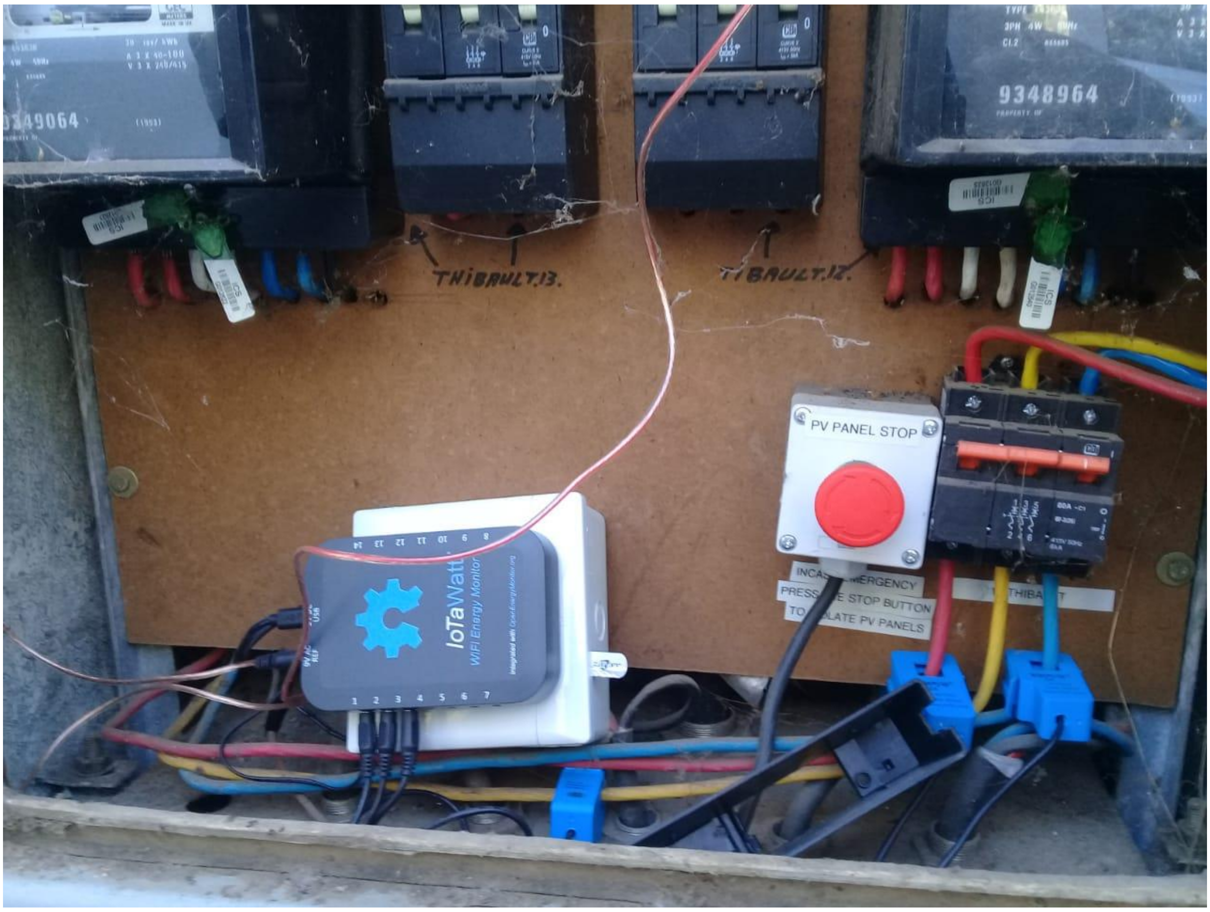


Figure 54: IoTaWatt Energy Wi-Fi meter set up in one of the kiosks metered

Appendix B Brief review of probability

This study proposes the use of a probabilistic concept to model a solution that can be used to estimate the amount of installed solar PV SSEGs on the network. The data used is stochastic. The variation and diversity in demand as well as the variability of the solar irradiance should be described through a stochastic model.

B.1 Probability theory

In this section we review briefly the concepts of probability, probability space and Monte Carlo methods.

Probability as a concept is a representation of chance, proportions or relative frequency of occurrence (Baron, 2007). Probability is therefore a finite measure which implies that the measure is derived from the largest possible value, referred to as a sample space. The set of possible outputs is called the range of probability while the domain refers to the set of inputs (Baron, 2007). A sample space defines the possible outcomes of a stochastic experiment while an event refers to a particular outcome. The occurrence of random events can be best described using probability distributions. As such, probability distributions describe the behaviour of random variables.

B.2 Probability Distributions

Probability distributions explain the statistical behaviour of data. In this case, to effectively explore the statistical behaviour of the load and irradiance, there is need to understand probability distributions.

A probability distribution is a pattern which characterizes the behaviour of random variable. Similarly, a random variable is defined as that variable generated through a random process(Reddy, 2001).

There exists a great number of distributions. These include but are not limited to:

i) Gaussian distribution

ii) Student t distribution

iii) Log-normal distribution

iv) Gamma Distribution

v) Beta distribution

vi) Exponential distribution

Table 16 shows the inherent characteristics of a few selected distributions

Table 16: Characterizing probability distributions (Reddy, 2001)

Distribution	**Description**	**Descriptive Equation**	**Standardized Descriptive equation**
Gaussian	Main parameters are µ and σ applicable for n>30	$N(x;\mu,\sigma)$ $= \frac{1}{\sigma(2\pi)^{\frac{1}{2}}} \exp\,[-\frac{x-\mu}{\sigma}]^2$	$N(x;0,1)$ $= \frac{1}{(2\pi)^{\frac{1}{2}}} \exp\,(-z^2/2)$ Where $z \equiv \frac{x-u}{\sigma}$
Student t	Described as $t = (\mu, s, v)$ where the parameters are mean, standard deviation and degrees of freedom n<30		$t = \frac{x-\mu}{s\sqrt{n}}$
Log-normal	For skewed data, with mostly non-negative outcomes. Defined by both standard deviation and the mean	$L(x;\ \mu,\sigma)$ $= \frac{1}{\sigma.x(\sqrt{2\pi}} \exp\left[-\frac{(lnx-\mu)^2}{2\sigma^2}\right]$	

B.3 Statistically significant parameters in assessing models

In a discrete random variable (DRV), the probability that an observation is made within the population is define by a frequency function $f_x(x)$ i.e. $f_x(x) = P(X = x)$.

For Continuous Random Variables, $f_{\mathrm{x}}(x)$ is referred to as the density function (Glasserman, 2003). $f_{\mathrm{x}}(x)$ represents the probability that a certain observation is within the range that is specified by the given integral.

The Probability Density Function of a random Variable describes the variables detailed behaviour. The following parameters are key in the describing a population(Ochi, 1976)

B.4 Arithmetic operations on random variables

Arithmetic operations on random variables take many forms. These include but are not limited to multiplication, addition, subtraction and even convolution. The sections below show the summary of this(Reddy, 2001).

B.5 Correlation between two random variables

Correlation between two random variables can be used as a measure of the independence of two or more variables. For two random variables X and Y, knowing the value of X will intuitively tell us nothing about the value of Y. such a relationship defines the independence of two variables (Yan, 2018). Correlation is a measure of linear dependence between variables, lack of correlation does not imply that the two variables are independent. This is because, if there exists a non-linear relation between two random variables, they could still be uncorrelated.

Parameter	**Description**	**Representation**	**Standardized Descriptive equation**
Expectation	Mean of all outcomes of a RV. Expressed as E[X]	f(x) is the density function	$E[X] = \frac{1}{N}\sum_{i=1}^{N} xi$ (Population) $E[X] = \int f\mathrm{X}(x)xdx$ (CRV) $E[X] = \sum f(x)x$ (DRV)
Variance	Deviation of the RV from the expectation value	$Var[X] = E[X - E[X]]^2 = E[X^2]\text{-}E[X]^2$ General Expression	$Var[X] = \frac{1}{N}\sum_{i=1}^{N}(xi - E[X])^2$ (population) $Var[X] = \int f\mathrm{X}(x)(x - E[X])^2 dx$ (CRV) $Var[X] = \sum f(x)(x - E[X])^2$ (DRV)
Standard Deviation	This is the root of variance; Measure of how far a RV is from the mean	$\sigma_{\mathrm{X.}} = \sqrt{Var\ [X]}$	
Covariance	Describes the relationship between two random variables X and Y. Shows the variance between the RV X and E[Y] and vice versa	$Cov[X,Y] = (E[X - E[X])(E[Y - E[Y]])$ $= E[XY] - E[X]E[Y]$ $Cov[X,Y] = Cov[Y,X]$ $Cov[X,X] = Var[X]$	$\begin{matrix} var[X1] & Cov[X1,X2] & Cov[X1,Xk] \\ Cov[X2,X1] & Var[X2] & Cov[X2,Xk\} \\ Cov[Xk,X1] & Cov[Xk,X2] & Var[Xk] \end{matrix}$ Covariance Matrix denotes the relationship between variables in a multivariate distribution
Correlation factor	Denotes the strength of relationship between two variables. Ranges between -1 and 1	$\rho_{\mathrm{X,\ Y}} = \frac{Cov[X,Y]}{\sqrt{Var[X]\, Var[Y]}}$	

Operation	Parameter	Expression	Explanation
Addition	Expectation	$E[Y1+Y2]=E[Y1]+E[Y2]$	
	Variance	$Var[Y1+Y2]=Var[Y1]+Var[Y2]+2Cov[Y1,Y2]$	
Subtraction	Variance	$Var[Y1-Y2]=Var[Y1]+Var[Y2]-2Cov[Y1,Y2]$	
Multiplication by a factor	Expectation	$E[aY]=aE[Y]$	
	Covariance	$Var[aY]=a^2Var[Y]$	
Convolution	CRV DRV	$f_X(x)=\int\limits_{-\infty}^{\infty} f_{Y1}(t)f_{Y2}(x-t)dt$ $f_X(x)=\sum\limits_{t} f_{Y1}(t)f_{Y2}(x-t)$	f x1+x2 (y) = $\int_{-\infty}^{\infty} fx1(y-\tau)fx2(\tau)d\tau$ $= \int_{-\infty}^{\infty} fx1(\tau)fx2(y-\tau)d\tau$ (General Case of convolution) $X\sim\aleph(\mu_1,\sigma_1^2)$, ,$Y\sim\aleph(\mu_2,\sigma_2^2)$ $X+Y\sim\aleph(\mu_1+\mu_2,\sigma_1^2,\sigma_2^2)$ (Normal distribution) $X\sim Bin(n1,p)$, $Y\sim Bin(n2,p)$, $X+Y\sim Bin(n1+n2,p)$ (Binomial distribution with same p value) $X\sim Poi(\text{ג}1)$ and $Y\sim Poi(\text{ג}2)$ $X+Y\sim Poi(\text{ג}1+\text{ג}2)$ (Poisson distribution, with$\text{ג}1\neq\text{ג}2$.)
Addition of two independent random variables is done through convolution (Yan, 2018). In turn, a new variable is created which may have a different probability distribution function (Mcmillan & Kohlberg, 2018). For proper use of convolution in the solution of this problem, the two variables must be independent. As such, if two random variables are independent, then the density of their sums equals the convolution of their densities. The density function of the sum of two independent random variables converges into a convolution integral if the two are statistically independent (Ochi, 1976). For X and Y that are independent, then the following holds: $E[X\ Y]=E[X]E[Y]$(Papoulis & Pillai, 1991).			

B.6 Monte Carlo methods

Monte Carlo methods are numerical methods that utilizes random variables to provide a non-analytical solution to mathematical problems (Terejanu, 2009). Equally, random variables are unpredictable in nature and it is not therefore possible to tell what value they assume. To be able to generate a set of random variables therefore, there is need to know the distribution of the variables. These distributions give the probability of a given value.

According to Pengelly, Monte Carlo methods find application in a a variety of applications in mathematics. As such, they provide a pathway to simulations that would otherwise be difficult and time consuming. Additionally, the use of random numbers makes Monte Carlo methods applicable in processes that are random or stochastic in nature (Pengelly, 2002). There are several types of Monte Carlo methods. The use of each method is determined by the required accuracy of the approximations.

The following list enumerates some of the Monte Carlo methods.

i) Crude Monte Carlo

ii) Acceptance -Rejection Monte Carlo

iii) Stratified Sampling

iv) Importance Sampling

B.7 Law of large numbers

Monte Carlo is anchored in several statistical principles. One of them is the law of large numbers. The law of large numbers states that for a random variable with an expectation [E], the sample average is as close to the theoretical mean (expectation value, E) when the number of the variables are large enough (Siddal, 1983)

For any large set of random variable Y with a known distribution, if we draw independent and identical set of N samples, we can estimate the same distribution using the samples. Additionally, we can compute the expectations of the same using the selected samples.

B.8 Central limit theorem

The central limit theorem states that for a sufficiently large number of random variables that are identically distributed, as the number of the variables increase, the random variable assumes a normal distribution(Reddy, 2001). The central limit theorem is significant in making statistical inferences for large numbers of samples. For example, it can be concluded that 68.7% of the values lie within one standard deviation and so on.

B.9 Monte Carlo simulation process

Monte Carlo simulations performs random sampling and is efficient in conducting large number of computations. The following is the general procedure that is followed from the model input to the model output for a Monte Carlo simulation (Siddal, 1983)

i) Possible values of the input random variables are sampled according to their distribution.

ii) A performance function is set, say Y =g(X).

iii) The output Y is determined by referencing the performance function.

iv) Statistical analysis is carried out on the output to determine the characteristic of the distributions.

Monte Carlo methods are mostly used in probability estimation, ascertaining of expectations and other distribution characteristics especially in computationally tasking problems (Baron, 2007). From the generated random variable sequence, it is possible to generate distributions of interest and estimate both probabilities and expectation values for a sample representing the whole population.