A BAYESIAN APPROACH TO ENERGY MONITORING OPTIMIZATION

by

Herman Carstens

Submitted in partial fulfillment of the requirements for the degree

Philosophiae Doctor (Electrical Engineering)

in the

Department of Electrical, Electronic and Computer Engineering

Faculty of Engineering, Built Environment and Information Technology

UNIVERSITY OF PRETORIA

June 2017

SUMMARY

A BAYESIAN APPROACH TO ENERGY MONITORING OPTIMIZATION

by

Herman Carstens

Promoter(s):	Professors Xiaohua Xia and Sarma Yadavalli
Department:	Electrical, Electronic and Computer Engineering
University:	University of Pretoria
Degree:	Philosophiae Doctor (Electrical Engineering)
Keywords:	Measurement and Verification, Retrofit, Bayesian, Errors-in-Variables, Generalised Linear Model, Longitudinal Sampling, Simulation Extrapolation, Machine Learning,

This thesis develops methods for reducing energy Measurement and Verification (M&V) costs through the use of Bayesian statistics. M&V quantifies the savings of energy efficiency and demand side projects by comparing the energy use in a given period to what that use would have been, had no interventions taken place. The case of a large-scale lighting retrofit study, where incandescent lamps are replaced by Compact Fluorescent Lamps (CFLs), is considered. These projects often need to be monitored over a number of years with a predetermined level of statistical rigour, making M&V very expensive.

M&V lighting retrofit projects have two interrelated uncertainty components that need to be addressed, and which form the basis of this thesis. The first is the uncertainty in the annual energy use of the average lamp, and the second the persistence of the savings over multiple years, determined by the number of lamps that are still functioning in a given year. For longitudinal projects, the results from these two aspects need to be obtained for multiple years.

This thesis addresses these problems by using the Bayesian statistical paradigm. Bayesian statistics is still relatively unknown in M&V, and presents an opportunity for increasing the efficiency of statistical

analyses, especially for such projects.

After a thorough literature review, especially of measurement uncertainty in M&V, and an introduction to Bayesian statistics for M&V, three methods are developed. These methods address the three types of uncertainty in M&V: measurement, sampling, and modelling. The first method is a low-cost energy meter calibration technique. The second method is a Dynamic Linear Model (DLM) with Bayesian Forecasting for determining the size of the metering sample that needs to be taken in a given year. The third method is a Dynamic Generalised Linear Model (DGLM) for determining the size of the population survival survey sample.

It is often required by law that M&V energy meters be calibrated periodically by accredited laboratories. This can be expensive and inconvenient, especially if the facility needs to be shut down for meter installation or removal. Some jurisdictions also require meters to be calibrated in-situ; in their operating environments. However, it is shown that metering uncertainty makes a relatively small impact to overall M&V uncertainty in the presence of sampling, and therefore the costs of such laboratory calibration may outweigh the benefits. The proposed technique uses another commercial-grade meter (which also measures with error) to achieve this calibration in-situ. This is done by accounting for the mismeasurement effect through a mathematical technique called Simulation Extrapolation (SIMEX). The SIMEX result is refined using Bayesian statistics, and achieves acceptably low error rates and accurate parameter estimates.

The second technique uses a DLM with Bayesian forecasting to quantify the uncertainty in metering only a sample of the total population of lighting circuits. A Genetic Algorithm (GA) is then applied to determine an efficient sampling plan. Bayesian statistics is especially useful in this case because it allows the results from previous years to inform the planning of future samples. It also allows for exact uncertainty quantification, where current confidence interval techniques do not always do so. Results show a cost reduction of up to 66%, but this depends on the costing scheme used. The study then explores the robustness of the efficient sampling plans to forecast error, and finds a 50% chance of undersampling for such plans, due to the standard M&V sampling formula which lacks statistical power.

The third technique uses a DGLM in the same way as the DLM, except for population survival survey samples and persistence studies, not metering samples. Convolving the binomial survey result

distributions inside a GA is problematic, and instead of Monte Carlo simulation, a relatively new technique called Mellin Transform Moment Calculation is applied to the problem. The technique is then expanded to model stratified sampling designs for heterogeneous populations. Results show a cost reduction of 17-40%, although this depends on the costing scheme used.

Finally the DLM and DGLM are combined into an efficient overall M&V plan where metering and survey costs are traded off over multiple years, while still adhering to statistical precision constraints. This is done for simple random sampling and stratified designs. Monitoring costs are reduced by 26-40% for the costing scheme assumed.

The results demonstrate the power and flexibility of Bayesian statistics for M&V applications, both in terms of exact uncertainty quantification, and by increasing the efficiency of the study and reducing monitoring costs.

OPSOMMING

'N BAYESIESE BENADERING TOT ENERGIEMONITERINGSOPTIMERING

deur

Herman Carstens

Promotor(s):	Professore Xiaohua Xia en Sarma Yadavalli
Departement:	Elektriese, Elektroniese en Rekenaar-Ingenieurswese
Universiteit:	Universiteit van Pretoria
Graad:	Philosophiae Doctor (Elektriese Ingenieurswese)
Sleutelwoorde:	Meting en verifieëring, Bayesies, retrofit, foute-in-veranderlikes, veralgemeende linieêre model, longitudinale proefneming, simulerings extrapolering, masjienleer

Hierdie proefskrif ontwikkel metodes waarmee die koste van energiemonitering en verifieëring (M&V) deur Bayesiese statistiek verlaag kan word. M&V bepaal die hoeveelheid besparings wat deur energiedoeltreffendheid- en vraagkantbestuurprojekte behaal kan word. Dit word gedoen deur die energieverbruik in 'n gegewe tydperk te vergelyk met wat dit sou wees indien geen ingryping plaasgevind het nie. 'n Grootskaalse beligtingsretrofitstudie, waar filamentgloeilampe met fluoresserende spaarlampe vervang word, dien as 'n gevallestudie. Sulke projekte moet gewoonlik oor baie jare met 'n vasgestelde statistiese akkuuraatheid gemonitor word, wat M&V duur kan maak.

Twee verwante onsekerheidskomponente moet in M&V beligtingsprojekte aangespreek word, en vorm die grondslag van hierdie proefskrif. Ten eerste is daar die onsekerheid in jaarlikse energieverbruik van die gemiddelde lamp. Ten tweede is daar die volhoubaarheid van die besparings oor veelvoudige jare, wat bepaal word deur die aantal lampe wat tot in 'n gegewe jaar behoue bly. Vir longitudinale projekte moet hierdie twee komponente oor veelvoudige jare bepaal word.

Hierdie proefskrif spreek die probleem deur middel van 'n Bayesiese paradigma aan. Bayesiese statistiek is nog relatief onbekend in M&V, en bied 'n geleentheid om die doeltreffendheid van

statistiese analises te verhoog, veral vir bogenoemde projekte.

Die proefskrif begin met ’n deeglike literatuurstudie, veral met betrekking tot metingsonsekerheid in M&V. Daarna word ’n inleiding tot Bayesiese statistiek vir M&V voorgehou, en drie metodes word ontwikkel. Hierdie metodes spreek die drie hoofbronne van onsekerheid in M&V aan: metings, opnames, en modellering. Die eerste metode is ’n laekoste energiemeterkalibrasietegniek. Die tweede metode is ’n Dinamiese Linieêre Model (DLM) met Bayesiese vooruitskatting, waarmee meter opnamegroottes bepaal kan word. Die derde metode is ’n Dinamiese Veralgemeende Linieêre Model (DVLM), waarmee bevolkingsoorlewing opnamegroottes bepaal kan word.

Volgens wet moet M&V energiemeters gereeld deur erkende laboratoria gekalibreer word. Dit kan duur en ongerieflik wees, veral as die aanleg tydens meterverwydering en -installering afgeskakel moet word. Sommige regsgebiede vereis ook dat meters in-situ gekalibreer word; in hul bedryfsomgewings. Tog word dit aangetoon dat metingsonsekerheid ’n klein deel van die algehele M&V onsekerheid beslaan, veral wanneer opnames gedoen word. Dit bevraagteken die kostevoordeel van laboratori-umkalibrering. Die voorgestelde tegniek gebruik ’n ander kommersieële-akkuurraatheidsgraad meter (wat self ’n nie-weglaatbare metingsfout bevat), om die kalibrasie in-situ te behaal. Dit word gedoen deur die metingsfout deur SIMulerings EKStraptolering (SIMEKS) te verminder. Die SIMEKS resultaat word dan deur Bayesiese statistiek verbeter, en behaal aanvaarbare foutbereike en akkuurate parameterafskattings.

Die tweede tegniek gebruik ’n DLM met Bayesiese vooruitskatting om die onsekerheid in die meting van die opnamemonster van die algehele bevolking af te skat. ’n Genetiese Algoritme (GA) word dan toegepas om doeltreffende opnamegroottes te vind. Bayesiese statistiek is veral nuttig in hierdie geval aangesien dit vorige jare se uitslae kan gebruik om huidige afskattings te belig Dit laat ook die presiese afskatting van onsekerheid toe, terwyl standaard vertrouensintervaltegnieke dit nie doen nie. Resultate toon ’n kostebesparing van tot 66%. Die studie ondersoek dan die standvastigheid van kostedoeltreffende opnameplanne in die teenwoordigheid van vooruitskattingsfoute. Dit word gevind dat kostedoeltreffende opnamegroottes 50% van die tyd te klein is, vanweë die gebrek aan statistiese krag in die standaard M&V formules.

Die derde tegniek gebruik ’n DVLM op dieselfde manier as die DLM, behalwe dat bevolkingsoorlewing-opnamegroottes ondersoek word. Die saamrol van binomiale opname-uitslae binne die GA skep ’n

probleem, en in plaas van ’n Monte Carlo simulasie word die relatiewe nuwe Mellin Vervorming Moment Berekening op die probleem toegepas. Die tegniek word dan uitgebou om laagsgewyse opname-ontwerpe vir heterogene bevolkings te vind. Die uitslae wys ’n 17-40% kosteverlaging, alhoewel dit van die koste-skema afhang.

Laastens word die DLM en DVLM saamgevoeg om ’n doeltreffende algehele M&V plan, waar meting en opnamekostes teen mekaar afgespeel word, te ontwerp. Dit word vir eenvoudige en laagsgewyse opname-ontwerpe gedoen. Moniteringskostes word met 26-40% verlaag, maar hang van die aangenome koste-skema af.

Die uitslae bewys die krag en buigsaamheid van Bayesiese statistiek vir M&V toepassings, beide vir presiese onsekerheidskwantifisering, en deur die doeltreffendheid van die dataverbruik te verhoog en sodoende moniteringskostes te verlaag.

ACKNOWLEDGEMENTS

Whatever is valuable in this thesis cannot be credited to me alone, as I received much help and support from so many different people. I hope that one day I can be a bit more like each of you.

First, Alvara. Thank you for your patient love and support, keeping me sane, fed, and able to focus, while you also have work and postgraduate studies. I cannot imagine how this could have been done without you. I love you very much. My and Alvara's families and friends have been incredibly supportive and understanding of my reclusivity. I wish that I could repay you, but I fear that I will not be able to do so.

Thank you also to my study leaders for sharing your expertise and time of your busy schedule, your kind assistance with the non-academic aspects, and for the myriad ways in which you have taught me to be a better thinker. There is still a long way to go, but you have shown me the way.

Then there is Arvind Rajan, a fellow PhD student at Monash University, Malaysia. Arvind and his colleagues developed the Mellin Transform Moment Method, of which I became aware through Dr. Mark Rawlins at Energy Combustion Services. I emailed Arvind for help, and found the most helpful and encouraging person I have ever met. Without him and his help on their method, the work in Chapters 6-8 would probably not have been publishable. Thank you, Arvind. People like you make the world a better place.

In my (long) career as a student, I have been fortunate enough to have had many good teachers, and it is difficult to single out any one in particular. However, I am especially grateful to Mr Thomas Hagspihl, who balanced his exceptional intellect with even greater kindness. He also taught us to love statistics rather than fear it, thereby gifting me with a lifelong fascination and a career trajectory.

This work was made possible through the financial support of the National Hub for the Postgraduate Programme in Energy Efficiency and Demand Side Management. An Innovation Scholarship from the National Research Foundation and the Department of Science and Technology also contributed to a fraction of the funding.

Soli Deo gloria

LIST OF ABBREVIATIONS

ADC	Analog to Digital Conversion
ADVI	Automatic Differentiation Variational Inference
AMI	Advanced Metering Infrastructure
ANN	Artificial Neural Network
ANSI	American National Standards Institute
ASHRAE	American Society of Heating, Refrigeration, and Air-conditioning Engineers
BEM	Building Energy Modelling
BLUE	Best Unbiased Linear Estimator
CCA	Correlation and Cluster Analysis
CRPS	Continuously Ranked Probability Score
CCC	California Commissioning Collaborative
CDM	Clean Development Mechanism
CEAC	Cost Effectiveness Acceptability Curve
CE	Cost Effectiveness
CFL	Compact Fluorescent Lamp
CI	Confidence Interval
CLT	Central Limit Theorem
CT	Current Transformer
CV	Coefficient of Variance
CV(RMSE)	Coefficient of Variation on the Root Mean Square Error
CV(STD)	Coefficient of Variation on the Standard Deviation
DAQ	Data AcQuisition (board)
DGLM	Dynamic Generalized Linear Model
DLM	Dynamic Linear Model
DMM	Digital MultiMeter
DSM	Demand Side Management
DSP	Digital Signal Processing
ECM	Energy Conservation Measure
EE	Energy Efficiency
EGF	Energy Governing Factor

EPC	Energy Performance Contract
ESCO	Energy Services Company
EUI	Energy Use Intensity
EV	Expected Value
FPC	Finite Population Correction
G14	ASHRAE guideline 14-2002 and 14-2014
GHG	Greenhouse Gas
GMM	Gaussian Mixture Model
GP	Gaussian Process
GUM	Guide to the expression of Uncertainty in Measurement
HDI	Highest Density Interval
HVAC	Heating, Ventilation, and Air Conditioning
IDM	Integrated Demand Management
IEC	International Electrotechnical Commission
IEEE	Institute for Electrical and Electronic Engineers
ICL	Incandescent Lamp
IPMVP	International Performance Measurement and Verification Protocol
ISO	International Standards Organization
LASSO	Least Absolute Shrinkage and Selection Operator
LBE	Linear Bayesian Estimation
LBNL	Lawrence Berkeley National Laboratory
LED	Light Emitting Diode
LP	Linear Programming
LR	Linear Regression
LRC	Lighting Research Centre
MAC	Marginal Abatement Cost
MARS	Multivariate Adaptive Regression Spline
MC	Monte Carlo
MC-LHS	Monte Carlo Latin Hypercube Sampling
MCMC	Markov Chain Monte Carlo
MEL	Miscellaneous Electrical Load
MEM	Measurement Error Model
MID	Measurement Instrument Directive

MLE	Maximum Likelihood Estimation
MTMC	Mellin Transform Moment Calculator
M&V	Measurement and Verification
NDB	Net Determination Bias
NIST	National Institute for Standards and Technology
NMBE	Normalised Mean Bias Error
NPV	Net Present Value
NREL	National Renewable Energy Laboratory
OLS	Ordinary Least Squares
ORNL	Oak Ridge National Laboratory
PD	Project Developer
PDF	Probability Density Function
PELP	Polish Efficient Lighting Programme
PF	Power Factor
PIR	Passive Infra-Red
PIT	Probability Integral Transform
RFVI	Random Forest Variable Importance
RS	Response Surfaces
SA	Survival Analysis
SANAS	South African National Accreditation System
SANS	South African National Standard
SARS	South African Revenue Service
SEE Action	State and Local Energy Efficiency action group
SEM	Stick-on Electricity Meter
SIMEX	SIMulation EXtrapolation
SNR	Signal to Noise Ratio
SRC	Standardized Regression Coefficient
SVR-GK	Support Vector Regression Gaussian Kernel
THD	Total Harmonic Distortion
TSDLM	Time Series Dynamic Linear Model
TUR	Test Uncertainty Ratio
UK	United Kingdom
UKAS	United Kingdom Accreditation Service

UMP	Uniform Methods Project
UNFCCC	United Nations Framework Convention on Climate Change
US	United States
UUT	Unit Under Test
VI	Virtual Instruments
VSD	Variable Speed Drive

TABLE OF CONTENTS

CHAPTER 1 INTRODUCTION

1.1 RESEARCH OUTPUTS

1.1.1 Journal Papers

[1] Herman Carstens, Xiaohua Xia, and Sarma Yadavalli, "Measurement Uncertainty in Energy Measurement and Verification: Present State of the Art", *Renewable and Sustainable Energy Reviews* (In print).

[2] Herman Carstens, Xiaohua Xia, and Sarma Yadavalli, "Low-Cost Energy Meter Calibration Method for Energy Measurement and Verification", *Applied Energy* (2017) 188 pp. 563-575.

[3] Herman Carstens, Xiaohua Xia, Sarma Yadavalli, and Arvind Rajan, "Efficient Longitudinal Population Survival Survey Sampling for the Measurement and Verification of Lighting Retrofit Projects", *Energy and Buildings Special Issue: Energy Efficient Lighting Strategies in Buildings* (2017) 150 pp. 163-176.

[4] Herman Carstens, Xiaohua Xia, and Sarma Yadavalli, "Efficient metering and surveying sampling designs in longitudinal Measurement and Verification for lighting retrofit", *Energy and Buildings* (2017) 154 pp. 430-447

1.1.2 Conference Papers

[1] Herman Carstens, Xiaohua Xia, and Sarma Yadavalli, "Measurement Uncertainty and Risk in Measurement and Verification Projects" **International Energy Programme Evaluation Conference**, August 2015, Long Beach, California.

1.2 CHAPTER OVERVIEW

This chapter introduces the concept of M&V and explains the problem and approach to the proposed solution. The most important terms and notation used in the rest of this thesis are also explained. Most references and rigorous motivation will be deferred to Chapters 2 and 3.

1.3 BACKGROUND

Measurement and Verification (M&V) is the process by which energy savings realised by energy efficiency and demand side management projects are independently and reliably quantified [1]. It is an established field in which the basic principles and practices are well defined, similar to auditing in finance. The crux of M&V is the fact that energy savings cannot be measured directly. A statistical model forecasting what the business-as-usual energy use in a given period *would have been* can be created, and this can be compared to the actual energy use. The difference between the forecast and actual consumption is the savings estimate. The uncertainty in the savings estimates should fall within certain statistical bounds, which are often set by regulators. For many real-world projects, this is a non-trivial problem. For other kinds of M&V projects, this counterfactual calculation is simple, but monitoring with satisfactory accuracy can be expensive. Because project payment often depends on the M&V report, and because M&V can be costly, the field is well suited to a statistical engineering investigation. Some projects are also only marginally feasible, especially if M&V needs to be added to the project cost. Efficient M&V methods can increase the feasibility and reduce the risk for others.

In the South African context, M&V was originally done by teams from various universities, and projects were administrated by Eskom's Integrated Demand Management (IDM) unit. In 2014 there were more than 700 active M&V projects. Recently, South African M&V was opened to private companies, subject

to accreditation by the South African National Standards Authority (SANAS). In a parallel development, the 12L tax incentive was promulgated by the South African Revenue Service (SARS) [2], and provides a rebate for savings certified by an accredited M&V company. These certifications take place according to the South African National Standard (SANS) 50010: Measurement and Verification of Energy Savings [3], which governs M&V in South Africa. The standard codifies minimum requirements and methods from guidelines such as the International Performance Measurement and Verification Protocol (IPMVP) [1], but does not place a numerical uncertainty constraint on M&V reporting. However, the process of identifying fair, quantified uncertainty reporting requirements is underway and contributes to the timeliness of this thesis. A parallel approach adopted by SANAS is to require energy meters to be calibrated before they may be used for accredited M&V projects. This requirement places no uncertainty requirement on reporting but limits the risk by quality control – at a cost – as will be discussed in this thesis. Regardless of whether firm uncertainty reporting limits are legislated in the South African context, the efficient quantification of uncertainty in M&V is relevant both locally and internationally.

A second factor compounds the problem of costly uncertainty quantification in M&V. The time horizons on M&V projects may be many years. For these projects to be eligible under the United Nations Framework Convention on Climate Change (UNFCCC)'s Clean Development Mechanism (CDM), the performance of lighting projects should be tracked for up to 10 years, while other projects may be tracked for up to 21 years [4]. For these studies, a longitudinal component is therefore present in M&V designs. This needs to be combined with the cross-sectional aspect to yield a savings estimate for a given year.

Lighting projects are frequently used as M&V case studies because they are relatively simple and do not detract from model development with case-specific model features. A lighting retrofit case was therefore selected for this thesis, as it allows the study to focus on the measurement of the energy use of the lighting population, and the sampling of such a population over time.

1.4 PROBLEM STATEMENT

M&V costs can be prohibitive for projects where uncertainty needs to be quantified with adequate statistical precision. Efficient M&V methods, that is, methods that achieve the required accuracy at

low cost, are therefore needed.

Specifically for lighting retrofit M&V studies, two kinds of data are needed to calculate the energy use in such projects: population survival data, and aggregated energy use data. Energy use data are obtained from electricity meters installed in a statistically representative number of households, while population survival data are collected through surveys.

Energy use data are usually obtained from calibrated energy meters or through spot metering and the use of lighting loggers. The former is a more accurate approach adopted as part of an M&V plan known as 'retrofit isolation with all parameter measurement'. However, installing an adequate number of meters to obtain a statistically representative sample can be expensive. Not only are meters to be bought and installed, but they also have to be calibrated periodically as a statutory requirement. This entails possible facility shutdown as well as sending a meter to a laboratory for calibration using special equipment. Ironically, energy metering uncertainty usually makes a small contribution to the overall uncertainty when compared to sampling uncertainty. A more cost-effective measurement calibration method may, therefore, reduce metering costs, and even if such meters are calibrated to a lower standard, may still make a small difference to the overall savings uncertainty.

The second aspect of a longitudinal M&V study is population surveillance. Considering multi-year savings rather than first-year savings only, can decrease the cost of electricity saved by up to 70% [5]. The savings reported in a given year should be a function of the number of units that have survived to that year. If only 50% of the units are still functioning in a given year, only 50% of the original savings can be claimed for that year, all else being equal. Determining this proportion to a given statistical precision is a longitudinal survey design problem.

Designing a longitudinal model for a single population has limited use in practice. It is often necessary to stratify a population by luminaire type, location, or application. Treating such strata as completely independent is statistically inefficient, but combining their population proportion estimates with other sources of uncertainty in a meaningful way has not been done to the author's knowledge.

The total savings uncertainty then needs to be calculated by combining results and uncertainties from the cross-sectional metering sample with those of the population survey.

1.4.1 Research gap

Basic frequentist statistics have been used on the majority of M&V problems and are recommended by all leading guidelines. However, these methods make restrictive assumptions on the data and do not reflect the complexities of real-world projects. They also fail to account for other information that is known about the project – especially for longitudinal studies where the same population is sampled repeatedly. Bayesian statistics provide a mathematically rigorous way of incorporating such data to decrease uncertainty and increase M&V monitoring efficiency. These methods have enjoyed much attention in other fields, but they have not been applied to M&V problems.

1.5 RESEARCH OBJECTIVE AND QUESTIONS

The objectives of this thesis are:

- To provide a systematic overview of the state of the art for measurement uncertainty in M&V.
- To develop a low-cost energy meter calibration method.
- To develop an efficient longitudinal cross-sectional meter sampling method that accounts for past data and quantifies uncertainty accurately.
- To develop an efficient longitudinal population survey sampling method that accounts for past data and quantifies uncertainty accurately.
- To develop an optimization method by which optimal sampling plans for stratified longitudinal studies may be determined.
- To combine metering and survey sampling into an overall longitudinal M&V plan.

1.6 HYPOTHESIS AND APPROACH

This thesis hypothesizes that by using Bayesian statistics, more efficient M&V sampling plans than those of standard frequentist M&V methods can be devised.

The approach is to use a lighting retrofit monitoring project as a case study and to develop efficient M&V methods using Bayesian statistics and various machine learning algorithms. The results of these methods will then be compared to the results of standard frequentist M&V methods.

M&V efficiency is improved on two fronts.

1. a) Lower the cost metering through low-cost calibration.

1. b) Improved metering sampling planning.

The second is lower cost sampling:

2. a) Improved metering sampling planning (overlapping with 1. b).

2. b) Improved survey sampling planning.

1.7 RESEARCH GOALS

The goal of the proposed research is to develop cost-effective monitoring plans by characterising lamp population survival with energy use while reporting energy use and savings with the required accuracy. To this end, Bayesian methods are employed in conjunction with other necessary statistical techniques.

1.8 OVERVIEW OF STUDY

This thesis is arranged as follows. In Chapter 2, relevant literature is reviewed and evaluated in the light of current needs and the present state of the art in M&V. This chapter informs all the other chapters

and includes a systematic study of measurement uncertainty in M&V to identify opportunities and research gaps, which has not been done before. It also includes an evaluation of the applicability of literature from related fields in applied statistics, such as Bayesian methods, survival analysis, and errors-in-variables research.

Chapter 3 introduces the Bayesian paradigm. After a discussion of the advantages of this paradigm relative to the standard frequentist one used for M&V, the theory is presented. Aspects relevant to M&V are discussed, and applications to measurement, sampling, and regression are made using simple examples.

In Chapter 4, a low-cost energy meter calibration method using Simulation Extrapolation (SIMEX) and Bayesian regression is given. SIMEX and naïve methods are compared to illustrate the effectiveness of the method in removing systematic bias from data measured with error. Bayesian optimization is added to reduce variance and bias even further.

In Chapter 5, cross-sectional and longitudinal metering sampling plans are discussed for lighting retrofit projects. This introduces the main case study of the thesis. A Dynamic Linear Model with Bayesian forecasting is used to quantify the meter sampling uncertainty for a large-scale project, and through forecasting to determine an efficient sampling plan for future sampling. The robustness of the plan to different future results is also explored.

In Chapter 6, an efficient longitudinal population survival survey sampling design method is presented. It uses a Dynamic Generalised Linear Model with Bayesian forecasting. A Mellin Transform moment calculation method and a genetic algorithm are then used for sampling optimization. Sampling optimization considers other sources of uncertainty and variance such as measurement error and variance in the hours of use of the luminaires as well.

In Chapter 7, the models from the previous two chapters are combined into an overall M&V sampling plan, where cross-sectional meter sample sizes are traded off against population survival survey sample sizes. Figure 7.1 illustrates the flow of the chapters and their contribution to the overall M&V plan.

Chapter 8 draws conclusions and recommends future research.

1.9 NOMENCLATURE

In this thesis, probability density functions are indicated by the notation $\sim [\cdot]$. In cases where specific density functions are used, an identifier is added, so that if x is normally distributed, for example, then it would be written as $x \sim N[\cdot, \cdot]$. The first term in the brackets is the first moment (the mean), and the second term the standard deviation. In other cases, square brackets will indicate a vector of data points or a range inclusive of the bounding values, but the difference should be apparent by the absence of the tilde ($\sim$) sign. The tilde sign is also used in the conventional manner to indicate "approximately", so that $\sim 5\%$ means "approximately 5%.

In a slight abuse of notation, $\Pr(\cdot)$ will indicate any probability density function. This is standard in Bayesian texts. It only indicates "the probability of...", and not a particular functional form.

The $|$ sign is used in the conventional manner to indicate a conditional probability, so that $\Pr(A|B)$ refers to the probability of A given, or conditional on, B.

The hat sign ($\hat{}$) is used to indicate an estimate of the true value so that $\hat{x}$ is the estimate of x. Similarly, the superscripted asterisk * indicates an observed value (measured with error), so that x^* is the observed value of x.

Boldface letters indicate multidimensional vectors, as opposed to normal letters which indicate scalars.

A few nuanced or subject-specific terms are used in M&V uncertainty quantification. In the interest of clarity, these are explained below. The terms are arranged by topic rather than alphabetically.

An M&V study is usually divided into two parts: the **baseline period**, and the **reporting period**. The baseline period is that period before an intervention is applied, when the pre-retrofit energy system is characterised. This information is used to create an adjusted baseline (a forecast) for the post-retrofit, or reporting period, against which the actual reporting period energy use is compared. It is called an **adjusted** baseline because the forecast baseline needs to be adjusted to the reporting period's Energy Governing Factors (EGFs). For example, a hot year should not be compared to a cold year for an air-conditioner retrofit study. Both years need to account for temperature, so that an accurate forecast

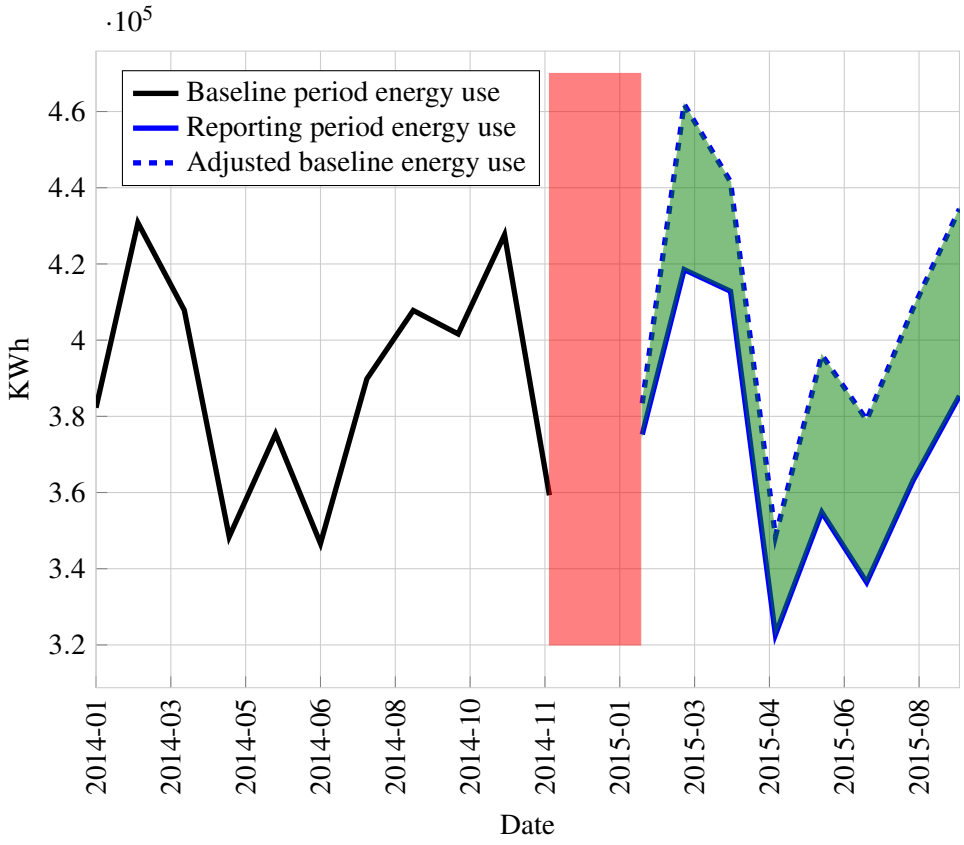


Figure 1.1. Illustration of M&V baseline and reporting periods, with adjusted baseline and savings indicated. Data from an actual University of Pretoria residence has been used and adapted. The red area is the installation period, and green represents the savings.

of what the energy use would have been, had no intervention taken place, can be calculated. This is illustrated in Figure 1.1.

An M&V **facility** is any energy system around which a boundary can be drawn. It can be a free standing building but may include more than one building, or only a wing of a building.

The definitive International Performance Measurement and Verification Protocol (IPMVP) [1] defines four methods for defining M&V project boundaries. Options A and B are **retrofit isolation** approaches. A boundary is drawn to include only the interventions or retrofits considered for a project. Option C is the **whole-facility approach**, where a facility's total energy use is considered – including non-project systems that might interact with the retrofits. Option C measurements usually rely on the utility meter or main incomer of the facility. Option D is **calibrated simulation**, which usually involves Building

Energy Modelling (BEM) software.

For lighting projects, incandescent lamps are often replaced with **Compact Fluorescent Lamps** (CFLs) during a retrofit. They are also known as 'energy savers'. Incandescent lamps can also be replaced with Light Emitting Diodes (LEDs), although CFLs will be considered for this study.

The **Project Developer** (PD) is usually an **ESCO** (Energy Services Company). They are responsible for identifying necessary EE and DSM interventions, and implementing them.

In electricity, the **power factor** is the ratio of real to apparent power. At a unity power factor, the real power in Watts is equal to the apparent power in Volt-Amperes, so that the $P = VI$ equation holds: power in Watts truly is equal to Volts multiplied by Amperes. However, when inductive or capacitive loads are present as the current and potential difference move out of phase, the power factor changes. This is because the power factor reflects the real-to-reactive power ratio. This phase difference is expressed in radians. Non-unity power factors are very common and are caused by electrical motors and power electronic circuits, which usually have inductive loads. Mismeasuring the power factor will have the net effect of changing the 'gain' of a meter.

Uncertainty reporting is usually done using the **expanded uncertainties** as per the International Standards Organisation (ISO)'s Guideline for the expression of Uncertainty in Measurement – the GUM. An expanded uncertainty expression reports a value with a given confidence and precision. For example "90% confidence that the value is within 10% of the mean" – known as the 90/10 criterion. Many guidelines require savings to be reported with expanded uncertainties such as 68/50, 80/20, or 90/10. Other guidelines use discount factors to ensure conservatism. These will be elaborated on in the literature review.

As mentioned in the previous section, lighting studies are usually simpler because there are usually no independent variables (also known as covariates or EGFs) to consider. Another way of describing it would be to say that in lighting studies there is **measurement** and **sampling uncertainty**, but little **modelling uncertainty**. These three kinds of uncertainty are usually mentioned when discussing M&V uncertainty quantification. While one usually associates measurement uncertainty in M&V with electricity or other meters, instruments measuring with error also include surveys and questionnaires [6], tracking databases, non-intrusive load monitoring, and inspection reports [7]. These instruments may

measure or record any number of variables such as occupancy [8], floor area, schedules, income, the proportion of Miscellaneous Electrical Loads (MELs) [9, 10], etc. Sometimes data such as plug load energy use are used as a proxy to measure occupancy [11]. More about this in Section 2.2.3.8. This thesis focusses on **energy metering uncertainty**, which is a subset of measurement uncertainty and is used to refer to the uncertainty in energy meters as opposed to the uncertainty in the measurement of other covariates in energy models.

Sampling uncertainty arises when the whole population of Energy Conservation Measures (ECMs) or facilities are not monitored. For example, when 100 000 Incandescent Lamp fixtures are retrofitted with CFLs in a residential mass roll-out programme, not all lamps can be tracked. In such cases, it is necessary to take a sample of the population to determine the energy saved by the project.

Modelling Uncertainty has to do with the fact that all models are merely an approximation of reality. Relevant EGFs may be omitted, or an irrelevant one included. Two EGFs may both be relevant, but may also be partially collinear. The correct EGFs may be considered, but they may be difficult to measure – occupancy is a classic example. Since modelling uncertainty inherits measurement uncertainty and adds other sources as well, various uncertainty typologies have been proposed [12–14].

The **CV** value, or coefficient of variance, is a normalised measure of the dispersion of the data and is calculated as $CV = \frac{\sigma}{\mu}$, where σ is the standard deviation, and μ is the mean. A CV of 0.5, therefore, denotes that the ratio of the standard deviation to the mean is 0.5. This is the traditional M&V assumption if nothing else is known about the data set [15].

Error is the difference between the actual and the measured value. **Random** errors are distributed symmetrically around the mean and usually follow a normal distribution. **Systemic** or non-random errors introduce **bias**. Bias "deprives a statistical result of representativeness by systematically distorting it" [16]. For example, biased data will consistently have a different mean to the true mean. Random errors usually do not have this effect, except in the case of **attenuation bias**, which will be discussed in Section 2.3.2.

Uncertainty is "the range or interval of doubt surrounding a measured or calculated value within which the true value is expected to fall with some degree of confidence" [17].

Precision relates to the "fineness of discrimination" [18] or "the closeness of agreement among repeated measurements of the same physical quantity" [17]. It is the uncertainty interval around a measured value, and should always be expressed with an associated statistical confidence. **Confidence** is a probability, whereas precision is a distance, or size, of the error band. Confidence and precision together usually define the broader term accuracy, which is "the capability of an instrument to indicate the true value of a measured quantity" [17]. Note that the above definition of confidence is popular although not technically correct [17, 19–21] unless Bayesian methods are used.

Homoscedasticity means that the variances in all data points are the same, or the variance on the residuals of a regression model is constant over the whole range of the input variables. Heteroscedastic datasets have non-constant variance.

This thesis uses **Highest Density Intervals** (HDIs), rather than standard equal-tailed Confidence Intervals (CIs). HDIs and CIs correspond exactly when symmetrical distributions such as the normal or Student's t-distribution are used. However, when asymmetrical distributions such as the beta distribution are used, there is a difference. The equal-tailed 90% CI (for example) cuts off the top and bottom 5% of the distribution range. However, when the distribution is skew (think about an exponential distribution), this will exclude the 5% most likely values, which are between zero and 5%, and include some unlikely values in the tail (between 90% and 95% of the range) even though they are far less likely. The HDI solves this problem by selecting the range by likelihood, not by the range of values. An illustration of the CI and HDI for an exponential distribution is shown in Figure 1.2.

Calibration is the process of comparing an instrument to a standard or reference (instrument) to characterise its errors and improve its accuracy. The range and kinds of values that should be compared are often codified in standards. **Disciplining** an instrument is a less complete calibration process where one only considers ranges and values expected to be encountered in a specific environment, and not the full range at which the instrument may be able to measure. Calibration is different from **qualification**, which ensures the quality of an instrument model range, because of its design and manufacturing process. For example, tests are done to ensure the stability of meter readings under different environmental conditions, specified by the International Electrotechnical Commission (IEC) [22–25]. Although a specific meter may be qualified because it is part of a model range and never lose this qualification, it may need to be calibrated periodically to compensate for drift.

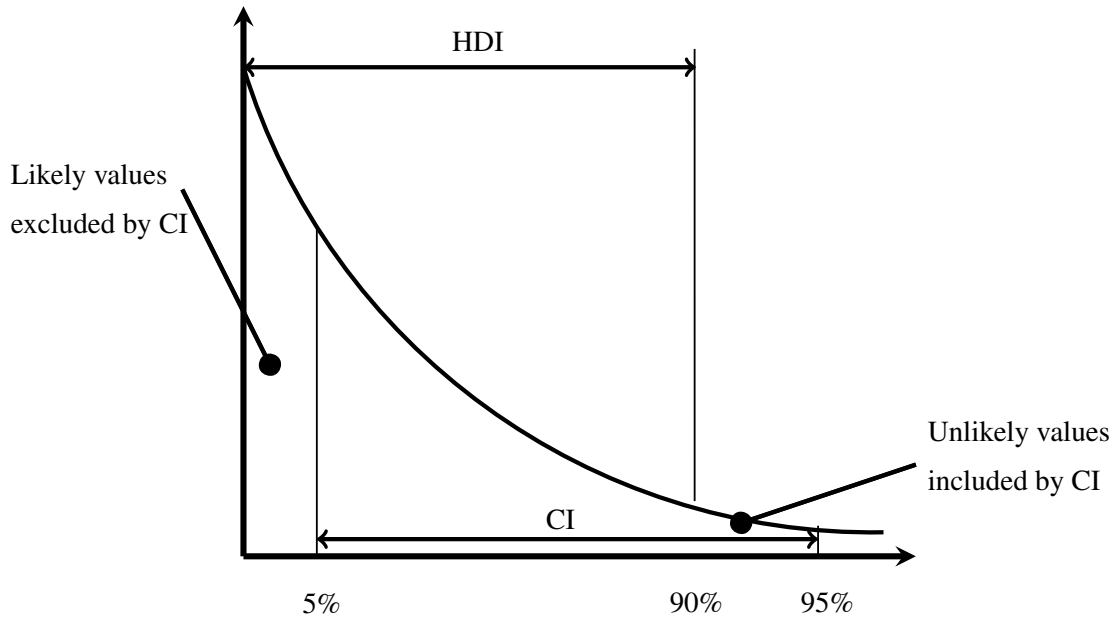


Figure 1.2. Illustration of difference between 90% equal-tailed Confidence Interval (CI) and 90% Highest Density Interval (HDI), for an exponential distribution.

CHAPTER 2 LITERATURE REVIEW

2.1 CHAPTER OVERVIEW

Since this thesis considers M&V uncertainty and how it relates to measurement and longitudinal sampling uncertainties, the literature survey and appraisal of the present state of the art is structured around these themes. The first section is a short introduction to the terms used in this thesis. The two main sections address relevant M&V literature, and general statistical methods applicable to M&V. Measurement uncertainty plays a critical role in calibration in later chapters. Therefore the sections addressing measurement uncertainty in M&V (specifically energy metering uncertainty), and statistical methods for addressing mismeasurement, are comprehensive. A discussion of Bayesian statistics is deferred to Chapter 3. [1]

2.2 MEASUREMENT AND VERIFICATION

This section will start by considering M&V guidelines, as these are foundational to the field. M&V uncertainty quantification will then be investigated, and the reasons and methods for uncertainty quantification assessed. Measurement uncertainty is then considered in detail, and finally persistence and longitudinal studies are investigated.

Note that a 'deemed savings' approach is sometimes adopted in M&V. In such an approach, each installed unit is deemed (or stipulated) to save a given (conservative) amount of energy per year. As such, no measurements are necessary - only installation verification. This merely is verification, and

[1] The current chapter is based on journal articles published as part of the author's PhD research [26–29].

will not be considered in this thesis unless the verification procedure is done to determine persistence as in Chapter 6.

2.2.1 M&V guidelines

Many M&V guidelines and protocols have been written, and each one reflects the different purposes of their context. This discussion will focus on the ones relevant to uncertainty quantification. The are the US Department of Energy's Uniform Methods Project (UMP) [30] and the American Society of Heating, Refrigeration, and Air Conditioning Engineers (ASHRAE)'s Guideline 14 (G14) [17, 31]. Both of these documents provide guidance on methods of data collection and validation. Both also prescribe methods for analysis, forecasting, and savings calculations with uncertainty quantification.

G14 has a 2002 and a 2014 version, the former being freely available. There is not much difference between the two, although the 2014 version does have critical typing errors in some of its formulae which have yet to be corrected. A useful summary by some of its authors has been written and applies to both [32]. G14 contains methods for considering correlated or uncorrelated residuals on the independent variables to determine savings uncertainty as a function of the fraction of total energy saved and other factors. These methods are based on Reddy and Claridge's fractional savings M&V approach [33], which has also been adopted by numerous other guidelines including the California Commissioning Collaborative [34]. G14 classifies projects according to the IPMVP schema described in Section 1.9, and provides uncertainty reporting requirements for each of these options. The Coefficient of Variance on the Root Mean Square Error (CV(RMSE)) and the Normalised Mean Bias Error (NMBE) or Net Determination Bias (NDB) are used to evaluate model goodness of fit. A survey of M&V professionals has found that these are considered the definitive goodness of fit measures for energy baseline models [35]. A method is then given for combining this (modelling uncertainty) with measurement uncertainty, sampling uncertainty, uncertainty in independent variables, and autocorrelation. It is a variation on the standard root-sum-of-squares method where $\sigma_{a+b} = \sqrt{\sigma_a^2 + \sigma_b^2}$, where σ denotes the standard deviation of the component uncertainties. G14 also provides excellent reference appendices detailing different regression approaches, instrument characteristics, and M&V plans. A few caveats should also be noted, however. The method is based on linear regression. Only lag-1 autocorrelation is considered, which is often not a realistic assumption [36]. Furthermore, the independent variable error

is added to the total uncertainty as a random error, and the errors-in-variables effect is not considered [2]. This could lead to significant bias and other effects, as explained in Section 2.3.2. Nevertheless, G14 is a valuable resource and one of the two leading M&V guidelines.

The UMP was commissioned by the US National Renewable Energy Laboratory (NREL), and its chapters were written by different experts. Each chapter creates a protocol for a different M&V technology or approach.[3] The chapter on sampling cross-cutting protocols was written by some of the same authors as the IPMVP's statistics and uncertainty guide [42], and is therefore similar. Although the UMP does contain calculations, it shows that accurate M&V is about more than applying the correct formula; the process leading up to the formula is often as important. This includes study design, data collection, validation, storage, and other aspects. The relevant chapters will not be discussed in detail here but will be referred to where applicable in the rest of this thesis.

Notable guidelines that do not provide requirements for uncertainty calculation but discuss uncertainty management generally include the IPMVP [1] on which the SANAS 50010 [3] is based, the State and Local Energy Efficiency Action Network's Energy Efficiency Program Impact Evaluation Guide [43], and the Federal Energy Management Protocol (FEMP) guideline [44]. The IPMVP does have a separate guideline for statistics and uncertainty that was published in 2014 [42]. It addresses sampling uncertainty but does not provide much detail on measurement and modelling uncertainties.

Greenhouse gas reduction programmes often require M&V. Vine *et al.* reported on different options considered for dealing with measurement uncertainty in such cases [45]. Although this was a work in progress in 2002, it is still relevant, since the debate around the advantages and disadvantages of different measurement approaches is explained well. Discount factors to compensate for the uncertainty of various methods are also listed. The scale of the United Nations Framework Convention for Climate Change's Clean Development Mechanism (UNFCCC CDM) methodology specifications dwarfs other M&V documentation. It contains over two hundred methodologies for different project scales and applications. Accuracy requirements vary, but the 90/10 criterion is most common, although Sonnenblick and Eto [46] have shown that this precision level is only necessary for projects where the

[2] See Section 2.3.2

[3] Chapter 2: Commercial and Industrial Lighting Evaluation, Chapter 6: Residential Lighting Evaluation, Chapter 9: Metering Cross-Cutting Protocols, Chapter 11: Sample Design Cross-Cutting Protocols, and Chapter 13: Assessing Persistence [37–41] are applicable to this thesis.

savings to cost ratio to be verified is small. In many cases, 90/50 is adequate for identifying project cost-effectiveness (that is, whether or not a project saved energy).

Shishlov and Belassen [47] provided a useful review of how monitoring uncertainty is handled in the CDM. For example, CDM AM0046 requires CFL Retrofit programmes to be monitored very stringently at the insistence of regulators, even requiring custom-made meters. Michaelowa, Hayashi, and Marr [48] who developed the methodology noted that no projects were completed under AM0046 until the alternative AMS II.C [49] was adopted. Later AMS II.J [50] was also adopted. In it, every CFL is deemed to operate for 3.5 hours/day, eliminating the need for measurement. Even so, they assert that there are still projects that would reduce emissions but are ineligible. These difficulties illustrate that measurement goals should always be construed in the larger project and social context. Achieving important individual statistical outcomes is never an end in itself. It may even hinder meeting overarching programme goals such as emissions reduction or development. Research on efficient sampling designs has been conducted to reduce the sampling burden as much as possible [51–53], although there is still much scope in this field. The CDM board is also working towards a stringency/cost trade-off system to replace the current system [54]. Table 2.1 summarises the comparison of these guidelines.

2.2.2 M&V uncertainty quantification

A mathematical description of M&V has been compiled [56], detailing the role of uncertainty in M&V. Uncertainty is usually quantified for three reasons, which will be considered in turn:

1. Compliance with a standard or reporting requirement.

2. Risk quantification and decision analysis when assessing project performance.

3. Efficient M&V study design.

Table 2.1. The treatment of measurement uncertainty in leading M&V guidelines.

Name	Year	Level of detail	Features	Reference
G14	2002, 2014	10	• Most comprehensive • Excellent methods • Instrument uncertainty database • Itemized measurement costs	[17, 31]
IPMVP	2012	5	• Introductory treatment • Sensitivity and Uncertainty Analysis	[1, 42] [1]
CDM	2015	8	• Approach varies between methodologies • Emphasis on being conservative • Discount factors for >95/5 measurement error • 95/10 for unknown measurement error • Deemed Savings also used	[4, 47] [48] [54] [54] [50]
UMP	2014	6	• Varies with authors of chapters • Errors-in-variables discussed in Chs 13, 23 • Metering error discussed in Ch. 9 • Survey error discussed in Ch. 11	[30] [41, 55] [39] [40]
SEE Action Guide	2012	4	• Practical guidance • Discussion of uncertainty and project risk	[43]
CCC	2012	6	• Appendix on uncertainty analysis • Adopts and simplifies fractional savings approach	[34]

2.2.2.1 Compliance and risk

Compliance calculations are those contained in the guidelines described above. From these, the risk of non-compliance may be calculated by an M&V team or PDs.

Besides the M&V guidelines, relevant research on this topic has also been conducted from a legal metrological perspective. Here measurement uncertainty and cost of non-compliance are traded off in a decision support framework. Crenna [57] and Pendrill [58, 59] used an MC method, while Fearn [60] used a more cumbersome analytical approach. Risk was viewed from a government perspective as a function of the cost of emissions to society, should a faulty meter be accepted. Sonnenblick and Eto also used this cost function in their report on the cost-effectiveness estimates of energy projects in the context of measurement precision [46], and Rysanek and Choudhary [61] used the marginal abatement cost: the ratio of net present value to GHG units saved. These metrics seem more rational than short-term financial risk measures when one considers the broader goals of energy research.

2.2.2.2 Project decision support through uncertainty quantification

Pendrill rightly observed that measurements are seldom made for their own sake, but rather in support of a financial decision [59]. Indeed, decision maker uncertainty about cost-effectiveness is the most frequently-cited barrier to the commissioning of energy projects [62]. However, the contribution of technical uncertainty in the performance of the ECM is usually smaller than economic uncertainty contributions [61, 63]. Project risk associated with measurement uncertainty has also been identified by both researchers and practitioners [64–66], but little M&V literature addresses this topic directly. In this section decisions under uncertainty for both M&V and Building Energy Modelling (BEM) are considered. Because these are difficult to separate, the distinction between the two is blurred at times, and BEMs or BEM data often form part of an M&V calculation. Economic project decision support literature will not be considered, as this is more concerned with economic aspects than is necessary for the current investigation. BEM studies are useful for M&V practitioners because the methods used and level of sophistication exceeds that of much of the M&V work being done, and can therefore serve as useful exemplars of uncertainty quantification in the buildings and energy field. It is especially relevant to IPMVP Option D (Calibrated Simulation).

Foundational work on uncertainty quantification for decision analysis in M&V was conducted in the US in the 1990s and is still relevant. In 1991 Violette presented an insightful discussion on statistical precision in DSM M&V projects and suggested that Bayesian statistics could be used to incorporate prior information from previous years [67]. In 1993 Violette *et al.* [68] presented a framework for cost-effective parameter determination for a lighting retrofit project. Sonnenblick and Eto's technical report for Lawrence Berkeley National Laboratory (LBNL) provided a definitive 'framework for the evaluation of the cost-effectiveness of utility DSM programs' [46]. It therefore considered only the uncertainty limits necessary for determining whether or not a project saved more money than it cost (levelised project cost vs. levelised savings), which is a less stringent requirement than quantifying those savings. Nonetheless, the report is insightful in the way that it applies decision theory (which hinges on uncertainty quantification) to M&V.

Goldberg [69] adopted a similar value of information approach to energy monitoring, weighing the cost of measurement against potential benefits when the buyer and seller have different perceptions about the value of the project. She was also one of the first researchers to present a coherent M&V sampling design framework, similar to what was adopted by the IPMVP [1] and UMP [30] in later years.

There is only one recent attempt to quantify the risk due to energy metering uncertainty [70, 71]. However, this calculation was much too simplistic and was presented by a marketing manager of a meter manufacturer calling for even-more-stringent standards to which the latest meters could be qualified. This standard is unnecessary since the current Class 0.2S energy meters are the smallest uncertainty sources in almost any conceivable project, and their uncertainties can already be neglected for risk calculation purposes in many cases [72].

Regarding project decisions under measurement uncertainty, research into uncertainty in BEMs has increased dramatically in the last ten years. This is because it has been recognised that considering model input uncertainty is essential to identifying which ECMs should be implemented.

Because uncertainties are naturally and easily quantified in Bayesian statistics, its applications have proven to be a fruitful area of research in recent years. Although most of it is not applied to M&V specifically but rather to project decisions, some of the methods are still relevant to the current discussion. Georgia University of Technology's research group led by Augenbroe has produced the

most notable research in this field under a risk-conscious design and retrofit research programme. Heo, now at Cambridge, started under Augenbroe and will also be referred to often for her work on uncertainty quantification using semi-Bayesian Gaussian Process (GP) methods.

Riddle and Muehleisen provided a useful introduction to building calibration with Bayesian models [73], and Heo has recently presented an overview of building simulation models under uncertainty, as well as an introduction to the Bayesian approach [74]. Note that in a Bayesian framework, measurement, sampling, and modelling errors are considered simultaneously, although they remain distinct [12].

Considering decision theory, an introduction in the context of energy projects was provided by Wang *et al.* [75]. An insightful cost-benefit trade-off for chilled-water system design in the context of uncertainty [76] influenced the G14 [31] approach, which also supplies elaborate tables for determining measurement costs for different instruments in various project scenarios, although it does not calculate risk adequately [77]. Research on financial decision support related to EPCs with project uncertainty and risk have been conducted from an economic perspective using MC analysis [78] and other techniques [79]. The US Department of Energy (DOE)'s *EnergyPlus* software is usually used [80]. Deng *et al.* [81] provided a useful summary of the design of EPCs under uncertainty and presented a relatively sophisticated EPC decision model [82]. Measurement uncertainty is not considered explicitly in these cases, although it can be incorporated without many extensions.

A full review of building simulation calibration literature is beyond the scope of this survey, which focusses on uncertainty quantification and related methods. For a broader view, a useful starting point is Reddy *et al.*'s research as part of ASHRAE's investigation into calibrated simulation in RP-1051 [83–86], and Coakley, Raftery, and Keane's more up-to-date review, considering uncertainty in detail as well [13]. Heo's PhD thesis also provided an in-depth discussion and case study of one approach [77].

Databases of parameter uncertainties have been compiled [87], and these, or results from the literature, are used for uncertainty analysis or quantification. The key problem however, is that doing an MC simulation considering all parameters simultaneously is infeasible due to the curse of dimensionality. Sensitivity analysis methods are thus needed to reduce the number of parameters to a feasible figure. Sun et al. provided one of the better discussions on this topic [88], and Tian also wrote an informative

review [89]. Several excellent examples of this process have been published, and are summarised in Table 2.2.

Most building simulation research accounts for varying input parameters through uncertainty and sensitivity analysis. However, much of this research concerns itself with how varying the input parameters changes the output, but not how *variance in* the input parameters affects the output. As will be demonstrated in Section 2.3.2, random zero-meaned noise in the input parameters do more than add uncertainty to the output, and this should be taken into account. It is possible that this is accounted for in Gaussian Processes (GPs), although it is uncertain.

Two related studies deserve mention. To alleviate the burden of MC computation for building simulation studies with large uncertainties and many options and combinations, Rysanek and Choudhary proposed a lightweight non-probabilistic decision approach [61]. These scenarios apply more to simulation (modelling) uncertainty rather than measurement uncertainty. On the other side of the spectrum, Sandal *et al.* reported a machine learning and supercomputer-based method to alleviate the modelling burden by pre-tuning simulation inputs to extant data for standard US buildings [90]. This speeds up model building significantly.

Heo and Augenbroe have built up a noteworthy body of work on building simulation covariate calibration and uncertainty analysis using semi-Bayesian GP methods [91, 92]. Quantitative risk analysis for decision support in retrofit project planning was then explored with a focus on the accuracy of the simulation rather than metering decision making [93]. Their latest research incorporates this into a scalable methodology whereby more optimal retrofit decisions can be made, given uncertainty in input parameters [94]. Along similar lines, a lightweight and reasonably accurate alternative to the GP has been proposed [95]. Another notable contribution has been made by Tian *et al.* who used sophisticated data analysis and Bayesian methods to show the relative importance of different data on building calibration, and the robustness of the Bayesian method to missing input data [96]. Bayesian methods have therefore been demonstrated to deliver very good estimates, but Heo notes that even if this were not the case, they could still be superior to deterministic models since they quantify model prediction uncertainty distributions [91].

Table 2.2 presents a summary of key uncertainty quantification publications in BEM. The techniques used can also be applied to M&V in many cases.

2.2.2.3 Optimal M&V study designs

The guidelines discussed above provide many examples and formulae for simple, once-off M&V designs. However, these sample sizes are not efficient for multi-year or longitudinal projects such as the ones considered in this thesis. That being said, the only research on optimal M&V sampling designs that the author is aware of, has been conducted in his research group. Some of it was on modelling and sampling uncertainty [97], although most was by himself and a colleague. Much of this work was on the longitudinal CFL retrofit M&V problem [51–53, 98, 99]. Although this research laid the groundwork for understanding the problem, there are significant limitations.

The method reduces planned future sample sizes in two ways. First, by aggregating results in different years. Second, by reducing sample sizes through the Finite Population Correction (FPC) factor, for later years where the population size declines because of failures. The first factor may be refined. Metering results from a meter installed in year one should not be added to the result from the same meter at the same facility in year two as if they were independent samples (or strata) from a larger population. R.T. Cox's simple definition of independence is that "knowledge about sample one should be irrelevant for reasoning about sample two" [100]. This is not the case for consecutive samples from the same facility: year one's energy use would be a good starting point for reasoning about year two's energy use. Therefore, 34 metering results from year one should not be added to 34 metering results from year two, so that the total sample size is 68. Due to serial correlation (autocorrelation), samples in year two will contain less information than samples in year one. G14 [17] suggests using an autocorrelation correction factor for lag-1 autocorrelation.

The second factor used previously to reduce meter sample sizes is FPC. However, FPC only becomes relevant for population sizes below 1 000 and is therefore not applicable to the large-scale studies considered.

The previous method also assumes that the means of the metering results for all years are stationary. The method proposed in this thesis does not make that assumption.

Low-cost meters with lower accuracies were also selected for low-CV populations [99]. However, high-accuracy meters only enhance the overall accuracy in low-CV cases, when process variability

Table 2.2. Recent and notable non-BEM energy project decision support literature considering input uncertainty. The abbreviations can be found in the list at the beginning of this thesis.

Author	Year	Application	Sensitivity Analysis	Uncertainty Analysis Method	Decision Analysis Metric	Reference
Sonnenblick, Eto	1995	Owner, ESCO profitability	-	MC	CE	[46]
Kammerud, Gillespie, Hydeman	1999	Chilled water system design	(Taylor Series Expansion)	Quadrature	Discounted cash flow CE	[76]
de Wit, Augenbroe	2002	Thermal comfort	-	Bayes	Expected Utility	[101]
Pendrill, Källgren	2006	Exhaust gas analysers	-	Analytical	Cost to society	[58]
Crenna, Rossi, Bovio	2009	Water meters	-	MC	Non-conformance cost	[57]
Jackson	2010	EE Investments	-	MC	Value at risk	[78]
Burhenne, Tsvetkova, Jacob, Henze, Wagner	2013	BEM	Sobol' Sequence	MC Filtering	NPV-CE	[79]
Lam, Yik, Chan	2013	EPC	Differential: Influence Coefficient	MC-LHS	Savings shortfall	[80]
Sun, Gu, Wu, Augenbroe	2014	HVAC sizing	MC-LHS, LASSO, ANOVA	MC	Unmet peak load percentage	[102]
Heo, Augenbroe, Graziano, Meuhleisen, Guzowski	2015	BEM	Morris, MC-LHS	Bayes	% savings, EV(savings), fifth quantile savings predictions	[94]
Lee, Lam, Lee, Chan	2016	Chiller replacement EPC	Correlation analysis	MC-LHS	EPC compliance PDF	[103]
Wang, Augenbroe, Kim, Gu	2016	Occupancy	MC-LHS, LASSO, Variance-based	Bayes	-	[8]

Table 2.3. Recent and notable BEM energy project decision support literature considering input uncertainty. Note that goodness-of-fit as an outcome is not included. These usually employ CV(RMSE) and NMBE. The abbreviations can be found in the list at the beginning of this thesis.

Author	Year	Application	Sensitivity Analysis	Uncertainty Analysis Method	Decision Analysis Metric	Reference
Reddy, Maor, Panjapornpon	2007	BEM	MC-LHS	MC	GOF	[85, 86]
Corrado, Mechri	2009	BEM	Morris, MC-LHS	MC	-	[104]
Heo (PhD Thesis)	2011	BEM	Morris, MC-LHS	Bayes-GP	EV, CE with payback time, Guaranteed Savings, Savings Curve Score	[77]
Tian, Choudhary	2012	BEM	MC-LHS, SRC, MARS	LP, Bayes	EUI	[105]
Booth, Choudhary	2013	BEM, Decision Support	Factorial Sampling	Bayesian regression	NPV-PDF, Multi-criteria decision utility, CE, CEAC	[106]
O'Niell, Eisenhower	2013	BEM	SVR-GK, RS, derivative-based	quasi-MC	GOF	[107]
Manfren, Aste, Moshkar	2013	BEM and M&V	DOE (min, max, mean)	Piecewise regression, Bayes, GP	GOF	[108]
Rysanek, Choudhary	2013	BEM	-	Non-probabilistic	CE; MAC: NPV vs. GHG emissions saved, Discounted payback period vs. required capital; Wald's Minimax, Hurcwiz's Maximin, Savage's Regret	[61]
Sun (PhD Thesis)	2014	BEM, HVAC	MC-LHS, LASSO, ANOVA	MC	CRPS, PIT	[109]
Li, Augenbroe, Brown	2016	BEM	MC-LHS, LASSO	Lightweight Bayes-GP with stepwise linear regression	GOF	[95]
Tian, Yang, Li, Wei, Pan, Li	2016	BEM	Sobol' Sequence, SRC, RFVI, CCA	Bayesian	GOF	[96]

plays a smaller role relative to energy metering uncertainty.[4] Energy meters with accuracies of 0.01% and 0.002% were used. Such meters do not exist. Class 1 meters (standard utility meters) have an accuracy of 1%, and high-precision Class 0.2S meters have an accuracy of 0.2%. However, as will be demonstrated in this thesis, there is no advantage in using a Class 0.2S meter rather than a Class 1 meter for sampling a population with a CV of 0.5. Also, the time resolution of the meter does not refer to how often the meter measures current and voltage, but rather the time period over which the meter integrates when storing a data point [110]. The measurement interval is shorter than the integration interval. The integration interval can also be set, and is not five minutes as is supposed for a Class 1 meter.

Furthermore, if meter accuracies are considered, Current Transformer (CT) accuracies should also be added, as these uncertainties can be more significant than the meter uncertainty itself. This is considered in Section 5.4.1.2.

Regarding optimization, gradient-descent methods were employed previously. However, the optimization function is an integer non-linear program (INLP) with discontinuities, as shown in the author's Master's work [111]. Heuristic methods will therefore be used to provide more reliable results, as discussed in Section 5.4.1.3. The plans devised in the previous studies should also not be called 'optimal', since optimality cannot be guaranteed by the gradient descent method or the heuristic used in this thesis. It is more accurate to refer to 'efficient' solutions, as many efficient solutions to a sampling problem may exist.

Last, the earlier method assumes that the proportion of lamps surviving at a given point in time is known with certainty, and does not combine this survey sampling uncertainty with the meter-sampling uncertainty. Survey sampling uncertainty was characterised in previous work [28], and will be incorporated in Chapter 7.

For these reasons, there is an opportunity to improve upon the current method using Bayesian statistics. This would incorporate prior information in a mathematically rigorous way to reduce the longitudinal monitoring burden.

[4] See Section 4.2.

2.2.3 Measurement uncertainty

As a subsection of M&V, measurement uncertainty is now considered in some detail, as it is the most neglected of the three kinds of uncertainty in M&V (the others being sampling and modelling). It therefore represents an opportunity for cost reduction if it is understood properly.

2.2.3.1 Meter uncertainty as a component of M&V uncertainty

In South Africa, measured and verified energy savings achieved by businesses are eligible for tax deductions according to the 12L tax incentive [2]. However, measurement devices used for such projects need to be calibrated by accredited laboratories. This is a sound principle and has been adopted by many other agencies as listed by Ahmad *et al.* [112]. However, it greatly increases measurement costs, which can make M&V prohibitively expensive and reduce the number of feasible projects significantly, as in the CDM case [47, 48]. Given the small contribution to overall uncertainty made by electrical meters, especially when sampling is done,[5] such requirements may be counter-productive. Overall accuracy requirements could be better served by spending the funds on obtaining a larger or more detailed sample, or measuring independent variables more accurately.

2.2.3.2 Measurement uncertainty in M&V literature

Measurement uncertainty is acknowledged in M&V literature, although firm guidance is seldom given. The IPMVP [1, 42] provides general guidance on uncertainty but does not address measurement uncertainty in much detail. The UMP [30] is the only guideline to discuss mismeasurement at all. ASHRAE Guideline RA96: *Engineering Analysis of Experimental Data* [113] also deserves mention. It is a general quantitative introduction to handling measurement uncertainty in engineering measurements and could be applied to some M&V cases.

[5] See Chapter 4.

2.2.3.3 Energy metering uncertainty

Energy metering uncertainty can be dominated by other uncertainties such as sampling or modelling [72], but can nonetheless be significant depending on the application. Five cases will be considered below: general energy metering uncertainty, sub-metering and its contribution to measurement uncertainty, how power quality affects energy metering uncertainty, virtual instrumentation, and the possibility of in-situ meter calibration.

Regarding energy metering uncertainty, static (solid-state) electrical energy meters used for reporting purposes have to be qualified to standards set by the IEC, or its national equivalents, such as ANSI C 12-20 [114] in the US. Metering classes indicate maximum allowable percentage errors over the majority of the measurement range, so that a Class 1 meter is 1% accurate, for example.[6] A graphic illustration of the accuracy requirements is shown in Figure 2.1. Close attention should be paid when acquiring meters, as accuracy class (mis)specification has also been abused as a marketing tool, as catalogued by Irwin [70]. M&V professionals should also note that influence quantities such as harmonics are tested for, but in a one-at-a-time fashion, with all other quantities held at default levels.

Even when meters are qualified to these standards, errors or bias can be introduced by environmental conditions. For example, even though temperatures in Saudi Arabia still fall within IEC specifications, systematic bias is introduced due to consistently abnormal values [115]. Even such small biases on revenue meters metering large installations can lead to significant billing errors.

The discussion above applies to the meter itself, but not to the CT which is often used to measure the current. In many cases, CT accuracies are lower than the meter accuracies. An example of CT accuracy specifications can be found in Figure 2.2. They need to be considered separately from energy metering uncertainty and added using the sum-squared error method. In many situations, the accuracy class of the CT and meter, together with their rated currents will suffice to determine the overall accuracy of the measuring system.

Rogowski coils have also become popular in M&V [110]. These devices are flexible wires that can be threaded around the conductor of interest and are more accurate than split-core CTs. They also

[6] IEC 62053-21 [22] refers to Class 1 and 2 (active), 62053-22 [23] to Class 0.2S and 0.5S (active), and 62053-23 [24] to Class 2 and 3 (reactive) meters.

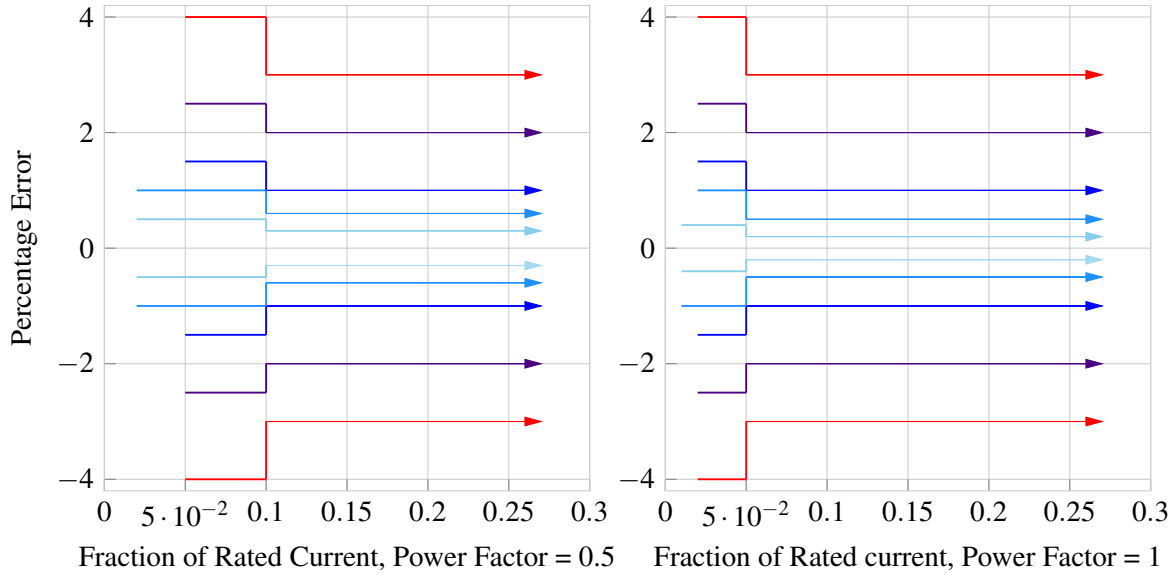


Figure 2.1. Comparison of different IEC accuracy class meters [22–24] for transformer-connected single or polyphase meters with balanced loads under sinusoidal conditions. The meter classes, from the inside to the outside, are Classes 0.2S, 0.5S, 1, 2, and 3.

do not saturate the way CTs do but have a multiplicative error proportional to the current in the conductor.

Although accuracy influences meter prices, the communication protocol used by the meter is also significant, as shown by Ahmad *et al.* in their review of energy and related sensors [112].

2.2.3.4 Sub-metering

Sub-metering an installation often provides valuable insight into the main load drivers but can be expensive if revenue-accuracy meters are used. One can consider less accurate and less costly options in these applications.

Plug-through meters are popular for metering Miscellaneous Electrical Loads (MELs). Polese *et al.* provided a comprehensive case study detailing the challenges in implementing such a solution at a large retailer, for an NREL project [116]. The study demonstrates the inaccuracy of such meters, as well as other factors that contribute to general measurement uncertainty. In this study, 41% of the

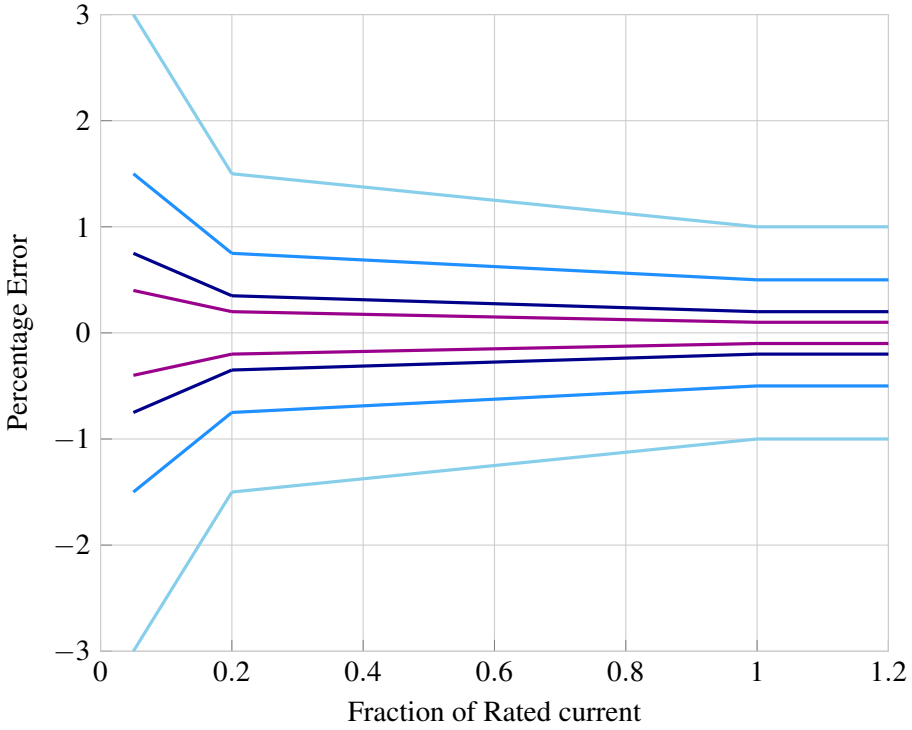


Figure 2.2. Instrument Current Transformer accuracies according to IEC 60044-8 [25]. From the inside, Classes 0.1, 0.2, 0.5, 1 are indicated. For Class 3 and Class 5, the limits are flat at 3% and 5% respectively.

meters had significant portions of the data series that were erroneous. Errors of 20% in the range 0-20 W were common, and 6% in the range 25-100 W. Given that 40% of the MELs operated below the 60 W level, these errors are significant.

Stick-on Electricity Meters (SEMs) represent an exciting new low-cost measurement or logging option [117]. These sensors are placed on the circuit breaker in the distribution board, and senses when current is drawn on the circuit. It is important to note that these do not work where relays are present.

Current-only meters are becoming a popular option for residential metering. They usually use split-core CTs and are much more affordable than revenue energy meters, but are not as accurate, or even qualified. In personal correspondence with a popular meter manufacturer based in the UK, the accuracy was quoted as 10% [118]. Given that they operate in a narrow environmental and electric range, this is usually not of great concern, provided that they can be verified in some way. However, they

can not be recommended as the sole meters used for projects. The voltage may vary due to supply-side fluctuations, or due to facility-level demand factors. On the demand side, current-only meters multiply their readings by a nominal voltage. The resultant power measurement is in Volt-Amperes: apparent power, not true power in Watts. The power factor is thus assumed to be unity. Inductive power electronic equipment found in most households will decrease the power factor to below one, biasing the measurement by the power factor. On the supply side, the utility voltage is seldom at the nominal level. It is regulated to be in a certain range [119]. In Europe, utility supply voltage is determined to be 230V$\pm$10% [120], and in the United States, 120V$\pm$5% [114]. However, certain asymmetrical tolerances may also hold. For ANSI C84.1 Range B [121], these tolerances are -13% and $+6\%$. These asymmetrical tolerances may skew the calculation since under-voltages are higher and possibly more likely than over-voltages.

For the symmetrical tolerance case, it may be argued that unmeasured variations would cancel out over time. However, a constant voltage offset may also apply. The supply voltage at a facility such as a house varies with a number of factors. These include the distance of its distribution transformer from the substation on the primary feeder, the distance between this house and the transformer on the secondary feeder, the number of facilities on the secondary feeder, and the load on the feeders. The average incomer voltage at a house on the edge of a distribution network may be at the lower end of the specification interval, while a facility closer to a transformer may be at the higher end of the interval. Therefore the distribution of voltage for a single facility may not be symmetric around the country's nominal voltage, biasing the measurements for which a nominal voltage was specified.

2.2.3.5 Power quality

Power Quality is an important consideration in energy metering uncertainty calculation, although M&V literature does not discuss it very much. The IEC standards qualify meters only for sinusoidal conditions, but on networks with modern power electronic equipment, this assumption is usually invalid. The harmonics which cause the non-sinusoidal condition may originate from some modern power electronics sources, such as Variable Speed Drives (VSDs), fluorescent lamps with electronic ballasts, switching power supplies, or controlled rectifiers [122]. These harmonics are generated by loads on the network but are observed as a supply-quality problem when measured. For certain cases where the customer pollutes the power network with large harmonic power flows, the presence of

harmonics may skew the reactive energy measurement to such an extent that a power factor greater than unity is indicated, even if this is not the case at all [122].

These conditions then lead to mismeasurement in static energy meters, especially when a non-unity power factor is present [123–125]. This does not apply to older electromechanical induction meters, but only to solid-state (static) smart meters [124, 126]. Berrisford provides an accessible and practical introduction to this problem [127]. Literature reviews of this field have been conducted [128, 129], although this thesis will focus on M&V applications.

The problem with measuring non-sinusoidal loads is that reactive power is calculated and defined in numerous ways [125, 130]. Although the different formulas give the same result under sinusoidal conditions, they differ when harmonics are present. Current magnitude and power factor are the main uncertainty drivers [122]. An example of this inaccuracy has been documented in the field [127]: an approved Canadian meter using Budeanu's power definition [131] was replaced by an approved Canadian meter using Fryze's power definition [132]. This resulted in a power factor penalty being added to the customer's bill when the meter was changed, even though the energy use did not change. Further investigation revealed non-sinusoidal conditions due to the harmonics generated by the client's VSDs, which the meters measured in different ways. Some of the inaccuracy noted by Polese *et al.* [116] in their metering of a retailer with many MELs may be due to such effects.

Because of these different definitions and different calculation methodologies among different meters, Cataliotti *et al.* [124, 130, 133] recommends that when calibrating a meter in-situ, a reference meter implementing the same metric as the Unit Under Test (UUT) should be selected, so as not to compound the errors. If the manufacturer does not state the metric used, methods for determining it experimentally have been devised. However, it was found that in such a case, the UUT only adheres to the accuracy limits set in the standard when compared with the reference meter adopting the same power definition, not with the true energy value.

There is, however, a course charted through the reactive power-definition confusion. The IEEE Standard 1459 (2010) [134] gives guidance on how reactive power should be defined and calculated. The consensus among most of the papers cited here is that this definition should be adopted. It is also endorsed by the IEC. Berrisford has demonstrated that reprogramming certain kinds of digital watt meters in minor ways can lead to them calculating power according to the IEEE 1459 definition [127].

Although utilities do not itemise harmonic distortion on the bill, preliminary work is being done to prepare the way for future considerations [135, 136].

M&V professionals should use meters measuring so-called 'fundamental' quantities, from which the true reactive power may be calculated according to the IEEE 1459. Meters with sampling rates adequate for including relevant harmonics should be selected, although increasing the sampling rate increases the price of the meter significantly in the range 0-80 μs [137].

2.2.3.6 Analogue to Digital Conversion (ADC) and Virtual Instrument measurement uncertainties

Most modern static meters employ ADC (also known as Digital Signal Processing, or DSP). ADC is also used in Virtual Instrumentation (VI), where a transducer is connected to a personal computer via a Data Acquisition (DAQ) board, for user-built DSP software to process [138]. Note that VIs can measure any analogue signal on which to perform ADC and that the general uncertainty principles remain the same. This field shows great promise for lower cost calibration and measurement of electrical signals for M&V purposes.

ADC technology is useful in electrical measurements as it has the potential for measuring true reactive non-sinusoidal power accurately, as discussed in Section 2.2.3.5. However, various standards specifying different parameters for ADC exist. Spataro [139] notes that ADC uncertainty has been quantified by the ISO GUM uncertainty propagation law (through a Fast Fourier Transform) [140], random-fuzzy variables [141], and MC approaches [138]. Due to the difficulty of convolving different uncertainty distributions analytically, such numerical methods make sense. This thesis would add MTMC [142–145] to the list. These require any number of different variables, depending on the standard and method employed. Spataro identifies that only offset (bias), gain, Total Harmonic Distortion (THD), spurious tones, and the Signal to Noise Ratio (SNR) are needed to quantify power quality. The details of such errors depend on the electronic components of the DAQ itself, but such systems can reach standard-level accuracies at a fraction of the cost [146]. They are thus expected to increase in popularity as they become commercialised [140]. In any event, the uncertainties introduced by ADC is usually much smaller than those of the transducers themselves [138]. The most recent

results in this field comprise a detailed theoretical model with experimental results for a DAQ-based sampling watt meter, based on the definitions set out in IEEE 1459 [128].

2.2.3.7 In-situ meter calibration

Due to the Measurement Instrument Directive (MID) ratified by the European parliament in 2004 [147], European meters (gas, water, electricity, etc.) need to be calibrated under actual conditions, interpreted as the actual meter installation location [148]. This has lead to various studies of how such a calibration may be achieved. Femine *et al.* [148] have devised a scheme for a field laboratory with a travelling standard. Power generated by the laboratory then allows a set of tests to be conducted at the facility. The directive has been viewed as impractical since not all plants can be shut down for such a procedure, metering cost increases drastically with a call-out for a portable metrology laboratory, and man-hours needed to test all Italian meters twice-yearly is unrealistic [149]. To offset this burden, Amicone *et al.* proposed a low cost, stable, 'add-on' calibrator that can be activated twice yearly to perform the necessary calibration [149]. Crenna *et al.* [57] considered the MID as a step toward the modernization of legal metrology. They considered water meters and proposed an MC approach based on statistical metrology and risk techniques, similar to Pendrill and Källgren's work on CO_2 meters [58,59] discussed in Section 2.2.2.2. This seems by far to be the simplest and most affordable proposal, although it relies on large quantities of manufacturer data and does not address all the concerns raised by the other authors. Meter ageing and water temperature are considered as influence factors similar to power factor and harmonic distortion for energy meters, although the analogy is not close enough to use the method as-is in electrical applications.

Measurement accuracy and its place in the smart grid are being investigated [150] and was proposed in rudimentary form a decade ago [151]. As smart meters become more common and interconnected, network cross-calibration to relieve the burden of calibrating every single meter may become a possibility, and represents an opportunity for future research.

2.2.3.8 Measurement uncertainty for non-electrical parameters

Often, non-electrical variables are also included in the energy model. Table 2.4 details typical errors for such cases. This is especially common when whole-facility regression models are constructed

using measurements of variables such as temperature [152], occupancy [8, 11] or flow rate [57]. Besides the error in the meter itself, poor meter selection, placement, or misestimation of independent variables may also contribute to unquantifiable errors in this case [42]. For example, the flow rate and temperature in a duct vary between the edge and the centre and features such as elbows impact flow and heat transfer characteristics for a non-negligible downstream portion of the duct. Because of these complex interactions, it is useful to work with general error estimates such as those found in G14 [31]. However, even these values should be used with caution. For example, CO_2 sensor accuracy was investigated [153] and the authors found that only seven of the eighteen sensors had errors of less than 20% at standard CO_2 levels for classrooms – a much higher value than that specified by G14.

Occupancy is a key factor in building energy use but is notoriously difficult to measure and model. Combinations of reed switches and passive infra-red (PIR) sensors seem to work well for offices [154], but these are very simple environments with single occupants per room. For more complex situations, proxies such as blind, fan, light, thermostat, door, or other sensors are used, although these are imperfect [155, 156]. Wang *et al.* [8] have shown in a sophisticated study that occupancy was not a significant energy use factor for their case study building. However, the building in question used a centrally controlled independent HVAC system, and this result is to be expected.

Occupancy models usually compare forecasts to data measured with error. However, as long as the measured variable predicts energy use well, the measurement error or true occupancy is not significant for energy models, unless occupant behaviour is being investigated.

Table 2.4. Instrument uncertainties for M&V Applications. Note that many of these values come from ASHRAE Guideline 14-2002 Appendix A5.6 [31], and are quoted at the 68% confidence level for this source. Guideline 14-2014 values are unchanged unless otherwise noted. Furthermore, Guideline 14-2014 stipulates these as minimum requirements, rather than typical values, but also recommends that they are used if no other values are available (Section 4.2.11.2). The confidence level for the other sources is unspecified or complex, and readers are referred to the original documents for more complete descriptions. FS denotes a percentage of full-scale.

Quantity	Type	Guideline 14	Other Source
Temperature	Ambient outdoor portable electronic	2-5%	

	Domestic water portable electronic	2%	
	Air ducts	5%	
	Pipes and ducts	2-5%	
Air velocity	Indoor: non-mechanical or blower door	5%	2-5% [112]
	Handheld anemometer	10%	
	Recording anemometer	5%	
	Meteorological grade anemometer	2%	
	Air ducts: array	2-5%	
Pressure	Gauge	0.25-2%	
	Ducts	1-5%	
	Pressurization/depressurization	3-5%	
Energy	Electrical Energy meter	1%	0.2-0.5% [22–24]
	Current Transformer	2-3%	0.2-3% [25]
	Portable Watt meter	1-5%	
	Current: low cost home energy		>10% [118]
	Stick-on Meter		5% [117]
	Plug-through meter		20% [116]
	Relative humidity	2-5%	4.5% [112]
	Energy meter (gas)	1%	
Flow rate	Bucket and stopwatch, portable meter/probe	5%	<1-5% [1]
	Domestic, accumulating	1-2%	
	HVAC inline or insertion meters	2%	<1% [1]
	Ultrasonic, flare		2.5-5% [157]
	Smokestack gas		5-20% [158]
Run-time	Permanent	1-5%	
	Portable	2-5%	
Light	Sensor / logger		8-10% [112]
Other	Pyranometer	2-5%	>10% [159]
	Door position	2%	
	RPM	1%	
	CO_2		>20% [153], 4% FS [112]
	Combustion	2%	∼0.5% [59]

2.2.4 Persistence and longitudinal M&V

In M&V, persistence refers to the effective useful life of an installed measure. In a 1991 M&V guideline by the Oak Ridge National Laboratory (ORNL), the authors noted that "Persistence is a genuine problem of undetermined scope. Its effects on cost-effectiveness, program planning, and resource reliability are clear. It is now time to address persistence in earnest" [160]. In a 2015 article [161] and a 2015 technical brief by the LBNL [162] similar comments were made.

For this thesis, only technical persistence, or equipment lifetimes, will be considered. The curves used may hold for overall persistence as well, but this is not proven. Laboratory tests and equipment lifetimes are not equivalent to actual persistence in the field [163], and studies should also account for human- and market-related factors [38, 164], although the UMP recommends that such factors not be taken into account for residential monitoring programmes [38]. For a foundational introduction to persistence study design, see Vine [165], and for updated treatments, see Hoffman *et al.*, Skumatz, and the UMP [41, 162, 163]. One engineering rather than statistical approach is to use technical degradation factors popular in the US [162, 166]. These are the lifetimes of the measures relative to the original equipment installed [43] – a single number, rather than curve characterisation or longitudinal studies as implemented in this thesis.

Since primary persistence research is expensive, secondary sources are often used [41, 163]. Primary research studies usually do not track populations, but try to provide a median measure life estimate [167, 168]. Two notable exceptions are the Polish Efficient Lighting Project (PELP) [169] and the Lighting Research Centre (LRC) at Rensselaer Polytechnic Institute's Specifier Report on CFLs [170]. These data sets will be used. Another reason for selecting CFLs as the application technology for this study is that CFL retrofits are often used as M&V case studies [1, 30, 68, 69] as it is a well-studied technology with relatively simple principles.

Regarding the curve shapes, The CDM recommends a linear decay curve [50]. Logistic decay curves similar to those used in survival analysis have also been introduced [51, 169, 171] and later improved upon [53] to the form used in this paper to fit the data sets referred to above. Logistic curves are widely used in reliability engineering and applied to many technologies besides CFLs [172]. The techniques in this thesis will, therefore, have broader applicability. More examples of linear and non-linear survival curve assumptions and study results for EE appliance models are listed by Young [173].

Persistence monitoring requirements range from 3.9 years [5] to 10 years for CDM lighting projects [4]. Pennsylvania and Texas require 15 years [162]. This thesis will consider 10 and 12-year studies, as this reflects both regulatory requirements and realistic CFL lifetimes.

2.3 STATISTICAL METHODS

2.3.1 General M&V statistics

Besides the methods contained in the guidelines and Reddy and Claridge's fractional savings method alluded to already [33], as well as general regression and statistical literature (such as Montgomery [174]), other recent M&V work should also be noted.

The Bonneville Power Administration has commissioned a regression guide [175] which is useful as an introduction to the subject. Walter, Price and Sohn [176] presented an uncertainty quantification method for baseline estimates based on Mathieu's time-of-week and temperature model [177] and cross-validation. A convincing case is made by Shonder and Im (G14 authors) using Bayesian regression and comparing it to the G14 method [178]. They point out that although linear regression is computationally efficient and robust, many energy monitoring problems require non-linear models. Furthermore, normalised savings calculations are done when the reporting period conditions do not reflect a standard operating period, and energy use from the baseline and reporting periods need to be 'normalised' to a third set of EGF conditions. There is currently no other way to do normalised savings uncertainty calculations other than by using Bayesian methods.

Walter, Price and Sohn's study does confirm that when finer graduations of data, such as half-hourly or hourly data points are used, the 'full operating cycle' of a year's data is not needed to characterise the system properly. Three to six months' data is adequate. This is echoed by Granderson *et al.*'s comparison of different M&V baseline methods [35, 179]. This finding applies to energy systems where rapidly fluctuating EGFs such as outside air temperature is present. However, it will not work for population decay or other slow processes.

An interesting comparison of baseline modelling techniques is given by Zhang *et al.* [180], who compare M&V models using linear regression, GPs, Gaussian Mixture Models (GMMs), and Artificial

Neural Networks (ANNs). When applied properly, all methods have similar fit characteristics, with the Gaussian models illustrating the uncertainties very well. Note that these methods specifically are non-parametric. That is, they do not specify an underlying system equation. As such, they are useful for fitting data in regression-type problems, but not for extrapolation as would be needed in time-series forecasting in longitudinal studies.

2.3.1.1 Measurement uncertainty in metrology

Metrology is the science of measurement and represents the larger field of which a large portion of M&V forms a part. Its guiding document is the GUM [181]. The GUM has standardised the expression of uncertainty across most quantitative scientific disciplines and is also applied to energy monitoring. Instructive tutorials have been written, most notably by the British [18, 182] and European [183] accreditation agencies. ISO/IEC 17025 [184] *General requirements for the competence of testing and calibration laboratories* has contributed to the GUM's popularity by stipulating that complying laboratories apply a procedure to estimate uncertainty in measurement.

The GUM distinguishes between measurement uncertainty calculated by statistical methods from measured data (Type A), and those measured or stipulated from prior information or judgement (Type B). It also standardised the expression of uncertainty as a *coverage interval*, also known as an *expanded uncertainty*. This is the confidence/precision format of expressing uncertainty, which should be familiar to most M&V professionals and is used in the IPMVP [1], RA96 [113], and CDM [4] documents. For example, when a measurement is expressed as 10 ± 1, the precision range (or *semi-range*) is $p = 1$. The interval from nine to eleven is expected to correspond to the 95% confidence interval if no more information is given [18, 113]. Since the standard score of the normal distribution $z_{95\%} = 1.96 \approx 2$, the *coverage factor* is 2. The rectangular/uniform distribution is recommended rather than the Normal distribution for digital volt meters and instruments where uncertainties are not stated [18]. Although this is conservative, it is not a realistic assumption for M&V. Energy data are usually aggregated or integrated over a time interval such as 30 minutes, and such errors would then be normally distributed. If an M&V practitioner opts for the uniform distribution assumption, and later convolves it with a normal distribution for sampling error, for example, the resultant coverage interval will be a statement about uncertainties, not probability density intervals [21]. Monte Carlo (MC) convolution or the Mellin Transform method (Section 6.4.2) is recommended for obtaining the probability distribution in such a

case. These will be discussed in the section on new directions in metrology.

The concept of dominant uncertainty components is also useful in M&V. As a rule of thumb, if one uncertainty component is two to three times larger than the next highest one, it may be considered to be the sole contributor to the overall uncertainty [18, p.17]. This is because of the sum-of-squares approach to adding standard deviations together allows larger standard deviations to dominate the final result. Commenting on the efficient allocation of measurement resources between Type A and Type B measurements, Birch, therefore, remarks that the

> [the] quantification of uncertainties in testing normally involves a large element of estimation of... uncertainty components. Consequently, it is seldom justifiable to expend undue effort in attempting to be precise in the evaluation of uncertainty for testing. [18, p.15].

This is relevant when trading metering calibration costs off against sampling costs, where meter calibration can be expensive for comparatively little gain.

New Directions in Metrology

Although acknowledged as very helpful, the GUM has drawn criticism, most notably from Bayesian statisticians [21].

One point of contention relevant to M&V is that the propagation of errors calculation is defined as a first-order Taylor series approximation, which does not always hold. Some physicists and statisticians are also uncomfortable with the frequentist approach to how confidence intervals are calculated in the GUM. It has been shown from first principles that this approach is invalid in many measurement cases [21]. This is explored more fully in Chapter 3.

In reaction to the criticisms above, the GUM was updated and a supplement describing a Monte Carlo (MC) alternative was published [185]. It is especially useful for non-linear cases, where any distribution other than the Gaussian or scaled-and-shifted T is used, or where the error propagation function is complex. It also delivers the final error estimation as a probability distribution rather than an uncertainty interval. Therefore it is all but recommended as the de facto method for uncertainty propagation calculation by the supplement. MC can be too computationally expensive for high-dimensional problems, and approaches such as MC-Latin Hypercube Sampling or Sobol' Sequences [186] are

then used. Respected Bayesian metrologists such as Lira have advocated analytical calculus-based approaches over MC methods where possible [187]. However, this is not a viable alternative in the energy M&V industry.

A second interesting approach is the MTMC [142, 143]. By this method, the moments of the posterior of such a convolution may be expressed in terms of the scale and shape parameters of the constituent distributions. Section 6.4.2 provides a practical example. The MTMC allows for exact expressions of the first four (or more) moments of the distribution: mean, variance, skewness, and kurtosis, at a fraction of the computational burden of an MC simulation. This work is made available through an online toolbox as the Mellin Transform Moment Calculator [188]. Although the first four moments of a distribution do not identify it uniquely for all cases, most metrological problems are unimodal, and therefore should not be a problem. A Johnson S_B (bounded) distribution [189] can then be fitted using these four moments. This distribution family was expressly designed for such flexibility and has been applied to skewed data in a variety of disciplines from econometrics [190] to quality [191]. For more information on uncertainty evaluation through moment-based distribution fitting, see Rajan *et al.* [144].

Regarding the Bayesian approach, the UK Accreditation Service (UKAS) noted that "Bayesian statistics is becoming recognised as being particularly useful in certain areas of testing" [182], and as of 2016 the GUM itself is also in the process of being extensively revised to accommodate the Bayesian paradigm [192]. This signals an interesting shift in metrology and the way in which uncertainty is viewed and calculated. Estler [193] provides a comprehensive tutorial of Bayesian theory in the context of measurement and the GUM, while shorter theoretical Bayesian frameworks for metrology have also been written [194, 195].

2.3.2 Mismeasurement

The measurement errors discussed thus far are mostly harmless. If random, mismeasurement of the *dependent* variable (usually energy) widens the confidence interval around the estimate but does not add bias to the parameter estimates. However, this is not the case when these noisy measurements are used as *independent* variables in a regression analysis. This *errors-in-variables* effect is seen in energy regression models when a covariate such as temperature or occupancy is measured with

error, and may also occur when one calibrates an instrument against a standard with some error. In such cases, the random variation is no longer in $\mathbf{y}$, but in $\mathbf{x}$. Random errors in $\mathbf{x}$ have two effects. First, all the regression parameters become biased due to the "flattening out" of the data points as they spread out on the x-axis (see Figure 2.3). This is called *attenuation*. Second, the confidence intervals on these estimates are narrower than they should be, giving misleadingly high confidence in biased values, also manifesting as a loss of statistical power [196, 197]. This is because as the measurement error (variance) increases, it becomes increasingly difficult to distinguish it from the process variance. This lack of power may then be misinterpreted as a lack of effect when pre- and post-retrofit measurements are compared [196]. To regain this power, much larger sample sizes are then required. Table 2.5 summarises the effect of mismeasurement on various statistics. Effects vary with error type and regression model type.

To illustrate attenuation, consider attempting to use one unbiased meter to calibrate another when the reference meter reading contains random error. Let the reference meter be $\mathbf{x}$, and the UUT be $\mathbf{y}$. If both the reference and the UUT are perfectly accurate, a regression line with a gradient of one should be drawn on the xy plane:

$$\mathbf{y} = a\mathbf{x} + b, \tag{2.1}$$

where $a = 1$ and $b = 0$.

If only the UUT has an error (thus an error in the response or dependent variable measurement), the dependent variable $\mathbf{y}^* = \mathbf{y} + \boldsymbol{\varepsilon}$ will be measured by the UUT, where the $\mathbf{y}^*$ indicates the *surrogate* reading and $\boldsymbol{\varepsilon}$ the error. $\mathbf{y}^*$ is observed in lieu of $\mathbf{y}$, where:

$$\mathbf{y}^* \sim N[\mathbf{y}, \tau\mathbf{y}] \tag{2.2}$$

The error will add noise, but will not bias the result, as illustrated in the left-hand graphs of Figure 2.4. These are Ordinary Least Squares (OLS) regression estimates for increasing values of the standard deviation multiplier τ. Increasing error does not bias the estimates. However, this does not hold for errors in $\mathbf{x}$ of the form

$$\mathbf{x}^* \sim N[= \mathbf{x}, \tau\mathbf{x}], \tag{2.3}$$

An illustration of one instance is shown in Figure 2.3. An illustration of the effect on parameter estimates for the straight line case over a range of error values is shown in Figure 2.4.

Mismeasurement is less of a problem for prediction, which is often the goal of M&V models. If one infers some function $\mathbf{y}^* = \boldsymbol{\theta}^*\mathbf{x}^*$ based on measurements of $\mathbf{x}$ made with random error, that relationship

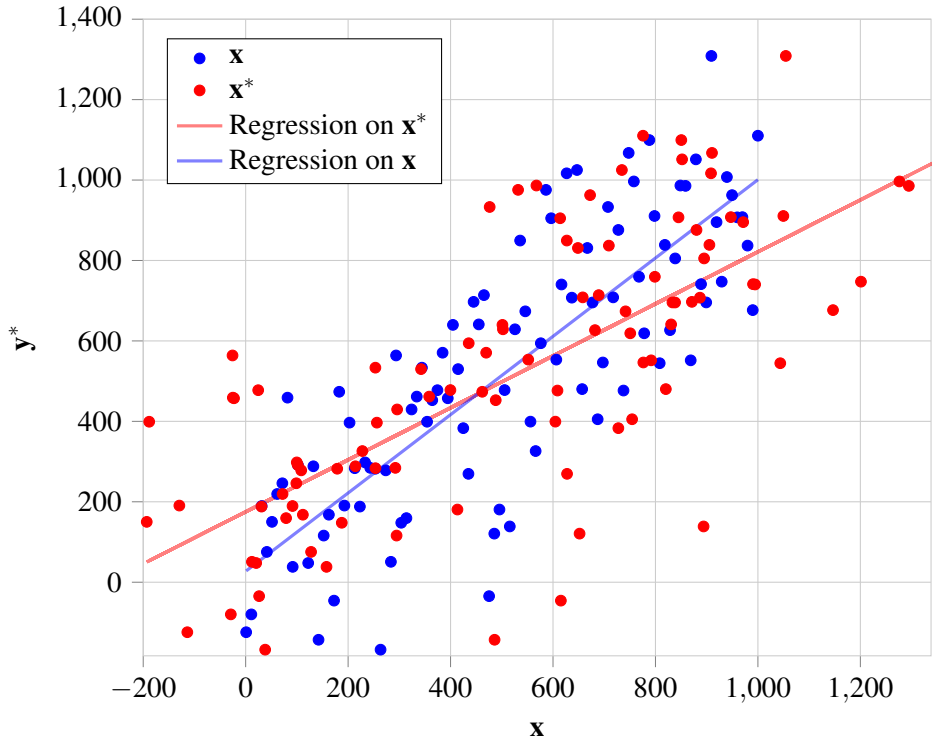


Figure 2.3. Demonstration of the effect of random error in the measurement of the independent variable on regression. For the regression on the observed values, the parameters are $a^* = 0.6$ and $b^* = 175.1$. However, the true parameters are $a = 1$, $b = 27.1$.

defined by $\boldsymbol{\theta}^*$ will continue to hold as long as you forecast and measure using $\mathbf{x}^*$ in lieu of $\mathbf{x}$. In such a case a Measurement Error Model (MEM) is unnecessary. This is part of the reason that measurement error is not a greater problem in M&V: often the baseline and reporting period measurements are made with the same instruments, and so the attenuation effect may 'cancel out', as long as inference about the physical meaning of the parameters (e.g. kWh/Heating Degree Day) is not attempted. Consider the 'time-of-week and temperature' M&V regression model [35, 176, 177], in a situation where the temperature is measured with error because the weather station is in a different microclimate to the facility [88, 198]. The relationship between energy use and temperature would be attenuated. This would cause certain elements of the time-of-week parameter vector to seem more influential than they actually are. However, this may not be a problem. Suppose that HVAC-related Energy Conservation Measure (ECM) is installed and the model is used for M&V. The forecast (adjusted baseline) energy use in the post-retrofit period will have the same attenuation as the baseline. It would, therefore, be accurate, assuming a calibrated model and same temperature data source. Therefore the total savings estimation will have a similar NMBE to the case with no measurement error, although the added noise

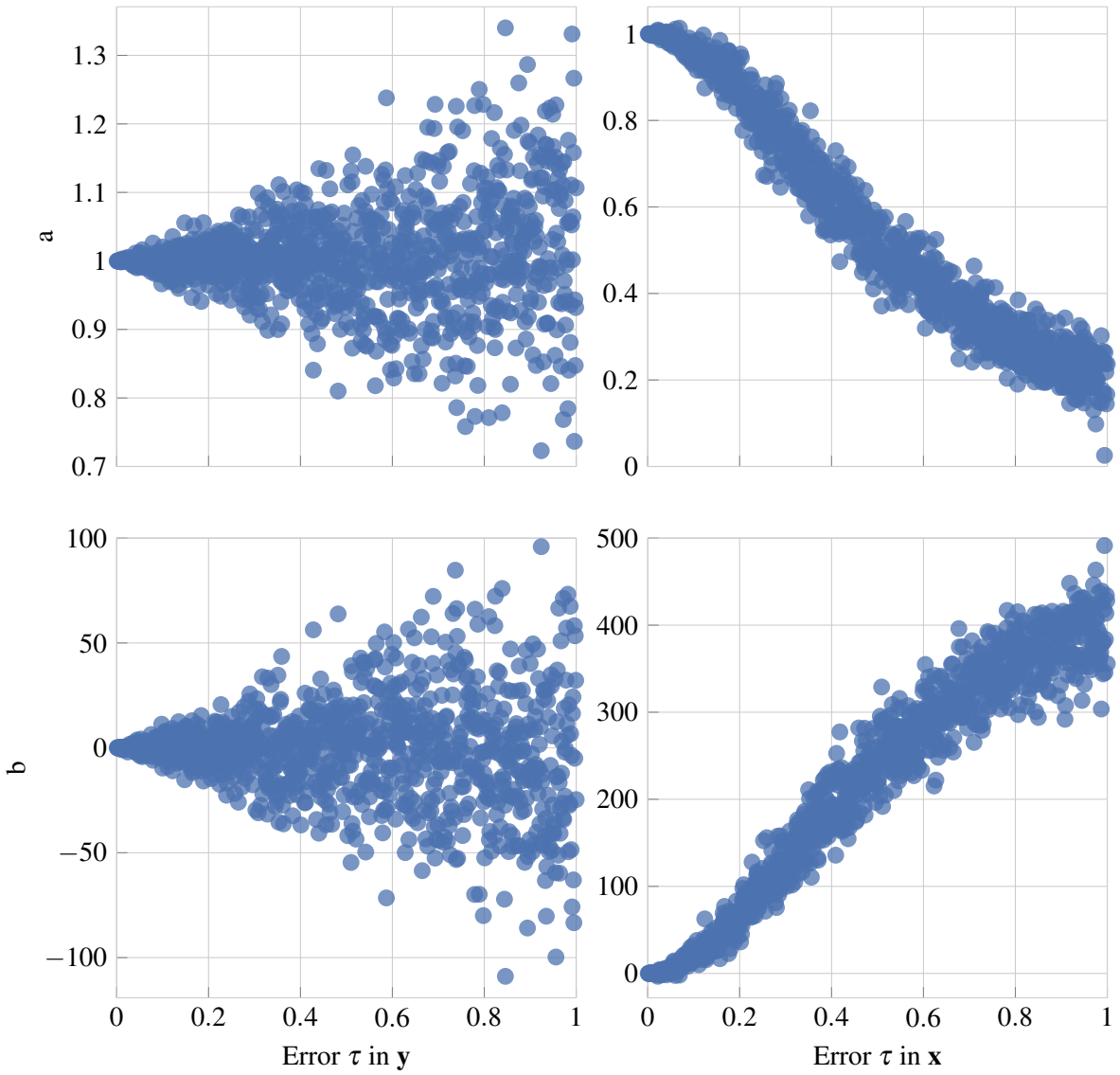


Figure 2.4. OLS parameter estimates for **y**=a**x**+b, where a=1 and b=0, given measurement error τ in the form (2.2) and (2.3).

may lead to a higher CV(RMSE) on the training set. This being said, one cannot regress energy use against temperature to infer the effectiveness of the ECM, nor can such a regression be transported for project decisions in other places. Furthermore, the confidence interval around the reported savings will also be too narrow.

From these results we can see that fitting a model to data $D = \{\mathbf{x}_i^*, \mathbf{y}_i^*\} = \{(x_1^*, y_1^*), (x_2^*, y_2^*) ... (x_n^*, y_n^*)\}$ where $\mathbf{x}^*$ and $\mathbf{y}^*$ are they surrogate values (ones measured with error), is a difficult problem not

adequately addressed by frequentist statistics [200]. Ordinary Least-Squares (OLS) and most other standard techniques can account for errors in **y** only, but have no way of dealing with simultaneous errors in **x**. This is because when presented with data, it is difficult to distinguish the contribution of measurement error in **x** from that of **y** for a specific data point.

From Appendix A5.6 of the ASHRAE Guideline 14-2002 [31] mentioned above it can be seen that covariate measurements have uncertainties many times greater than that of the energy metering system. By the rule of thumb mentioned in Section 2.3.1.1, measurement uncertainty is therefore often dominated by non-electrical covariate measurements. The recalibration of energy meters to high accuracy standards may therefore be an unnecessary expense for large-sample applications. The question then arises: can meters be verified to a lower specification, using less precise calibrators or power supplies, and still be useful for M&V purposes?

Table 2.5. Spurious effect of mismeasurement in **x** on various statistics assuming classical additive errors [196, 197, 199].

Statistic	**Effect**
Mean	None
Variance	Increases
Covariance	None
Regression, single predictor, slope	Decreases
Regression, single predictor, intercept	Increases
Regression, multiple predictors	Complex
Confidence on regression coefficients	Increases
Statistical power for detecting relationships	Decreases
Correlation	Decreases
Partial correlation	Increases
Non-linear features (such as $\mathbf{y} = sin\mathbf{x}$)	Masked

2.3.2.1 Mismeasurement in M&V literature

Although attenuation bias due to mismeasurement has been documented in M&V, the effect is not well-known. With the exception of the UMP Chapters 13 and 23 [41,55], all M&V guidelines discussed so far, as well as M&V regression guides [175] do not mention attenuation, even when measurement errors are discussed. The UMP Chapters 11 and 12 (Sample and Survey Design) [6, 40] state that random measurement error does *not* lead to bias, even though survey measurement error is one of the most common MEM test cases [201]. G14-2014 stipulates that the total span of the additional uncertainty created by errors in independent variables shall be determined by biasing the variables to their maximum and minimum values [17]. Attenuation is unaccounted for.

Regarding literature, an MC analysis was done by Sonnenblick and Eto from LBNL in 1995. They found this bias effect for measurement precision of energy programmes [46, Fig. EX-2], and identified it as the errors in variables effect. The measurement of operating hours was considered to be the most sensitive to this effect.

Ridge [202] presented an informative paper on mismeasurement in M&V in 1997. He relates how the Californian utility Pacific Gas and Electric's 1992-1993 Commercial New Construction Program and the 1994 Commercial HVAC program realisation rate estimates were unreasonably low. The realisation rate is the ratio of expected to actual savings. He traced the problem back to random errors in independent (explanatory) variables that led to attenuated estimates. This was corrected for in subsequent studies by the use of dummy variables.

A more recent example of mismeasurement is found in the case where Canadian economists Rivers and Jaccard published a study which found that Demand Side Management (DSM) interventions made no statistically significant impact on energy demand when viewed at a national level [203]. This generated some controversy. Rivers and Jaccard proposed that measurement error in the independent variable (DSM spending proportion vs EE spending proportion) may have played a role in attenuating the DSM-effect parameter estimate. However, although Violette *et al.* [204] also acknowledged this errors-in-variables possibility, they proposed that other features of the original Rivers and Jaccard model were more influential.

2.3.3 MEMs and calibration techniques

There are two main bodies of research addressing measurement errors relevant to energy models. First, commercial electrical metrological techniques have been honed over the last half-century. These methods usually employ Test Uncertainty Ratios (TURs), which is the ratio of the precision of the calibrator to that of the UUT. They have had to be revised recently as the accuracy of calibrators and digital multimeters (DMMs) has converged to 8.5 digits (one part in 10^8). Second, trans-disciplinary academic investigations have been conducted using a variety of approaches. These have advanced significantly in response to the stringent and complex requirements of medical fields such as epidemiology, coupled with the relatively poor accuracy of the instruments measuring certain human epidemiological variables.

2.3.3.1 Electrical calibration techniques

These techniques are applicable mainly to calibration. They are commercial techniques usually using indirect, empirical, conservative methods, and cannot be classified as true MEMs. A TUR of 4:1 is generally required. This means that an instrument accurate to $p\%$ may be used to calibrate an instrument accurate to $4p\%$ (called the Unit Under Test, UUT). This may reflect the other rule of thumb proposed in Section 2.3.1.1. However, since DMMs such as the 8.5-digit Fluke 8508A do not allow for a TUR$>$4 between the UUT and the calibrator, other techniques had to be developed. The simplest and most accurate is to characterize the long-term drift of the instrument by plotting the change in measurement errors over time, and then drawing a regression line through the successive measurement points [205, 206]. This regression line has been shown to be more accurate than the individual calibrations [207]. Within limits, and with a large enough calibration history, this technique may be used to accurately quantify an instrument's error without recent calibration. This technique has also been proposed for characterising the stability of a calibrator that may not meet the TUR$>$4 nominally but does meet it practically. This is possible as the calibrator's stability specifications are usually lower than what an individual instrument's stability may be when measured with a more accurate DMM.

On the other hand, if one wants to test an instrument with no history, and one can not achieve the required TURs, alternative methods also exist [208]. For true calibration, the only option is

'disciplining' the calibrator by using an additional, more accurate DMM to measure the calibrator output in real time [206].

In cases where an accept/reject decision has to be made rather than full calibration, there are three options: lower the confidence level of the test, invest in a more accurate standard, or analyse and document the measurement points for which inadequate TURs exist. The first option (lowering the confidence level) is called *guard banding* [209–211]. A guard band is a test limit stricter than the instrument specification limit [212]. In other words, by employing guard bands, one can use a calibrator with a TUR of 2 instead of 4. The price one pays is that the UUT may still be rejected, even if the test result falls between the Lower Confidence Limit and the Upper Confidence Limit of the calibrator. This is because to compensate for our lower TUR, the test limits are narrower than the instrument specification limits. Thus guard banding keeps the consumer's risk constant even though a less accurate calibrator is used, but increases the producer's risk for such a case. When considering this approach, one must remember that at a certain level, testing becomes uneconomical. For example, for a TUR of 2 and specification limit of 2σ, the consumer's risk is as significant as it would be if no testing at all took place, and the consumer simply accepted the probability of the unit being outside of specification (probability=1.2%) [208]. In such scenarios, the expected value of the test, or the cost/benefit trade-off between testing and not testing, should be considered.

Rossi and Crenna [213] provided a good example of setting test limits lower than specification limits for in-house testing at the producer side to minimise risk, which they applied to water meters [57]. To this end, they have developed a software package called UNCERT – essentially an automated MC approach. Researchers from the US National Institute of Standards and Technology (NIST) have also shown that a Bayesian approach to the accept/reject decision rule of ISO 14253-1 (inspection of workpieces) [214] delivers superior results in cases where it is applicable [215].

2.3.3.2 Trans-disciplinary MEM techniques

Not all uncertainty analysis models (also known as uncertainty quantification models) considering measurement error are MEMs. On the other hand, some probabilistic models using MC methods could well be incorporated into MEMs, although their function in most literature is exploratory what-if analysis, sensitivity analysis, or forecasting (see Table 2.2). Other methods are simply robust:

insensitive to outliers. This section discusses statistical techniques for unbiasing regression estimates. These will include some techniques common to general statistics (such as MCMC and MLE), as well as mismeasurement-specific methods.

There is a notable amount of literature on MEMs, although much of it is too technical to be useful to the M&V practitioner without a strong background in statistics. For linear problems Fuller [216] is popular, and his method-of-moments is straightforward and recommended for OLS regression with additive measurement errors (cf. Carroll *et al.* [196]). The non-linear case presents a greater challenge, but may also be more relevant to M&V and instrument calibrations as shown by Carobbi *et al.* [217]. The most appropriate (and readable) treatments are by Carroll *et al.* [196], and Gustafson [197].

MEMs can be divided into functional and structural approaches. Functional approaches make no assumptions about underlying distributions (thus avoiding model misspecification) and include Regression Calibration and SIMulation EXtrapolation (SIMEX). These techniques are specific to the mismeasurement sub-field. Structural approaches make assumptions about the underlying distributions and relations governing the measurement system and include Maximum Likelihood Estimation (MLE) and Markov Chain Monte Carlo (MCMC) techniques: ones used in other areas of statistical inference as well. All four of these techniques are powerful and can yield useful results if applied well. The choice of method depends on its appropriateness to the data and ease of implementation.

The **SIMEX** concept is simple and powerful. Suppose one knows that the variance $\mathrm{VAR}(\mathbf{x}^*|\mathbf{x}) = \tau$. Current parameter estimate $\boldsymbol{\theta}^*|\mathbf{x}^*$ are also known, that is, $\boldsymbol{\theta}^*|\tau_0^2$. The true parameters $\boldsymbol{\theta}|\mathbf{x}$ are sought. If one now *increases* the error τ in the dataset, the parameter estimates will start drifting away from their true values due to attenuation. In this way, one can obtain values for $\boldsymbol{\theta}^*|\tau_1^2$, $\boldsymbol{\theta}^*|\tau_2^2$, $\boldsymbol{\theta}^*|\tau_3^2, \ldots$ A trend will be observed, and a curve can be fitted to these points. Extrapolating backwards will then yield $\boldsymbol{\theta}^*|(\tau = 0)$, which is $\boldsymbol{\theta}|\mathbf{x}$. See Figure 4.5 for a graphical illustration. The disadvantage is that SIMEX is difficult for cases where there are combined multiplicative and additive errors and that it can be expensive for non-linear higher dimensional models. It has also been found that in certain cases MLE methods yield considerable smaller variances [218], although for most applications SIMEX is simple and effective.

Regression Calibration methods essentially trade an exposure model for a validation (calibration) sample: a sub-sample measured without error, using a 'gold standard'. From the information gleaned

from the sub sample, values for $\mathbf{x}$ are imputed instead of the $\mathbf{x}^*$ values measured. Repeated measurements may also be used. It is not susceptible to bias due to model misspecification since the exposure models do not need to be specified. Regression Calibration is useful for trials where extensive, precise, or repeated testing is only feasible for a small sub-sample.

One potential weakness of the Regression Calibration method is that it maps $\mathbf{x}^*$ onto $\mathbf{x}$ in a one-to-one fashion, where methods such as Bayes-MCMC consider all reasonable values for $\mathbf{x}$ given the data. Therefore the uncertainty is specified as fully as possible. This avoids the effect of not considering the uncertainty contribution of imputing $\mathbf{x}$ values for the first step of the Regression Calibration procedure.

Two structural approaches will now be discussed. It is important to note that although these techniques have solid analytical foundations, they are solved numerically. Different kinds of inference algorithms could be used, with MCMC and MLE being the most popular.

Maximum Likelihood Estimation has become a very powerful structural approach in many areas of statistics. It produces a likelihood distribution on parameters of interest, and can account for measurement error by specifying such errors in the structure of the model. MLE techniques have the potential of producing better estimates than functional approaches if the model is well specified, although this is often difficult [196]. MLE methods are advanced empirical Bayesian methods using non-informative priors (more on this in Chapter 3). Full Bayesian methods provide some advantage since the models are easily specified and solved, no approximations are necessary, and standard errors on the estimates are more easily calculated [197]. Stopping or convergence criteria are a concern for both approaches [219]. Gelman [220] also notes that EM algorithms with multivariate normal approximations are not ideal for small data sets as convergence is only asymptotic, and the normal distribution not ideal for describing such cases.

Much literature on the technical merits and application of **Bayesian methods** exists, as it is the natural structural MEM approach [196].

Bayesian approaches with non-informative priors provide MLE estimates of data [220]. However, they are more flexible since they do not require ad hoc techniques dealing with special cases, as with most

frequentist statistics. This allows rapid model development and greater ease in specifying and building complex, realistic models.

The disadvantages of the Bayesian-MCMC techniques are that they can be computationally expensive, susceptible to model misspecification, and require more thinking on the part of the practitioner. The computational expense becomes a problem when many variables (or data points) have uncertainties in them which need to be modelled using MCMC. The model then suffers from the curse of dimensionality. Thus, for problems such as the real-time calibration of thermal network parameters, Bayesian techniques have been found to be too computationally expensive even though they are more robust than lightweight 'grey-box' techniques [221]. Variational inference may alleviate this concern, and although the technique is relatively new it has been implemented in popular software [222]. Model misspecification arises when the true error structure is different from the one specified in the model. Investigating the robustness or sensitivity of the model to such assumptions becomes necessary. Last, there are few simple 'recipes' in Bayesian statistics. There is no t-test or F-test blanket equivalent, although Kruschke provides alternatives [19]. However, Bayesian solutions are more problem-specific than popular frequentist tests.

Two non-technical reasons for the application of Bayesian approaches to M&V should be noted. First, a Bayesian MEM is similar to a standard, well-specified Bayesian model. The model's ability to deal with measurement errors follows from the nature of the Bayesian mathematics itself. Second, the development of Markov Chain Monte Carlo (MCMC) techniques has allowed for the previously intractable integration involved in most non-trivial Bayesian calculations to be done efficiently and accurately.

2.3.4 Longitudinal studies

Two methods are directly applicable to the longitudinal sampling problem: Survival Analysis (SA) and regression. SA is used for time-to-event data and can account for censoring (where exact failure times are unknown) as well as for measurement error. For an introduction, see Clark and Bradburn *et al.* [223–226], and for an application to EE and DSM persistence studies, a commercial study where this was implemented [168]. As with logistic regression, the focus of the method is on identifying the effect of covariates, and not on time-series forecasting, although such applications have been

made [227]. Most SA models use the 'proportional hazards' assumption of fixed hazard or failure rates. This is not accurate for CFLs, although alternatives do exist and are mentioned below. SA is not used in this study but is a promising approach for future persistence research.

The second approach is regression. Various regression methods exist in energy monitoring [35, 177, 180, 228]. A suitable regression method should weigh points according to sample size and account for the binomially distributed nature of the samples. It should also quantify uncertainty accurately. This was achieved in West *et al.*'s seminal work on Bayesian Forecasting and DGLMs [229, 230], building on McCullagh and Nelder's GLM work [231]. Triantafyllopoulos [232] provided a useful comparison of these and related methods such as particle filters and extended Kalman filters with posterior mode estimation. Gamerman and others have applied these models to survival analysis [233–236] and hierarchical models [237]. These models work with parametric distributions that do not describe the complexities of the energy savings calculations discussed in Section 6.4, but more research in this area is warranted. A model similar to West, Harrison, and Migon's advertising awareness study [229] has been adopted for this thesis, which uses a DGLM with Bayesian forecasting to model binomial survey response data in Chapter 6. This model uses the conjugate prior property of the beta-binomial distribution pair to incorporate information from past surveys into current estimates, even when those surveys found the population proportion to be higher than the current proportion due to decay. This is an implementation of Violette's proposal of using a Bayesian framework for longitudinal M&V studies [67].

CHAPTER 3 THE BAYESIAN PARADIGM FOR MEASUREMENT AND VERIFICATION

3.1 CHAPTER OVERVIEW

In this chapter, no novel theory or method will be proposed. Instead, an overview of some aspects of the Bayesian approach relevant to M&V practitioners is presented. Those familiar with the Bayesian paradigm may wish to go directly to the next chapter, as the examples below are well known. After the introduction, a motivation for selecting the Bayesian approach is given by comparing it to existing frequentist methods. The theory of the Bayesian approach and theorem is presented, and a discussion of subjectivity, information and entropy, and numerical methods is offered. Applications are then made by considering sampling, measurement, and regression.

3.2 INTRODUCTION

This chapter proposes a Bayesian approach to M&V, but cannot be a full exposition of Bayesian theory. Many useful texts have been written on the philosophy and application of Bayesian statistics, and these will be alluded to below. The author also does not intend to settle the Bayesian-vs.-frequentist (classical statistics) debate in this chapter. Both have advantages and disadvantages, and in the hands of a skilled statistician both can be useful for many (but not all) estimation problems. Instead, a pragmatic approach is adopted. Engineers prefer simple, effective techniques, and it will be argued that the Bayesian option makes sense in theory, and can be powerful, flexible, and simple to implement in practice. Bayesian techniques have become popular in fields closely related to M&V: industrial

machine learning and metrology. Although it has been recommended for error analysis of energy measurement and verification, especially for cases where errors are financially significant [65], it is underutilised in M&V. One recent exception is Tehrani, Khan, and Crawford who have used recursive Bayesian regression in a novel way for baseline forecasting in M&V [228].

The Bayesian approach derives its name from a posthumously published article by Reverend Thomas Bayes [238], although Pierre-Simon Laplace developed the modern notation of the theorem. The approach relies on the logic of conditional probabilities, developed from first principles via various axioms. Frequentism has been the dominant approach for the last hundred years, however. In the frequentist approach, data are assumed to be realisations or "snapshots" of long-run processes, or frequencies. This approach does have some appeal in many cases: flipping a coin ten times, whatever the result, can be seen as a sample of results of flipping a coin millions of times. Probability is therefore equated to frequency-calculations, which often simplifies the mathematics. However, such long-run frequency approximations are not always valid for problems in energy measurement, as will be argued below.

3.3 MOTIVATION

When is an M&V plan efficient? Why should one M&V study be preferred over another? Besides quality control measures, an M&V study should report savings accurately, at low cost. [1] A small sample, taken with low cost, inaccurate equipment, will still yield a result, albeit a possibly biased one with high variance around the estimate. A large sample obtained with accurate equipment would be preferable if cost were not a concern (assuming this implies low bias and variance). M&V is therefore an uncertainty quantification exercise, in the context of cost. 'Optimal', or 'efficient' M&V will yield savings estimates with low bias and accurately quantified uncertainty, at low cost. The author adopted the Bayesian paradigm for this thesis because he realised during his Master's work [111] that it might satisfy the above argument in a way that standard frequentist methods cannot.

The Bayesian-vs frequentist debate is not as simple as "Bayesians are right and frequentists are wrong". The frequentist paradigm is not as deficient as some Bayesian texts argue, and the Bayesian paradigm does not seem to be the panacea it is made out to be by some of its exponents (on which,

[1] Accuracy can be measured in terms of bias (not being consistently higher or lower than the true value), and variance (having a narrow range of possible values around the estimate).

see the discussion of Lindley's paper [239]). Kruschke does make a convincing case, motivated by the underlying theory as well as numerical comparisons [240], that for the comparison of samples, the Bayesian approach provides richer information, and often different conclusions, to standard Null-Hypothesis Significance t-tests. This should not be generalised to all methods, however. In all cases, careful modelling is still necessary in the Bayesian case, and the expert frequentist statistician who is aware of the shortcomings of the methods, can still draw valid (but more limited) conclusions. Besides the theory, there is a practical advantage to the Bayesian approach as well. Most users of statistics (including M&V engineers) are not professional statisticians. The simpler, more reliable method should, therefore, be preferred: the Bayesian one. This statement will be motivated by considering some of the shortcomings of the frequentist paradigm for M&V.

At first, the two paradigms may seem very similar. Both provide a way of making sense of data and provide a method of inference about the world. However, they do diverge significantly. For example, in the Bayesian paradigm, the data are fixed, and the parameters are viewed probabilistically: as uncertain values described by probability density functions. 'Uncertainty' therefore describes our state of knowledge of reality, or our degree of belief, codified in mathematics. In the frequentist paradigm, the data are viewed as random realisations of a reference set which has fixed parameters. Probability is described in terms of the frequency of the data arising from the hypothesised, but unknown, fixed-parameter reference set. If one is sampling items from a production line, the meaning is clear. However, if one is calculating failure probabilities of nuclear reactors or the population from which the energy measurement made at 14:15 today at a specific site, given specific weather conditions, it is not clear. Kruschke [19, 240] mentions illuminating cases. For example, a coin is flipped twenty times, and seven heads are observed. What is the probability of the coin being fair? The answer depends on the reference set from which one *intends* to sample, since this set includes the data which might have happened, but did not [241]. If one intended to stop after twenty flips, $p = 0.032$. If one intended to stop after seven heads, $p = 0.017$. If one intended to stop after two minutes of flipping, $p = 0.024$. If one intended to compare it to a second coin, $p = 0.103$. The probabilities obtained from the calculation therefore depend on the hypothetical reference set, and not only on the data observed, as frequentists often claim.

Probably the most significant problem with the frequentist paradigm in M&V is the use of confidence intervals. According to Neyman, who devised these intervals, they do not convey a degree of belief, or confidence, as is often thought. They are a product of a process that produces an interval which

contains the true value a given percentage of the time [20]. Montgomery and Runger explain the difference as follows in their textbook *Applied Statistics and Probability for Engineers* [242], under "Interpreting a Confidence Interval". The bold and italic emphases are theirs:

> How does one interpret a confidence interval? In the impact energy estimation problem in [the notch impact test] Example 8-1 the 95% CI is $63 \leq \mu \leq 65.08$J , so it is tempting to conclude that μ is within this interval with probability 0.95. However, with a little reflection, it's easy to see that this cannot be correct; the true value of μ is unknown and the statement $63 \leq \mu \leq 65.08$ is either correct (true with probability 1) or incorrect (false with probability 1). The correct interpretation lies in the realization that a CI is a *random interval* because the probability statement defining the end-points of the interval L and U [lower and upper] are random variables. Consequently, the correct interpretation of a ... CI depends on the relative frequency view of probability. Specifically, if an infinite number of random samples are collected, and a [95%] confidence interval for μ is computed for each sample, [95%] of these intervals will contain the true value of μ.
>
> ...
>
> Now in practice, we obtain only one random sample and calculate one confidence interval. Since this interval either will or will not contain the true value of μ, it is not reasonable to attach a probability level to this specific event. An appropriate statement is that the observed interval $[l,u]$ brackets the true value of μ with **confidence** [95%]. This statement has a frequency interpretation; that is, we do not know if the statement is true for this specific example, but the *method* used to obtain the interval $[l,u]$ yields correct statements [95%] of the time.

Therefore, a frequentist confidence level is the probability of the interval including the parameter, while a Bayesian credible interval (or probability) is the probability that the parameter is included in the interval. To be sure, the frequentist and Bayesian confidence intervals agree for common problems such as linear regression, provided that the linear regression assumptions hold and that a non-informative prior (explained below) is appropriate for the Bayesian analysis [21]. In such cases, the linear regression is easier and adequate, and the Bayesian interpretation can be appropriated for the frequentist interval for the calculation of risk. However, this is not guaranteed outside of these special (though common) cases. In general, the limits of the frequentist interval should be viewed as random numbers, and the interval is simply an interval, not a probability density function. This makes

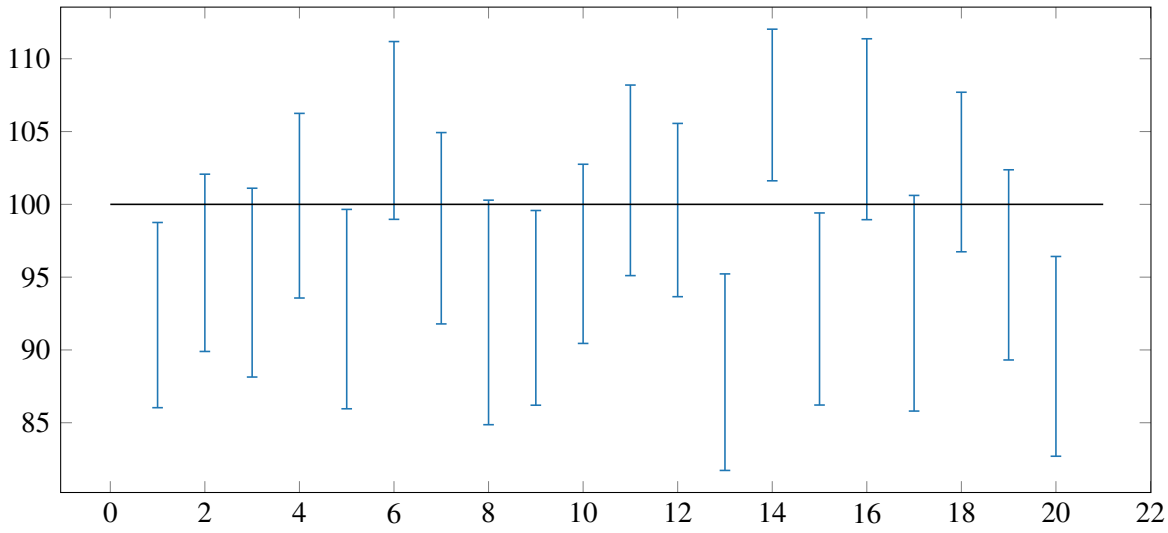


Figure 3.1. Illustration of frequentist 90% confidence interval for 68 samples drawn from the distribution $N[100, 50]$, as per standard M&V practice, repeated for twenty different realisations.

the calculation of risk, or the quantification of uncertainty, problematic. Shonder and Im [178] also point out that describing the uncertainty by a PDF rather than a point estimate enables other kinds of calculations that a standard confidence interval does not. For example, the probability of savings being above any given value may be determined. The PDF shape may also indicate where most of the probability mass lies, in the case of skew distributions.

The second, well-known problem with frequentism is the interpretation of p-values used for null-hypothesis significance testing. Besides the difficulty in explaining to clients that one "cannot reject the null-hypothesis", p values have proven to be unreliable measures of statistical significance. This was already a well-known problem when Berger and Delampady discussed it, and Bayesian alternatives, in 1987 [243]. However, for reasons of convenience and institutional inertia, p-values remained popular in medicine, until Ioannidis [244] and others [245] found that most medical studies resulting from its use contain false conclusions and exaggerated effects. This is because data may reach 'statistical significance', but the study is still underpowered. In other words, $p = 0.05$ for a small sample is less reliable than $p = 0.05$ for a large sample. The significance may still be spurious, and $p = 0.05$ does not mean that there is a 95% probability of the effect being true, as is commonly thought (and taught). As Button *et al.* [245] note, this is because the prior odds of an effect being present should also be considered. The probability of a false positive should consider the data (which may show a 'positive'),

as well as the odds of an experimental treatment actually working. If only one in five experimental treatments that are investigated work, the odds of the 'positive' being false is higher than if four out of five experimental treatments investigated, worked. This then is exactly the Bayesian paradigm, which will be explored in more detail below. Before continuing, it should be noted that M&V may not be as affected by this phenomenon as medical studies since it is assumed that most projects where M&V is done, do save energy. The prior odds of an effect being present is therefore much higher than for the medical field. However, the probability of exaggerating this effect due to low statistical power is still of concern.

Other shortcomings are less serious. For example, it is often noted that frequentist methods are ad-hoc. That is, different methods and measures are used for different problems, or indeed for the same problem, and all have certain assumptions. There is no unified theory. This contributes to the confusion non-statisticians have about the statistics, since one might not know which test to apply to one's current problem, or what a desirable result for the chosen test would be. The Bayesian approach is relatively standard and its interpretation more intuitive. Its application is admittedly difficult in some cases, but the use of numerical methods such as MCMC has simplified the task. As such, the Bayesian framework makes it easier to combine the different sources of uncertainty typically quoted in M&V. For the frequentist paradigm, ASHRAE's G14 [17] does provide a way of combining these into a single figure for cases satisfying linear regression assumptions, although it is at best approximate [77].

The Bayesian paradigm is well-suited to M&V analysis specifically. As Estler [193] notes, measurements always imply inference, since one reasons from incomplete information to make rational decisions in the context of uncertainty. In the Bayesian paradigm, all unknowns are approached *probabilistically*: as probability distributions. In fact, computer programs used for Bayesian calculations are called probabilistic programs. This is one of the most powerful features of Bayesian analysis in practice. Instead of working with unknown, fixed variables, one works with PDFs. Since M&V is mostly concerned with uncertainty quantification rather than hypothesis testing, Bayesian methods are well-suited. The second feature of M&V that makes it suitable for full Bayesian analysis is that it has well-defined utility functions. The cost of a study and the benefit derived from it can often be described in monetary terms, as savings realised, and penalties for non-compliance. Sampling planning can then proceed according to established principles [246–248].

3.4 BAYESIAN THEORY

The derivations of Bayes' theorem is probably the simplest part of the whole paradigm and is also instructive.

Let $\Pr(SD|I)$ be the joint probability of events S and D given information I. S and D are left as generic signifiers for now. The product rule of probability states that

$$\Pr(SD|I) = \Pr(D|I)\Pr(S|DI) \tag{3.1}$$

This means that the probability of S and D simultaneously, is equal to the probability of D, multiplied by the probability of S given D, both conditional on the other information (the 'reasoning environment') I. This accords with intuition. For instance, the probability of savings being realised at a given household by an energy savings programme is the probability of the household participating in the energy savings program, multiplied by the probability of savings being realised, given that they participate in the program. The term I is implicit in all probability calculations since all probabilities are conditional on a set of assumptions [193]. By rearranging the equation and dropping I (it is implicit in all further calculations), we find:

$$\Pr(S|D) = \frac{\Pr(SD)}{\Pr(D)} \tag{3.2}$$

expanding the $\Pr(SD)$ term in the numerator by the product rule, but this time the other way around than before, we find that:

$$\Pr(S|D) = \Pr(S)\frac{\Pr(D|S)}{\Pr(D)}. \tag{3.3}$$

S and D are usually chosen to represent the hypothesis and the data. Since this thesis considers energy savings, let the hypothesis be "energy is saved", denoted S, and the data be denoted D. Then this reads: the probability of savings, given the data, is proportional to the prior probability of savings, multiplied by the likelihood of the data arising if savings were indeed realised, divided by the global probability of the data. The resultant term $\Pr(S|D)$ is called the posterior distribution, since it describes our state of knowledge after combining our prior probability that savings have been realised, $\Pr(S)$, with the likelihood of observing the data, had the savings been realised, $\Pr(D|S)$. The latter term is called the likelihood and corresponds to the frequentist idea of inference from the data alone. The numerator $\Pr(D)$ is the most troublesome, as it rarely has a physical meaning. However, it should be included in a normalising factor so that the right-hand side of the equation integrates to one, to make it a proper PDF. It may, therefore, be replaced by a normalising factor. Obtaining this factor and convolving the prior and likelihood PDFs analytically can be difficult or impossible for many kinds of problems. However,

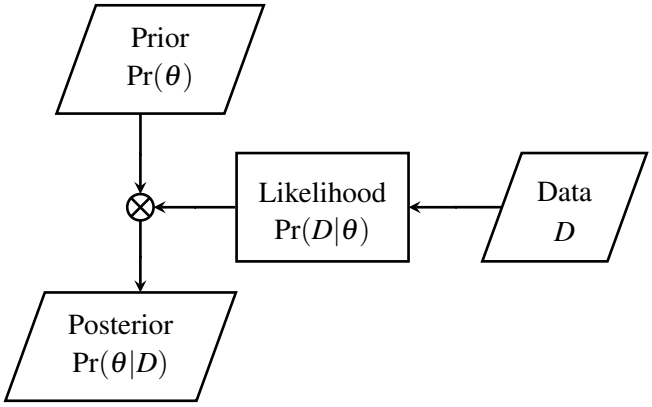


Figure 3.2. Flow chart illustrating Bayes' theorem for a single sampling step.

what has made Bayesian statistics so popular since the early 1990s is that it happens automatically in MCMC numerical computations of the posterior. Otherwise, Bayes' theorem is often written in unnormalised form as

$$\Pr(S|D) \propto \Pr(S)\Pr(D|S). \tag{3.4}$$

In general in statistics, certain model parameters θ are sought, given data D. Therefore Bayes' theorem is often written as

$$\Pr(\theta|D) \propto \Pr(\theta)\Pr(D|\theta) \tag{3.5}$$

which is, the probability of the parameters θ, conditional on the data D, is proportional to the prior on the parameters, $\Pr(\theta)$, multiplied by the probability of the data D given parameters θ. A diagram illustrating this graphically is shown in Figure 3.2.

The Bayesian statistician's task, therefore, is specify credible PDFs for the prior and likelihood functions. If we are mostly ignorant about what we are investigating or want to see the results from the data alone, we can specify a diffuse, also called a non-informative or ignorance prior. If this prior is locally uniform in the region of the more peaked likelihood, and a normal distribution is assumed, then Bayes' theorem reduces to maximum likelihood estimation. If we have a firm prior belief that the hypothesis is false, and set the prior to zero, then no matter how much data we collect, it will not alter our posterior. The same is true for cases where we set the prior to 1 (on which see Estler [193, Eq. 14]). For values of the prior in between zero and one, the posterior is a weighted product of the prior and likelihood, with the weights based on the variances of the two distributions.

3.4.1 Subjectivity, objectivity, and the selection of priors

Since questions regarding the objectivity of the prior are the most common objection to the Bayesian paradigm, they will be addressed briefly here. It should be stated that priors are useful for making implicit assumptions explicit. Implicit assumptions are a danger in both frequentist and Bayesian analysis, although the Bayesian paradigm makes it more difficult to gloss over such assumptions. How the priors are applied is where the controversy arises, though (no matter what is selected). Recently, Gelman and Hennig offered a helpful philosophical discussion on this question [249], which is recommended for further reading.

The first point to note is that the prior is not implicitly subjective or objective. It simply *is*. Subjective Bayesians such as Lindley [239] and de Finetti [250] have used the prior to encode their prior beliefs (although this is not the only interpretation). They reason that this simulates the human decision process by which our prior beliefs are (or should be) updated to incorporate new data. Since the (subjectivist) Bayesian posterior is seen as describing our state of knowledge about a system, it means that if two people differ in their assessment of the data, they are reasoning from different priors. Although the following quotation from an interview with Tom Redman for the *Harvard Business Review* does not refer to subjective Bayesian analysis specifically, it illustrates the point well:

> "You always lay your intuition on top of the data," he explains. Ask yourself whether the results fit with your understanding of the situation. And if you see something that doesn't make sense ask whether the data was right or whether there is indeed a large error term... And, he says, never forget to look beyond the numbers to what's happening outside your office. "You need to pair any analysis with a study of the real world. The best scientists – and managers – look at both." [251]

There are methods for eliciting such personal distributions [252, 253], but it is not always easy or possible since decision makers are less rational than we would like to think [254]. The subjective prior is also strict in the sense that it may not be changed, regardless the outcome of the model. This can be frustrating when one runs a model and finds that a diffuse prior allows the model to explore regions of the solution space (combinations of model parameters) that are physically impossible or unlikely. For these, and other reasons contained in the discussion of Lindley's paper [239], a subjectivist Bayesian approach is not recommended for M&V.

Objective Bayesians such as Jaynes [255] and Berger [256], try to make the prior as non-informative or diffuse as possible, so that the data "speaks for itself", and the result is not altered by the analyst's beliefs. This seems like a better idea than the subjectivist approach and is the more popular choice. Much research has been conducted on representing ignorance (or using neutral priors) in Bayesian statistics. A normal distribution with a large variance, or a uniform distribution, is often used. For cases where the scale may span several orders of magnitude, a scale-invariant prior is needed, and the Jeffreys prior [241] of $\Pr(\theta) = 1/\theta$ is useful since $d(\ln\theta) = constant$. However, non-informative priors also influence the outcome and may have a negative impact in cases where pertinent information is omitted [256, 257]. Not all non-subjectivists would ignore it. In energy studies, for example, informative priors based on data from previous studies have often been used and enjoy a strong precedent [74, 77, 87, 106]. Ignoring such prior information, if available, may not be wise.

A falsificationist (or empirical) Bayesian approach would adapt the prior to constrain the model, as is done in Section 4.3.4.3. Care must be taken when selecting data-dependent priors, however, since these can lead to a case of "data reinforcing data". This results in misleadingly high confidence on posterior estimates. Nevertheless, when such techniques are used correctly, they do have precedent [257], and are mathematically defensible in some instances, as was shown by Darnieder in his PhD thesis on the topic [258].

In this thesis, the author hopes that the chosen priors are not controversial since they will be derived rationally from repeated measurements in time-series data. Bayes' theorem works so that the posterior for a single analysis can be used as a prior for the subsequent analysis. In other words, the result of drawing n samples and analysing them simultaneously will be the same as the result of drawing one sample, inserting the result into the likelihood, calculating the posterior, using this posterior as a prior for the next draw, and repeating n times. This property is used for time-series measurements, so that the prior contains information from previous sampling results, as shown in Figure 3.3. When the posterior distribution is of the same form as the prior distribution, it is called a *conjugate prior*. This is a useful property which will be exploited in this thesis. Specifically, the normal-normal and beta-binomial prior-likelihood pairs will be used. These are all of the exponential (e^x) family of distributions. In simplified terms, if the prior and likelihood are both e^x, their product is e^{2x}, which can easily be used as the prior for another calculation with a likelihood e^x, and so on. Estler [193, Eqs 50-56] and Lira [192, Eqs 5-8] give clear examples of this using the normal distribution.

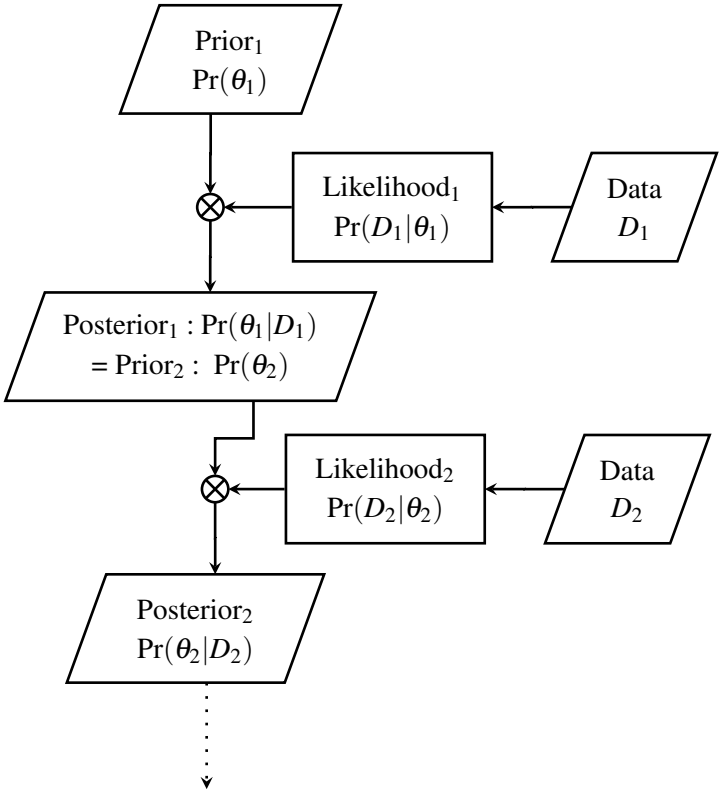


Figure 3.3. Diagram of Bayes' theorem, where the posterior of the first step is used as the prior for the second step. This diagram can be compared to the single-step case in Figure 3.2.

3.4.2 Information and entropy

Since uncertainty quantification is one of the main themes in this thesis, it is worthwhile to mention how uncertainty is described in the Bayesian paradigm. Uncertainty is usually measured by the variance of the distribution in question. More data increases the 'information' about the parameter of interest and decreases uncertainty by decreasing the variance in the estimate. In the Bayesian paradigm[2] 'information' is not just a figurative term, but refers to an actual quantity. As discussed up to now, it is important to specify correct distributions in the calculations, since one can unwittingly add more information than is justified by the data, or omit information that is available. For unimodal distributions, uncertainty is represented by the variance in the distributions specified. However, this only describes *location* uncertainty, and would not hold for bimodal distributions, for example. In such cases, there is a high likelihood of the parameter being in two distinct areas and a low likelihood of it being in the range between. Variance is therefore an incomplete description of uncertainty.

[2] and some aspects of the frequentist one as well

A more general opposite for information is *entropy*, as described by Shannon in his information theory [259]. Shannon's description is applicable in many different Bayesian settings, for example in sample size determination [247]. The upshot of the current discussion is that it is also useful for finding least-informative distributions. Specifically, using the theory one can prove that the maximally entropic distribution for an uncertainty stated as a confidence/precision value (popular in M&V and the GUM [181]), is the normal distribution. Therefore assuming the normal distribution for such cases is not only convenient but also has the mathematical property of being 'maximally non-committal with regard to missing information' [260].

3.4.3 Numerical and non-parametric calculations

Were it not for the advent of numerical methods, Bayesian statistics would have remained a theoretical field, applicable only to simple problems. However, in the early 1990s, a method developed for statistical physics [261] was applied to the problem, and the MCMC Bayesian revolution began. The simplified explanation of MCMC is that it is a random-walk algorithm with a Markov Chain function whose stationary distribution is the posterior distribution of the Bayesian model. The algorithm visits different parts of the solution space in direct proportion to their probabilities in the posterior, and converges reliably on the analytical solution [262]. The more samples (steps) are taken, the higher the resolution of the posterior distribution, similar to Monte Carlo simulation.[3] Different samplers choose the next steps in their random walks more or less intelligently, but the process remains the same. The most popular algorithms at the time of writing are the Gibbs Sampler [263], NUTS (No U-Turn Sampler) [264] and ADVI (Automatic Differentiation Variational Inference) [222], which is used in this thesis.

MCMC algorithms have allowed Bayesian data analysis to flourish, since many different kinds of distributions and functions can be used in simple and hierarchical models, with the MCMC resolving all of the underlying mathematics. As with all machine learning algorithms, some knowledge of the algorithm is useful for specifying and debugging models efficiently. However, for better or for worse, the modeller needs to know only the basics of Bayesian theory to solve many different kinds of problems.

[3] Note that this is the simulation sample size, based on the sample data. It does not "manufacture" data, but approximates the sample data with increasing accuracy as the simulation size increases.

The most popular software for doing MCMC is JAGS (Just Another Gibbs Sampler) [265], Stan [266], and PyMC3 [267].

Although a whole section could be devoted to non-parametric Bayesian regression, it will be mentioned only briefly. Bayesian data analysis as described in this thesis usually refers to *parametric* models. In parametric modelling, the exact functional relationship between the data and the output is described, and the distributions on the model parameters are obtained. However, in many cases it is not possible to specify such models. Non-parametric Bayesian regression models, specifically GPs and GMMs, have proven very useful for energy studies, specifically in the work of Heo *et al.* [77,91,93] . GPs and GMMs allow for Bayesian uncertainty quantification in the context of heteroscedasticity and models that cannot be described parametrically. Although these methods can be very useful, they do not extrapolate well, as far as the author is aware. This is a disadvantage for time series forecasting as is done in this thesis, and will therefore not be pursued further. However, for models where the range of the independent variables is not wider in the reporting period than in the baseline period, these methods allow for accurate uncertainty quantification and regression without needing to worry about specifying change points or a functional relationship.

3.5 APPLICATION

In this section, some of the theory discussed so far will be applied to the M&V case, with some simplifications.

Many guides to Bayesian metrology have been written, but the most helpful ones for M&V are by Estler [193] and M.G. Cox *et al.* [195]. The notation does differ significantly between different authors, however. This can be confusing to someone new to the field but helps in the sense that the principles rather than the notation are learned.

3.5.1 Regression

Since chapters 5 and 6 discuss specific regression methods in depth, and Shonder and Im [178] have described Bayesian M&V regression well, only general remarks about regression will be made below.

In M&V, one often uses the baseline data ($\mathbf{D}_b$) to infer the baseline (pre-retrofit) model parameters $\boldsymbol{\theta}$ through an inverse method:

$$\boldsymbol{\theta} = f^{-1}(\mathbf{D}_b, t), \tag{3.6}$$

Where $f(\cdot)$ is a function relating the independent variables to the energy use of the facility, and t is time. The model parameters describe the sensitivity of the energy model to the independent variables such as occupancy, outside air temperature, or production volume. For some inverse methods such as linear regression, uncertainty quantification is reasonably straightforward. However, linear regression seldom captures the different states of a facility's energy use, and piecewise regression is then implemented. Such regression techniques for M&V have been proposed [33, 65, 176, 268, 269], and they tend to work reasonably well if their assumptions are satisfied, but they are not stable in all cases, are approximate [178], and the assumptions are often restrictive.[4] Once the parameters have been obtained, they may be substituted into the function f so that the 'adjusted baseline' energy use in the reporting (post-retrofit) period can be predicted, given the reporting period data $\mathbf{D}_r$. The adjusted baseline energy use E_{ab} can be represented by the predicted value

$$\hat{E}_{ab} = \int_R^T f(\boldsymbol{\theta}, \mathbf{D}_r, t)dt. \tag{3.7}$$

where R is the time of the retrofit and T is the end of the study. If the energy use during the reporting period is E_r, the energy saved is simply $E_{ab} - E_r$. However, this is a deterministic description of M&V. Since all quantities are uncertain, it would make sense to treat them probabilistically, as in the Bayesian paradigm.

Because Shonder and Im have written a detailed paper on Bayesian regression in M&V [178], and because work in later chapters is similar, only a short conceptual discussion will be given here. Kruschke [19] and Gelman [220] have also given clear, detailed accounts of general Bayesian regression.

Bayesian analysis proceeds according to (3.5). Suppose one has a simple regression model where the energy use of a building $\mathbf{E}$ is correlated with the outside air temperature $\mathbf{T}$. Let the intercept coefficient be θ_0, and a slope coefficient θ_1. One could then write

$$\mathbf{E} = \theta_0 + \theta_1 \mathbf{T}. \tag{3.8}$$

[4] If other techniques such as genetic algorithms or particle swarm optimization [270] are used for parameter estimation, uncertainty quantification may not be possible.

In standard linear regression, one would write $\hat{\boldsymbol{\theta}}$ as the vector of two coefficients and do some linear algebra to obtain their estimates. There would be a standard error on each, which would indicate their uncertainties, and if the assumptions of linear regression, such as normality of residuals, independence of data, homoscedasticity, etc. hold, then it would be accurate. In Bayesian regression, one would describe the distributions on the parameters as

$$\Pr(\boldsymbol{\theta}|\mathbf{D}) \propto \Pr(\mathbf{D}|\boldsymbol{\theta}) \Pr(\boldsymbol{\theta}) \sim N[\hat{\boldsymbol{\theta}}, \boldsymbol{\sigma}], \tag{3.9}$$

where $\boldsymbol{\sigma}$ is the standard deviation on the estimates. Generating random pairs of values from the posterior, at a given value of T, according to the appropriate distributions, will yield the posterior predictive distribution. This is the distribution of energy use at a given temperature, or over the range of temperatures. Overlaying such realisations onto the actual data is called the posterior predictive check.[5]

For a problem such as the one above, Bayesian regression is hardly necessary. However, in modern Bayesian software such as PyMC3 [267], the code used to describe and run the Bayesian model is not much longer than that for linear regression and runs in a few seconds on a standard computer. Therefore both could be used. The difference becomes more apparent when the Bayesian model is extended. A simple way to change the model to be more robust to outliers is to use a Student's T-distribution [272] rather than a normal distribution. The heavier tails accommodate outliers better so that they have a smaller leverage or influence on the regression line. Non-linearity, or even generalised linear models are also very easily described by simply changing the functional relationship in the model specification. One disadvantage of Bayesian regression is that it does not scale well. For high-dimensional problems it would take very long to run, and may not explore the solution space fully unless much care is taken with the sampler.

A further advantage in the Bayesian paradigm is the use of hierarchical models. This is due to the model structure rather than the Bayesian calculation itself (it also works for MLE) [19], but it is nevertheless useful in M&V. Suppose that multiple measures are installed at multiple sites so that the IPMVP Option C: Whole Building Retrofit is used for M&V. The UMP Chapter 8 [273] reports that there are two ways to analyse such data. The two-stage approach involves first analysing each facility separately and then using these results for the overall analysis in stage two. The fixed effects approach analyses all buildings simultaneously but assumes that the effects are constant across them. This then

[5] Note that strictly speaking one can specify more scale-invariant priors for regression coefficients than simply using normal or uniform distributions in $Pr(\boldsymbol{\theta})$ [271]. However, the author has not seen this done in practice.

uses an average effect for all buildings. Hierarchical modelling considers both the individual facility's energy saving and the overall effect, simultaneously. It does this by assuming that the group effects are different realisations of an overarching distribution with a mean and variance, which is used as a prior. This can lead to 'shrinkage' (of the variance, similar to shrinkage in stratified sampling), because the group effects are mutually informative. For groups with little data, the overarching effect distribution plays a larger role, and for groups with more data, a smaller role. Also, the overall variance is reduced because the sources of inter-facility variance are isolated from that of inter-measure variance. The result for a hierarchical model is that the effect estimation for an individual facility is influenced by the overall estimate of the measure effect, as well as by the data for the facility. As another example, consider a program that retrofits air conditioning units in different provinces in South Africa. One could fix the savings effect across all facilities, but this will underestimate some and overestimate others. Or one could analyse by facility, then by province, and then overall. The hierarchical model provides a better alternative in these cases and comprises the bulk of many Bayesian data analysis texts [19,220]. Booth, Choudhary, and Spiegelhalter have provided an excellent study on using hierarchical Bayesian models in M&V [274]. For comment on the errors-in-variables model used by them, see Section 4.3.3.3.

3.5.2 Sampling

The goal of measurement is to reduce our uncertainty about a parameter θ (which is our prior $\Pr(\theta)$) by taking measurements. When this is done our state of knowledge is represented by $\Pr(\theta|\mathbf{D})$, which we obtain via Bayes' theorem. Stated the other way around, we enhance our state of knowledge by obtaining data. That is, as long as the data is of such quality as to outweigh the prior distribution.

If we are certain that the true value lies within 5% of some known value, and we take measurements with an instrument that has an accuracy of 20%, we are not increasing our certainty about the true value, but calibrating our instrument. No matter how much data we collect, our prior $\Pr(\theta)$ will be more sharply peaked than the likelihood $\Pr(\mathbf{D}|\theta)$, so that the posterior is dominated by the prior.

To illustrate how such intuition is reflected mathematically by Bayes' theorem, a well-known calculation is demonstrated. More comprehensive descriptions can be found in Kruschke [19] and Estler [193]. Normal distributions are assumed, as these are both common and simple for illustration purposes. Let

the prior estimate of the energy use be denoted E_p centred at E_0, which follows the distribution

$$Pr(E_p) \sim N[E_0, \sigma_p] = \frac{1}{\sigma_p\sqrt{2\pi}} \exp\left[\frac{-(E-E_0)^2}{2\sigma_p^2}\right]. \tag{3.10}$$

if E is some measured value. The system is then measured by an instrument with some error. Assume that the instrument is unbiased, so that

$$\Pr(E_m) \sim N[0, \sigma_m] = \frac{1}{\sigma_m\sqrt{2\pi}} \exp\left[\frac{-(E_m-E_0)^2}{2\sigma_m^2}\right]. \tag{3.11}$$

If a measurement with result E is taken, then the uncertainty in the value is $\Pr(E_m|E) = N[E, \sigma_m]$. However, usually one is not interested in E_m itself, but in the true value E. Using Bayes' theorem, this can be obtained by

$$\Pr(E|E_m) \propto \Pr(E_m)\Pr(E_m|E) \tag{3.12}$$

so that

$$\Pr(E|E_m) \propto N[E_0, \sigma_p] \times N[E, \sigma_m]. \tag{3.13}$$

It simplifies the mathematics to work with the *precision* of the distribution, rather than the variance or standard deviation, where the precision is the reciprocal of the variance. The precision is often denoted τ. [6] Therefore

$$\tau = 1/\sigma^2. \tag{3.14}$$

By convolving the two distributions in (3.13), and with some algebraic manipulation[7] one finds that:

$$\Pr(E|E_m) \propto \frac{\bar{\tau}}{\sqrt{2\pi}} \exp\left[\frac{-(E-\bar{E})^2\bar{\tau}}{2}\right]. \tag{3.15}$$

where

$$\bar{\tau} = \tau_p + \tau_m \tag{3.16}$$

and

$$\bar{E} = \frac{\tau_p E_0 + \tau_m E_m}{\tau_p + \tau_m}. \tag{3.17}$$

The resultant distribution in (3.15) is $N[\bar{E}, 1/\bar{\tau}]$, and by substituting the values of the above equations into this formula, the posterior of the mean of E can be described. Its mean is the weighted average of the prior and measurement, where the weights are determined by the respective precisions. If one can increase the precision τ_m by taking repeated measurements with accurate equipment, these measurement data will dominate the posterior distribution, and the prior will have a small effect. If n

[6] It should not be confused with its notation in Chapters 5-7 where it denotes 'present time'.

[7] Detailed by Kruschke [19, Ch. 16].

measurements are taken, then

$$\bar{E} = \frac{\tau_p E_0 + n\tau_m E_m}{\tau_p + n\tau_m} \tag{3.18}$$

and

$$\bar{w} = \tau_p + n\tau_m. \tag{3.19}$$

The main point of this calculation is to illustrate how the Bayesian paradigm makes mathematical sense of repeated measurement and the associated decrease in uncertainty. These results, although useful, are limited by the fact that only the mean energy use E is estimated, and it is assumed that the variances are known. Allowing both to vary simultaneously causes the mathematics to become more involved. In Section 3.5.4 the power and simplicity of Bayesian numerical methods are illustrated for such a case.

3.5.3 Measurement

The Bayesian approach also makes sense of how one should approach nuisance parameters. A nuisance parameter is one whose value is uncertain and influences the outcome of a process but is not of interest itself. A good example of this is measurement error in M&V. The way the Bayesian paradigm deals with nuisance parameters in the context of conditional probability is called *marginalisation*. The example below is a very simplistic treatment of mismeasurement, and references to more complex methods are discussed in Section 2.3.2. Gregory [275, p.68] also presents a worked out example similar to the one below.

Suppose that one wants to infer a parameter θ from the data: $\Pr(\theta|D)$. The parameter could be energy use. However, one measures the data with an instrument containing an error a, so that the joint distribution

$$\Pr(\theta, a|D) \tag{3.20}$$

is produced. This is illustrated graphically in Figure 3.4. However, one is only really interested in

$$\Pr(\theta|a, D). \tag{3.21}$$

To marginalise a out of the equation means to integrate over a. This collapses the vertical axis of Figure 3.4 by summing all of the columns into the "margin", to obtain (3.21): a probability distribution over θ only, but considering a. This is expressed mathematically as

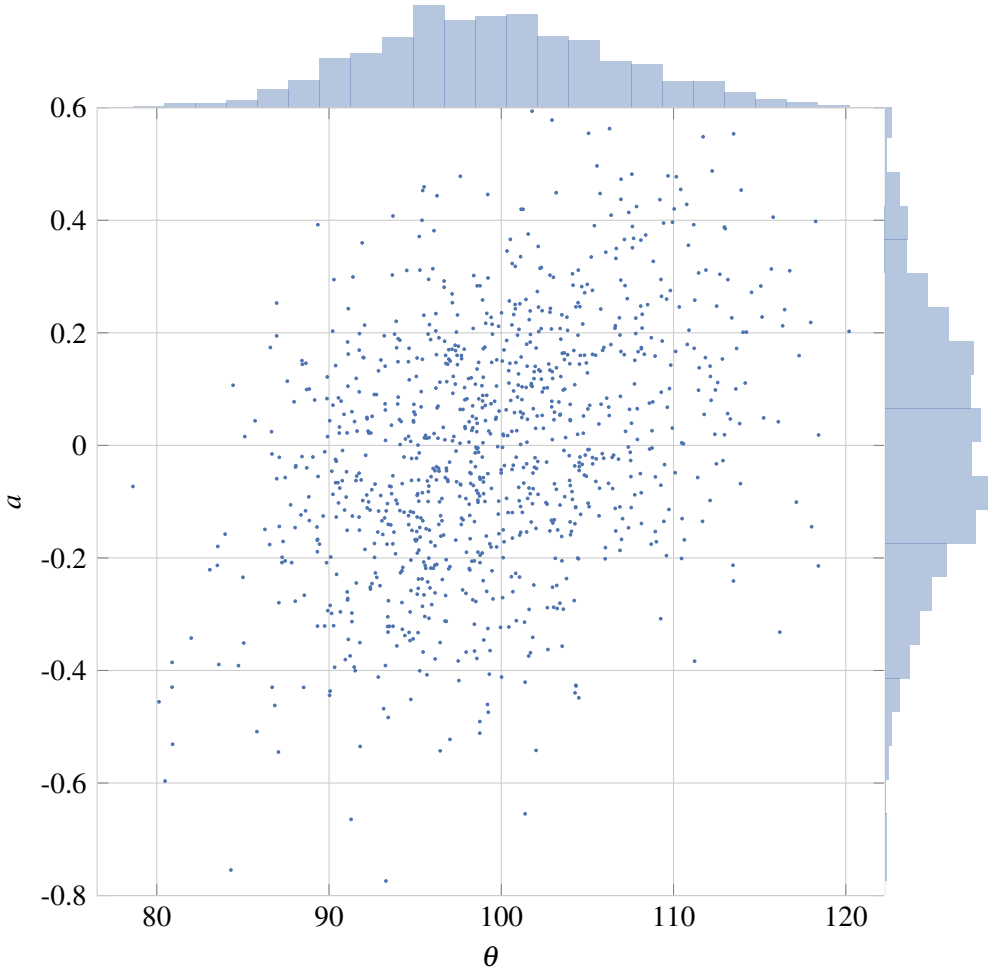


Figure 3.4. Joint distribution of a and θ.

$$\Pr(\theta|a,D) = \int_a \Pr(\theta|a)\Pr(a|D)da. \tag{3.22}$$

In practice, however, θ needs to be inferred from the data, and for this Bayes' theorem is needed. Let $a \sim N[a,\sigma_m]$ where σ_m is the standard deviation. Also, let the measured data fall in a distribution $D \sim N[m,s]$. Implementing Bayes' theorem, we obtain

$$\Pr(\theta|a,D) = \int_a \Pr(\theta)\Pr(a|D,\theta)\Pr(D|\theta,a)da. \tag{3.23}$$

Since the prior is not a function of a, it may be removed from the integral:

$$\Pr(\theta|a,D) = \Pr(\theta)\int_a \Pr(a|D,\theta)\Pr(D|\theta,a)da. \tag{3.24}$$

If one assumes for the sake of simplicity that the measurement error is additive and therefore independent from the measurand, θ may be neglected from the metering accuracy term:

$$\Pr(\theta|a,D) = \Pr(\theta)\int_a \Pr(a|D)\Pr(D|\theta,a)da. \tag{3.25}$$

By substituting the normal distribution formula into each conditional probability term, the probability may be written as:

$$\Pr(\theta|a,D) = \Pr(\theta) \int_a \underbrace{\frac{1}{\sqrt{2\pi}\sigma_m} \exp\left(\frac{-(a-1)^2}{2\sigma_m^2}\right)}_{\text{marginalised}} \underbrace{\frac{1}{\sqrt{2\pi}s} \exp\left(\frac{-(m-a\theta)^2}{2s^2}\right)}_{\text{likelihood}} da. \tag{3.26}$$

In this way, the measurement error term is 'marginalised' or integrated out. The two-dimensional joint probability distribution of (3.20) is collapsed to only the θ dimension: the variable of interest, as a function of the data and the nuisance parameter (measurement uncertainty). In very simple cases, the mathematics reduce to adding the variances of convolved independent normal distributions according to $\sigma_{total}^2 = \sigma_1^2 + \sigma_2^2$. However, in practice, Bayesian software packages do this calculation numerically on the user's behalf. As with most Bayesian modelling, the modelling decisions relate to which distributions to specify, and how to relate them to the parameters (in a hierarchical model), rather than doing integration.

3.5.4 IPMVP example

To illustrate a practical Bayesian M&V model, consider the following example from the IPMVP [1]. Twelve readings are taken by a meter. These are reported as monthly readings, but are assumed to be uncorrelated with any independent variables or other readings, and are therefore construed to be random samples. The values are

$$\mathbf{D} = [950, 1090, 850, 920, 1120, 820, 760, 1210, 1040, 930, 1110, 1200]. \tag{3.27}$$

The units are not reported, and the results below are therefore left dimensionless, although kWh would be a reasonable assumption. These data were carefully chosen, and have a mean $\mu = 1\,000$, sample standard deviation $s_s = 150$.

3.5.4.1 IPMVP solution

The standard error is $SE = 43$. The confidence interval on the mean is calculated as

$$CI = \mu \pm t \times SE \tag{3.28}$$

Since $t_{90\%,11} = 1.80$, the 90% confidence interval on the mean was calculated as $1000 \pm 1.80 \times 43 = (933,\ 1\ 077)$, or a 7.7% precision. Energy metering uncertainty is not considered in this calculation.

3.5.4.2 Bayesian solution

The Bayesian estimate of the mean is calculated as follows. First, prior distributions on the data need to be specified. Vague priors will be used:

$$\Pr(\mu) \sim Uniform[0, 2000] \tag{3.29}$$

$$\Pr(\sigma) \sim Uniform[0, 1000] \tag{3.30}$$

A t distribution will be used for the likelihood below, and the degrees of freedom parameter (ν) of this distribution will therefore need to be specified. One could fix ν for the t-distribution at 12. However, if outliers are present or if the data has more or less dispersion than the standard t-distribution, this would not be realistic. It is therefore warranted to indicate the uncertainty in the data by specifying a prior distribution on ν. Kruschke [240] recommends an exponential distribution with the mean equal to the number of data points. This allows an equal probability of ν being higher or lower than the default value:

$$\Pr(\nu) \sim Exponential[1/12]. \tag{3.31}$$

If $\boldsymbol{\theta} = (\mu, \sigma, \nu)$, the likelihood is:

$$\Pr(\mathbf{D}|\boldsymbol{\theta}) \sim StudentT\left[\Pr(\mu),\ \Pr(\sigma),\ \Pr(\nu)\right]. \tag{3.32}$$

Note that the t distribution is not used because of the t-test, but because its heavier tails are more accommodating of outliers. Any distribution could have been specified if there was good reason to do so. The posterior on μ is plotted in Figure 3.5. It was simulated in PyMC3 using the ADVI algorithm with 100 000 draws, which is stable and converges on the posterior distribution in 10.76 seconds on a middle-range laptop computer.

It is important to note that no probability statements about the values inside the frequentist interval can be made, nor can one fit a distribution to the interval. The distribution indicated is strictly a Bayesian one. The Bayesian (highest density) interval is slightly wider than the frequentist confidence interval, at a precision of 8.5%. If ν were fixed at 12, (indicating that we are certain that the data does indeed reflect a t distribution with 12 degrees of freedom exactly), Bayesian and frequentist

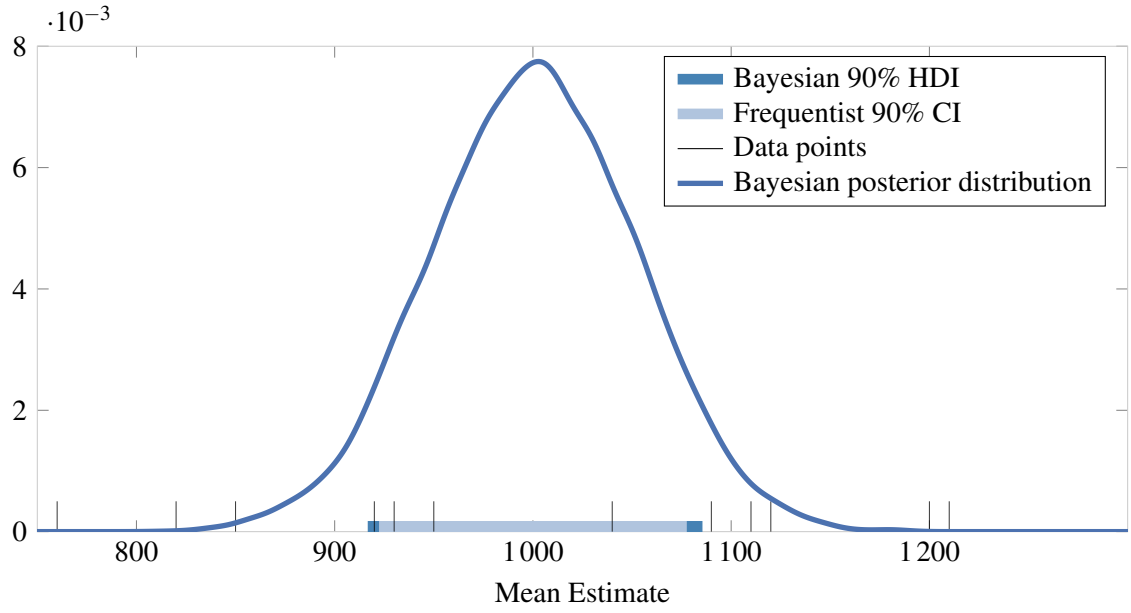


Figure 3.5. Illustration of Bayesian posterior density $\Pr(\mu|\mathbf{D})$, 90% HDI, and frequentist 90% CI.

intervals correspond exactly. However, the Bayesian alternative allows for a more realistic value. With comparisons between two groups (two-sample t-tests), the effect of uncertainty in the priors becomes even more pronounced [240].

The posterior distribution can now be used to answer many interesting questions. For instance, what is the probability, given the data at hand, that the true mean is below 900? Alternatively, is it safe to assume that the standard value of 950 is reflected by this sample, or should the null hypothesis be rejected? (If previous data to this effect is available, it could be included in the prior, maybe using the equivalent prior sample size method [253]). The frequentist may say that there is not enough evidence to reject the null, but cannot accept it either. In the Bayesian paradigm, 950 falls comfortably within the 90% confidence range, and can therefore be accepted at that level. As a further question, if this is an energy performance contracting project, and we assume that the data points are different facilities rather than different months, would it be worthwhile taking a larger sample to increase profits, if we believe that the true mean is at 1 100? (On which see Lindley [246], Bernardo [247] and Goldberg [69]).

3.6 CONCLUSION

The Bayesian approach provides an intuitive and coherent framework by which M&V uncertainty can be assessed. It treats all unknown parameters as random variables and codifies the interactions of their probability distributions given the data. Although most M&V problems will not be solved analytically, the analytical solutions provide a logical foundation for how measurement and sampling uncertainties are treated intuitively. Numerical solvers and modern software have greatly expanded the range of application of the Bayesian approach, and have precluded the need for analytical solutions. In many cases, the Bayesian approach is preferable to standard frequentist methods regarding theory and has become simple to implement in practice.

CHAPTER 4 ENERGY METERING UNCERTAINTY AND CALIBRATION

4.1 CHAPTER OVERVIEW

This chapter will consider two aspects of the cross-sectional energy metering uncertainty problem. The first is the contribution of energy metering uncertainty to overall metering-and-sampling uncertainty. It considers the difference that installing less accurate meters make on the overall uncertainty. Since it will be shown that energy metering uncertainty plays a small part in overall project uncertainty, the next question is whether this discovery can be used to minimise project cost. A portion of a large project's metering costs is due to meter calibration. A low-cost meter calibration method is therefore devised. To devise such a method, the nature of mismeasurement is explained, and different mismeasurement mitigation techniques are considered. A technique is then decided upon, and after modification is tested on a real-world data set.

There is a third aspect to the cross-sectional sampling problem, which relates to how many meters should be installed to estimate the population's energy use accurately. For single-year studies, optimal simple random or stratified sampling formulae could be used. But for multi-year studies, the information from previous years could be used to reduce planned future sample sizes. The mathematical framework for this calculation is left to Chapter 6. [1]

[1] This chapter is based on a conference paper presented at the International Energy Programme Evaluation Conference [72] and journal articles published as part of the author's thesis [27–29].

4.2 ENERGY METERING UNCERTAINTY IN THE CONTEXT OF OVERALL PROJECT UNCERTAINTY

The summation of simple measurement and sampling uncertainties is reasonably well understood, both in statistics and in M&V. Note that this does not include attenuation bias due to mismeasurement as described in Section 2.3.2 and later in this chapter. For now, the discussion will focus on simple energy metering uncertainty. Although G14 [17, 31] provides the best description in terms of formulae and calculation methods, to understand the relative contribution of measurement uncertainty to overall uncertainty it is necessary to rearrange some of the known formulae and plot them graphically.

Before a detailed investigation of measurement uncertainty can be made, the sampling distribution should be carefully defined. There are three distributions relevant to sampling: The *population distribution* is the true distribution of the population, and is unavailable to the engineer unless he samples the total population with perfect measurement equipment. The *sampling distribution* is the idealised distribution for samples of a given size. With perfect measurement equipment, the sample distribution will be equal to the sampling distribution. The *sample distribution* is the observed distribution on the sample that was actually taken, with the measurement equipment actually used. This is the only distribution accessible to the engineer.

The calculations below are only valid under the standard statistical assumptions of independent, normally distributed data. We also assume that although the measurement instrument may be inaccurate, a large population of such instruments will be unbiased. This implies that the measured sample mean will tend to the true mean as the sample size tends to infinity. It is also assumed that measurement errors are normally distributed around the mean.

Let the subscript $_s$ denote the (theoretical) sampling distribution, and the subscript $_m$ denote measurement parameters. Furthermore, let σ_m be the measured standard deviation of the sample and z_m be the standard score of the known confidence level α on the measured data. Since only measured data is available, consider s_m as the sample standard deviation and $\bar{x}$ as the sample mean, and p_m as the precision or error bound. The upper limit of this error bound should be equal to the upper confidence limit:

$$\bar{x} + p_m\bar{x} = \bar{x} + s_m z_m, \tag{4.1}$$

$$\therefore p_m\bar{x} = s_m z_m, \tag{4.2}$$

$$\therefore s_m = \frac{p_m}{z_m}\bar{x}. \tag{4.3}$$

The standard deviation, and therefore the variance and distribution on the measurement data has now been characterised by writing the standard deviation in terms of the known precision level, desired confidence level, and the sample mean.

The error in a measurement system may be expressed statistically as a standard deviation from the mean, or it may be expressed as a maximum error. The maximum error approach is popular and conservative. However, it represents a highly unlikely and unnecessarily strict case where all the individual errors are assumed to be at their maxima simultaneously. Instead, the statistical approach will be considered here. The total error is calculated as a root mean square, which is the way in which standard deviations are added. It should also be stated with a certain confidence level.

Errors can also be expressed in absolute or relative terms. 200 kWh$\pm$10 kWh has an absolute error of 10 kWh, but a relative error of 5%. The expressions for adding and multiplying uncertain values differ according to which expression is used. Relative errors will be used in this thesis.

When combining two independent normal distributions, the means are added arithmetically. However, the total variance of the combined distribution should be

$$s^2_{combined} = s^2_s + s^2_m, \tag{4.4}$$

where s^2_s is the sampling variance. But from (4.3),

$$s^2_m = \frac{p^2_m}{z^2_m}\bar{x}^2. \tag{4.5}$$

Therefore,

$$s^2_{combined} = s^2_s + \frac{p^2_m}{z^2_m}\bar{x}^2. \tag{4.6}$$

It is useful to define these relations in terms of the coefficient of variance(CV), since this makes the calculation independent of the size of the mean and variance:

$$\text{CV} = \frac{s}{\bar{x}}. \tag{4.7}$$

Also, since sample size required for M&V reporting is proportional to the CV value, the relative contribution of energy metering uncertainty to $\text{CV}_{combined}$ is an indication of size of the effect of energy metering uncertainty on overall project cost. Substituting (4.6) we can now define the combined CV as

$$\text{CV}_{combined} = \frac{s_{combined}}{\bar{x}} = \frac{\sqrt{s^2_s + \frac{p^2_m}{z^2_m}\bar{x}^2}}{\bar{x}}. \tag{4.8}$$

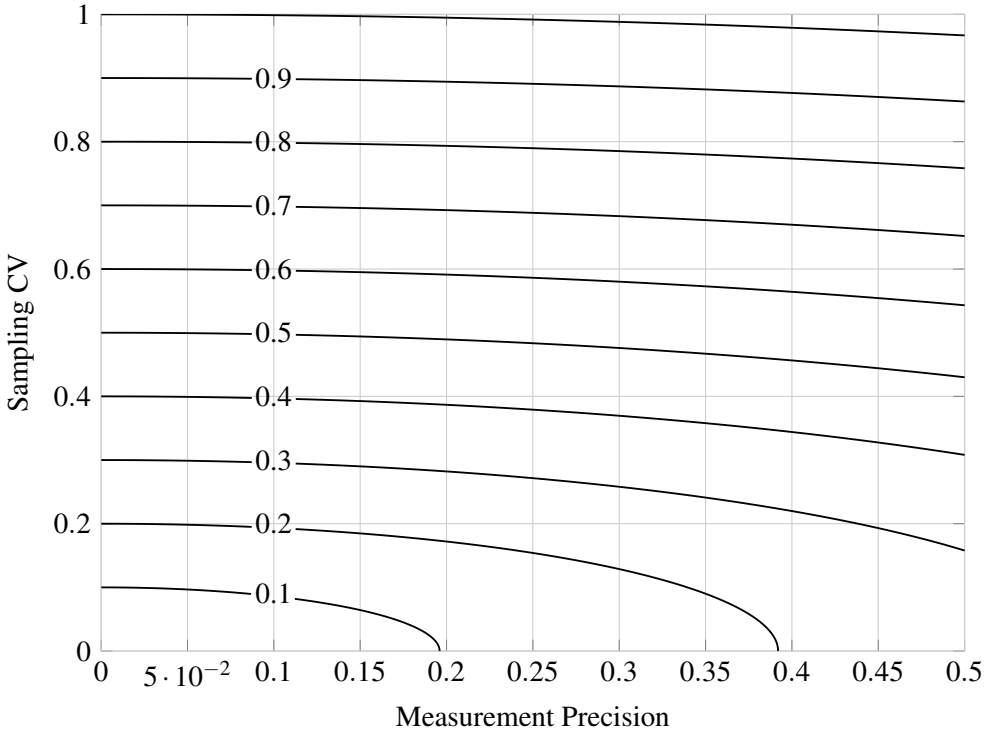


Figure 4.1. Contour plot of combined CV as a function of sample CV and measurement precision in equation (4.9), for a measurement confidence of 95%.

This may be simplified to

$$\mathrm{CV}_{combined} = \sqrt{\mathrm{CV}_s^2 + \frac{p_m^2}{z_m^2}}, \tag{4.9}$$

The combined CV has now been reduced to a formula needing only values that are readily available (meter accuracy), and widely estimated (CV_s [15]). An example of (4.9) is plotted in Figure 4.1 at the 95% confidence level, which is the most common one used in metrology [113]. This corresponds to a "coverage factor" of $k = 2$, or 2σ. We can see that for $p_m \leq 0.1$ and $\mathrm{CV}_s \geq 0.2$, the overall uncertainty is dominated by sampling uncertainty, and energy metering uncertainty can be safely neglected.

The formula for the sample size n required to report with a given confidence α_r at z_r, and a precision p_r, is:

$$n = \frac{z_r^2 \mathrm{CV}^2}{p_r^2}. \tag{4.10}$$

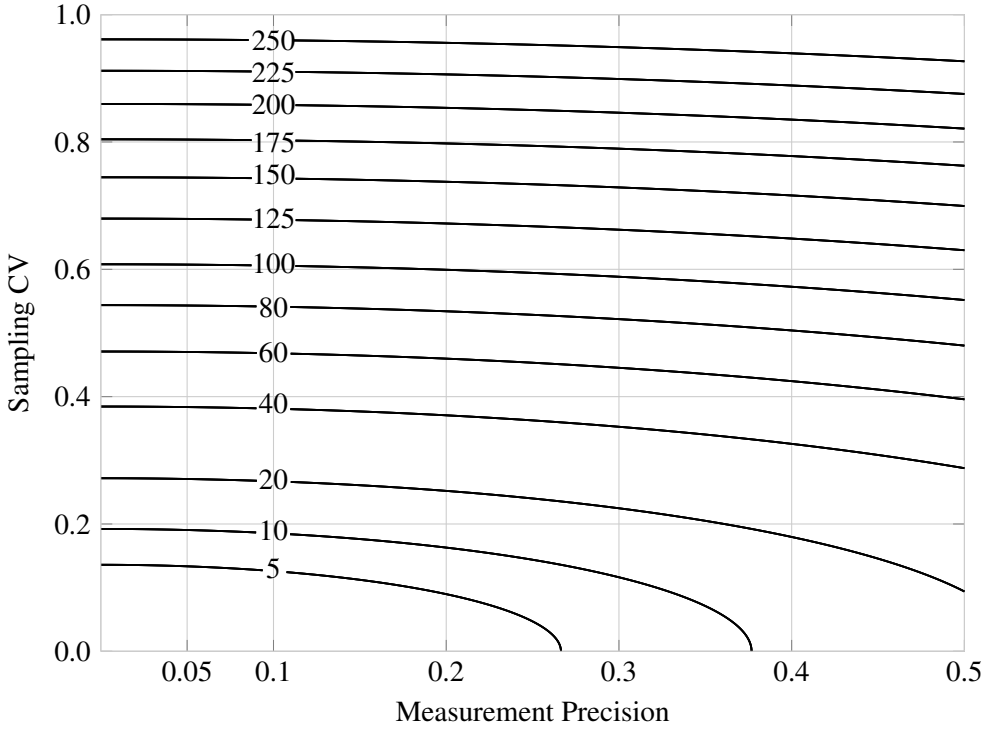


Figure 4.2. Contour plot of sample size as a function of sample CV and measurement precision.

By substituting (4.9), one can write the required sample size as a function of sampling CV, and measurement accuracy, and required reporting precision:

$$n = \left(\mathrm{CV}_s^2 + \frac{p_m^2}{z_m^2} \right) \frac{z_r^2}{p_r^2}. \tag{4.11}$$

An example of (4.11) is plotted in Figure 4.2.

4.2.1 Practical implementation

Consider the case of energy meters which conform to the IEC 62053-22 [23]. These standards specify that electricity meters should have an accuracy of 0.5% for class 0.5S and 0.2% for class 0.2S during normal operation. However, for the 0.5S class, precision may be up to 1% for low power factors. ASHRAE 14-2002 Technical note #7 of A5.6.2.1 [31, p.91] gives the instrument system error as 2%, which includes the CT accuracy. The standards do not specify a confidence interval on these values. One may therefore select the 2% value as a realistically low precision, and $z_m = 1.96$, which

corresponds to a 95% confidence level [113]. One may also assume CV$=0.05$: that is, a stable process with a coefficient of variance of 5%. This reduces the contribution of sampling uncertainty to overall uncertainty.

Using the above-mentioned figures, by (4.9), the ratio between $\mathrm{CV}_{combined}$ and CV_s is:

$$\frac{\mathrm{CV}_{combined}}{\mathrm{CV}_s} = \frac{\sqrt{\mathrm{CV}_s^2 + \frac{p_m^2}{z_m^2}}}{\mathrm{CV}_s}. \tag{4.12}$$

Therefore, by substituting our assumptions above,

$$\frac{\mathrm{CV}_{combined}}{\mathrm{CV}_s} = \frac{\sqrt{0.05^2 + \frac{0.02^2}{1.96^2}}}{0.05} = 1.021. \tag{4.13}$$

Thus in a worst-case scenario, energy metering uncertainty would add 2.1% relative to the sampling uncertainty. In other words, if the uncertainty on the savings is 10%, sampling uncertainty comprises roughly 9.8% of this figure, and energy metering uncertainty comprises the other 0.2%. It can be seen that in such cases, energy metering uncertainty may be neglected in most practical applications.

4.2.2 G14 uncertainty formula sensitivity analysis

As further confirmation, the G14 uncertainty summation formula 4-8 was considered. This formula combines metering, sampling, and modelling uncertainties and is widely used in industry. It is a more complicated formula than that described above, with many input variables. To consider the importance of energy metering uncertainty on the total G14 uncertainty, a sensitivity analysis was conducted. A Sobol' sequence [276, 277] with Saltelli *et al.*'s sampling improvement [278] was implemented in Python via the SALib module [279]. It was run with 50 000 points, and the overall sensitivity was considered.[2] Energy metering uncertainty proved to be one of the least influential factors considered. The results of the sensitivity analysis are shown in Table 4.1. From this, and the result of the calculation in the previous subsection, it is clear that energy metering uncertainty plays a small role in overall M&V reporting uncertainty for cases where sampling is done.

[2]The Morris method [280,281] could also be used and has proven popular among energy researchers as shown in Table 2.2, because it is more computationally efficient for the calculation of higher-order effects. This becomes important for expensive BEMs. Menberg, Heo, and Choudhary [282] have shown that it gives similar results to the Sobol' sequences for such cases.

Table 4.1. Sensitivity Analysis of ASHRAE G14 [17, 31] overall M&V uncertainty equation, in descending order of overall influence.

Parameter	Range	Overall Sensitivity Score
Fraction Saved	(0, 0.5)	1.449667
Lag-1 Autocorrelation Coefficient	(0, 1)	0.263855
Model CVRMSE	(0, 0.5)	0.040357
Reporting Confidence	(68%, 95%)	0.023823
Sampling CV	(0, 1)	0.019965
Number of Reporting Period Points	(12, 60)	0.019742
Proportion of Population Surveyed	(10%, 100%)	0.011008
Population Size	(2, 100)	0.008594
Number of Baseline Period Points	(12, 60)	0.007507
Energy Meter Uncertainty	(0%, 5%)	0.000029
Independent Variable Uncertainty	(0%, 5%)	0.000026
Number of Parameters in Regression Model	(2, 5)	0.000005

4.2.3 Sampling power

The standard sampling formula (4.10) recommended by leading guidelines is not robust, or in statistical terms, yields underpowered study designs. To illustrate, the following simple example is presented.

According to the formula, one needs 68 samples for 90/10 precision, if the CV of the sampling population is 0.5. Randomly generate 68 data points (sampling results) from such a distribution:

$$\mathbf{D}_{0-68} \sim N[1000, 500] \tag{4.14}$$

As in Section 3.5.4.1, the standard error (SE) is calculated as $\sigma/\sqrt{n}$, and the 90% confidence interval as $\mu \pm 1.645 \times SE$. By repeatedly sampling $n = 68$ points according to the distribution above, and checking whether the true mean does, in fact, lie in the interval, and whether these 90% bounds are less than 10% from the mean, it can be verified whether the interval does, in fact, satisfy the 90/10 requirement.

This simulation was repeated 10 000 times. It was found that the interval does not contain the true mean 10.75% of the time. This accords with the 90% frequentist interval specification. The 0.75% additional violation can be reduced to 0.21% if the t-score for 67 degrees of freedom is used, rather than the z-score (which assumes an infinite sample size). The precision is where the problem lies, however. The lower bound on the interval is more than 10% away from the sample mean in about 44% of cases. In more than 50% of cases, the interval either does not contain the true value, or the precision bound is violated. In other words, when using (4.10), there is a 50/50 chance of not reaching the desired precision level. This demonstrates that the sample sizes yielded by (4.10) are underpowered. If a sample size of $n = 80$ is used instead, the precision constraint is violated in only 16.5% of cases, and the total violation rate drops to 24.5%. If $n = 100$ is used, these are reduced to 0.83% and 11% respectively.

These values do not change noticeably when measurement errors between 0.2% (Class 0.2S meter) and 3% (Class 3 meter) are added.

4.2.4 Metering vs sampling uncertainty conclusion

The practical implication of the results above is that more accurate and expensive meters do not provide an advantage over more cost-effective meters in cases where sampling is done, all else being equal. There is no increase in risk for the project developer or client when using more cost-effective metering for such projects [72]. As long as meters are properly calibrated and suitable for the environmental conditions of the application, more valuable information will be gained from installing a larger number of standard meters, rather than a smaller number of high-accuracy meters.

4.3 LOW-COST CALIBRATION

4.3.1 Introduction

This section builds on the background given in Section 2.3.2.

The first question to be answered is, “is it really necessary to develop a low-cost calibration method when the chapter up to now has shown that energy metering errors do not contribute meaningfully

to overall uncertainty for cases with standard sampling and modelling uncertainties?". It is indeed necessary. The results up to now have shown that more expensive meters with very high accuracies are not necessary in such cases. However, any meter still needs to be calibrated. While reducing the purchasing cost of a meter may reduce the capital outlay for a project or an M&V company, reducing calibration costs will reduce the operational overheads of such a project or company. Note that only meter models qualified to international standards are considered.

South Africa's 12L tax incentive programme [2] requires that M&V meters be calibrated by an accredited laboratory at fixed intervals, and other international programmes adopt similar approaches [112]. This is a sound principle from a regulatory point of view. It minimises the consumer's risk, that is, the risk of using an inaccurate meter and paying for savings that did not occur. However, a significant opportunity cost is incurred because many projects are never implemented due to monitoring, laboratory, and plant shut-down costs. An example of this has been recorded for the CDM lighting retrofit project specifications [47, 48]. Striking a balance between calibration costs and monitoring accuracy is, therefore, an essential but non-trivial consideration for policymakers.

Furthermore, the European Measurement Instrument Directive (MID) [147] requires that meters be calibrated in-situ, that is, in the environment in which they will be installed [148]. Besides regulatory compliance in European countries, a method capable of doing this is also convenient and practical.

One of the reasons imprecise reference instruments are avoided is because it will lead to an error-in-variables effect, requiring Measurement Error Models (MEMs).[3] To the best of the author's knowledge, MEMs have not been applied to electrical meter calibration before. The Bayesian approach will be used below. The method proposed in this chapter is therefore novel for a number of reasons. Calibration is usually done in a laboratory, using highly accurate and expensive laboratory equipment, whereas this method will use a commercial-grade meter as a calibrator. Calibration usually does not account for errors in the calibrator, whereas this method will do so. To the author's knowledge, Simulation Extrapolation has not been used for meter calibration and has also not been combined with Bayesian regression as is done in this paper. Finally, the proposed approach provides a more practical solution to in-situ calibration than those proposed in the literature.

[3] See Section 2.3.2.

It is recognised that calibration is about more than having access to an accurate reference instrument and that quality and traceability procedures as set out in ISO 17025 [184] should also be in place. However, even energy meters calibrated to lower accuracies than the current classes should be sufficient for most M&V applications, where uncertainties are dominated by other factors.[4]

The cost saving from using the method proposed in this paper will vary with the number of meters disciplined instead of being sent to a calibration laboratory. The cost saving for the client will also vary with the cost of facility down-time needed to install and remove meters. The meters needed when using the proposed method are not more or less accurate than standard energy meters, and their accuracy will normally be determined by other factors than the method proposed.

The commercial meter-as-calibrator will measure with a non-negligible error, and therefore the error-in-variables effect should be taken into account. A range of scenario-specific MEMs has been developed to account for how the measurement errors may arise. The nature of the errors needs to be classified accurately to apply the correct MEM to a problem. In some cases, certain simplifying assumptions may restrict the model's applicability. In others, incorrect assumptions may lead to erroneous results. Mismeasurement in M&V is treated more fully in Section 2.3.2. Carroll *et al.* [196] and Gustafson [197] have also written excellent textbooks on the topic.

The notation $\mathbf{x}$ will be used to denote the true values of the independent variable (reference instrument or calibrator) and $\mathbf{y}$ the true values of the dependent variable (UUT). To differentiate between the true values and the observed values which are measured with error, an asterisk (*) is used for measured values.

Before considering the errors themselves, two related concepts need to be mentioned. An *exposure model* is often needed when specifying an MEM. Although we often have a model of how errors arise in the form $f(\mathbf{x}^*|\mathbf{x})$, we cannot work backwards to infer $\mathbf{x}$ from the observed $\mathbf{x}^*$. An exposure model describes this function: $f(\mathbf{x}|\mathbf{x}^*)$. This is often done through a third variable $\mathbf{z}$. The exposure model then takes the form $f(\mathbf{x}|\mathbf{z})$, where $\mathbf{z}$ is some covariate measured without error.

Model identifiability is another concern. Sometimes a key piece of information is missing, and the data are not enough to identify all the model parameters uniquely. Carrol *et al.* [196] and Gustafson [197]

[4] compare Section 2.3.1.1.

adopt complementary approaches. Briefly, Gustafson found that non-identifiability is not always detrimental, and Carroll *et al.* found that formal identifiability is not always good enough, especially for threshold cases. Gustafson also found that specifying uncertainty (Bayesian priors) on some parameters may even lead to better results than fixing those parameters at slightly incorrect values for the sake of identifiability.

4.3.2 Error taxonomy

Errors may vary in a number of ways. First, errors can be **correlated or uncorrelated**. This is not in the same category as the classifications that follow but is an important distinction nonetheless. Errors that are uncorrelated with other variables are the simplest to model. Consecutive errors may also be autocorrelated in a time series. This sequentiality is hidden in scatter plots and regression analyses, although it still affects the estimates.

Errors can be **classical or Berkson**. If ε is a generic error term, classical errors take the form $\mathbf{x}^* = \mathbf{x} + \boldsymbol{\varepsilon}$, and are more common. This is when the error is in the instrument itself. Berkson errors take the form $\mathbf{x} = \mathbf{x}^* + \boldsymbol{\varepsilon}$. This occurs when the actual value of the measurand varies around the assigned or measured value, because the source of the error is external to the instrument.

Errors are classified as **multiplicative or additive**. Multiplicative errors are of the form $\mathbf{x}^* = \mathbf{x}\boldsymbol{\varepsilon}$, whereas additive errors take the form $\mathbf{x}^* = \mathbf{x} + \boldsymbol{\varepsilon}$. The additive error assumption is a popular one as it greatly simplifies MEM mathematics: additive errors are usually associated with constant variance throughout the measurand range. This is called *homoscedasticity* and is a critical assumption when performing Linear Regression (LR). The majority of techniques have been developed to describe this kind of model. However, this assumption is not always valid. For example, it has been demonstrated that energy meter measurement errors are non-linear and multiplicative [217], and are thus *heteroscedastic*. This has been acknowledged to produce problems in econometric energy analyses [283], and frequentist methods to account for some cases in regression analysis has been developed [201]. It may be mitigated by assuming a log-normal distribution and working with $log\mathbf{x}^*$, since $log\mathbf{x}\boldsymbol{\varepsilon} = log\mathbf{x} + log\boldsymbol{\varepsilon}$, transforming the error model to an additive one. However, the assumption of a log-normal distribution on ε (so that $log\varepsilon \sim$ Normal), although mathematically convenient, is not always valid or preferred [196]. Heteroscedasticity can be present even for additive errors when they have non-constant

bounds over the measurement range, such as energy meters and CTs [22–25]. These bounds are shown in Figure 2.2.

Errors may be **differential or non-differential**. Non-differential errors mean that $\mathbf{x}^*$ contains no more information about $\mathbf{y}$ than $\mathbf{x}$ does. The response does not change due to measurement. Differential errors may occur when the response $\mathbf{y}$ is measured before the covariates $\mathbf{x}^*$ and $\mathbf{z}$, and these variables are liable to change. For example, the diet ($\mathbf{x}$) of women with breast cancer may be measured only after their diagnosis $\mathbf{y}$. It is possible that the test subjects change their diet as a result of the diagnosis [196]. Another example is when $\mathbf{x}^*$ is a proxy for $\mathbf{x}$, not simply a mismeasurement. For example, plug loads are sometimes used as a proxy for occupancy [11]. Differential errors may also occur in ex-post energy use surveys for residential retrofit programmes where the response (purchasing of certain equipment, for example) is measured before other variables of interest are measured.

Last, the function $\mathbf{y}(\mathbf{x})$ may be **linear or non-linear**. This is not an assumption about the errors themselves but does affect the kinds of errors that are permissible. The linear assumption is popular as it allows LR to be used if one assumes normally distributed additive errors. For many models, this is a valid assumption. However, Carobbi, Pellicci, and Vieri [217] have shown that the standard $P = VI$ electrical power equation, where P is Power in Watts, V is potential difference in Volts, and I is current in Amperes, can be modelled as

$$P_n = (1+\alpha)VIcos(\phi+\phi_c)+\varepsilon, \tag{4.15}$$

when an energy meter measures with error. In this equation, α is the gain error, ϕ_c is the phase error, and ε is the bias error. The gain error α changes the amplitude of measured power fluctuations, but does not affect the mean. In other words, the larger the energy reading, the larger the error. The bias error ε offsets the measured power, changing the mean power read by the meter, but not the amplitude of the fluctuations. This error may bias the power and energy reading upwards or downwards. The phase error ϕ_c has a similar net effect to the gain error, but changes according to the power factor error of the meter. Carobbi, Pellicci, and Vieri's contribution [217] was to show that (4.15) is a statistically adequate model, capturing the real error behaviour of energy meters without specifying too many parameters.

Although this error is multiplicative, the error bounds in the IEC meter qualification standards [22–24] are additive. The meter may still have a multiplicative error, but this error is always smaller than the

additive error bound. In cases where these are the only data available, additive errors may have to be assumed. Furthermore, the error model is only non-linear if the phase error term ϕ_c is of interest.

4.3.3 Meter calibration

The method below focusses on energy meters but can be used for instruments measuring other parameters as well. The most analogous cases are flow measurement [57], and possibly exhaust gas analysis [58]. Occupancy measurement may also benefit from thoughtful application [8, 11], but temperature measurements are often biased due to spatial variations [152], and will require more careful application.

The proposed approach is to discipline a meter (the UUT) using another relatively low-specification commercial-grade metering system. This could be done by installing the meters in parallel in-situ at the facility for a short period, such as 24 hours if both measure at a resolution of 15 minutes. The data from the calibrator are then used to correct (discipline or calibrate) the data from the UUT. Although the UUT is not calibrated, we assume that it is of reasonable quality. For example, the model range to which the UUT belongs should be qualified to an IEC specification. This is necessary to ensure that readings will remain stable under different operating conditions such as winter and summer temperatures.

For high-accuracy laboratory multimeters measuring to six or eight decimal places, various additional factors should be considered during calibration. These include thermoelectric voltages, cable impedance, and performance at different frequencies [182]. However, these fluctuations are small enough to be negligible for commercial energy measurement applications.

4.3.3.1 Errors in x

The calibrator data is selected for the x-axis, rather than the UUT. This is because the calibrator should have smaller errors than the UUT. In this way, attenuation bias is minimised as much as possible before MEM adjustments are made.

Table 4.2. Accuracy specification for IEC Class 3 meter [24]. P_n denotes the rated power, I_n rated current, and I_{max} the maximum current. See also Figure 2.1

Value of Current	Power Factor	Error limit
$0.02I_n \leq I \leq 0.05I_n$	1	$\pm 0.04P_n$
$0.05I_n \leq I \leq I_{max}$	1	$\pm 0.03P_n$
$0.05I_n \leq I \leq 0.1I_n$	0.5	$\pm 0.04P_n$
$0.1I_n \leq I \leq I_{max}$	0.5	$\pm 0.03P_n$

To be conservative, the highest (least accurate) IEC class meter and Current Transformer (CT) combination will be used as a reference instrument. This would be a Class 3 meter [24] with a Class 5 CT [25]. The meter accuracy limits are shown in Table 4.2. For power factors between ± 0.5 and ± 1, the accuracy limits were linearly interpolated. The CT has a flat accuracy limit of 5% of the rated current. These are additive error bounds relative to the rated, or full scale, current. It is assumed that this meter is calibrated. The true errors may still be multiplicative but will fall within these additive bounds.

Metrology guidelines often recommend that a uniform error distribution between the error bounds be assumed [18]. However, this is too conservative. Instead, errors bounds are assumed to be the 95% confidence limits on a normal distribution [18, 31]. The readings are also assumed to be unbiased. Errors are assumed to be classical, non-differential, and uncorrelated. Even though errors are additive, they are heteroscedastic (having non-constant variances) due to the stepwise nature of the error bounds as described by Table 4.2 and Figure 2.1. The total error would be the root sum of squares of the meter and CT error bounds at a given point:

$$p_{combined} = \sqrt{p^2_{meter}(x) + p^2_{CT}(x)}. \tag{4.16}$$

Let $p_{combined}(x)$ be the combined error bound at x, and z be the standard score (or coverage factor). The standard deviation on the a given reading can then be written as

$$\sigma_u = \frac{p_{combined}(x)}{z}. \tag{4.17}$$

The rated power of the meter is assumed to be 200 kW, and the rated current for the CT is assumed to correspond to this value.

The measured values on the calibrator $\mathbf{x}^*$ can then be defined as

$$\mathbf{x}^* \sim N[\mathbf{x}, \sigma_u] \tag{4.18}$$

4.3.3.2 Errors in y

For errors in our UUT ($\mathbf{y}$) more detailed assumptions may be made. Following Carobbi *et al.* [217], the characteristic function for the UUT is assumed to be

$$\mathbf{y}^* = (1+\alpha)\mathbf{x}cos(\phi+\phi_c)+\varepsilon, \tag{4.19}$$

where α is the gain error, ϕ is the phase difference between voltage and current, ϕ_c is the phase error, and ε is the bias error. The errors are classical, with multiplicative and additive components. They are also homoscedastic, and the function is non-linear. Since these errors will not cause attenuation bias, the MEM is not selected on their basis. However, they are built into the overall measurement model.

4.3.3.3 MEM selection

Since ϕ_c is one of the variables of interest, this is a non-linear function, and standard LR techniques such as Fuller's method of moments [216] are not valid unless the $cos(\phi+\phi_c)$ term in (4.15) and (4.19) is neglected.

Although $f(\mathbf{x}^*|\mathbf{x})$ is available by (4.18) in the form of a distribution function, $f(\mathbf{x}|\mathbf{x}^*)$ is not. To obtain this, an exposure model would be needed, which is not available.

One approach would be to specify a naïve Bayesian model on the data using (4.18). By specifying a distribution on $\mathbf{x}^*$, the noisy independent variable is taken into account, mitigating the attenuation effect to some degree. If errors were Berkson rather than classical, this would be accurate. However, this is not the case for measurements under investigation.[5]

[5]The author believes that this mistake was made in an otherwise excellent previous Bayesian errors-in-variables investigation for M&V [274].

Since the availability of an exposure model, repeated measurements, or a sub-set of gold-standard measurements is not assumed, MEMs like Regression Calibration, Maximum Likelihood techniques, and the Bayesian approach are not available. Instead, a hybrid SIMEX solution is proposed

4.3.3.4 SIMEX

SIMEX is a simple, powerful algorithm that compensates for measurement error using only $f(\mathbf{x}^*|\mathbf{x})$ in the form of σ_u explained in Section 2.3.3.2. It was first proposed by Cook and Stefanski [284], and a useful summary can be found in Carroll, Stefanski, *et al.* [196]. The premise is that although the biased parameter estimates $\{\alpha^*, \phi_c^*, \varepsilon^*\} = \boldsymbol{\theta}^*|\mathbf{x}^*$ cannot be unbiased directly, they can be biased even more by adding more noise to $\mathbf{x}^*$. By repeating this biasing for increasing noise levels, the relationship between noise in $\mathbf{x}$ and bias in $\boldsymbol{\theta}$ is found. A trend can be observed from these successive noise levels, and the noise-free state $\boldsymbol{\theta}|\mathbf{x}$ can then be inferred by backwards extrapolation. Figure 4.5 illustrates this graphically. The SIMEX procedure can be defined more rigorously as follows:

1. Describe the variance σ_u due to mismeasurement.

2. Describe the UUT function $\mathbf{y} = f(\mathbf{x})$.

3. Specify the vector of noise multiples to obtain a vector $\boldsymbol{\zeta}$ of length n at which simulation will be done. Values for $\boldsymbol{\zeta}$ can start at zero and could go up to five.

4. Calculate $\mathbf{x}^*_{\boldsymbol{\zeta},\mathbf{n}} = \mathbf{x}^* + (1+\sqrt{\boldsymbol{\zeta}})\sigma_u$. The reason for the square root on ζ is explained by Carroll *et al.* [196], but is beyond the scope of this study.

5. Solve $\mathbf{y}^*_{\boldsymbol{\zeta},\mathbf{n}} = f(\mathbf{x}^*_{\boldsymbol{\zeta},\mathbf{n}})$ to find $\boldsymbol{\theta}(\boldsymbol{\zeta})$. If $f(\mathbf{x})$ is linear, this can be done by LR. For non-linear problems, an appropriate function should be specified, and an optimization algorithm is needed to solve for the function parameters.

6. For every element of $\boldsymbol{\theta}$ (that is, $\alpha, \phi_c, \varepsilon$), a vector of n solutions in $\boldsymbol{\zeta}$ is now available. Consider the gain error α. If the function $\hat{\alpha}(\boldsymbol{\zeta})$ were linear, one could now solve

$$\hat{\alpha}(\boldsymbol{\zeta}) = a_\alpha \boldsymbol{\zeta} + b_\alpha. \tag{4.20}$$

Carroll *et al.* [196] divided $\boldsymbol{\zeta}$ into discrete levels with many samples per level. They then used the mean of every level of $\boldsymbol{\zeta}$. However, since this is not an expensive step, one would rather regress against the full data set than assume that the distribution is symmetric. Also, rather than using discrete levels, a linear spacing of points between the maximum and minimum values of $\boldsymbol{\zeta}$ was used.

7. The unbiased parameter estimate $\alpha|\mathbf{x}$ is found by solving (4.20) for $\zeta = -1$. This is illustrated graphically in Figure 4.5.

8. Repeat Step 7 for ϕ and ε.

4.3.4 Case study: SIMEX application

The SIMEX algorithm was modified slightly and applied to the meter calibration problem at hand. Initially, the algorithm was tested with an energy data set of linearly interpolated points between 0 and I_n, at three different power factor levels. This simulates a laboratory set-up. However, to simulate in-situ calibration, real load profile data was needed. The actual energy consumption of a university residence at the University of Pretoria, on 2 February 2016 was used. The data are plotted in Figure 4.3. The power factor was converted to a phase angle by $\theta = cos^{-1}(Power\ Factor)$.

A graphical representation of the process used to produce the values in this case study is shown in the flowchart of Figure 4.4.

One problem with such data is that power factor and energy use are correlated. High power factors occur at high loads, and low power factors occur at lower loads. This could be due to heavy loads such as geysers having unity power factors and forcing the overall power factor upwards during peak times. Such a correlation has a confounding effect on parameter estimation, of ϕ especially. Using larger calibration data sets such as a one-week rather than a one-day period helps only marginally since the system still has the same correlation characteristics.

For the experiment, the (unknown) parameter values are set as shown in Table 4.3, and altered the data using (4.18) and (4.19) to produce the observed data $\mathbf{x}^*$ and $\mathbf{y}^*$. The SIMEX algorithm was

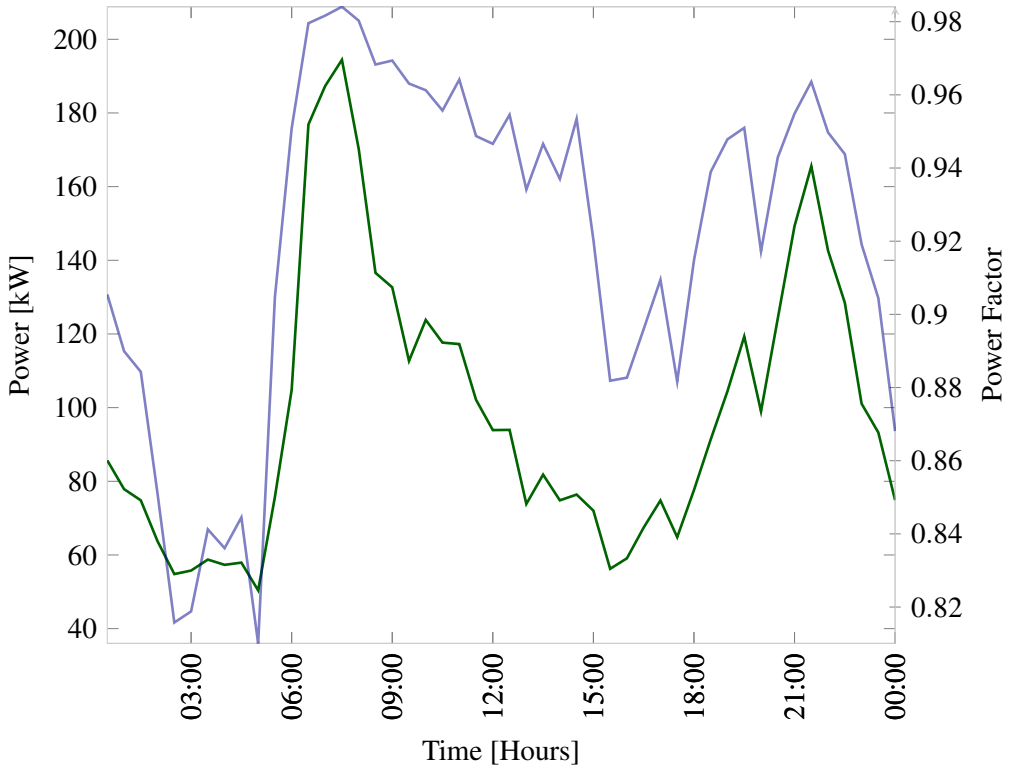


Figure 4.3. Load (green) and power factor (blue) profiles for the period used for calibration.

Table 4.3. Parameter values

Parameter name	Symbol	Value
Gain Error	α	0.2
Phase Error	ϕ_c	0.2
Bias Error	ε	$\sim$ Normal[5, 2.5]

implemented in the following manner, according to the steps described in Section 4.3.3.4:

1. The variance σ_u is described by (4.16).

2. The UUT function $\mathbf{y}^* = f(\mathbf{x})$ is described by (4.19).

3. The SIMEX graphs were found to be non-linear, especially for ζ values above 2. Therefore, $n = 300$ points between $\zeta = 0.5$ and $\zeta = 5$ were selected. Points between 0 and 0.5 were not included because in this region the data converge asymptotically to $\zeta = 0$, which is an artefact

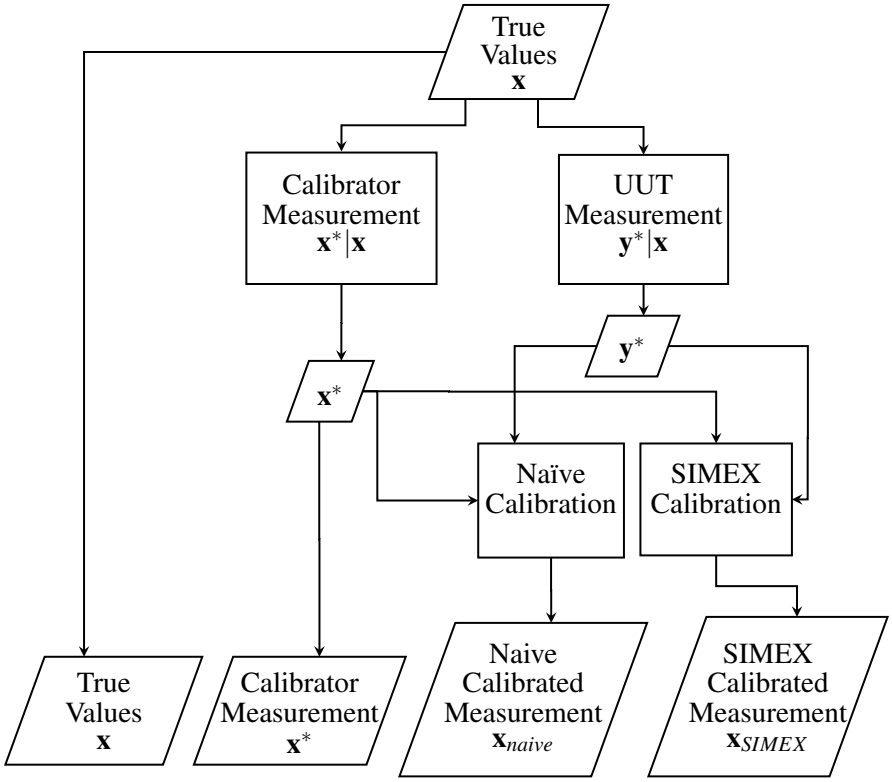


Figure 4.4. Flow chart demonstrating the process of derivation of various values in calibration simulation procedure.

of the algorithm rather than a real trend.

4. These n realisations were generated using Python's `numpy` library [285] and the `numpy.random.normal` pseudo-random number generator for

$$\mathbf{x}^*_{\boldsymbol{\zeta},\mathbf{n}} \sim N[\mathbf{x}^*, \sigma_u]. \tag{4.21}$$

The variance σ_u was defined by (4.18).

5. In this case, Python's `scipy` [286] module was used to find the least-squares solution of (4.19) for $\boldsymbol{\theta}(\boldsymbol{\zeta})$. The library implements the Broyden *et al.* quasi-Newton method [287] by default. Non-default optimization algorithms were also tried but showed poorer convergence and efficiency.

6. A non-linear model was assumed to solve for $\boldsymbol{\theta}(\boldsymbol{\zeta})$. The data exhibit a sigmoid shape, and various sigmoid-shaped functions such as piecewise linear, hyperbolic tangent, sinusoid, and

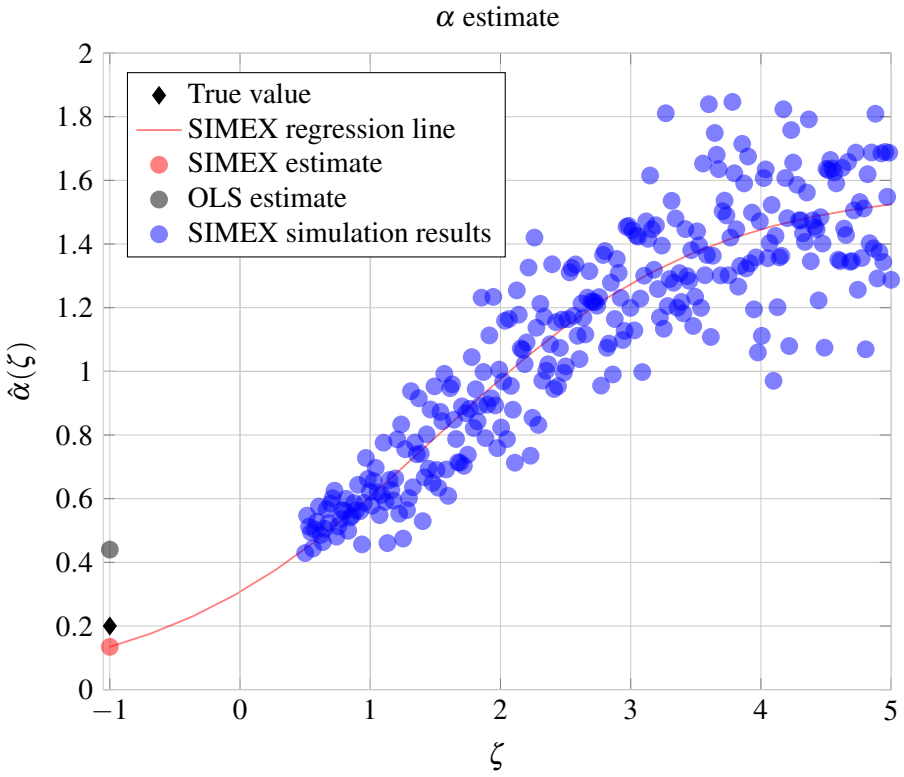


Figure 4.5. Illustration of the SIMEX procedure of Section 4.3.3.4. The error added to the measured data is indicated by the factor ζ, with $\zeta = -1$ indicating the error-free state towards which simulation is extrapolated. This figure illustrates one realisation of the simulations for α.

logistic functions were tested. The standard logistic function below delivered the most reliable results. For α, for example, one would solve

$$\hat{\alpha}(\boldsymbol{\zeta}) = \frac{L_\alpha}{1 + e^{k_\alpha(\boldsymbol{\zeta} - \zeta_{0,\alpha})}} \tag{4.22}$$

for L_α, k_α, and $\zeta_{0,\alpha}$. L determines the curve's maximum value, k determines the slope, and ζ_0 determines the x-value of the midpoint. The data and resultant fit for one realisation can be seen in Figure 4.5. The same optimization algorithm as the previous step was used.

7. Once the unbiased parameter estimates $\hat{\boldsymbol{\theta}}(\zeta = -1)$ were found by substitution into equations such as (4.22), the errors relative to Table 4.3 were calculated as

$$Error = \frac{\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}(\zeta = -1)}{\boldsymbol{\theta}} \times 100. \tag{4.23}$$

Table 4.4. Summary of distributional characteristics of parameter estimate errors for 300 random error realisations. These data are presented graphically in Figure 4.6.

Method	α			ϕ_c			ε		
	2.5%	**Mean**	**97.5%**	**2.5%**	**Mean**	**97.5%**	**2.5%**	**Mean**	**97.5%**
Naïve	-188	-91	-9.21	-245	-162	-58	-459	-286	-123
SIMEX	-23	39	73	-108	-16	57	-173	54	26
Bayes	-62	-3	39	-111	-24	59	-175	-56	25

The author recommends that calibration for M&V purposes only be done using IEC-qualified meters. The overall accuracy of such a system, over the majority of the measurement range, is $\sqrt{0.03^2 + 0.05^2} = 5.8\%$. One can see that the CT error dominates the overall uncertainty [18]. Replacing the meter in this system with a more accurate one will have little effect, reducing uncertainty to 5.4% for a Class 2 meter. However, replacing the Class 5 CT with a Class 3 CT will reduce the overall uncertainty to approximately 4.24%.

Initially, LR was used on a smaller, approximately linear subset of the data, namely $\zeta \in [0,2]$. This worked well for α and ε estimates, but consistently overestimated ϕ_c. The sigmoid shape was also partially hidden while the discrete $\boldsymbol{\zeta}$ approach described in Step 6 of Section 4.3.3.4 was used. If this approach is followed, the mean or mode of each ζ should be plotted rather than the full set, to show the shape of the data more clearly for regression model selection. However, it was found that a linearly spaced $\boldsymbol{\zeta}$ illustrates the shape of the function the best, as is seen in Figure 4.5.

Selecting the right calibration period is important. If calibration is done over a weekend, for example, the proper power and power factor ranges will not be observed. Selecting a good calibration period is easy for a simulation study such as this one where all the data are available. However, it is more difficult in real situations when the data have not been observed yet. Therefore, the in-situ meter calibration period should be selected with care and in consultation with the facility manager. The IPMVP's recommendation for whole-building measurement, that "all operating conditions be represented fairly" during the baseline measurement period, should be followed. Furthermore, if ECMs are installed after the baseline period in an M&V project, meter recalibration may be necessary, depending on the changes. The installation of Power Factor Correctors, which would decouple the power and power factor profiles, is an example of a case where baseline period parameter estimates may not hold during

the reporting period.

4.3.4.1 Discussion of results

Although SIMEX is viable for this case, it does not un-bias parameter estimates perfectly: for certain realisations of random noise, such as where most points happen to be biased in the same direction, the starting data set for $\zeta = 0$ is misleading, and SIMEX estimates will be imperfect. Therefore, to evaluate the reliability of the different methods, the process above was repeated for various realisations of $\mathbf{x}^*$ and $\mathbf{y}^*$ in (4.18) and (4.19). Altogether 300 realisations were simulated, and a summary of the results are shown in Table 4.4 and in a violin plot in Figure 4.6. This figure also shows the SIMEX-Bayes result for comparison. The SIMEX-Bayes method will be introduced and discussed in the next section.

A violin plot is similar to a box-plot in that it shows the probability distributions of the parameters. Where a box plot indicates the quartiles with a box and whiskers, a violin plot shows the full probability density function in mirrored form around a vertical axis. The dashed line indicates the median, and the dotted lines the quartiles. Long, slender shapes such as for the Naïve bias estimate in Figure 4.6 indicate a large variance and thus uncertainty in the estimate. Short, wide shapes like the SIMEX gain estimate indicate low variance and concentrated probability mass. Symmetric shapes such as for the SIMEX phase estimate indicate a symmetric probability distribution around the mean. Asymmetry such as for the SIMEX bias estimate indicates that the parameter estimates are skewed, in this case towards zero.

For Figure 4.6, estimates with zero (error) means will, on average, be error-free, although some variance is expected. This is the desirable result. The first notable observation is that the naïve estimates are further away from the zero line than the SIMEX estimates. This is to be expected: the naïve method should be more biased, and this feature confirms the errors-in-variables theory. It is also observed that the SIMEX estimate errors have smaller variances. This means that the SIMEX method converges on its less biased estimates more reliably. It is, therefore, more robust to the random effects of sampling than the ordinary least squares regression. The error in the ε estimate is the largest. However, to put it in perspective, a 100% error in ε means that $\hat{\varepsilon} = 10$ for $\varepsilon \sim N[5, 2.5]$, given data in the range $(0, 200)$. A 100% error is therefore only a 2.5% error relative to the data range. A 100% error in the gain α

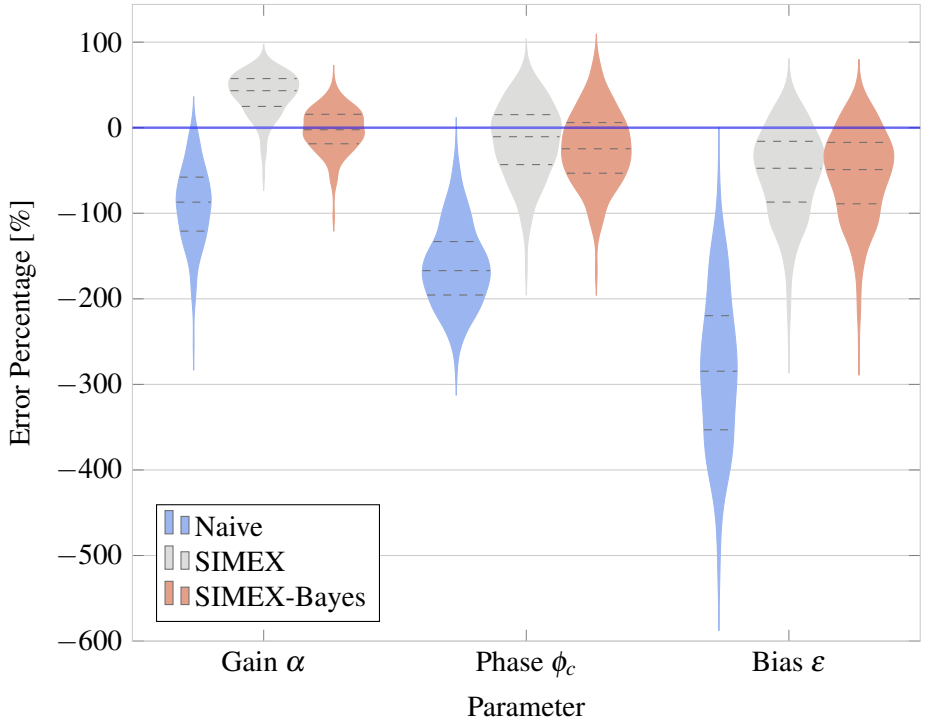


Figure 4.6. Violin plot showing probability distribution shapes for Naïve, SIMEX, and SIMEX-Bayes parameter estimates, with quartiles indicated. A discussion of this figure can be found in Section 4.3.4.1.

could be much more significant (representing a 100% error relative to the data range), although a caveat to this assertion is discussed in Section 4.3.4.2.

From these results it is demonstrated that the SIMEX procedure produces superior estimates to naïve regression, although they are not perfect. However, even if SIMEX produces better estimates on average, the quality of the prediction will depend on the specific combination of estimates in a specific set, and not only on the means across sets. A discussion of this result would be premature in this section, and the reader is referred to Point 4 of Section 4.3.4.2, as in the next section, this interactive effect will be evaluated.

4.3.4.2 Application to meter calibration

The three meters used above will now be compared based on how accurately they predict a longer measurement period than the calibration period. Three cases are considered. The first is a laboratory-

calibrated Class 3 meter with a Class 5 CT. This case is simply the readings of the reference instrument (calibrator) used for disciplining the other two meters. The second is a meter disciplined using the naïve procedure; assuming that the calibrator readings contain no error. The third is a meter disciplined using the SIMEX procedure, with Bayesian refinement. The parameter estimates obtained by disciplining the meter using the data from 2 February 2016 are then used to predict the energy consumption for the period 1 January 2016 - 3 August 2016.

Two goodness of fit metrics were selected to evaluate how well the predictions correspond to the true values for each of these 300 data sets. The NMBE measures whether the model consistently over-predicts or under-predicts energy use. The CV(RMSE) measures how closely the model tracks the actual data up and down: similar to its variance. An NMBE of 0% would indicate no difference between the prediction and actual mean energy use, and a CV(RMSE) of 0% would indicate no variance in the prediction relative to the actual.

For the calibrator, the CV(RMSE) happens to correspond to its combined precision of 5-6%. However, the two metrics express uncertainty in slightly different ways and do not always correspond. Since it is assumed that the meter is unbiased, and specify it in that way for the calibration, its NMBE is close to 0%.

This goodness of fit was evaluated in the following way:

1. Generate observed energy use for the UUT ($\mathbf{y}^*$), for the full data set, by (4.19).

2. Generate observed energy use for the calibrator ($\mathbf{x}^*$), for the calibration period, using (4.18).

3. Using only the 24-hour calibration data set, employ SIMEX and the naïve regression to estimate parameters α, ϕ, and ε.

4. Refine SIMEX estimate through Bayesian regression, discussed further in Section 4.3.4.3.

5. Generate predicted energy use for the full data set by inverting (4.19) using the parameter estimates, so that:

$$\mathbf{x}_{predicted} = \frac{\mathbf{y}^* - \hat{\varepsilon}}{(1+\hat{\alpha})cos(\phi + \hat{\phi}_c)} \tag{4.24}$$

Table 4.5. Summary of distributional characteristics of two goodness of fit metrics for the methods under investigation: the Coefficient of Variation on the Root Mean Square Error (CV(RMSE)), and the Normalised Mean Bias Error (NMBE). These results are presented graphically in Figure 4.7.

Method	**CV(RMSE)**			**NMBE**		
	2.5%	**Mean**	**97.5%**	**2.5%**	**Mean**	**97.5%**
Naïve	3.03	5.8	9.91	0.33	3.08	6.34
SIMEX	4.59	8.87	12.49	-10.344	-6.79	-2.33
Bayes	2.27	2.96	4.35	-2.05	-0.09	2.03

6. As with the calibration procedure in Section 4.3.4, repeat Steps 1-5 300 times to account for different random realisations of $\mathbf{x}^*$ and $\mathbf{y}^*$. The summary statistics of the goodness of fit metrics from these simulations are given in Table 4.5, and plotted in Figure 4.7.

Before the results are discussed, an explanation of the Bayesian refinement is given.

4.3.4.3 Bayesian refinement

Although the parameter estimates of the SIMEX method are clearly superior to the naïve method, as shown in the previous section, Figure 4.7 shows that the resultant CV(RMSE) and NMBE on the rest of the data set are *worse*. The reason is plotted in Figure 4.8.

Although the naïve estimates of the parameters are much worse than the SIMEX estimates, the prediction quality (goodness of fit) is dependent on their combination. Thus α may be overestimated and ϕ_c underestimated, but they cancel each other out in such a way that the final result is close to the true value, especially with noise in ε adding some tolerance to the results. Neglecting ε for a moment, one can visualise this in Figure 4.8. Gain error is the x-coordinate on the map, phase error is the y-coordinate, and CV(RMSE) is the height, indicated by colour. Low CV(RMSE) values form a valley running north-west to south-east. Although there is only one coordinate that is “correct" in the sense of corresponding to the true values, this valley indicates the combinations of gain and phase error values that will also yield a low CV(RMSE). Now, because the sum-squared error is a major component of the CV(RMSE) calculation, a low sum squared error will lead to a low CV(RMSE).

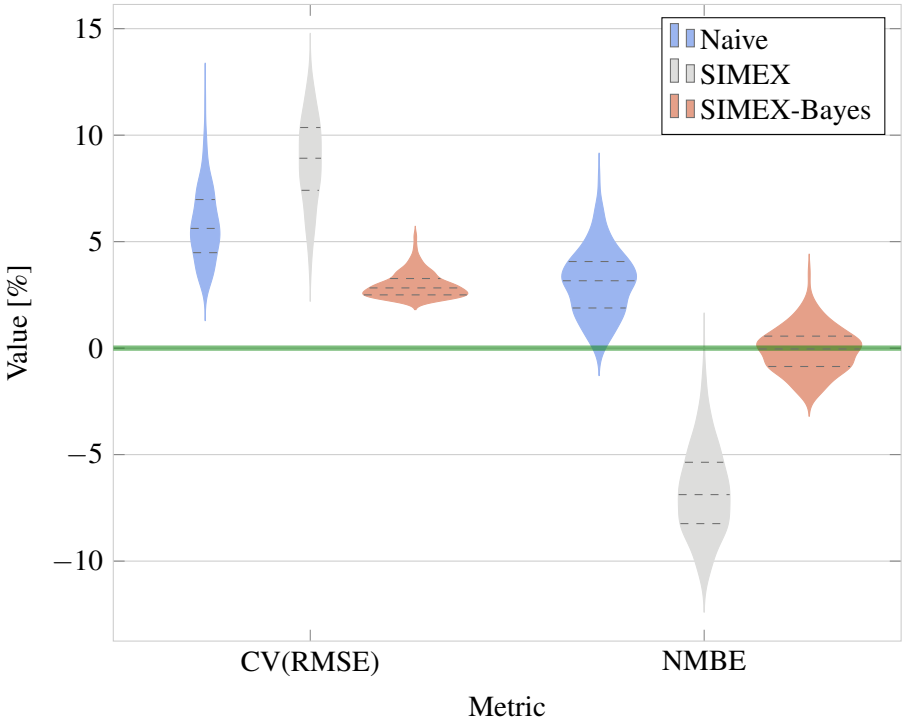


Figure 4.7. Violin plot showing probability distribution shapes of goodness of fit metrics using parameter estimates of Naïve and SIMEX methods. Quartiles and median indicated by dashed lines. Two outliers were removed from the SIMEX plots to improve the vertical scale. A discussion of this figure can be found in Section 4.3.4.4.

Least Squares regression finds a solution with the least sum of squares error. In other words, the naïve method effectively optimizes for CV(RMSE), and we are therefore not surprised that it produces results with low CV(RMSE)s, even if the individual parameter values themselves are not accurate. This lack of convergence on the true values shows a parameter identifiability problem between the gain and phase errors α and ϕ_c in (4.19). Another confounding factor is that the power factor ϕ is correlated with energy use as referred to earlier. This correlation, as well as the small range for ϕ, do not help identifiability.

Because the SIMEX method improves the parameter estimates independently of each other, it does so without considering their combined effect on the sum squared error of the fit. This results in more accurate estimates of the parameters, but slightly higher CV(RMSE) values when they are combined. It was therefore decided to refine SIMEX estimates using Bayesian regression. This changes the SIMEX estimates slightly to serve the double purpose of improving the goodness of fit metrics and providing probability distributions on the parameter estimates. These distributions can be used for risk

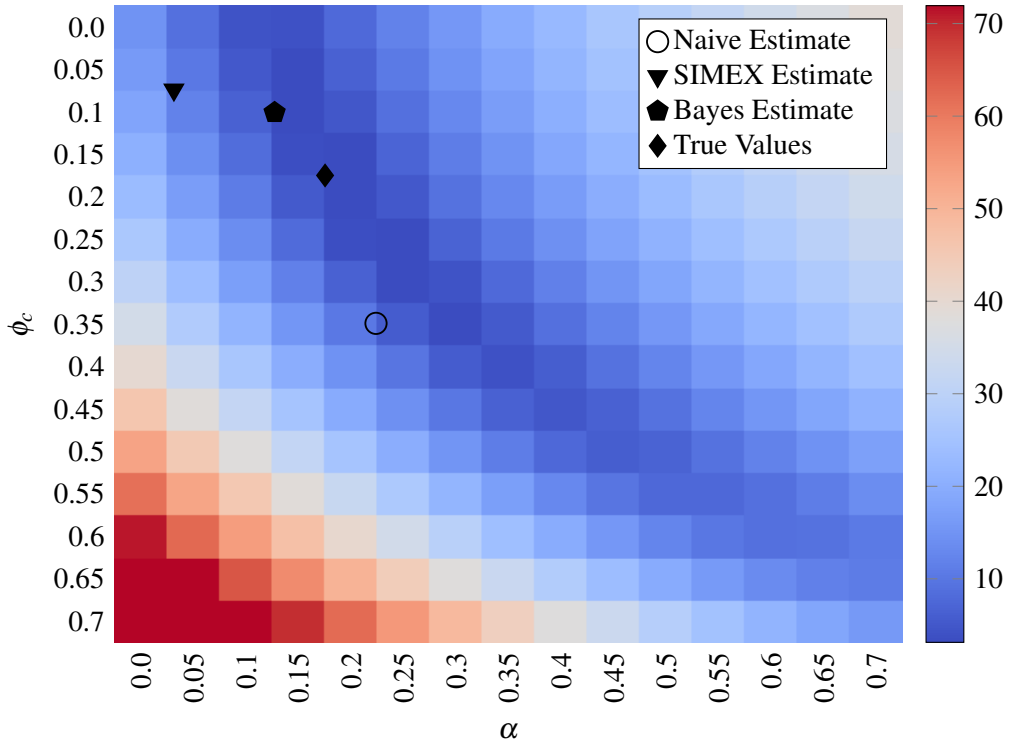


Figure 4.8. CV(RMSE) (indicated by colour) for different combinations of parameters α and ϕ_c. The parameter combinations plotted are for single instances of solutions. This plot assumes a bias error $\varepsilon = 5$. The positions of the SIMEX and Bayes estimates relative to the true values varies from realisation to realisation. A discussion of this figure can be found in Section 4.3.4.4.

and uncertainty quantification calculations, both on the parameter estimates and also on the predicted energy use, although a careful use of the data-dependent prior [258] will be necessary for uncertainty values to be valid. As shown in Figure 4.8, the Bayesian method does not interpolate linearly between the SIMEX estimates and true values. However, it does converge on parameter estimates in the SIMEX region while yielding improved CV(RMSE) and NMBE values. The method is explained more fully in Section 4.3.4.3. Using the Bayesian method on the naïve estimates, or using the naïve optimization algorithm with the SIMEX estimates as its starting position, did not improve on the original naïve estimates.

In conditional probability terms, we observe

$$\Pr(\mathbf{D}|\boldsymbol{\theta}) = \Pr(\mathbf{x}^*, \mathbf{y}^* \mid \alpha, \phi_c, \varepsilon, I) \tag{4.25}$$

where I is the prior information at our disposal through the SIMEX result, and α, ϕ_c, and ε are unknown. By Bayes' theorem in (3.5), through a numerical algorithm, this can be inverted so that the posterior

conditional probability estimates of the parameters

$$\Pr(\boldsymbol{\theta}|\mathbf{D}) = \Pr(\alpha, \phi_c, \varepsilon \mid \mathbf{x}^*, \mathbf{y}^*, I) \tag{4.26}$$

are found. The modes of the posterior distributions for α, ϕ_c, and ε will correspond to their maximum likelihood estimates given the data observed. To do this, the priors and likelihood function need to be defined, and the model solved. This is discussed below.

Prior selection

The empirical or falsificationist Bayesian approach is adopted here, so that $\hat{\boldsymbol{\theta}}_{SIMEX}$ obtained from the SIMEX algorithm can be used to constrain the MCMC. Further discussion of this topic, see Section 3.4.1 and Gelman and Hennig [249].

For this case, specifying vaguely informative priors is justified because the SIMEX parameter estimates do not arise naturally from the data itself. The priors are used to 'constrain' the algorithm to the solution space around the SIMEX solution. If overly vague priors are specified, the algorithm tends to converge on low CV(RMSE) solutions far away from the SIMEX estimates, and thus far away from the true values. The priors on the parameters are specified as follows:

$$\Pr(\alpha) \sim N[\hat{\alpha}_{SIMEX}, 5], \tag{4.27}$$

$$\Pr(\phi_c) \sim N[\hat{\phi}_{c,SIMEX}, 1], \tag{4.28}$$

$$\Pr(\varepsilon) \sim N[\hat{\varepsilon}_{SIMEX}, 5]. \tag{4.29}$$

A prior is also specified on $\mathbf{x}^*$. If the meter errors were Berkson, this prior would be perfectly representative. However, since the errors is located in the meter itself, they are classical. Therefore the prior below is not perfect but does allow for variation in $\mathbf{x}^*$ so that the model does not consider the observed values for $\mathbf{x}^*$ as fixed. The prior on $\mathbf{x}^*$ is specified as

$$\Pr(\mathbf{x}^*) \sim N[\mathbf{x}^*, \sigma_u]. \tag{4.30}$$

The likelihood function $\Pr(\mathbf{D}|\boldsymbol{\theta})$ is defined as a multivariate Student-T distribution. The heavier tails of this distribution allows for more robust inference, since outliers have a smaller effect on the posterior mean [272]. In this case, the data are the values observed from the reference and the UUT meters, and the priors are the SIMEX parameter estimates. Therefore:

$$\Pr(\mathbf{y}^*|\mathbf{x}) \sim StudentT[\mathbf{y}^*|\boldsymbol{\mu} = \boldsymbol{\mu}_p, \boldsymbol{\sigma} = \Pr(\sigma_p),\ \boldsymbol{\nu} = \Pr(\nu_p)] \tag{4.31}$$

where

$$\boldsymbol{\mu}_p = (1 + \Pr(\alpha))\Pr(\mathbf{x}^*)cos(\phi + \Pr(\phi_c)) + \Pr(\varepsilon), \tag{4.32}$$

as in (4.19) and the hyper-priors are defined as

$$\Pr(\nu_p) \sim Exponential[48^{-1}] \tag{4.33}$$

and

$$\Pr(\sigma_p) \sim HalfCauchy[1]. \tag{4.34}$$

Hyper priors add a second layer of variation by allowing uncertainty in the parameters of the uncertainty distributions used in the Bayesian models. The choice of '48' as the inverse scale parameter for the exponential distribution relates to the number of data points in the calibration period [19]. For the scale parameter σ, we follow Gelman's recommendation of a half-Cauchy distribution [288].

Solving the model

Although a full Bayes-MCMC is standard, Automatic Differentiation Variational Inference (ADVI) [222] is a new and much faster alternative to standard MCMC algorithms. It has comparable accuracy and is useful for batch runs where the different approaches are compared for different error realisations on the same data set. The model is solved using 50 000 runs of the ADVI algorithm. The starting points are specified as the SIMEX estimates. The analysis is performed in Python via the PyMC3 [267] library. Because only point estimates of the parameters are of interest for the current problem, the full Bayesian capability of eliciting full posterior probability distributions for each of the runs was not exploited.

4.3.4.4 Discussion

The resultant CV(RMSE) and NMBE for the naïve and SIMEX calibrated meters are shown in Table 4.5 and Figure 4.7. In these, it can be seen that the Bayesian refinement improves the CV(RMSE) SIMEX estimates substantially, from 8.87% to 2.96%. The average NMBE improves from -6.79% to -0.09%. A CV(RMSE) of 2.96% seems lower than the original 5.8% noise in the data. However, one should bear in mind that although CV(RMSE) is the appropriate metric to use, it cannot be compared to the way in which the noise is expressed originally. From Equation 4-4 of G14 2014 [17] for α,

$$\text{CV(RMSE)}_\alpha = \frac{\sqrt{\frac{\sum(\alpha_i - \hat{\alpha})^2}{n-par}}}{\bar{\alpha}} \tag{4.35}$$

where y_i is the true value, $\hat{y}_i$ is the model estimate, $\bar{y}$ is the mean, n is the number of data points, and par is the number of parameters. As the name suggests, it is, therefore, the mean of the sum squared

error, normalised with respect to the mean of the data. This is a different value to the relative precision of the meter.

Figure 4.7 shows that the Bayes-SIMEX procedure produces predictions with superior goodness of fit, both in terms of bias and in terms of CV(RMSE). Besides the violin plot, it is also graphically illustrated in Figure 4.8, where the SIMEX-Bayes coordinate approaches the true coordinate. The distributions are also tighter than for the other procedures, indicating improved consistency compared to SIMEX and naïve regression. Figure 4.6 indicates that Bayes-SIMEX does not do this at the cost of individual parameter estimates. On the contrary, superior and more consistent parameter estimates are also obtained.

To put these values in perspective, the G14 requires an NMBE below 5% for monthly data and 10% for hourly data [17]. CV(RMSE) requirements are 15% and 30% respectively. As this is half-hourly data, the requirements are in effect even more generous. However, it should be kept in mind that the G14 metrics do not refer to the calibration of measured energy data, but to building energy modelling requirements *relative to* measured energy data. The calibration figures in this paper are therefore baselines to which traditional M&V modelling uncertainty is added, before being compared to G14 requirements. Nevertheless, the calibration procedure is so effective, even with low accuracy meters and only 24 hours of calibration, that building models on energy use data obtained from this calibration method should still be acceptable. With longer calibration times or more accurate calibrators, these figures would also improve.

It is noted again that valid calibration requires more than simply having a reference instrument available. An adequate quality system needs to be followed to ensure that results are traceable and repeatable. However, we may conclude that from a technical point of view, the calibration itself does not require exceptionally accurate instruments for practical M&V purposes, and can reduce monitoring costs significantly through in-situ calibration.

4.4 CONCLUSION

Energy metering uncertainty makes a relatively small contribution to overall M&V reporting uncertainty when sampling is also done, and therefore presents an opportunity for M&V cost reduction, because

laboratory calibration of energy meters for monitoring projects can be expensive, and may not be cost-effective in terms of the gains in accuracy. A method is presented for disciplining or verifying a qualified, uncalibrated meter in-situ by using another calibrated commercial-grade metering system, in this case, a Class 3 meter and a Class 5 CT. By using the SIMEX MEM and refining parameter estimates using a Bayesian approach, the verified meter is shown to report energy use accurately and with low error variance compared to naïve OLS methods. For the data set under investigation, the CV(RMSE) was reduced from 8.87% to 2.96%, and the NMBE from -6.79% to -0.09%. To be conservative, the most inaccurate meter-CT combination for IEC-qualified instruments was selected and has been demonstrated to have acceptable accuracy. For any other combination of IEC-qualified meters and CTs, more accurate results should be obtained if calibration period data is representative. The general method proposed may also be applied to instruments other than energy meters.

CHAPTER 5 EFFICIENT METER SAMPLING

5.1 CHAPTER OVERVIEW

The previous chapters have dealt with the literature and theory of M&V and the Bayesian paradigm, as well as low-cost calibration of energy meters. The second half of the thesis will apply these principles to an M&V case where retrofitted lamps are monitored over a number of years using the retrofit isolation with all parameter measurement approach [1]. In this chapter, the metering aspect of such a longitudinal M&V study is considered. After an introduction to the problem and brief remarks about existing methods, a Dynamic Linear Model with Bayesian Forecasting is presented in Section 5.3. The method is demonstrated and verified using a minimal working example and is also compared to previous methods. A more realistic case study is then presented in Section 5.4. This involves both the design of an efficient sampling plan, as well as an evaluation of its execution and robustness. Finally, conclusions are drawn. [1]

5.2 INTRODUCTION

The rest of this thesis will consider the lamp retrofit longitudinal monitoring problem. In such an M&V case, the energy savings resulting from a lamp retrofit project is monitored over a number of years. There are two aspects to this monitoring problem: metering to determine the average annual energy use of a lamp in a given year, and population survival survey sampling, to determine how many of the original lamps are left in a given year. The metering aspect, which includes cross-sectional considerations (how many to install in a given year) and longitudinal ones (how many to install this year, given the results from the previous years), will be considered in this chapter. The longitudinal

[1] This chapter is based on a journal article written by the author as part of his PhD research, published in *Energy and Buildings* [29].

population survival survey sampling component, which considers persistence, will be addressed in Chapter 6.

Meters often need to be installed over a wide geographic area spanning many facilities or circuits, such as different parts of a factory or different homes. Since it is not practical to meter all facilities or circuits, only a sample is metered. The size of this sample is determined by the interaction between the meter's accuracy, the variance in the energy use between different lamp circuits, and the reporting requirements regarding statistical precision. These are the cross-sectional aspects of the meter sampling problem.

The longitudinal aspect of the meter sampling problem considers how previous sample sizes and sampling results influence the choice of the current and future sample sizes. A regression or time series model can be implemented to determine these effects. Since OLS regression is a special case of Bayesian regression, OLS or Bayesian regression may also be used in a leveraged sampling design [40,69]. However, to enhance the flexibility of the model, a Dynamic Linear Model (DLM) with Bayesian forecasting will be used. The Bayesian forecasting component allows for exact uncertainty quantification which may then be used for optimal or robust sampling design. Furthermore, the informative prior and updating step are useful for forecasting and sampling planning. This is because although past data can be incorporated into a regression model, future data need to be simulated. For small sample sizes, simulating draws from the distribution will not reflect the distribution from which they were drawn accurately, for most cases. It is therefore desirable to specify the distribution from which they were sampled, rather than a random draw of samples. However, this distribution will vary with the number of samples planned, making the model heteroscedastic and thus violating the assumption of OLS regression. It is allowed in the DLM, however, and the constant variance (V_t) can be scaled by a factor, in this case the sample size $n_{m,\,t}$, to obtain the standard error on the mean. This variance can be added to the prior variance to produce the posterior variance on the regression estimate, as a function of the sample sizes at various points in time. The V_t/n_t method can also be used for modelling past samples in a simulation such as this one, so that the same data effect (not being representative of the underlying distribution) is mitigated.

This chapter is only part of the greater M&V approach, but can be described in M&V terms as follows: **M&V measurement option:** Retrofit isolation with key parameter measurement (measuring energy consumption over time, but not measuring population survival over time).

Project boundary: The lighting circuit(s) under investigation.

Baseline and baseline adjustment approach: The baseline is assumed from the metered data, assuming a constant energy consumption difference between the retrofitted units and the original units.

Savings determination approach: Standard energy efficiency savings (as opposed to normalised savings) is assumed. This is done in Chapter 7.

5.3 EFFICIENT CROSS-SECTIONAL METERING SAMPLING

5.3.1 Modelling assumptions

It is assumed that meters are placed on circuits containing only one kind of luminaire, as per the retrofit isolation approach of the IPMVP [1]. The circuits may contain one or many fixtures and may contain switches with sub-circuits so that not all fixtures are on at the same time. The average annual energy use per lamp is modelled by dividing the annual energy use of a circuit by the number of lamps on the circuit. Seasonality can be built into the model to increase model granularity to monthly or hourly levels [230], but is not considered here.

The aggregated meter results are normally distributed. That is, if n meters are placed on different circuits, the distribution of the n average luminaires is approximately normal. This assumption seems reasonable by the Central Limit Theorem, but warrants further investigation in future research.

It is assumed that the average annual luminaire energy use varies linearly over time. A straight-line linear model is used, although other linear functions may also be specified.

It is assumed that samples are independent in time. This means that the same facilities cannot be sampled repeatedly in consecutive years, unless by chance. A new random selection of facilities needs to be made in each sampling year. Although this was not done in previous works on this problem [51, 52, 98, 99] it is necessary for the validity of the study design, and is used in other longitudinal energy use studies such as the US Commercial Buildings Energy Consumption Survey (CBECS) [289]. If the same meters are used in the same buildings, the independence assumption is violated, and normal distribution statistical and linear models will probably be invalid. It is often argued that this makes the proposed method very expensive and laborious in comparison to the approach of

metering a specific cohort for the duration of the longitudinal study. However, if such a cohort study is done rigorously, the costs should be similar. The reasons are the following:

For standard sampling statistics to be applicable, samples must be independent. Should one opt for the alternative, a longitudinal randomised control trial should be designed. For a longitudinal study such as the one under consideration, but where the same sample is monitored continuously for many years, certain practicalities also need to be considered:

1. **Subject dropout.** How large should one's sample be to ensure that enough study units are left at the end of ten years for the 90/10 accuracy to hold on the savings? How does one deal with censoring (individuals who terminated early or started late)?

2. **Uniformity.** Bearing in mind that a control group also needs to be monitored and that this control group must be similar to the treatment group in almost all respects (matched), but must not install energy efficient lights for the next ten years. The control group and treatment groups must also have sample sizes much greater than 68 (assuming CV=0.5, and 90/10) for their difference to be determined with 90/10 accuracy, to which the dropout factor should be added. Other longitudinal factors also need to be controlled. For example, persons in certain homes may have children in the next ten years, changing their occupancy and usage patterns. Older children may leave home, or the family may move. A circuit may be lost to follow-up due to renovations or accident. If the same individuals are monitored continuously, these changes will be imputed as trends of the general treatment population.

3. **Self-selection.** Are those who agree to have their lighting circuits monitored for the next ten years representative of the population as a whole, or are they self-selected? It would seem that such a radical position is not representative.

4. **Free ridership.** Would the treatment group not have made the changes independent of the project? In such a case the savings cannot be attributed to the project.

These are important questions in practical longitudinal M&V study design (addressed to some extent in the UMP Ch. 8 [273]), but they are difficult to quantify for the sake of comparing the current approach to a longitudinal RCT.

Then there is the question of autocorrelation. If the same meters in the same homes are used in consecutive years, the sample size will be affected by autocorrelation. Consecutive samples are deemed independent if the first sample contains no information about the second or third, and so on. Of course, this is very unrealistic when dealing with the annual energy consumption of a facility or lighting circuit. One could use this year's energy use as a very good approximation for next year's estimate: they are very highly correlated on an annual level. With enough data the extent of this serial correlation can be determined, but probably only halfway through the study. The implication of autocorrelation is that the individual samples do not count as much (in terms of information), as they would have had they been independent. Therefore *the sample size grows with the size of the serial correlation factor*, to compensate for the lack of information. For the annual energy use case, the required sample size would almost double for every year of the longitudinal study: If this year's figure for one building presents almost no new information beyond last year's figure, then one needs to sample roughly two buildings to obtain the same amount of information: one for information about this year, and one for information about last year. This can be modelled with an autocorrelation correction factor. In previous works, an exponential windowing function was used [53, 111], but a sample size adjustment factor as per G14 [17] is better. These are just adjustment factors, though, and may not be accurate enough for uncertainty quantification. Best practice dictates that if the meters monitor only a sample of the population and cannot be moved, an unbiased comparison group needs to be found and monitored as for a randomised control trial, which is an expensive and challenging task in itself. As Violette [67] has shown, the means of both groups then need to be determined with much higher accuracy than 90/10 for the savings estimate to achieve that level. Chapter 8 of the UMP discusses such designs as applied to M&V [273].

As smart meters and load disaggregation techniques become more common, it is unlikely that project-specific meters will have to be installed in every new facility in every year, especially for long time-horizons. It has been costed it in this way to be conservative, but realistically the amount of annual follow-up required in the alternative case would rival that of the assumed method.

Due to the considerations discussed above, a clear financial advantage in continuously sampling the same facilities in a longitudinal study, over the method assumed in the paper, is not apparent.

Even though it was shown in Section 4.2.3 that the standard sampling formula is underpowered, it will be used as a general benchmark in this study due to its popularity.

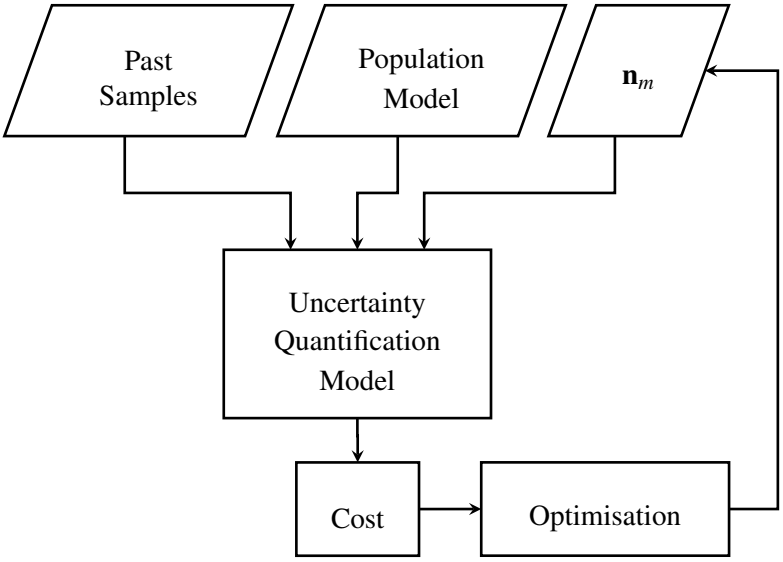


Figure 5.1. Flow diagram illustrating existing methods [51–53, 98], where $\mathbf{n}_m$ denotes the metering plan.

5.3.2 Dynamic Linear Model with Bayesian Forecasting

The proposed solution to the problem described above uses Dynamic Linear Models (DLMs). These can be thought of as adaptive models in which the new information that becomes available at each time step changes not only the estimates of the mean, but also the parameter estimates and variance matrix of the underlying model. For non-adaptive or static models, the model parameters would be fixed before calculation, and the process data would only update the state of the system. For example, in previous works, the average annual energy use measured by the meters was fixed at the beginning of the study [51–53, 98]. For models taking population decay into account, the population decay rates were fixed at study inception, and not updated as new information became available. These differences are illustrated in Figure 5.1 vs. Figure 5.6. In a dynamic modelling framework, new data alters both the parameters and the estimates of the system state in real-time.

The sequential updating and filtering aspects of Bayesian forecasting used with the DLM are the same as Kalman filtering [290, 291], applied to time-series analysis rather than control. However, according to West and Harrison [230]:

> To say that "Bayesian forecasting is Kalman filtering" is akin to saying that statistical inference is regression.

The function of Bayesian forecasting is, therefore, broader than only fitting models and making forecasts. Furthermore, where Kalman filters assume normality and use least squares and minimum variance methods, Linear Bayesian Estimation (LBE) is more general. Kalman filters are therefore a special case of general LBE where normality is not assumed. The disadvantage of LBE is that the solution is linearised (similar to extended Kalman filters), and that only the first two moments of the distribution are used. For normal distributions the first two moments define the distribution, but for other kinds it may not do so. A more complete explanation of LBE in the context of DLMs is given by West and Harrison [230].

For simple special cases, the DLM estimate at a given point in time would be equal to the weighted OLS regression estimate. For example, the DLM estimate (and forecast) given three data points would be the same as the OLS regression estimate and forecast, given that OLS regression assumptions hold. If a fourth point is added, redoing the OLS regression on all four data points (offline estimation) would yield the same value as the DLM updated "online" only for the fourth point. In such cases, the DLM would not yield a better 'Best Linear Unbiased Estimator' (BLUE). Both reduce modelling uncertainty to the best weighted OLS estimate for this case. However, DLMs with Bayesian forecasting have other desirable properties and capabilities that will be explored below.

The Bayesian forecasting component allows for exact uncertainty quantification, which is not always available for OLS regression. These uncertainty results may then be used for efficient or robust sampling design, without resorting to computationally expensive bootstrapping or cross-validation approaches [176, 292].

The informative prior and updating steps of the DLM are useful for forecasting and sampling planning. This is because although past data can be incorporated into a regression model, future data also needs to be simulated for sampling planning. Consider two scenarios. In the first case, a sample of 50 meters is planned. In the second case, a sample of 20 meters is planned. Only their means are used in the regression model. How should the model distinguish between these two plans? It is therefore desirable to specify the variance of the distribution from which they were sampled. However, the sample variance will vary with the number of samples planned or taken, making the model heteroscedastic and thus

violating a key OLS regression assumption. Unequal variances are allowed in the DLM, however. The constant variance (V) can be scaled by a factor, in this case the sample size n_t, to obtain the standard error on the sample mean. This variance can be added to the prior variance to produce the posterior variance on the regression estimate as a function of the sample sizes taken or planned for different points in time.

Turning to the method now, for the univariate case, the observation equation is

$$Y_t = \mathbf{F}'\boldsymbol{\theta}_t + \nu, \qquad \nu \sim [0, V] \tag{5.1}$$

where Y_t is the observed value at time t, $\mathbf{F}$ is called the regression vector, $\boldsymbol{\theta}$ the state vector at t, V is the population variance as defined before, and $'$ denotes the transponent. The state equation is

$$\boldsymbol{\theta}_t = \mathbf{G}\boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t, \qquad \boldsymbol{\omega}_t \sim [0, \mathbf{W_t}] \tag{5.2}$$

where $\mathbf{G}$ is the evolution matrix and $\mathbf{W}_t$ is the evolution variance. For the Time-Series Dynamic Linear Model (TSDLM) under investigation, $\mathbf{F}$ and $\mathbf{G}$ are constant in time, although for many other models (e.g. [28]) this may not be the case.

During M&V modelling and sampling planning, there are two cases that need to be considered. The first is step-ahead forecasting into the future given the current data, but no new data. The second is updating parameters to the current time-step, given new data at time t. For sampling planning in future years, these two steps happen simultaneously: a forecast to $t+k$ is made and using the forecast value and the planned sample size, the uncertainty in Y_{t+k} is determined.

5.3.2.1 Variable definitions

Since it is assumed that the annual average energy use after the retrofit, $E_{r,\,t}$ can vary linearly from one year to the next according to the gradient β_t, it can be described as

$$\hat{E}_{r,\,t} = \beta_t t + constant \tag{5.3}$$

The state vector for this system is then

$$\boldsymbol{\theta}'_t = (\hat{E}_{r,\,t},\ \beta_t), \tag{5.4}$$

where the regression vector is

$$\mathbf{F} = (1, 0), \tag{5.5}$$

so that (5.1) is satisfied by yielding $Y_t = \hat{E}_{r,\,t}$. The evolution matrix is defined as

$$\mathbf{G} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \tag{5.6}$$

so that (5.2) is satisfied by yielding $\boldsymbol{\theta}'_t = (E_{r,\,t-1} + \beta_{t-1},\ \beta_{t-1})$. In this way, the linear regression line is extended to time t through forecasting, given $\mathbf{D}_{t-1}$.

5.3.2.2 Forecasting

Forecasting is done when no data are available for that time step. The joint forecast distribution can be described as follows. Let f_t be the forecast mean, $\mathbf{a}_t$ the prior on $\boldsymbol{\theta}_t$, Q_t the variance on the mean in (5.10), $\mathbf{R}_t$ the prior variance in (5.11), and $\mathbf{A}_t$ the adaptive vector in (5.16) (not used explicitly in forecasting). Let the data up to the previous time step be $\mathbf{D}_{t-1}$. In the LBE scheme only the first and second moments are specified. The joint distribution on Y_t and $\boldsymbol{\theta}_t$ is then

$$\begin{pmatrix} Y_t \\ \boldsymbol{\theta}_t \end{pmatrix} \Bigg| \mathbf{D}_{t-1} \sim \left[\begin{pmatrix} f_t \\ \mathbf{a}_t \end{pmatrix}, \begin{pmatrix} Q_t & Q_t\mathbf{A}'_t \\ \mathbf{A}_t Q_t & \mathbf{R}_t \end{pmatrix} \right]. \tag{5.7}$$

In this study, the equation above describes a normal distribution, although other kinds can also be described this way. Again, West and Harrison [230] provide a full explanation of the DLM and distributions on all parameters. For this study and its application to M&V, the updating, forecasting, and filtering equations will be given in an applied format useful to M&V.

The step-ahead forecast mean f_{t+1}, which corresponds to the energy use $E_{r,\,t+1}$ is defined as

$$(\hat{E}_{r,\,t+1}|\mathbf{D}_t) = f_{t+1} = \mathbf{F}'\mathbf{a}_t. \tag{5.8}$$

Since there is no posterior in the forecast case, the prior for $\boldsymbol{\theta}$ is simply updated by evolving it according to

$$\mathbf{a}_t = \mathbf{G}\mathbf{a}_{t-1}. \tag{5.9}$$

Updating the variance is more involved. The variance on the mean, Q_{t+1}, is calculated as

$$Q_{t+1} = \mathbf{F}'\mathbf{R}_{t+1}\mathbf{F}. \tag{5.10}$$

The prior variance $\mathbf{R}$ is evolved according to

$$\mathbf{R}_{t+1} = \mathbf{G}\mathbf{R}_t\mathbf{G}' + \mathbf{W}_t. \tag{5.11}$$

The evolution variance $\mathbf{W}_t$ can be static, but as explained more fully in the next chapter, it can also be updated according to

$$\mathbf{W}_t = \mathbf{G}\mathbf{U}_t\mathbf{G}' \tag{5.12}$$

where, using a discount factor δ and covariance matrix $\mathbf{C}_t$,

$$\mathbf{U}_t = \delta\mathbf{C}_t. \tag{5.13}$$

$\mathbf{W}_t$ has a small effect on the uncertainty at times steps where data are available, but becomes prominent during forecasting periods. Since δ is subjective, it should be chosen carefully if it is non-zero.

5.3.2.3 Calculation

The equations below apply to the time steps in which data are available, so that $\mathbf{D}_t = \{Y_t, \mathbf{D}_{t-1}\}$. They combine calculations from the updating or filtering steps in the standard method. The values $f_t, \mathbf{a}_t, \mathbf{R}_t$, and $\mathbf{W}_t$ are updated according to (5.8), (5.9), (5.11), and (5.12) respectively.

In the calculation step, $\mathbf{a}_t$ and $\mathbf{R}_t$ in the forecasting calculation are replaced by $\mathbf{m}_t$ and $\mathbf{C}_t$ respectively, so that

$$(\boldsymbol{\theta}_t|\mathbf{D}_t) \sim StudentT[\mathbf{m}_t, \mathbf{C}_t]. \tag{5.14}$$

These are calculated as follows. Because data are available, rather than using (5.10), the variance on E_t is updated according to

$$Q_t = \mathbf{F}'\mathbf{R}_t\mathbf{F} + k_t V \tag{5.15}$$

where V is the observational variance and k_t is a weight, or variance divisor. If one assumes the variance to be constant throughout the process, it may result in a non-constant CV if the mean estimate $\bar{x}$ changes, since CV$= \sqrt{V}/\bar{x}$. It is therefore preferable to define $V = f_t$CV. Furthermore, the term added in (5.15) refers to the *observational* variance, and should therefore be scaled according to the sample size at t: $k_t = 1/n_t$.

The adaptive vector $\mathbf{A}_t$ translates the forecasting error from the previous step into an adjustment when new data become available. It is calculated as

$$\mathbf{A}_t = \mathbf{R}_t\mathbf{F}Q_t^{-1}. \tag{5.16}$$

The state is updated by

$$\mathbf{m}_t = \mathbf{a}_{t-1} + \mathbf{A}_t e_t \tag{5.17}$$

where

$$e_t = Y_t - f_t, \tag{5.18}$$

and

$$\mathbf{C}_t = \mathbf{R}_t - \mathbf{A}_t \mathbf{A}_t' Q_t. \tag{5.19}$$

5.3.3 DLM demonstration

To demonstrate how (and verify that) the DLM works, a hypothetical case is considered, and is illustrated in Figure 5.2. Sampling is done at $t = 0, 1, 2, 3, 4, 5$, and the mean of $E = 12.49$ kWh is set for every sampling result. According to standard theory for normal distributions, the sample size n is calculated from (4.10). Therefore, 68 samples are needed for a 90% confidence interval ($z = 1.645$) at 10% precision when CV = 0.5 [42]. The metering sample size at time t is denoted by $n_{m,\,t}$. The demonstration sampling plan (the vector containing the sample sizes for future years) $\mathbf{n}_m$ is

$$\mathbf{n}_m = [68, 68, 68, 68, 200, 68, 0, 0, 0, 68, 0]. \tag{5.20}$$

It is evident that the 90% confidence interval narrows as more information becomes available between $t = 0$ and $t = 2$. When a large sample of $n_{m,\,4} = 200$ is taken, there is a more dramatic change in the interval, but it widens again, when a smaller sample of $n_{m,\,5} = 68$ is taken. This widening occurs because of the inherent process variation specified through the CV. For other CV-to-sample size ratios, no widening may take place. The narrowing of the confidence intervals over the first three years ($t = 0$ to $t = 2$) is also considerably more dramatic for smaller CVs. After $t = 5$, no samples are taken for three years, and the confidence interval on the forecast widens, but is reduced again at $t = 9$ when a sample is planned. Another realisation is shown in Figure 5.3. In this case random sampling results were drawn from the sampling distributions defined by the sample sizes and process variances. Multiple results are overlaid to demonstrate the randomness inherent in each individual sampling realisation. It can be seen that DLM estimates also follow an approximately normal distribution, with a greater density of predictions close to the mean. A large sample is planned for $t = 9$ rather than $t = 4$ as in the previous example. Such a sample "filters" the estimate, forcing subsequent estimates to be much closer to the true mean, and forecasting an approximately constant energy use, which is accurate.

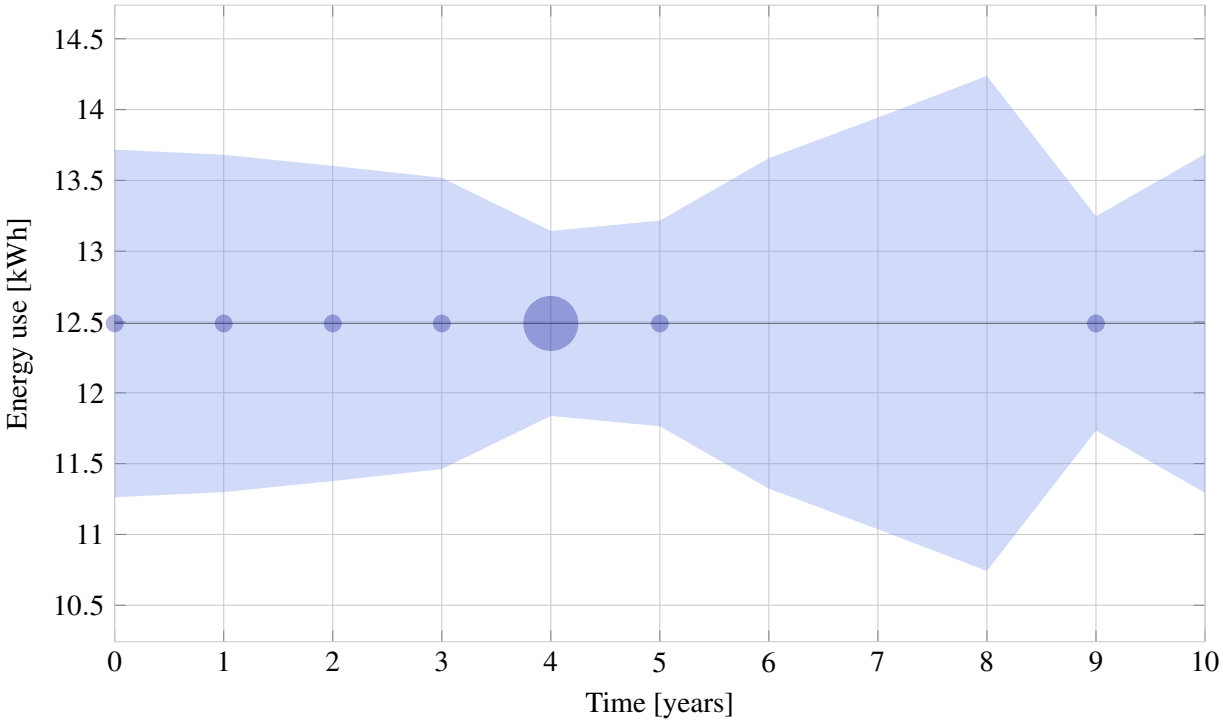


Figure 5.2. DLM demonstration. Hypothetical case where all sampling results fall on the mean. Sample sizes are $\mathbf{n}_{m,\,0-3,\,9} = 68$, and $n_{m,\,4} = 200$.

5.3.4 Comparison to previous method

In this section, the DLM will be compared against the earlier method [51–53, 98, 99], using the case study from [53]. However, a direct comparison can be misleading because of the differences between the two approaches. Some of these differences can be addressed by restricting the capability of the current model. For example,

- The old method assumes a stationary mean. A comparison can therefore only be made if the DLM is restricted to a horizontal line, no matter the trend in the data. To do this, prior on the slope is set to zero.

- The old method uses FPC to compensate for population decay. FPC cannot be included in the DLM without significant changes. However, for models such as those under investigation, FPC is only applicable to populations smaller than about 1 000, or 0.16% of the installed population in the benchmark study [53]. Therefore it does not affect the calculation and may be neglected

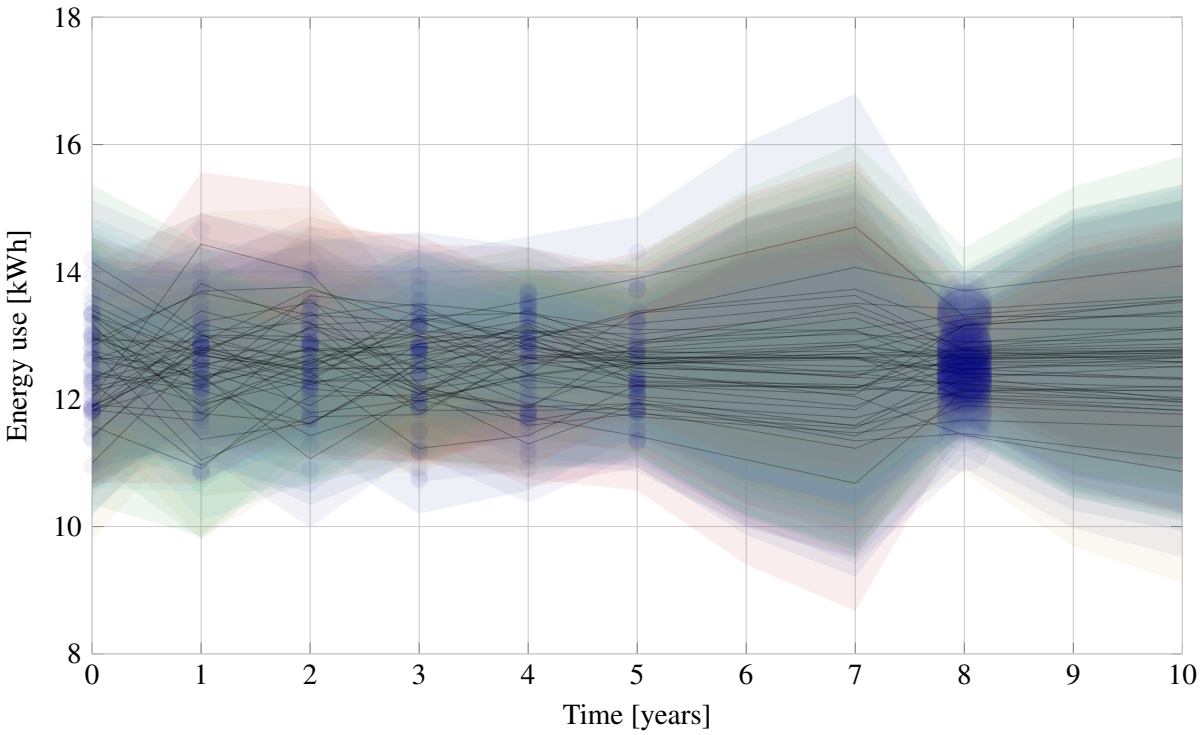


Figure 5.3. DLM demonstration reflecting true sampling results. Multiple realisations shown.

in the DLM.

Other differences are not as easy to address, and indicate fundamental differences of approach:

- The previous approach uses frequentist confidence intervals. The difference between these and Bayesian intervals is discussed in Section3.3.

- The improvements to the old model [53] include an exponential windowing function. This decreases the influence of prior data points exponentially, to compensate for the autocorrelation present in taking repeated measurements from the same study units. It transforms the method into a moving average function. Exponential windowing is mathematically convenient for the way the model was set up and is better than nothing. However, it does not address autocorrelation satisfactorily because such correlation is the strongest between consecutive measurements, while the windowing function reduces the influence of less recent samples. The discount factor in (5.13) is a similar mechanism in the DLM but increases the estimated variance. The problem

with choosing a discount or windowing factor is that the figure is arbitrary. When this is done, uncertainty quantification is no longer objective.

- In the old model, confidence and precision levels are undefined for years in which no sample is taken. The result is that the precision of the model is unaffected by non-sampling. If sampling is done at $t = 1$ and then again at $t = 4$, the increase in uncertainty is equivalent to sampling at $t = 1$ and $t = 2$. It is not possible to adapt the DLM accordingly since forecasting increases uncertainty.

With these caveats in mind, a case study for the old method [53] is analysed by the DLM, using the optimal sample sizes determined using the previous method. In this case study, 607 559 CFLs rated at 20W were distributed to households in the Northern Cape, Free State, Gauteng, Limpopo, and Mpumalanga, to replace 100W ICLs, as part of a CDM project. It was assumed that they burn for an average of 4.5 hours per day, but no uncertainty on this value was specified.

Exponential windowing (for the earlier method) is neglected, as is the discount factor for the DLM, to avoid confusion about their functions. The earlier method disregards autocorrelation from consecutive measurements of the same facility, while the DLM assumes random sampling. This will narrow the apparent uncertainty bounds resulting from the DLM calculation using those results, but is left as-is. Other changes in the bulleted points above also apply. The average annual energy saving for that study was 131.4 kWh. The sampling plan $\mathbf{n}_m$ was

$$\mathbf{n}_m = [68, 68, 28, 16, 8, 8, 6, 6, 4, 4, 2]. \tag{5.21}$$

The result is shown in Figure 5.4. The red error bars represent the 10% precision limits. The figure indicates that (had the samples been independent), there is slight oversampling in years two, four and six, and under-sampling in years eight and ten. However, since the model is simplified to a case where there is zero inter-sample variance, it becomes sensitive to the DLM priors on the mean energy use and slope. Increasing the prior on the slope of the regression line to a number above zero results in under-sampling for all years, for example. Such changes do not affect models with inter-sample variance as strongly.

When decreasing the effective sample size by using an autocorrelation factor of 0.25 [17], it is found that year six is also undersampled. However, when the exponentially windowed sampling plan is used,

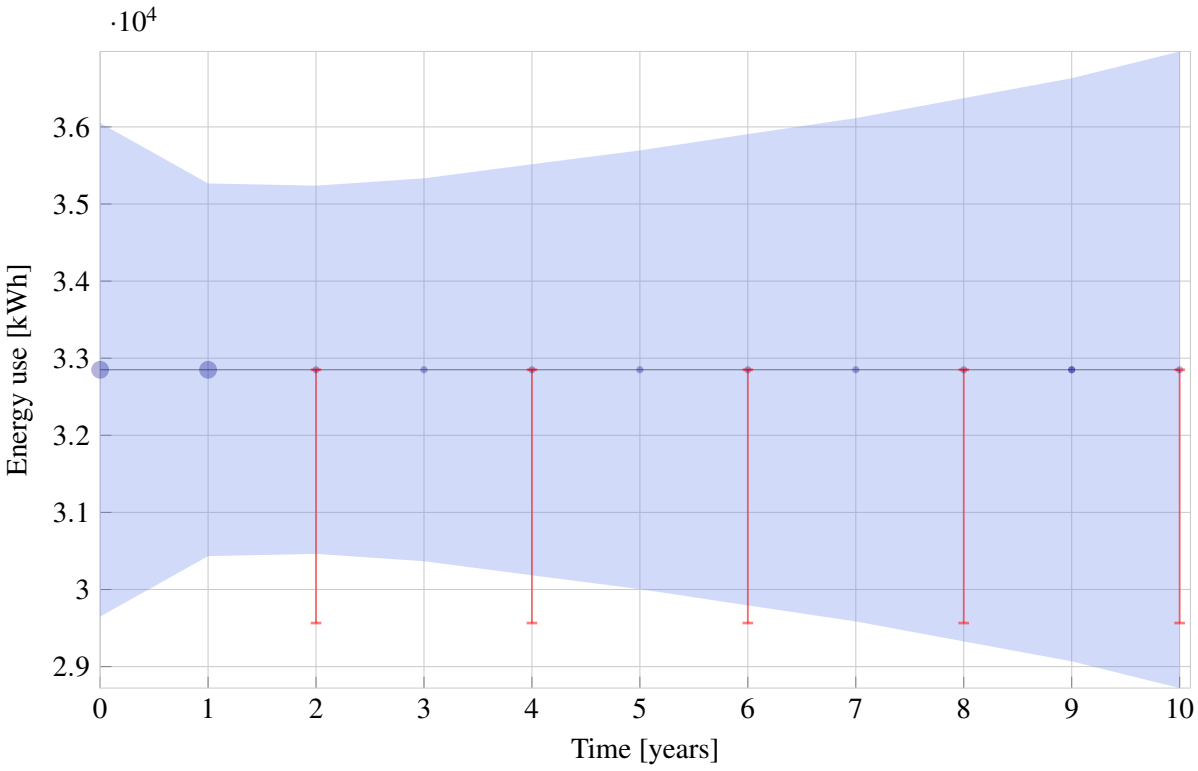


Figure 5.4. Confidence bounds (shaded) compared to precision limits (red error bars) for previous sampling plan using earlier method [53, 98], analysed by DLM.

the confidence bounds are much closer to the precision limits for all years.

When this is optimised with the DLM as described in the next section, an efficient sampling plan is found to be

$$\mathbf{n}_m = [68, 68, 65, 0, 61, 0, 55, 0, 27, 0, 31], \tag{5.22}$$

plotted in Figure 5.5. It is difficult to compare the two because of the assumptions as mentioned above. However, it can be seen that the DLM satisfies the reporting precision constraints and that the number of meters required also decreases over time.

5.4 CASE STUDY: EFFICIENT CROSS-SECTIONAL METERING DESIGN

In this case study, the DLM is used in an optimization routine to design an efficient sampling plan, given past data. It is assumed that the luminaires are 11W CFLs that operate for an average of

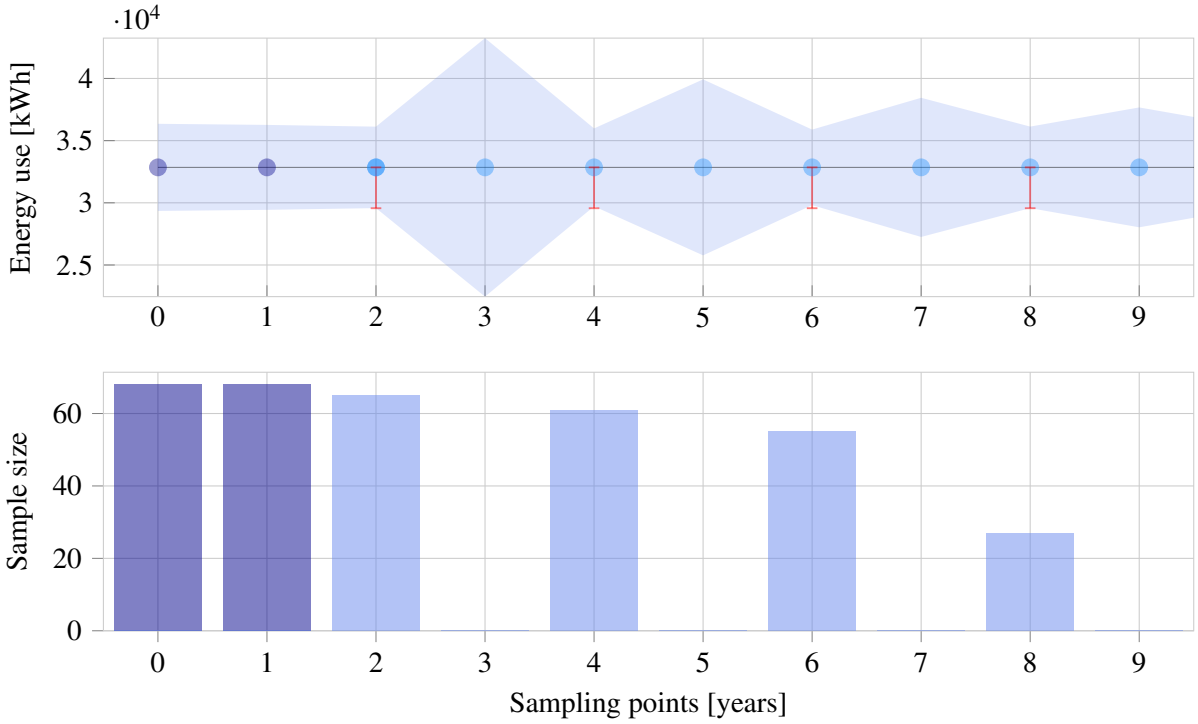


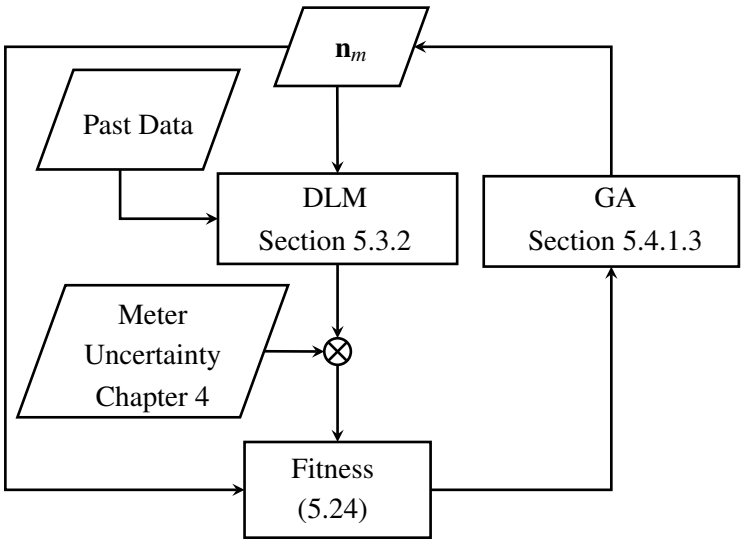
Figure 5.5. Efficient sampling plan using the DLM for the case study in Section 5.3.4.

3.11 hours per day [293], or $E = 12.49$ kWh per year. The CV in the sample is set to 0.5; a standard M&V assumption [15]. This implies that the distribution on the estimate of the annual energy use per luminaire is $\hat{E} \sim N[12.49, 6.24]$ kWh. Assuming CV $= 0.5$ is reasonably conservative and dominates the priors. At lower CV values, the information contained in the prior becomes dramatically more significant. For this case study, it was assumed that CV is constant. However, if sampling results from the first few years justify it, the CV value may be decreased. The Bayesian model can easily be updated in any year to adjust the CV values – another useful feature of the DLM.

Note that the study commences at $t = 0$.

The true energy use is modelled as being constant in time (thus a straight line with zero gradient). However, the estimate for a specific year will fall in the probability distribution described above. It may therefore seem as if there are short-term trends, depending on the realisations of the data from the underlying distributions. It is assumed that three years' data are available and that the remainder of the 11-year study is to be planned. Let the vector defining the reporting points be $\mathbf{M}$. For this study, $\mathbf{M} = \{3, 5, 7, 9\}$.

Figure 5.6. Flow diagram of cross sectional metering sampling designs as in Section 5.4.1.



The priors are defined as follows. It is assumed that the average annual energy use can be approximated reasonably well from previous case studies. It is assumed that there is a 99% chance that the energy use is within 25% of the prior. The same numbers hold for the expected change in energy use: not more than 25% per year, at 99% confidence. Therefore $3\sigma = 12.49/4$, with the prior variance specified as σ^2.

5.4.1 Cross-sectional metering sampling designs

In this section the case of cost-efficient sampling design using the DLM is considered. The basic flow is illustrated in Figure 5.6.

5.4.1.1 Robust and efficient designs

The design with the smallest sample size that still adheres to the reporting precision requirement is not necessarily the most cost-efficient design when uncertainty is present. It is only optimal in the best-case scenario, where the forecast is perfectly accurate. This is because installing just enough meters in future years, based on a forecast, runs the risk of not controlling variance adequately, since the forecast may be inaccurate. A meter may malfunction, or the sampled result may differ from the forecast to increase the variance in the estimate enough to violate the reporting precision constraint.

By the end of the measurement period, it is too late to install more meters for measuring the energy use of that period. Insufficient reporting precision would render the project ineligible, or incur a penalty from the regulator. Therefore these are referred to as *naïve* efficient designs. A robust design with more meters, on the other hand, will prove to be more cost-efficient over the whole range of possible scenarios (thus lowest expectation cost), even though the metering cost may be higher than the most efficient design for the most likely scenario would be. However, determining such a robust cost-efficient sampling design will depend on assumptions made about the penalty incurred for not complying to the reporting precision constraint, which may vary significantly between programmes. In the more common case where projects are rendered ineligible, the cost of non-compliance may be very high. For these reasons, as well as for brevity, the current investigation is limited to the narrow sense of the meaning of efficiency (except for Section 5.4.1.6) and robust efficiency is recommended for future research.

5.4.1.2 Adding energy metering uncertainty

Modelling and sampling uncertainty are combined automatically in the Bayesian framework described above. However, meters also have inherent uncertainty. It has been shown [72] that energy metering uncertainty makes a small contribution to overall uncertainty for sampling designs with standard variance assumptions. We assume Class 1 meters [22] are used with Class 1 Current Transformers (CTs) [25], as these are common for revenue metering. Since no load profiles are assumed for the study, a flat error rate of 3% is assumed. (For plots showing the change in error rate as a function of the rated current of the instruments, see [26]). The 3% figure allows for the combined meter-CT accuracy, as well as for low-cost calibration [27]. However, at this level, it can be shown [72] that the difference made by metering error is so small that the required sample sizes do not change due to the additional uncertainty.

5.4.1.3 Optimization

Thus far a model has been created that determines the overall uncertainty at a specific point in time, given the sampling regime and certain modelling assumptions. Such a model can be used to determine an efficient sampling regime, given past sample times, sizes, and results. These are combined with a forecast of future energy use and associated uncertainties. Planned (future) sample sizes can then

be used to control the reporting precision at future reporting points. Sampling is not constrained to reporting years only, however. If it is advantageous for the algorithm to sample in a non-reporting year, it may do so.

Optimization can be done in one of two ways. If the present time is τ, the first is to forecast one step ahead to $\tau+1$, and then determine an efficient sample size. This can be repeated for all time steps. The other option is to consider all future sample sizes simultaneously, given the forecast from the present time. This will produce a multi-year sampling plan in which earlier future samples may be traded off against later future samples. The latter approach is adopted.

Since only a discrete number of meters can be installed, an integer program is needed. Although the DLM is linear, the behaviour of the uncertainty bounds is not linear. The optimization algorithm will, therefore, need to be able to solve an integer non-linear program (INLP). Gradient search methods are therefore not appropriate choices for optimization, and a Genetic Algorithm (GA) was selected. The constraints are discontinuous [53], and will, in this case, be represented by very large stepwise changes rather than invalid regions, as this is more efficient for the GA. Similar optimization programs have been described in previous works [28, 53]. The GA was implemented via the `DEAP` (Distributed Evolutionary Algorithms in Python) library [294].

The details of the GA can be summarised as follows. Each solution $\mathbf{n}_s$ is a vector of numbers representing the sample sizes in the different years. These are called individuals, and the sample size for a specific year is called a gene. The algorithm starts with a population of different individuals, which are different sampling plans. A portion of the population is mated or hybridised to produce offspring by selecting genes from the parents according to certain rules. In this case, the “uniform crossover" rule was used to determine how offspring inherit traits from the parents. To ensure genetic diversity, a given proportion of the population is mutated by altering random genes. The optimality or fitness of the individuals in the population is evaluated according to the fitness function (known as the objective function in standard optimization). This evaluation takes place in the form of a tournament. The fittest individuals are kept for the next generation. This process repeats for a predefined number of generations and rapidly converges on excellent solutions. Because of the random nature of the solution generation process, it is not feasible to constrain a GA in the conventional manner. Rather, the fitness function is programmed to penalise infeasible solutions to such an extent that they are too costly to propagate to future generations, as shown in (6.44).

The parameters used to tune the GA for this case will need to be the same as for the optimization in Section 6.3.3, since these are combined in Chapter 7, and the survey sampling model of Chapter 6 is more computationally expensive and therefore dictates how large the GA population and other parameters may be. The parameters are described in Table 5.1. The mutation function was set so that the genes that are selected for mutation are altered by adding a number from the distribution

$$\text{Mutation factor} \sim N[-10,\ 500]. \tag{5.23}$$

Previous cross-sectional efficient metering studies have considered installation, maintenance, and meter removal costs separately for each meter [51–53, 98]. This cost structure is based on the assumption that the same facilities are monitored throughout the study, and that these individuals are representative of the whole population. However, as discussed in Section 5.3.1, the least problematic and most consistent solution would be to draw random individuals from the population at each sampling point, as is assumed in this study. The costing structure for such a sampling plan would be a simple fixed rate per meter per sampling point. This fixed rate would possibly include purchasing costs, subscription to an Advanced Meter Reading (AMR) telemetry service for accessing the data online, as well as installation and removal costs. Since the rate is fixed, the optimization function will simply reduce the total number of meters installed over the duration of the study. The price is therefore irrelevant. It does become a factor when metering is traded off against surveying as in Chapter 7, however. From industry experience, this is set at R3 000 (South African Rand) per meter per sampling point, although it may vary significantly by contract and supplier.

Mathematical formulation

From the notation on the following page, the fitness function can be defined as

$$\min \sum_{t=\tau}^{N} n_{m,\,t} w_m + r(\mathbf{n}_m), \tag{5.24}$$

where

$$r(\mathbf{n}_m) = \sum_{t \in \mathbf{M}} \left(10^5 w_m (e_t - \varepsilon) + 10^7 + 5 w_m n_{m,\,benchmark,\,t}\right) \forall t \in \chi \tag{5.25}$$

and

$$e_t = \frac{\hat{E}_{r,\,t} - LCL_{m,\,t}}{\hat{E}_{r,\,t}}. \tag{5.26}$$

Notation

Let:

χ	Number of sampling points where $e_t > \varepsilon$
$n_{m,\ benchmark}$	Non-DLM solution at time t
$n_{m,\ t}$	Decision variable. Sample size at time t.
	$n = \{\tau, \tau+1, ..., N\}$
w_m	Cost per meter in Rand/sample
τ	Present time, where $\tau \in \{1, 2, ..., N\}$
N	Last year of study
e_t	Precision of reported average annual energy use
	at time t, where $e_t \in [0, 1]$
ε	Given precision limit, where $\varepsilon \in [0, 1]$
$\mathbf{M}$	Required reporting points (years), where
	$\mathbf{M} \subset \{\tau+1, \tau+2, ..., N-1\}$
$\hat{E}_{r,\ t}$	Estimate of average annual energy use at t
$LCL_{m,\ t}$	Lower Confidence Limit at t

Description

The decision variable is the metering sampling plan $\mathbf{n}_m$, the individual elements of which are written as $n_{m,\ t}$ in (5.24).

The fitness function (objective function) for the model is reasonably simple. There is a cost to metering and a cost to violating the reporting precision requirement. The first term in (5.24) describes the metering cost, and the second term describes a penalty function for violating the precision constraint. Setting a hard constraint for a GA is not efficient due to the randomness inherent in the optimization process [28]. The penalty $r(\mathbf{n}_m)$ is therefore invoked only for sampling plans which violate the precision constraint. The shape of this penalty function is designed so that solutions that do incur a penalty are directed into the feasible region, rather than away from it [28]. It reduces to zero when the precision requirement is satisfied. Consider Figure 5.7. If there were no constraint, the cost would increase with $n_{m,\ t} w_m$ along line *ab*, and the GA would optimize to zero, violating the actual constraint. A penalty function could be specified simply as a constant added to the cost function if the confidence/precision bounds are violated: line *dcb*. However, this is not efficient. If a solution (or

Table 5.1. GA parameter values. These values have been used in all case studies.

Parameter	Value
GA Algorithm	`MuPlusLambda`
Crossover Rule	Uniform Crossover
Crossover proportion	45%
Crossover exchange probability	75%
Mutation Proportion	40%
Individual gene mutation probability	30%
Number of Generations	35
Population Size	100

population of solutions) violate the constraint (placing it on d), the algorithm would tend to optimize *away* from the constraint boundary in the wrong direction towards the local minimum at the y-intercept of d. Mutation could transport an individual to b, but it is inefficient to rely solely on this mechanism. Therefore line ef is needed to direct the algorithm *towards* the constraint rather than away from it. This is what the $10^5 w_m(e_t - \varepsilon)$ term does. The 10^5 term increases the gradient of the line (or 'gain' of the error size), and therefore encourages the algorithm to optimize downwards. The threshold value n_ε at which the penalty occurs is unknown — that is why the GA heuristic is needed. A step is built into the model to ensure that adhering to the constraint is always preferred over violating the constraint. However, since the exact number of samples at which this occurs is unknown, and a larger required sample size would also increase the constraint violation cost, a step of $10^7 + 5w_m n_{m,\ benchmark,\ i}$ is built in to ensure that constraint violation is always costly, where $n_{m,\ benchmark,\ i}$ is defined by (4.10). This step is represented by line ce.

Regarding (5.26), only the lower bounds are considered when calculating precision. For a normal distribution where these bounds are symmetric about the mean, this makes no difference. However, for asymmetric distributions as will be encountered later, there may be a difference. The reason the lower bounds are considered rather than the upper bounds is that reported savings should always be conservative in M&V [1]. This means that although the post-retrofit savings value may be higher than the reported value, it should not be lower.

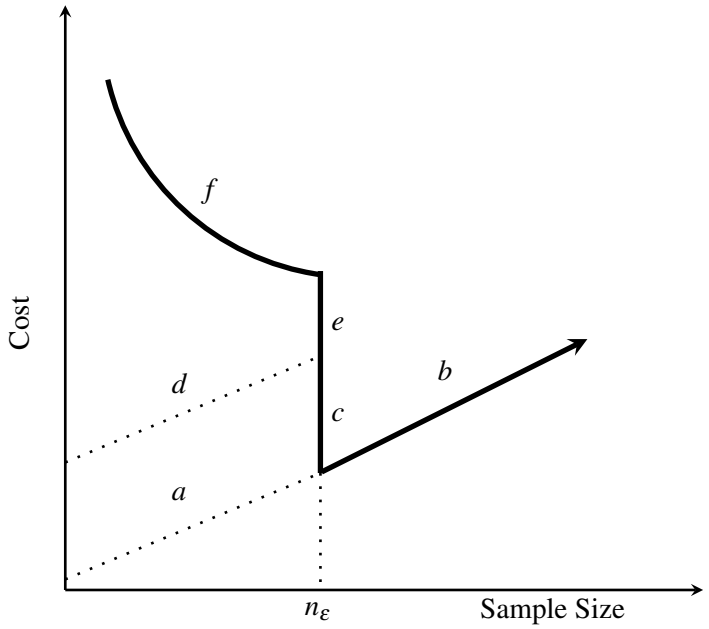


Figure 5.7. Genetic Algorithm constraint function $r(\mathbf{n})$ in (6.45), where n_ε represents the threshold sample size.

5.4.1.4 Benchmark

The DLM model with Bayesian forecasting should be benchmarked against current best-practice efficient sampling designs. It has been suggested that for cases involving weighted or normal regression, the sample size may be reduced by a factor of $(1-R^2)$ [69]. R^2 is the coefficient of determination, which is the square of the Pearson moment correlation coefficient. This is similar to 'ratio-estimation' [295] , where the additional information contained in the known ratio or regression line can be used to reduce the sample size [296]. However, for cases where the process is supposed to be stationary, the regression line will have a slope coefficient equal to zero. It should, therefore, be "uncorrelated" even if the regression line exhibits high goodness of fit. This means that the correlation coefficient and thus R^2 will be zero, even if all the sampled points fall exactly on the straight (horizontal) line. In fact, for a stationary process, any other (erroneous) slope estimate would increase the R^2 value spuriously and thus decrease sample size.

A more reliable and popular measure of goodness of fit in M&V is the CV(STD) or CV(RMSE) [17,35], which does not reduce to zero for stationary processes. However, these are not ratios bounded by zero and one like R^2. How they relate to a sample size reduction factor can be the topic of future research

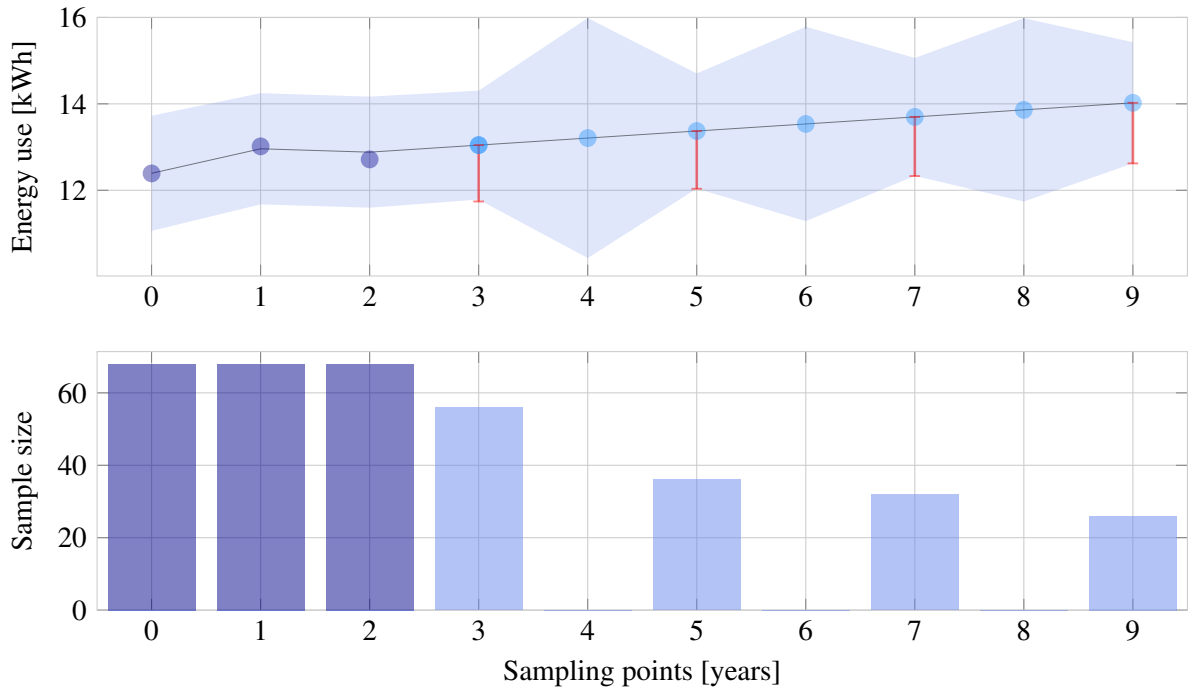


Figure 5.8. Efficient sampling plan using the DLM for one random model realisation

as an extension of G14 [17] and Reddy and Claridge's work [33].

Therefore the method is benchmarked against the standard M&V approach of (4.10). Since metering error has been determined to not affect sample size, it may be neglected.

5.4.1.5 Results and discussion

The values generated for the first three points are $\mathbf{D}_{0-2} = [12.39, 13.02, 12.71]$. One efficient sampling plan for one realisation of results is shown in Figure 5.8. The planned sample sizes $\mathbf{n}_m$ are

$$\mathbf{n}_{m,\ DLM} = [56, 0, 36, 0, 32, 0, 26], \tag{5.27}$$

while standard sampling theory yields

$$\mathbf{n}_{m,\ Benchmark} = [68, 0, 68, 0, 68, 0, 68]. \tag{5.28}$$

The total number of meters under the DLM plan is 147 at a cost of R450 000, while under the standard plan 272 meters are installed at a cost of R816 000. A saving of 66% is achieved.

The red error bars in Figure 5.8 indicate the reporting precision limits. Should the uncertainty bounds (light blue area) fall outside these limits, the reporting precision requirement will have been violated, and $r(\mathbf{n}_m)$ in (6.45) invoked. Efficient sampling plan precisions tend to be in the range 0.97 to 0.99. If a certain year has a precision of 0.97, sample sizes can be reduced to so that the precision is closer to 0.1 (being more efficient), but doing so usually results in precisions in later years violating their constraints, requiring more samples in those years.

Since the full solution space is not known and convergence is not guaranteed mathematically, the solution cannot claim to be 'optimal'. It may be the case that the solution is only a local minimum, or that one or two samples may still be removed from the solution, resulting in an even more efficient sampling plan. That is why the solution is presented as 'an efficient solution' rather than 'the optimal solution', although the GA does converge reliably to very efficient solutions. This consideration has been noted before [53, 111], but has not always been adopted [51, 52, 98, 99].

This model illustrates certain crucial characteristics that M&V study designers should take into account. The first is that although this is a stationary process, random realisations from the distribution could indicate a trend. In this case, it appears as though energy use is increasing, although it is not the case. Another realisation may show the opposite with equal probability. The larger the sample size, the less pronounced this trend should be, but the sampling error effect will not be mitigated completely.

As in Figure 5.2, the uncertainty decreases over time as more samples are taken and the prior information of the Bayesian method becomes more prominent. This results in smaller sample sizes in later years. The CV of the process plays a significant role in this narrowing effect.

An interesting relationship emerges when solving the optimization model for different energy use realisations in years zero to three (sampling results drawn from the relevant distributions). It is plotted in Figure 5.9. The sum of all future (efficient) sample sizes are related to the gradient of the energy use line (least-squares regression line) plotted through these three data points. From this relationship, an estimate of future sampling costs may be obtained, even before a GA is used to determine exactly how these samples should be spread over the remaining years. This can be done by simply calculating the gradient of the weighted regression line drawn through the past sampling points. The caveats for using the graph are that it is specific to the parameters used for this model, since many variables may affect this relationship. These include past sampling points and sample sizes, CV, future reporting

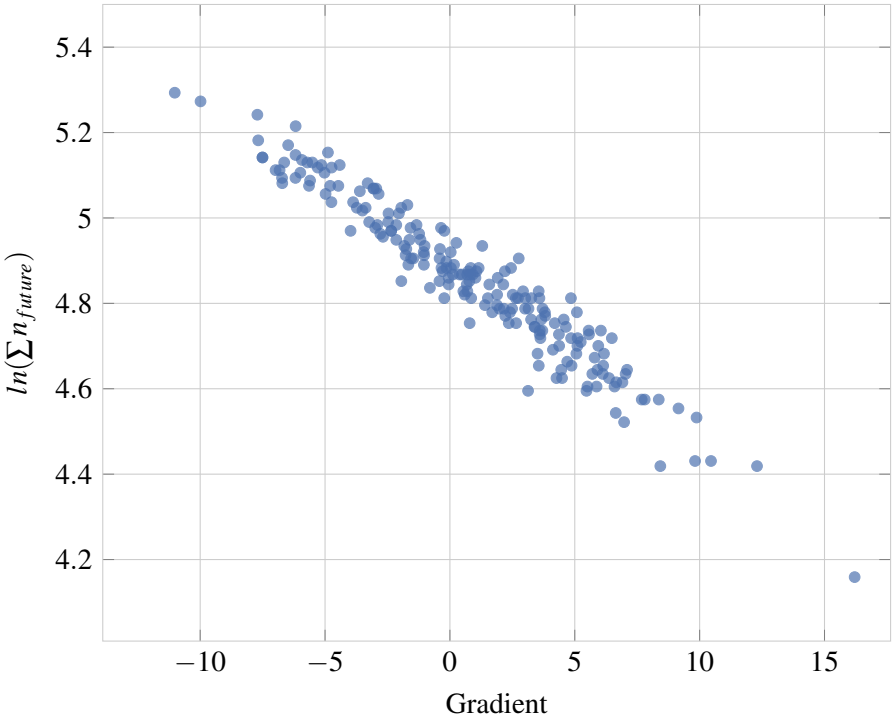


Figure 5.9. Natural logarithm of the total number of future samples under efficient sampling plans, as a function of the gradients of the regression lines on past samples, e.g. in Figure 5.8.

points, reporting precision, and others. The model also assumes that such an increasing or decreasing relationship apparent in the past sampling results, does exist. However, all the points on the graph were generated from realisations of what is in fact a stationary process (gradient = 0). One should therefore be very careful about interpreting low future sample sizes from a positive gradient-model, especially with few past sampling points. The algorithm may recommend small future sample sizes when such sample sizes will yield inadequate precision. The forecasting uncertainty bounds should certainly be considered.[2] Nonetheless, the relationship shown in Figure 5.9 is true in the sense that if that relationship is correct, the required future sample sizes do follow that curve.

5.4.1.6 Sampling plan execution

After an efficient sampling plan has been designed, it should be executed. In this section, the reliability of efficient sampling plans is investigated, in terms of compliance to reporting precision requirements.

[2]Note that the forecasting uncertainty bounds in Figure 5.8 are instantaneous future sample sizes which include results from planned future samples.

Since the sampling plan needs to be updated every time new data becomes available, only the next time step beyond the sampling plan already devised is investigated. It is supposed that three years' data are available ($\mathbf{D}_{0-2}$), and that the fourth year is forecast, planned, and executed. Such scenarios are simulated and the result analysed. The investigation proceeds as follows:

1. Generate $\mathbf{D}_{0-2}$ from the distribution $\sim N\left[12.49, \frac{12.49CV}{\sqrt{\mathbf{n}_{m,\ 0-2}}}\right]$.
2. Fit DLM to data points, forecasting $t = 3$.
3. Find minimum sample size $n_{m,\ 3}$ that adheres to the reporting uncertainty limit.
4. Instead of assuming that D_3 will correspond exactly to the most likely forecast value, generate a random realisation of D_3, given the planned sample size $n_{m,\ 3}$: $D_3 \sim N\left[12.49, \frac{12.49CV}{\sqrt{n_{m,\ 3}}}\right]$.
5. Update the DLM to include $D_3|n_{m,\ 3}$
6. Calculate reporting precision at $t = 3$.
7. Repeat steps 1-6 10 000 times to examine the adequacy of the sample size for different random realisations of the sampling distribution.

As discussed in Section 5.4.1.1, a naïve efficient design is not necessarily efficient when all possible scenarios are considered. For this case study, if only the best-case scenario is considered and sampling is planned accordingly, the reporting precision requirement will be met in only 48% of cases, as shown in Figure 5.10. Meeting the reporting precision requirement means that the lower confidence limit (LCL) at time $t = 3$, given the data at time $t = 3$, $(LCL_{m,\ 3,\ 90\%}|D_3)$ is within 10% of the most likely value. Taking only a naïvely efficient (or 'optimal') number of samples has a 50/50 chance of being inadequate, according to the simulation described above. Note that this lack of power is not due to the DLM or regression generally, but due to the sample size produced by the standard M&V sampling formula (4.10) recommended by the leading guidelines [1, 17, 40], on which see Button *et al.* [245] and Senn [248]. As shown in Section 4.2.3, the interval produced includes the true value and satisfies the 90/10 criterion in only 50% of cases.

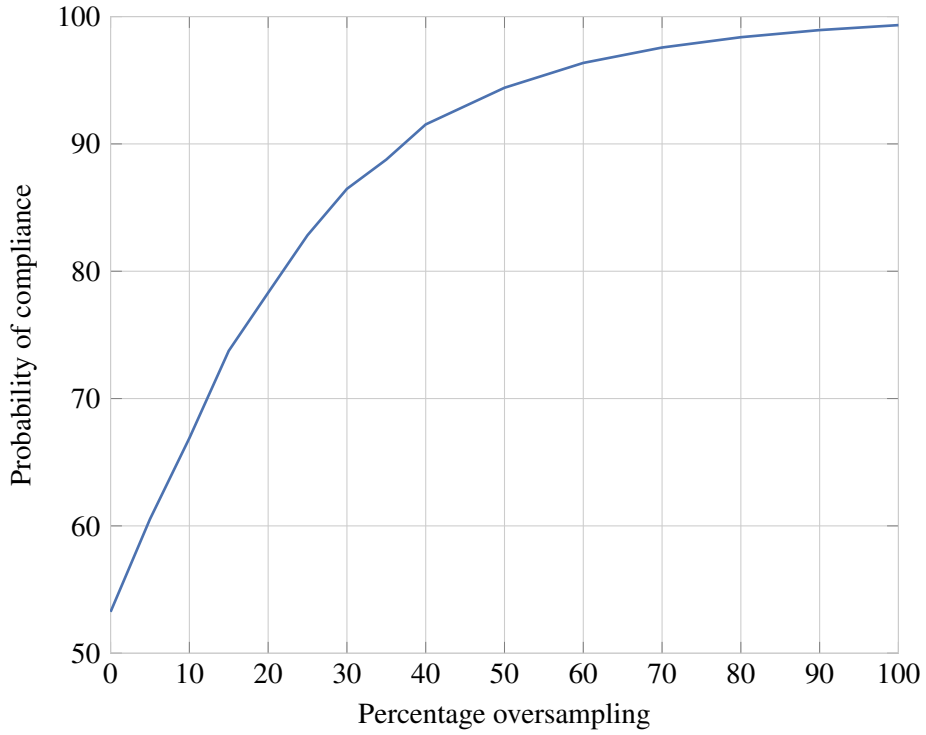


Figure 5.10. The effect of oversampling on the probability of compliance, as per Section 5.4.1.6.

In the first approach, the algorithm oversamples by 10-70% and plot the results in Figure 5.10. (The UMP recommends 10-30% oversampling [37].) This relationship depends on past sample sizes, CV, reporting uncertainty requirements, and other factors. It can be seen that the probability of compliance increases as the percentage of oversampling increases, but there is also a diminishing return on investment.

The second approach is to determine a robust sampling design based on the DLM. In this approach, Step 3 above is planned not according to D_3 taking the most likely value of the forecast, but according to the value at the forecast lower confidence limit $LCL_{m,\ 3,\ 90\%}$. Instead of blindly oversampling, this result leverages the capabilities of the DLM to decrease the likelihood of non-compliance. It was found that when this is done, the probability of compliance reaches 100%. It comes at a cost, however. Robust designs have larger samples, following the curve illustrated in Figure 5.10.

From these results, it is evident that naïve efficient M&V designs have an inherent risk in cases where metering is done. The risk is compounded by the fact that the sampling plan cannot be amended or expanded at a later date, as survey designs could be.

It may seem as though robust sampling is much more costly than naïve efficient sampling. However, this is only true if cost is narrowly defined as metering cost. In a robustly efficient sampling plan, on the other hand, the cost of metering is traded off against the cost of non-compliance to uncertainty reporting requirements. Considering non-compliance can make naïve efficient plans be as costly as they actually are, because such penalties may be incurred in all but the best-case scenarios. Furthermore, a robust sample size in the next year will decrease the sample sizes needed in the years after that. One should not expect the robust plan to have the same overall sampling cost as a naïve efficient plan, however.

The analysis above represents a very simple robust plan, and future works may develop a more complete, robust framework, similar to that of Rysanek and Choudhary [61] or Lindley [246] and Bernardo [247].

5.5 CONCLUSION

A DLM with Bayesian forecasting is shown to provide superior uncertainty quantification and sampling designs compared to standard and previously proposed methods. The current method combines the three significant M&V uncertainty sources, namely metering, sampling, and modelling uncertainty, into a coherent energy model which can be used for quantifying uncertainty and designing other types of M&V studies. It is applied to a multi-year M&V lighting retrofit study and found to reduce metering costs by 40%. However, an investigation into the robustness of efficient sampling plans is also conducted. It is found that efficient plans yield valid results for only half of possible scenarios, given the assumptions in the case study.

DLMs are recommended as a useful alternative to standard linear regression for M&V, should reliable uncertainty quantification be required.

CHAPTER 6 EFFICIENT POPULATION SURVIVAL SURVEY SAMPLING

6.1 CHAPTER OVERVIEW

This chapter addresses the second part of the longitudinal M&V problem: population survival survey sampling. After an introduction to the nature of the problem, modelling is discussed in Section 6.3. A dynamic decay model for CFL populations is presented, as well as a mathematical technique for using this model. A method of including this technique in an optimization function is also discussed. In Section 6.4 and Section 6.5, two cases studies are presented: the first is for a homogeneous population, and the second for a stratified population. [1]

6.2 INTRODUCTION

The longitudinal M&V problem has two components: cross-sectional metering, and population survival survey sampling. The metering component was addressed in the previous chapter. The longitudinal sampling component, which considers persistence, will be addressed in this one.

Persistence describes how annual savings are preserved over time, the literature of which was discussed in Section 2.2.4. Population survival is measured using surveys. Survey design is a science in itself, on which the UMP Chapter 12 is a useful introduction [6]. This chapter will focus on the data analysis and modelling method, and not consider factors such as non-response, self-selection, and

[1] This chapter is based on a journal article written by the author as part of his PhD research, published in *Energy and Buildings* [28].

other measurement errors, for which Gustafson presents an excellent Bayesian treatment [197], and Carroll *et al.* present other important methods [196].

After considering various methods for modelling population survival with quantifiable uncertainty, it was decided to use a Dynamic Generalised Linear Model (DGLM) with Bayesian Forecasting. The DGLM is very similar to the DLM used in the previous chapter. The difference is that the *Generalised* Linear Model allows for the modelling of non-normal distributions, which are needed to describe the binomial nature of survival data.

This chapter will consider the survey sampling problem only, and not use the meter sampling method developed in the previous chapter. In the next chapter, these two methods will be implemented simultaneously.

This chapter can be described in M&V terms as follows:
M&V measurement option: Retrofit isolation with key parameter measurement (Measuring population survival, estimating energy use).
Project boundary: The lighting population(s) under investigation.
Baseline and baseline adjustment approach: The baseline is assumed as the business-as-usual case where original-type failed luminaires would have been replaced by identical original-type luminaires.
Savings determination approach: Standard energy efficiency savings (as opposed to normalised savings) is assumed. This is done in Chapter 7.

6.3 MODELLING

6.3.1 General remarks

The purpose of the DGLM in this chapter is to define the remaining population proportion and quantify the uncertainty with which population survival can be reported at a given point in time. This is done by producing a probability distribution on the proportion of the population left at time $t = k$. Once the DGLM has been defined, it will be used as a constraint in an optimization routine in Section 6.3.3 to find an efficient sampling plan that accounts for past data and sample sizes and adheres to future reporting requirements.

In this section the modelling assumptions will be stated before presenting the mathematical model.

1. Energy savings, and not population survival, is the key performance indicator of the EE project.

2. Energy use is the product of energy meter data and population survival data.

3. Incandescent lamps have been retrofitted with energy-saving CFLs.

4. Only the CFLs installed by the original project are considered part of the population, although efficient maintenance approaches for such projects have been developed [52, 98, 297, 298]

5. CFL electricity consumption is constant over time. The decay in performance (reduced lumen output) is not considered.

6. Minimising monitoring cost is not necessarily equivalent to minimising the number of sampling points, or the number of samples per sampling point.

7. M&V reporting frequency and sampling frequency are different. Reporting can be required every second year (or every fifth year in some cases), but sampling may take place annually if there is an advantage in doing so.

8. Survival data can be described as a time series of population proportions (fractions).

6.3.1.1 Dynamic model for population decay

Survival data are time-dependent and have an autoregressive or dynamic relationship [299], which is also useful in control applications [300]. This means that the state at time $t = k+1$ can be inferred from $t = k$, provided that some model parameters are known. This attribute is useful for predicting the population proportion at some time in the future, given current data. In autoregressive form this population survival relationship is

$$\Phi_{k+1} = \beta\gamma\Phi_k^2\Delta t - \beta\Phi_k\Delta t + \Phi_k, \tag{6.1}$$

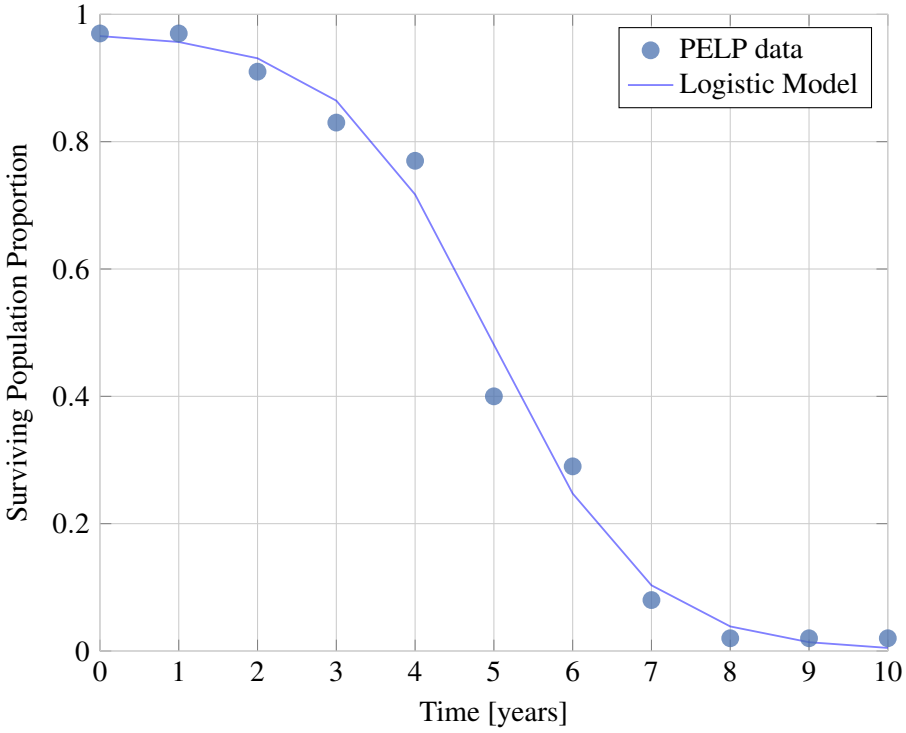


Figure 6.1. PELP data and best fit line of (6.2), as reported by (6.47).

where Φ_k is the proportion of the population surviving at step k, and β and γ are model parameters [111]. Their meanings become apparent when (6.1) is written in its standard form as a logistic equation:

$$\Phi_t = \frac{1}{\gamma + e^{\beta t - L}}. \tag{6.2}$$

The slope of the logistic curve is determined by β, and the starting population proportion is determined by γ. Theoretically $\gamma = 1$, but this is not the case for real data: some measures are removed immediately because of customer dissatisfaction, for example [165]. The model is sensitive to fixing this value, and therefore it is best left as a variable. The L-term falls away for the autoregressive model in (6.1) as it shifts the curve left or right, determining the median life of the population, but is not needed to predict $t_{k+1}|t_k$. To visualise this logistic curve, data from the PELP [169] has been plotted with a best fit line of (6.2) in Figure 6.1. The PELP was a large-scale CFL retrofit study undertaken in the late 1990s where over a million luminaires were installed and tracked over many years. It is a reliable data set and was adopted for use in the South African case [171], but represents only one instance of such data. For Section 6.5 on stratified sampling, other data sets will be used.

If the points on the logistic curve were known with certainty, predicting future population survival would need only simple regression. However, the current state of knowledge of the system is imperfect. Uncertainty is introduced because the whole population is not surveyed. Credible intervals around current and future estimates should, therefore, be determined, but are dependent on the confidence with which previous estimates were determined. This, in turn, is dependent upon the sizes of previous samples. Furthermore, the population proportion estimate $\hat{\Phi}_t$ at t is dependent on the regression model and the sample size $n_{s,\,t}$ at t. A larger sample size at t should have a greater influence on $\hat{\Phi}_t$ than a smaller sample size. The sampling schedule and recency of the previous sample should also play a role in the current estimate. Therefore it is desirable to use a weighted regression technique to characterise the population survival curve. Fortunately, such models do exist in the form of Generalised Linear Models (GLMs). However, GLMs do not usually take sample sizes into account, and do not provide uncertainty bounds around predictions. To do so, Bayesian forecasting techniques need to be added. The model is referred to as a Dynamic GLM because the parameter values are updated at each time step.

6.3.1.2 Bayesian forecasting

An introduction to Bayesian theory has been given in Chapter 3. A brief remark on the use of priors is in order, however. As discussed in Section 3.4.1, the priors used in Bayesian calculations can be informative (and often subjective), or non-informative (as is often used in the objective approach). The DGLM allows for the use of informative, objective priors. This is done by using the weighted regression forecast derived from previous samples and forecasting according to the DGLM and the population decay model in (6.1).

An informative prior is used because longitudinal population survival sampling models can be solved analytically, since the likelihood and prior are naturally conjugate. This condition is satisfied if convolving the likelihood distribution type with the prior distribution results in a posterior which is of the same distribution type as the prior. This is the case for the binomial survival data set, since it is known that binomial likelihood distributions and beta prior distributions form a conjugate pair resulting in a beta posterior. It is, therefore, possible to solve a repeated Bernoulli trial sampling problem without using Bayesian MCMC. The conjugate prior property is used instead.

Let $\hat{\Phi}_t$ be the estimate of the true population proportion and $\mathbf{D}_{t-1}$ be the data available up to but not including t, Y_t the number of lamps in the sample functioning at t, and $n_{s,\,t}$ the survey sample size at t. Also, let r_t and s_t be the first and second moments defining the Beta distribution. Then:

$$\mathbf{D}_t = \{\mathbf{D}_{t-1}, Y_t\}, \tag{6.3}$$

and

$$(\hat{\Phi}_t|\mathbf{D}_t) \sim \text{Beta}(r_t + Y_t, s_t + n_{s,\,t} - Y_t), \tag{6.4}$$

given that the conjugate prior is Beta:

$$(\hat{\Phi}_t|\mathbf{D}_{t-1}) \sim \text{Beta}[r_t, s_t], \tag{6.5}$$

and the likelihood function is binomial:

$$(Y_t|\hat{\Phi}_t) \propto \begin{pmatrix} n_{s,\,t} \\ Y_t \end{pmatrix} \hat{\Phi}_t^{Y_t}(1-\hat{\Phi}_t)^{n_{s,\,t}-Y_t}. \tag{6.6}$$

This means that if all prior knowledge of the system up to $t = k-1$ can be summarised by a Beta distribution describing the uncertainty around the population proportion, Bayesian statistics may be used to update this estimate with the new data, to provide a new estimate of the population proportion. This will be demonstrated below.

6.3.2 Dynamic Generalised Linear Model

The model proposed below is very similar to an extended Kalman filter, which also relies on Bayesian statistics. The state (population proportion and model parameters) at time t is estimated using previous data $\mathbf{D}_{t-1}$. Data D_t are then collected, and the state estimate for time t is updated. If need be, the state at time $t = k$ in the future can then be forecast. A summary of these steps is given in Table 6.1.

This model was derived from West, Harrison, and Migon's model for Television Viewer Ratings [229, 230]. Their model estimated and forecast television viewer ratings based on sampling a population of viewers and asking them if they were aware of the particular product that was advertised that week. By correlating this awareness to the number of advertisements shown (the perturbation in the system), the system could be characterised. The model used for population decay is somewhat simpler as it is assumed that there are no inputs to the system.

Table 6.1. Notation

Variable	Symbol	Information	Updating	Forecasting
Time	t	$t-1$	t	$t+k$
Data	D	$\mathbf{D}_{t-1}$	$\mathbf{D}_t$	$\mathbf{D}_t$
State Vector	$\boldsymbol{\theta}$	$\sim [\mathbf{m}_{t-1}, \mathbf{C}_{t-1}]$	$\sim [\mathbf{m}_t, \mathbf{C}_t]$	$\sim [\mathbf{a}_t(k), \mathbf{R}_t(k)]$
Dist. on P	P	$\sim$Beta$[r_t, s_t]$	$\sim$Beta$[r_t^*, s_t^*]$	$\sim$Beta$[r_t(k), s_t(k)]$
P "mean"	E[P]	f_t	f_t^*	$f_t(k)$
P "variance"	VAR[P]	q_t	q_t^*	$q_t(k)$

The state of the population decay model from (6.1) can be described by the parameters:

$$\boldsymbol{\theta}_t' = (\beta_t, \gamma_t, \Phi_t). \tag{6.7}$$

The estimated population proportion $\hat{\Phi}_t$ can then be described by

$$\hat{\Phi}_t = \mathbf{F}\boldsymbol{\theta}_t \tag{6.8}$$

where the regression vector is

$$\mathbf{F} = (0,0,1). \tag{6.9}$$

The state vector evolves according to

$$\boldsymbol{\theta}_t = \mathbf{G}_t \boldsymbol{\theta}_{t-1}, \tag{6.10}$$

however, in this special case

$$\boldsymbol{\theta}_t = \mathbf{g}_t(\boldsymbol{\theta}_{t-1}). \tag{6.11}$$

Letting $'$ denote a transponent, so that column vectors can be written as row vectors, $\mathbf{g_t}(\mathbf{z})$ and $\mathbf{G_t}(\mathbf{z})$ are defined as follows. Using (6.1),

$$\mathbf{g}_t(\mathbf{z})' = (\beta, \gamma, \beta\gamma\Phi^2 - \beta\gamma + \Phi) \tag{6.12}$$

and

$$\mathbf{G}_t(\mathbf{z})' = \begin{bmatrix} 1 & 0 & \gamma\Phi^2 - \Phi \\ 0 & 1 & \gamma\Phi^2 \\ 0 & 0 & 2\beta\gamma\Phi - \beta + 1 \end{bmatrix}. \tag{6.13}$$

6.3.2.1 Information step

The moments of the parameter estimate distributions for the information step $\boldsymbol{\theta}_{t-1}$ may be defined as

$$(\boldsymbol{\theta}_{t-1}|\mathbf{D}_{t-1}) \sim [\mathbf{m}_{t-1}, \mathbf{C}_{t-1}], \tag{6.14}$$

where $\boldsymbol{m}_{t-1}$ is the mean vector estimate of $\boldsymbol{\theta}$, taken from the previous time step's m_t, defined in (6.37). $\mathbf{C}$ is the covariance matrix. The current state given past data is defined as:

$$(\boldsymbol{\theta}_t|\mathbf{D}_{t-1}) \sim [\mathbf{a}_t, \mathbf{R}_t] \tag{6.15}$$

with the mean vector of the prior as

$$\mathbf{a}_t = \mathbf{g}_t(\mathbf{m}_{t-1}), \tag{6.16}$$

as in (6.11). The variance matrix of the prior is defined as

$$\mathbf{R}_t = \mathbf{G}_t(\mathbf{C}_{t-1})\mathbf{G}_t' + \mathbf{W}_t, \tag{6.17}$$

where the evolution variance matrix $\mathbf{W}_t$ is

$$\mathbf{W}_t = \mathbf{G}_t\mathbf{U}_t\mathbf{G}_t', \tag{6.18}$$

and the discounted covariance matrix $\mathbf{U}_t$ is

$$\mathbf{U}_t = 0.03\mathbf{C}_{t-1}, \tag{6.19}$$

and

$$\mathbf{G}_t = \mathbf{G}_t(\boldsymbol{\theta}_t)_{\boldsymbol{\theta}_t=\mathbf{m}_{t-1}}. \tag{6.20}$$

This is the prior. Translating this to the moments of the prior distribution at time t is done as follows:

$$\begin{pmatrix} \mu_t \\ \boldsymbol{\theta}_t \end{pmatrix} \Bigg| D_{t-1} \sim \left[\begin{pmatrix} f_t \\ \mathbf{a}_t \end{pmatrix}, \begin{pmatrix} q_t & \mathbf{F}_t'\mathbf{R}_t \\ \mathbf{R}_t\mathbf{F}_t & \mathbf{R}_t \end{pmatrix} \right], \tag{6.21}$$

where the forecast mean of the information step is

$$f_t = \mathbf{F}_t'\mathbf{a}_t, \tag{6.22}$$

in accordance with (6.8), and the forecast variance of the information step is

$$q_t = \mathbf{F}_t'\mathbf{R}_t\mathbf{F}_t. \tag{6.23}$$

f_t and q_t are the mean and variance terms of the distribution of μ_t given D_{t-1}. However, in the binomial case

$$(\mu_t|D_{t-1}) \sim \text{Beta}[r_t, s_t], \tag{6.24}$$

which implies that

$$f_t = \mathrm{E}[\mu_t | D_{t-1}] = \frac{r_t}{r_t + s_t}, \tag{6.25}$$

and

$$q_t = \mathrm{V}[\mu_t | D_{t-1}] = \frac{f_t(1-f_t)}{r_t + s_t + 1}. \tag{6.26}$$

r_t and s_t are the first and second moments of the beta distribution, and are related to the number of successes and failures obtained in a certain set of Bernoulli trials. For binomial linear regression, West and Harrison [230] invert (6.25) and (6.26) so that:

$$r_t = f_t \left[\frac{f_t(1-f_t)}{q_t} - 1 \right] \tag{6.27}$$

$$s_t = (1-f_t) \left[\frac{f_t(1-f_t)}{q_t} - 1 \right]. \tag{6.28}$$

However, this is found to be inaccurate for the current case. An alternative is given as

$$f_t = \psi(r_t) - \psi(s_t) \tag{6.29}$$

and

$$q_t = \dot{\psi}(r_t) + \dot{\psi}(s_t) \tag{6.30}$$

where ψ and $\dot{\psi}$ are the digamma and trigamma functions respectively. However, optimizing for these in Python does not lead to very accurate estimates of r_t and s_t. The following formulae were used:

$$r_t = \left(\frac{1-f_t}{q_t} - \frac{1}{f_t} \right) f_t^2, \tag{6.31}$$

$$s_t = \frac{\left(q_t + f_t^2 - f_t\right)(f_t - 1)}{q_t}. \tag{6.32}$$

This then is the prior used for the Bayesian analysis.

6.3.2.2 Updating step

Updating Φ_t for data sampled at t, the posterior may be written as

$$(\Phi_t | D_t) \sim [f_t^*, q_t^*], \tag{6.33}$$

with the update step forecast mean and variances defined as

$$f_t^* = \mathrm{E}[\Phi_t | D_t] = \frac{r_t + Y_t}{r_t + s_t + n_{s,\,t}}, \tag{6.34}$$

and

$$q_t^* = \mathrm{VAR}[\Phi_t | D_{t-1}] = \frac{f_t^*(1-f_t^*)}{r_t + s_t + n_{s,\,t} + 1}. \tag{6.35}$$

The posterior moments can then be updated as follows:

$$(\boldsymbol{\theta}_t|D_t) \sim [\mathbf{m}_t, \mathbf{C}_t], \tag{6.36}$$

where

$$\mathbf{m}_t = \mathbf{a}_t + \mathbf{R}_t\mathbf{F}_t\left(\frac{f_t^* - f_t}{q_t}\right) \tag{6.37}$$

and

$$\mathbf{C}_t = \mathbf{R}_t - \mathbf{R}_t\mathbf{F}_t\mathbf{F}_t'\mathbf{R}_t\left(\frac{1-q_t^*}{q_t^2}\right). \tag{6.38}$$

6.3.2.3 Forecasting Step

For forecasting to year k given the data at t,

$$(\boldsymbol{\theta}_{t+k}|\mathbf{D}_t) \sim [\mathbf{a}_t(k), \mathbf{R}_t(k)] \tag{6.39}$$

with $\mathbf{a}_t(k)$ as in (6.16). $\mathbf{R}_t$ is calculated as follows:

$$\mathbf{R}_t(k) = \mathbf{G}_t(k)\mathbf{R}_t(k-1)\mathbf{G}_t'(k) + \mathbf{W}_t(k) \tag{6.40}$$

where $\mathbf{W}$ is the evolution variance matrix, defined as

$$\mathbf{W}_t(k) = \mathbf{G}_t(k)(\mathbf{C}_t + \mathbf{U}_t)\mathbf{G}_t'(k), \tag{6.41}$$

Similarly, $f_t(k)$ and $q_t(k)$ are calculated according to (6.22) and (6.23). The parameters of the posterior (forecast) distribution can be calculated through (6.24), (6.27), and (6.28).

6.3.2.4 Confidence/credible interval estimation

Confidence (or credible) intervals may also be calculated using the posterior Beta distribution. Because the probability distributions are asymmetric, an equal-tailed confidence interval will not capture the true nature of the system, and the HDI is used, as explained in Section 1.9.

As in (6.5), the distribution on μ_t is beta. However, in (6.5) $\mathbf{D}_t$ is not taken into account, and r_t and s_t should be considered in the light of f_t^* and q_t^*. Thus, following (6.27) and (6.28),

$$r_t^* = f_t^*\left[\frac{f_t^*(1-f_t^*)}{q_t^*} - 1\right] \tag{6.42}$$

$$s_t^* = (1-f_t^*)\left[\frac{f_t^*(1-f_t^*)}{q_t^*} - 1\right]. \tag{6.43}$$

However, the HDI on the population proportion estimate is not of specific interest for this study. Instead, the HDI on the total savings, which incorporates other uncertainty sources as well, is sought. This is solved from the Johnson PDF in Python. More on this point in Section 6.4.2.

6.3.3 Optimization

The DGLM described above computes the confidence with which population proportions can be reported at a given point in time for a given sampling plan. It does not produce an optimal sampling plan. For that, a cost or fitness function should be specified and a sampling plan devised using an optimization model. In this section, such a model will be formulated. The notation will be the same as for Section 5.4.1.3, with the subscript $_s$ indicating survey sampling, instead of the subscript $_m$ used before to indicate metering.

6.3.3.1 Mathematical formulation

From the notation above and in Section 5.4.1.3, the fitness function can be defined as

$$\min \sum_{t=\tau}^{N} d_t v + n_{s,\,t} w + r(\mathbf{n}), \tag{6.44}$$

where

$$r(\mathbf{n}) = \sum_{i \in \mathbf{M}} \left(10^5 w(e_t - \varepsilon) + 10^7 + 5w n_{s,\,benchmark,\,i}\right) \ \forall t \in \chi \tag{6.45}$$

and

$$e_t = \frac{\hat{\Phi}_t - LCL_t}{\hat{\Phi}_t}. \tag{6.46}$$

6.3.3.2 Model explanation

The proposed fitness function evaluates a particular sampling regime $\mathbf{n}$ based on the cost w per sample, and a cost v of initiating sampling for a given time point, as well as a penalty constraint discussed in the next paragraph. For the cost, if $w = 10$ and $v = 1000$, the cost of taking one sample at t would be R1 010, and for 5 samples it would be R1 050. This is the standard survey costing scheme used in literature [301, 302]. Levy and Lemeshow do not consider sampling initiation costs [303], while Hansen describes a more thorough approach where costs may vary per stratum [304]. Barnett [301] adds a term for travelling required between samples.

Table 6.2. Case Study 1 Model Parameters. $\sim N[\cdot]$ indicates a normal distribution.

Description	**Symbol**	**Value**
Initial Covariance Matrix	$\mathbf{C}_0$	$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$
Confidence	α	0.9
Discount factor		0.03
Precision	ε	0.1
Study duration	N	12
Fixed cost	v	1000
Variable Cost	w	10
Baseline Power	P_b	$\sim N[60, 1.5]$
Reporting Period Power	P_r	$\sim N[11, 0.275]$
Hours of use	HOU	$\sim N[3.11, 0.15]$
Number of units	$n_{retrofitted}$	10^5

A GA and constraint function as described in Section 5.4.1.3 was used. The parameter values for the DGLM were set as indicated in Table 6.2.

6.3.4 Risk-conscious sampling design

The optimization described above is for the lowest cost sampling plan. Under this strategy, it is assumed that the outcomes of future surveys fall precisely on the forecast population survival curve and that no data are lost. Assuming the most likely future survey outcome is the natural approach to sampling planning. However, taking only an optimal number of samples is risky as discussed in Sections 5.4.1.1 and 5.4.1.6. Survey outcomes which differ from this forecast value will result in worse confidence intervals than those forecast by the model. The result could be that the optimal sample size does not meet the required accuracy level for M&V reporting. Unlike the metering case, it would be simple to enlarge the sample size by surveying additional units, and therefore this scenario will not be detrimental to the M&V study in the same way metering under-sampling would. Furthermore, the concern is only for the present year of the study, not future years. Under- or oversampling in past

years is accounted for as a matter of course by the DGLM, and future sample sizes will be adjusted accordingly by the optimization routine. Nevertheless, the robustness of the current year's required sample size is still of concern.

6.4 CASE STUDY 1: SIMPLE RANDOM SAMPLING DESIGN

6.4.1 Data

The first case study considered is for a single unstratified population of CFLs which were installed in a lighting retrofit project. The survival data used for this case study is the PELP dataset [169] discussed above. The parameters for (6.2) were determined to be $[\gamma, \beta, L] = [1.030, 1.056, 5.233]$, so that

$$\Phi_{t,sim} = \frac{1}{1.030 + e^{1.056 \times t - 5.233}}. \tag{6.47}$$

A realistic way to use this data would be to generate data points according to

$$\mathbf{D}_{sim,t} \sim Binomial[n = n_{s,\,t},\ p = \hat{\Phi}_{t,sim}]. \tag{6.48}$$

However, what often happens is that due to random variation, a dataset will have a sequence such as [0.91, 0.9, 0.96]. Fitting a logistic curve to these data results in a monotonically increasing function. This often happens with optimal (small) sample size allocations for early project years where there is little change in the population. The sample sizes may satisfy the 90/10 criterion but are inadequate for trend determination. Furthermore, if sampling points are not exactly the same as the PELP data, the fitted line will be higher or lower than the PELP best-fit line. This is realistic, but makes benchmarking difficult as it changes the (relative) accuracy limits and therefore the required sample sizes. The PELP data points are therefore used. Smaller sample sizes will still have larger variances in the DGLM; the way real data would.

The PELP data are:

$$\mathbf{D}_{\text{PELP}} = [0.97, 0.97, 0.91, 0.83, 0.77, 0.4, 0.29, 0.08, 0.02, 0.02, 0.02]. \tag{6.49}$$

For this study, it is assumed that data has been collected for the first three years, and supposed that $\mathbf{n}_{0-2} = [250, 250, 250]$ lamps were surveyed. We suppose further that according to the contract, reporting must be done in years 3-6. Therefore $\mathbf{M} = \{3, 4, 5, 6\}$. The contract stipulates that reports

should have a 90% confidence and 10% precision on the estimate, as per CDM [49] and IPMVP [1] guidelines. It is usually assumed that savings are distributed normally and symmetrically around a mean value. However, this will not be the case with beta-distributed estimates. For asymmetrical distributions the mean does not represent the most likely value – the mode does. This is referred to as the Maximum A-Posteriori or MAP estimate. Furthermore, because the distribution of $\hat{\Phi}_t$ is asymmetrical, it may happen that the lower confidence limit of the HDI falls within 10% of MAP estimate of the distribution on $\hat{\Phi}_t$, but the upper limit does not. Additional samples will then be needed to constrain the upper limit of the 90% HDI. However, the upper limit on the savings estimate is not of interest since the conservatism principle of M&V dictates that savings may be underestimated, but should not be overestimated [1]. Therefore the 10% precision bound is considered to apply to the lower limit of the 90% HDI only. If the lower limit of the 90% HDI is more than 10% away from the MAP, the accuracy constraint is violated, and the function is penalised. The GA will then attempt to find a solution for $\hat{\Phi}_t$ that has a tighter HDI around the MAP by increasing the sample size in that or a previous year.

The uncertainty limits do not apply to the population survival estimate alone, but to the overall estimate of the energy saved by the project during the monitoring period. This means that the population proportion estimate $\hat{\Phi}_t$ should have an accuracy greater than 90/10. How much greater will depend on the variance of the estimate of the energy saved per unit, as well as the variance in the baseline energy use.

This chapter considers the retrofit isolation with key parameter measurement approach. The next chapter will expand this to retrofit isolation with all parameter measurement by including meter sampling. However, to keep the current model focussed on survey sampling, the metering sampling DLM will not be considered here.

If one supposes that 99% of the lamps were working at the time of the retrofit, $_b$ represents baseline and $_r$ represents reporting, then the savings may be calculated as:

$$E_{saved,\,t} = n_{retrofitted} HOU\,(0.99P_b - P_r)\hat{\Phi}_t, \tag{6.50}$$

Where HOU is the distribution of daily hours of use and P is the distribution of the power drawn by

the old (P_b) and new (P_r) units. The number of lamps retrofitted is $n_{retrofitted}$. Interactive heating and cooling effects and the in-service rate [38] are not considered.

For the power P drawn by the unit, the uncertainty may be relatively small, although estimates vary. Some tests report a mean value of about -5.8% compared to the labelled power in laboratory tests of CFLs [170], others 1.75% [305], while others report in the order of $\pm 0.5\%$ in actual operating conditions of fluorescent lamps and pre-retrofitted fixtures generally [306]. These are only the active power (in Watts) measurements. CFLs also have significant power factor and harmonic distortion effects, but considering these will take us too far afield for the current study. Usually, 60W incandescent lamps are replaced with 11W CFLs. The energy saving is therefore 49W per fixture, according to the recommended lumen-equivalence savings method [38]. Therefore a 2.5% error was selected.

The values used for these distributions are summarised in Table 6.2. Typically the hours of use are the most uncertain factor in lighting retrofit projects [1], and this uncertainty should be taken into account. Vine and Fielding [293] conducted a meta-study of CFL HOU studies. For the 25 CFL HOU estimates listed by them for summer interior fixtures, the mean of was 3.11 hours per day, and the median 3.00 hours per day.[2] This accords with the CDM assumption [50]. Their data are distributed as

$$\text{HOU study estimates} = 2+ \sim Exponential[0.893]. \tag{6.51}$$

Few of the studies listed by Vine and Fielding mention uncertainty. Those that do list them as 17%, 10%, 4%, and 3% at the 68% confidence level. The last two will still be adequate at the 90% confidence level (assuming normally distributed data). For this study, a 4% uncertainty on the HOU is assumed. It is also assumed that the HOU stay the same between the baseline and the reporting period. Should snapback or rebound [38] be proven, two different HOU terms could be defined for the baseline and reporting periods.

6.4.2 Distribution convolution

Many of these values have been assumed to be normally distributed, but need to be convolved with the beta estimates calculated by the DGLM. This is difficult to do analytically, and therefore numeric

[2] Another meta-study disaggregated lamp HOUs and found lower numbers [307]. For information on CFL HOUs disaggregated per installation location, see [308].

Monte Carlo convolution would be the standard way of calculating (6.50). However, MC simulation has two related disadvantages in the context of threshold optimization.

The first is inconsistency. MC is usually considered very accurate, provided that enough trials have been conducted. This is because the shapes and means of the posterior MC distributions are usually of interest, not the low-mass tails as in this case. The problem manifests when MC simulation is done for each individual in each generation of the GA. There is some inter-simulation variation between the MC realisations of the same parameters. Because the GA seeks an optimal solution, certain individuals which are slight MC outliers due to noise, are evaluated to be the fittest individuals because they *seem* to adhere to the constraints, when this is just an artefact of MC noise. They only seem to conform because that specific MC realisation is not a perfect reflection of the convolution. On most other runs the same solution (sampling plan) would not conform to the accuracy requirements. Although these 'false positives' happen relatively rarely, they mislead the algorithm by incorrectly altering the ranking of good solutions. For 10^6 MC trials, 20 out of 20 of the best GA solutions were such false positives. For 10^7 MC trials, 19 out of 20 were false positives, violating the constraints by 0.01-0.05. Although increasing the MC trial size could help, it is very expensive to convolve such large datasets for each individual in each generation, and the GA approach then becomes impractical. For example, the case study in this chapter runs for 33 minutes for an MC trial size of 10^6, and 5 hours, 20 minutes for trial size of 10^7. Although these speeds may be improved by using specialised hardware or faster software, this would reduce the appeal of MC, which is its wide applicability and ease of implementation.

The solution to this problem is to calculate the E_{saved} in (6.50) analytically. This is usually thought to be very difficult or impossible. However, recent work by Kuang and Rajan et al. [142, 143] have produced a method by which the moments of the posterior of a convolution of a polynomial expression of distributions may be expressed in terms of the scale and shape parameters of the constituent distributions. This allows for exact expressions of the moments of the resultant distribution, at a fraction of the computational burden of an MC simulation. This work is made available through an online toolbox as the Mellin Transform Moment Calculator (MTMC) [188], on which see Section 2.3.1.1. By using the first four moments (translated to mean, variance, skewness, and kurtosis), a Johnson S_B (bounded) distribution [189] can then be fitted. Although the first four moments of a distribution do not identify it uniquely for all cases, the distribution on E_{saved} is unimodal and will be adequately described, since this distribution family was expressly designed for such flexibility. For more information on uncertainty evaluation through moment-based distribution fitting, see Rajan *et al.* [144]. The Johnson

distribution is fitted using Hill's algorithm [309], implemented in Matlab/Octave [310] and then linked to the DGLM in Python.

To illustrate, let (6.50) by expressed as a polynomial of distributions:

$$E_{saved,\,t} \sim n_{retrofitted} \cdot N[m_1, s_1] \cdot (\Phi_b \cdot N[m_2, s_2] - N[m_3, s_3]) \cdot Beta[b_{4a}, b_{4b}, lb_4, ub_4], \tag{6.52}$$

where Φ_b indicates the proportion of functioning lamps during the baseline period, and lb_{4a} and ub_2 are the upper and lower bounds of the beta distribution, which for this case are zero and one respectively.

The subscript t is omitted from the following equations for notational clarity, but the moments do still apply to distributions at specific time points. The first four raw moments of the resultant distribution obtained via the MTMC are:

$$\mathrm{E}[y_1] = \frac{n_{retrofitted} b_{4a} m_1 (\Phi_b m_2 - m_3)(lb_4 + ub_4)}{b_{4a} + b_{4b}}, \tag{6.53}$$

$$\mathrm{E}[y_2] = \frac{n_{retrofitted}^2 b_{4a}(1 + b_{4a})\left(m_1^2 + s_1^2\right)\left((m_3 - \Phi_b m2)^2 + b^2 s_2^2 + s_3^2\right)(lb_4 + ub_4)^2}{(b_{4a} + b_{4b})(1 + b_{4a} + b_{4b})}, \tag{6.54}$$

$$\mathrm{E}[y_3] = \frac{n_{retrofitted}^3 b_{4a} m_1 (1 + b_{4a})(2 + b_{4a})(\Phi_b m_2 - m_3)\left(m_1^2 + 3s_1^2\right)\left((m_3 - \Phi_b m_2)^2 + 3\Phi_b^2 s_2^2 + 3s_3^2\right)(lb_4 + ub_4)^3}{(b_{4a} + b_{4b})(1 + b_{4a} + b_{4b})(2 + b_{4a} + b_{4b})}, \tag{6.55}$$

and

$$\mathrm{E}[y_4] = \frac{AB}{(b_{4a} + b_{4b})(1 + b_{4a} + b_{4b})(2 + b_{4a} + b_{4b})(3 + b_{4a} + b_{4b})}. \tag{6.56}$$

where

$$A = n_{retrofitted}^4 b_{4a}(1 + b_{4a})(2 + b_{4a})(3 + b_{4a})\left(m_1^4 + 6m_1^2 s_1^2 + 3s_1^4\right) \tag{6.57}$$

and

$$\begin{aligned} B \;=\; & \left[(m_3^4 - 4\Phi_b^3 m_2 m_3\left(m_2^2 + 3s_2^2\right) + \Phi_b^4\left(m_2^4 + 6m_2^2 s_2^2 + 3s_2^4\right) + 6m_3^2 s_3^2 + 3s_3^4 + \right. \\ & \left. 6\Phi_b^2\left(m_2^2 + s_2^2\right)\left(m_3^2 + s_3^2\right) - \Phi_b m_2 m_3\left(m_3^2 + 3s_3^2\right)\right](lb_4 + ub_4)^4. \end{aligned} \tag{6.58}$$

They could also be expressed as hypergeometric ${}_2F_1$ functions, but the algebraic expressions evaluate faster on a computer. The mean, variance, skewness, and kurtosis can be calculated as

$$, \text{mean} = \mathrm{E}[y_1] \tag{6.59}$$

$$\text{standard deviation} = \sqrt{\mathrm{E}[y_2] - \mathrm{E}[y_1]^2}, \tag{6.60}$$

$$\text{skewness} = \frac{\mathrm{E}[y_3] - 3\mathrm{E}[y_1]\mathrm{E}[y_2] + 2\mathrm{E}[y_1]^3}{\text{standard deviation}^3}, \tag{6.61}$$

$$\text{kurtosis} = \frac{\mathrm{E}[y_4] - 4\mathrm{E}[y_1]\mathrm{E}[y_3] + 6\mathrm{E}[y_1^2\mathrm{E}[y_2] - 3\mathrm{E}[y_1]^4}{\text{standard deviation}^4}. \tag{6.62}$$

These in turn are used in the Johnson distribution:

$$E_{saved,\,t} \sim Johnson\,[\text{mean}, \text{standard deviation}, \text{skewness}, \text{kurtosis}] \tag{6.63}$$

to yield very accurate (and consistent) representations of the true distribution.

6.4.3 Specification of initial estimates for DGLM optimization

The initial conditions for $\boldsymbol{\theta}'$ in (6.7) were specified using the known PELP data. Weighted least-squares regression could also be used if enough sampling points were made available. For practical problems with realistic sample sizes, the model is insensitive to the covariance matrix and discount factor specification, as long as the variances are not set to zero. It is sensitive to the specification of γ_0 and β_0, however. If less than three sampling points are available, and these are from the early years of the study, it is recommended that these values be specified as equal to one rather than doing a least-squares regression to determine the parameters.

6.4.3.1 GA tuning

Usually GAs are not very sensitive to the initial population (or starting point). However, there are significant stepwise discontinuities in the solutions space. Therefore certain steps were taken to

Table 6.3. GA Parameter values.

Parameter	Value
GA Algorithm	`MuPlusLambda`
Crossover Rule	Uniform Crossover
Crossover proportion	50%
Crossover exchange probability	95%
Mutation Proportion	50%
Individual gene mutation probability	10%
Number of Generations	30
Population Size	50

improve the GA effectiveness. The first is that the starting population was populated with known good solutions – either from non-DGLM benchmarks or from previous GA results. The mutation proportion and probability was increased, and the mutation function was also adapted to yield negatively-biased mutations (see (5.23)), since the GA was found to converge on good sampling patterns, but not to optimize those patterns to the precision limits.

The `DEAP` Python library [294] was used for this calculation, as it includes many of the standard methods and allows for rapid prototyping. The parameters used are reported in Table 6.3.

A tournament size of ten supplies a severe selection pressure and homogenises the population within a few generations. The mutation function was then set in such a way to make incremental improvements. The large mutation proportion ensures a steady incremental improvement rate

6.4.4 Benchmark

The proposed method needs to be benchmarked against a realistic alternative method for solving the problem at hand. Goldberg [69] and the UMP [40] proposed leveraged sampling for M&V designs, where regression or prior estimates are used to reduce the variance in the estimated mean, under the assumption of normality. However, as normality is not assumed this is not applicable.

First, a note on confidence intervals. Various confidence intervals for binomial proportion sampling have been suggested, on which Brown, Cai, and DasGupta provide an illuminating study [311]. Of these, Jeffreys sampling has been shown to be the most accurate and least conservative [311] and is used here, although less accurate methods are often used in M&V [40]. The Jeffreys interval is derived from a Beta distribution with a (0.5, 0.5) prior. However, it yields an equal-tailed confidence interval and not an HDI. Also, it does not account for the other uncertainty components. To be consistent with the approach in the rest of this study, the benchmark is defined as the smallest lamp population survey sample size for which the lower confidence limit of the 90% HDI is less than 10% away from the mode. The difference between this sample survey plan and the DGLM plan would then be the cost saving contribution made by the predictive power of the prior information used by the DGLM.

To determine the benchmark sample size, the estimate population proportion $\hat{\Phi}_t$ is needed. One could use the PELP best-fit line, but this has misleading results: if the DGLM line is higher than the PELP line for some future point, the 10% precision limits will be larger than the PELP limit. A smaller sample size will therefore be required, making it appear as though the DGLM approach is superior when it is only the population proportion forecast which is higher. Like would not be compared to like in such a case case. A more fair comparison would be to use the same population proportion value $\hat{\Phi}_{\mathbf{M},\,DGLM}$ that was used by the DGLM. Then, using the GA, an optimal sampling plan can be devised using conventional sampling theory (therefore not including the prior information as the DGLM does).

6.4.5 Results and discussion

The model takes a few minutes to run on a laptop computer, with the majority of this time being spent in the GA. The MTMC convolution of the different distributions needed to determine the HDI also has a noticeable effect on performance. This could be because the Johnson distribution is not evaluated natively in Python, but called as an external function. A plot of the minimum and average population fitness vs the generations of the GA (not shown) exhibits the classic concave-up shape, although the incremental-improvement mutation function does lend a more linear quality to it. The average population fitness decreases rapidly in the first ten generations and then approaches a minimum asymptotically in the next 20.

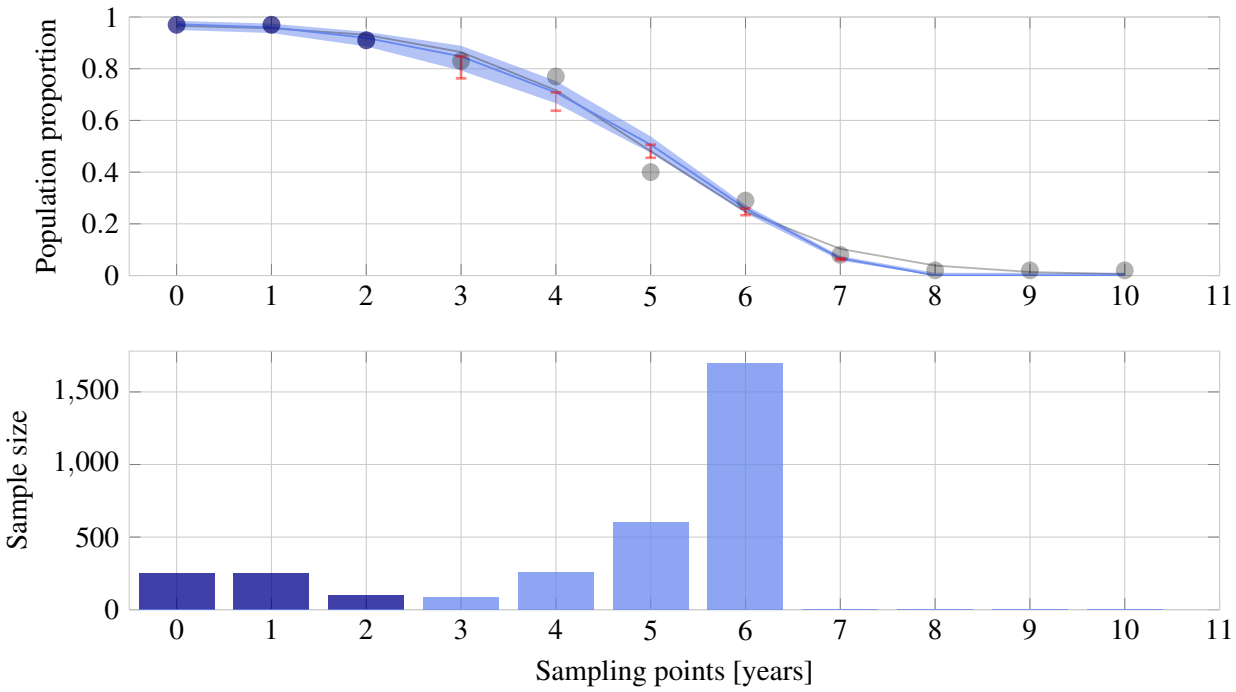


Figure 6.2. Population proportion inferred from results and sample size data from DGLM for Case Study 1, with forecasting to future years, and optimal sample sizes. The grey lines and points are the PELP data and fits, dark blue past sampling results, and light blue future sample sizes. The 90% confidence area is shaded, and the 10% precision limits indicated by the red error bars.

It is assumed that adequate sampling is done within the first few years. Characterising a logistic function with two sampling points close to $\Phi = 1$ (for the first two years) will not be adequate. Therefore the DGLM will not yield accurate sampling plans with such data. From simulations, it is recommended that samples greater than 150 be taken for the first few years. This figure is affected by the error in the data points which translate to a modelling error in the DGLM regression.

An efficient sampling plan is found to be

$$\mathbf{n}_s = [83, 260, 601, 1696, 0, 0, 0], \tag{6.64}$$

at a cost of R30 400. The benchmark for this realisation is

$$\mathbf{n}_{s,\ benchmark} = [153, 303, 707, 2101, 0, 0, 0], \tag{6.65}$$

at a cost of R36 640. The DGLM reduces the sampling cost by 17%. If only every second year is sampled, the savings reduce to single digits, depending on the configuration. This is because the prior

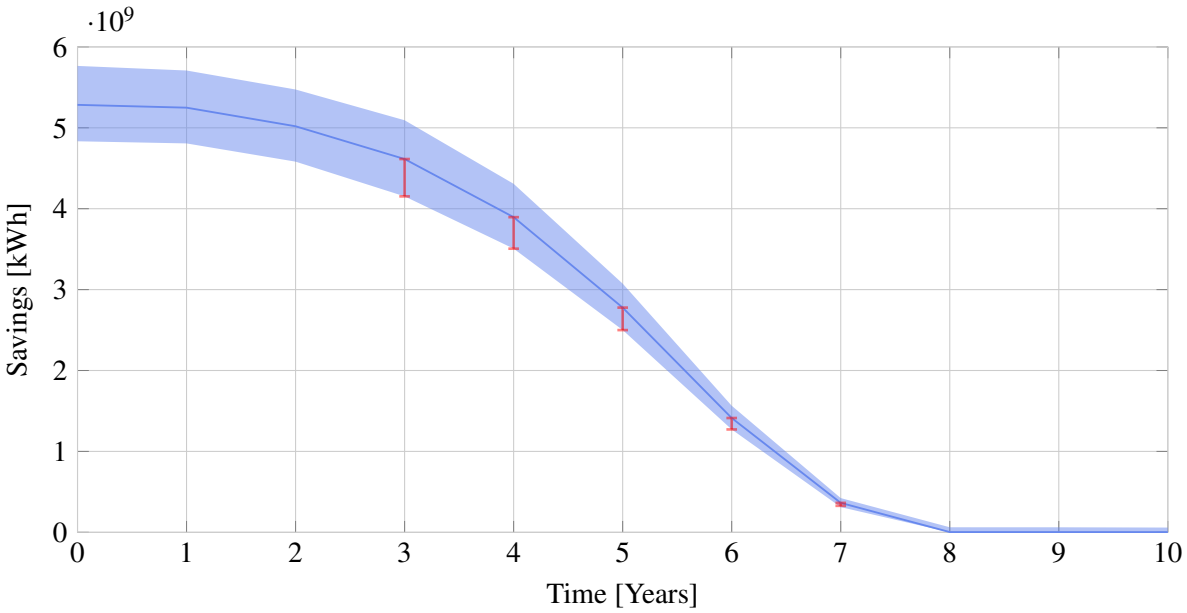


Figure 6.3. Savings inferred by DGLM using (6.50) for Case Study 1.

moments (r_t and s_t) decrease during the forecasting step, as forecasting without annual data increases uncertainty.

At year six only about 25% of the original population of lamps are left. This drops to less than 10% for year seven, greatly increasing the sampling burden at diminished returns, although the DGLM then saves 26% relative to the non-DGLM method.

The results plotted in Figure 6.2 and Figure 6.3 show instantaneous confidence levels. That is, the confidence around year t in year t: $LCL_t|\mathbf{D}_t$. For future sampling years $t+k$, the confidence levels shown assume the samples taken between and including t and $t+k$, that is, $LCL_{t+k}|\mathbf{D}_{t+k}$. The forecast confidence intervals for future years given only the samples taken up to the present time is also possible, but not shown. The DGLM also allows for the calculation of a retrospective confidence level. This is the confidence level for some $t-k$ time in the past, given all the sampling done up to the present time, including the sampling done after $t-k$. This allows for the updating of past estimates, should that be necessary. These are not shown.

Some features of the results plotted in Figure 6.2 warrant attention. It is clear that the actual population proportion result from the survey sample may be different from the true population proportion. This

has been illustrated by plotting both the blue (inferred) and grey (actual or base-case) line. Of course, as the sample size increases this discrepancy tends to disappear. This is also a function of the sample sizes, as well as the number of past sampling points. As the study progresses and real samples are taken, the shape of the true curve emerges.

The red error bars in Figure 6.2 and Figure 6.3 indicate the 10% precision limits around the reported value, and are only plotted for the reporting years. The algorithm does not adhere to these for non-reporting years, which may have confidence intervals wider than the reporting accuracy requirements. Because there is still uncertainty in the other parameters of (6.50), it is expected that the algorithm constrains the population proportion estimate to less than the 90/10 bound, to meet the overall savings 90/10 bound. The overall uncertainty is plotted in Figure 6.3. Less detail is shown in this curve, but it is clear that the 90/10 bound is adhered to for the overall savings estimation, at least in years three to six.

The errors bars show that the uncertainty reporting requirement becomes more stringent as the savings decreases. This leads to a situation where monitoring smaller savings require greater resources [69]. This is illustrated by the relatively large sample sizes required for the later years. This all assumes that the accuracy requirement holds for every year individually, rather than the total projects savings aggregated over the project lifetime, which would be a more efficient policy requirement from an M&V point of view.

6.5 CASE STUDY 2: STRATIFIED SAMPLING DESIGN

The second case study considered is for a heterogeneous population: one in which where there are sub-populations that have different energy use and survival characteristics. In such a scenario sampling should be approached with the overall savings uncertainty in mind. This means that the uncertainty in each sub-population need not adhere to the prescribed uncertainty reporting bounds, but the combined uncertainty of all populations should do so. Practically, this means that one sub-population may be under-sampled and another oversampled if it is justified by the sampling cost and the population's overall uncertainty contribution.

The optimum allocation of sample sizes across different strata of a stratified survey sampling design is

Table 6.4. Case Study 2 Model Parameters. Only those parameters that differ from Table 6.2 are shown. $\sim N[\cdot]$ indicates a normal distribution.

Description	Symbol	Value
Baseline Power 1	$P_{b,\,1}$	$\sim N[60, 1.5]$
Baseline Power 2	$P_{b,\,2}$	$\sim N[60, 1.5]$
Baseline Power 3	$P_{b,\,3}$	$\sim N[100, 2.5]$
Retrofit Power 1	$P_{r,\,1}$	$\sim N[11, 0.275]$
Retrofit Power 2	$P_{r,\,2}$	$\sim N[11, 0.275]$
Retrofit Power 3	$P_{r,\,3}$	$\sim N[14, 0.7]$
Hours of use 1	HOU_1	$\sim N[.11, 0.15]$
Hours of use 2	HOU_2	$\sim N[2, 0.1]$
Hours of use 3	HOU_3	$\sim N[4.11, 0.21]$
Population 1	$n_{retrofitted,\,1}$	5×10^4
Population 2	$n_{retrofitted,\,2}$	2×10^4
Population 3	$n_{retrofitted,\,3}$	3×10^4

well-studied. For example, Barnett [301] listed formulae for different cases, and the UMP Chapter 11 discussed cost-optimal Pearson allocation for M&V in some detail [40]. However, such formulae are not applicable to this case because although an allocation may be optimal for a given year, given certain population proportions for various strata, it may not be optimal in the context of the larger multi-year sampling model. It also does not account for the possible non-normality of the overall savings equation (6.50). Therefore the optimum allocation formulae will not be used. Instead, the GA will be used to find an efficient allocation.

To generate different realistic population survival curves, the data published in the LRC's Specifier Report on CFLs [170] were used. The LRC laboratory tests monitored test bench mounted lamp populations and reported the 5% population decrease intervals. In other words, the population survival interval was fixed, and the time between recordings variable. However, the model described in this paper is more suited to real-world studies in which the observation interval is fixed (once per year), and the population survival figures are variable, such that the proportions of the population surviving after one, two, or three years are reported. To convert the LRC data to a suitable format, least-squares

logistic curves were fitted to the data. The data sets which fitted (6.2) with the sum-squared errors less than 0.05 were then selected (19 of the original 20 sets). Three of these curves were used for the simulation. Slow, medium, and rapid decay rate curves were selected. The binomial uncertainty as discussed in Section 6.4.1 was used to reflect sampling variation and uncertainty. This could be done since the CFLs used for this case study last longer than the PELP ones, and thus more preliminary data points could be collected.

The total energy saving is the sum of the three populations described by (6.50):

$$E_{saved,\,total} = \sum_{i=1}^{3} E_{saved,i}. \tag{6.66}$$

The MTMC can be used to evaluate the moments in the same way as Section 6.4.2, for every individual distribution. The overall savings distribution will then be the sum of the three strata's distributions. The four moments can be calculated as:

$$\mathrm{E}[y_{1,\,T}] = \mathrm{E}[y_1] + \mathrm{E}[y_2] + \mathrm{E}[y_3], \tag{6.67}$$

$$\mathrm{E}[y_{2,\,T}] = \mathrm{E}[y_1^2] + 2\mathrm{E}[y_1]\mathrm{E}[y_2] + \mathrm{E}[y_2^2] + 2\mathrm{E}[y_1]\mathrm{E}[y_3] + 2\mathrm{E}[y_2]\mathrm{E}[y_3] + \mathrm{E}[y_3^2], \tag{6.68}$$

$$\begin{aligned} \mathrm{E}[y_{3,\,T}] \;=\; & \mathrm{E}[y_1^3] + 3\mathrm{E}[y_1^2]\mathrm{E}[y_2] + 3\mathrm{E}[y_1]\mathrm{E}[y_2^2] + \mathrm{E}[y_2^3] + \mathrm{E}[y_1^2]\mathrm{E}[y_3] + \\ & 6\mathrm{E}[y_1]\mathrm{E}[y_2]\mathrm{E}[y_3] + 3\mathrm{E}[y_2^2]\mathrm{E}[y_3] + \mathrm{E}[y_1]\mathrm{E}[y_3^2] + 3\mathrm{E}[y_2]\mathrm{E}[y_3^2] + \mathrm{E}[y_3^3], \end{aligned} \tag{6.69}$$

and

$$\begin{aligned} \mathrm{E}[y_{4,\,T}] \;=\; & \mathrm{E}[y_1^4] + 4\mathrm{E}[y_1^3]\mathrm{E}[y_2] + 6\mathrm{E}[y_1^2]\mathrm{E}[y_2^2] + 4\mathrm{E}[y_1]\mathrm{E}[y_2^3] + \mathrm{E}[y_2^4] + 4\mathrm{E}[y_1^3]\mathrm{E}[y_3] + \\ & 12\mathrm{E}[y_1^2]\mathrm{E}[y_2]\mathrm{E}[y_3] + 12\mathrm{E}[y_1]\mathrm{E}[y_2^2]\mathrm{E}[y_3] + 4\mathrm{E}[y_2^3]\mathrm{E}[y_3] + 6\mathrm{E}[y_1^2]\mathrm{E}[y_3^2] + \\ & 12\mathrm{E}[y_1]\mathrm{E}[y_2]\mathrm{E}[y_3^2] + 6\mathrm{E}[y_2^2]\mathrm{E}[y_3^2] + 4\mathrm{E}[y_1]\mathrm{E}[y_3^3] + 4\mathrm{E}[y_2]\mathrm{E}[y_3^3] + \mathrm{E}[y_3^4]. \end{aligned} \tag{6.70}$$

which can be used in the Johnson distribution in (6.63).

For a three-stratum, twelve-year monitoring project, the sampling solution $\mathbf{n}_s$ is a 3×12 vector.

Table 6.5. Stratified sampling plans for Case Study 2. Benchmark (top), Efficient (bottom)

Years	6	7	8	9	10	11	12
Stratum 1	66	0	178	0	280	0	758
Stratum 2	11	0	11	0	0	0	557
Stratum 3	105	0	301	0	841	0	9338
Stratum 1	481	0	347	0	101	257	601
Stratum 2	0	0	0	0	0	0	0
Stratum 3	356	0	62	0	1010	0	6470

Regarding past samples, for stratum one, the sample sizes were $\mathbf{n}_{s,\,0-4} = [100, 100, 100, 100, 200]$. For strata two and three, the sample sizes were $\mathbf{n}_{s,\,0-4} = [50, 50, 75, 75, 100, 150]$. The reason that the sample sizes increase for the survey sampling is that it is critical to identify the point at which the population curve changes from the plateau to the transition phase. Small sample sizes during these years add disproportionate noise which leads to inaccurate forecasts.

The GA fitness function also needed to be adapted to account for the expanded sampling term and the three separate E_{saved} distributions. Aside from these differences the algorithm works in the same way as for the simple random sampling case, and can easily be expanded to accommodate more strata, or different sampling or initialisation costs for each stratum.

The benchmark for this case study was calculated in a similar way to Section 6.4.4 A GA was used to find the optimal sample size allocation across the different strata, given the uncertainty reporting requirements, while considering the combination of population survival and energy use distributions as specified in Table 6.4.

6.5.1 Results and discussion

Much of the discussion in Section 6.4.5 is also relevant here. The benchmark cost was found to be R116 625, while the DGLM study cost is R59 370. This represents a saving of 49%. The sampling plans are shown in Table 6.5. It is graphically illustrated in Figure 6.4, and the savings curve is shown in Figure 6.5. The DGLM method, therefore, presents a significant advantage over simple

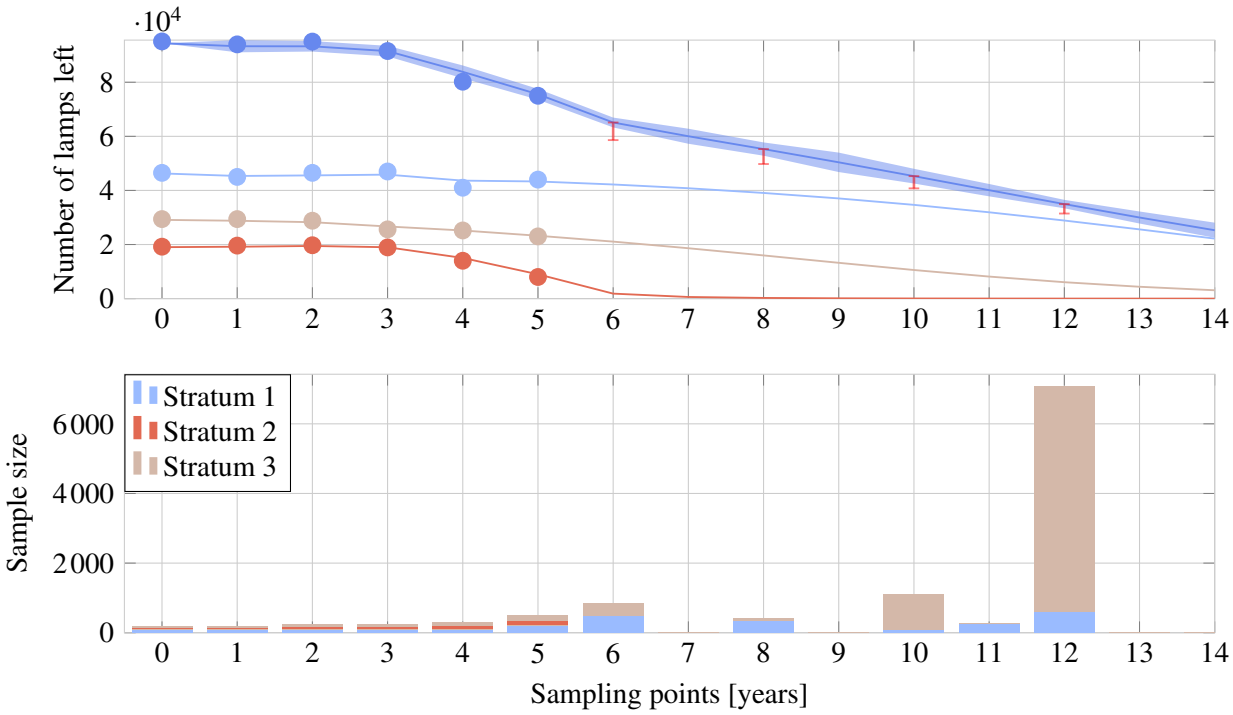


Figure 6.4. Case Study 2 population proportion inferred from results and sample size data from DGLM, with forecasting to future years, and optimal sample sizes. Sampling is done up to and including $t = 5$.

regression methods, even if these methods still use the Mellin Transform Moment Calculator and Genetic algorithm for overall efficient sampling design.

In this case, it was found that accurately determining the point at which the population curve enters the transition phase from the plateau phase is critical for sampling planning. Consider year 5 of stratum 1 in Figure 6.5. If the sampling error makes the population proportion in year 5 appear too high, the DGLM curve fit will predict very little population decay: essentially a horizontal line. This is to be expected: the full decay characteristics of a population can only be determined once the population starts decaying. This is the reason for increasing the sample size of stratum 1 in year 5: to reduce variability and ensure a more accurate estimate. Prior information does help, but in this case γ_0 and β_0 were determined from a weighted ordinary least squares regression on the known data points.

It is evident that in years where no sampling takes place (year 7, for example), the uncertainty bounds widen. In years where large samples are taken (e.g. year 12), the uncertainty bounds are 'pinched' as the DGLM accounts for the increased certainty.

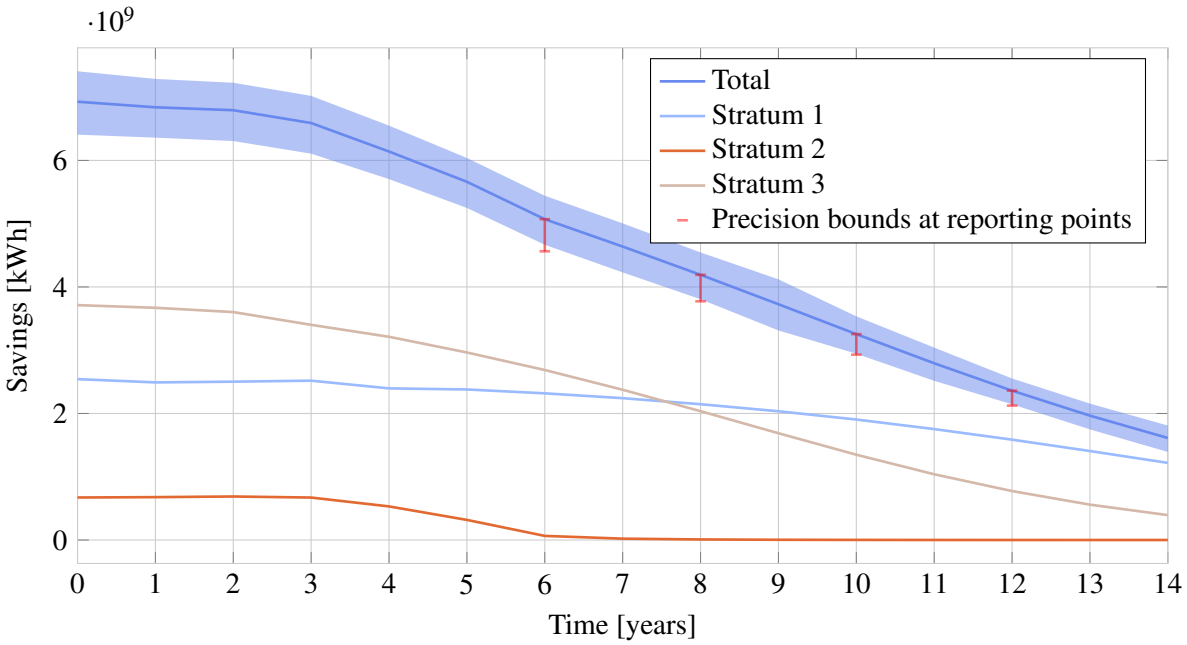


Figure 6.5. Savings inferred by DGLM using (6.50) for Case Study 2.

Also, note that no samples of stratum two are taken after year 5 (the last past-sampling year). In the simple random sampling case (Section 6.4), the smaller the surviving population proportion, the more stringent the sampling requirement. However, this is because the total savings are also small. In year six the total savings are still relatively large due to the other two populations. It is therefore unnecessary to determine stratum 2's (small) contribution accurately.

6.6 CONCLUSION

DGLMs with Bayesian forecasting provide an advantage over traditional regression approaches for longitudinal measurement and verification study designs. This is because they incorporate information about past sample sizes. The GA and MTMC combination allows for efficient sampling design, based not only on the population sampling distribution but also on other uncertainties in the savings calculation, such as hours of use and luminaire power consumption. The flexibility of this approach allows for both simple random sampling, as well as stratified sampling designs to be devised. Sampling cost savings are in the order of 17-49%, depending on stratification and how the study costs are calculated.

CHAPTER 7 COMBINED METERING AND POPULATION SURVIVAL SURVEY SAMPLING

7.1 CHAPTER OVERVIEW

This short chapter demonstrates the combination of metering and survey sampling into a comprehensive M&V planning model. After introducing the structure of the combined model and the moment equations for the combined uncertainty distribution, two case studies are presented. The first considers a simple random sampling model for a homogeneous population. The second considers a stratified model where three different groups (by lamp type or usage) are surveyed and metered.[1] This case study is very similar to previous stratified longitudinal M&V case studies [99], but obtains hours of use from published sources (see Section 6.4), and includes population survival survey sampling from Chapter 6, which is novel.

7.2 INTRODUCTION

Instead of combining the survey result uncertainty from Chapter 6 with estimates for energy consumption (hours of use and power consumption), it will be combined with more accurate meter sampling results from Chapter 5.

[1] This chapter is based on a journal article written by the author as part of his PhD research, published in *Energy and Buildings* [29].

This chapter can be described in M&V terms as follows:

M&V measurement option: Retrofit isolation with all parameter measurement (measuring energy, and measuring population survival).

Project boundary: The lighting circuit(s) under investigation.

Baseline and baseline adjustment approach: The baseline is assumed from the metered data, assuming a constant energy consumption difference between the retrofitted units and the original units. (See (7.1) below).

Savings determination approach: Standard energy efficiency savings (as opposed to normalised savings) is assumed.

For this case, metering and survey sample sizes need to be traded off against one another to ensure adherence to the overall uncertainty reporting bounds, at low cost. Note that measurement, sampling, and modelling uncertainty are considered simultaneously in this model. A diagram illustrating how the various components discussed so far fit into the overall plan is shown in Figure 7.1. This is different to previous combined sampling designs (Figure 5.1), where only meter sampling was optimized, assuming that population decay was known with certainty and with no adaptive population decay model considered.

The vector of the saved energy distributions in this combined model may be calculated by element-wise multiplication of vectors as

$$\hat{\mathbf{E}}_{saved} \sim \hat{\mathbf{\Phi}} \cdot \mathbf{n} \cdot \Delta\hat{\mathbf{E}}, \tag{7.1}$$

where $\Delta\hat{\mathbf{E}}$ is the difference in annual energy use between an original and a retrofitted luminaire. The power difference between these luminaires can be taken from the product specification, but G14 [17] recommends that this difference be measured in situ. A simple measurement may therefore be done in the retrofitting year by measuring the pre- and post-retrofit energy use on the lighting circuit. Let P_b be the baseline lamp power draw, P_r the retrofitted lamp power draw, and s_b and s_r their respective standard deviations. Assuming that there is a measurement error in the meter of 2.52% as described in Section 5.4.1.2, the uncertainty distribution on the ratio of the power draws P_b/P_r can be described by the distribution

$$P_b/P_r \sim N\left[\frac{P_b}{P_r}, z\frac{P_b}{P_r}\sqrt{\left(\frac{s_r}{P_r}\right)^2 + \left(\frac{s_b}{P_b}\right)^2}\right], \tag{7.2}$$

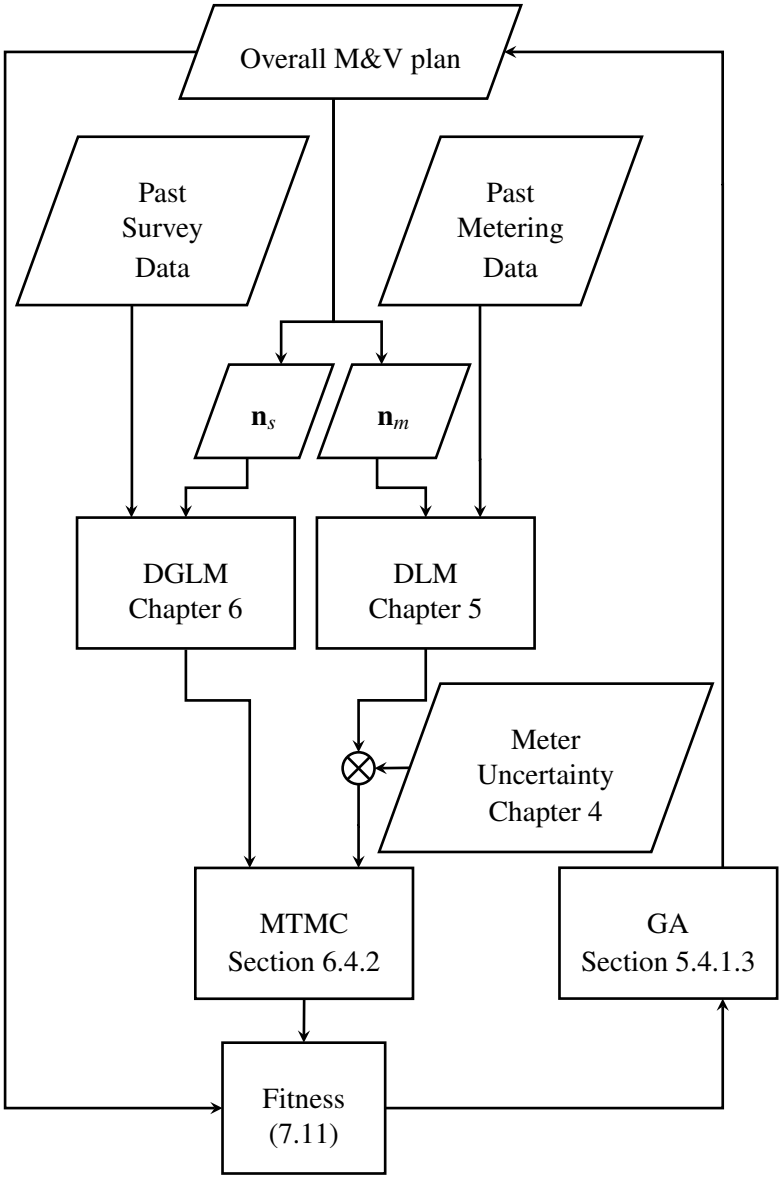


Figure 7.1. Flow diagram illustrating proposed method for combining metering and surveying data. The metering plan is denoted $\mathbf{n}_m$, and the sampling plan $\mathbf{n}_s$.

as per the ASHRAE's guideline RA96 [113]. The annual energy saving per luminaire given this ratio can then be expressed as

$$\Delta\hat{\mathbf{E}} \sim \hat{\mathbf{E}}_r(P_b/P_r - 1). \tag{7.3}$$

The MTMC method of Section 6.4.2 can then be used to calculate the first four moments of E_{saved} in (7.1) and (7.3), which can be rewritten as:

$$\hat{\mathbf{E}}_{saved} \sim n_{retrofitted} \cdot Beta[b_{1a}, b_{1b}, lb_1, ub_1] \cdot N[m_2, s_2] \cdot N[m_3, s_3]. \tag{7.4}$$

The first four moments of $\hat{\mathbf{E}}_{saved}$ can be calculated as

$$E[y_1] = \frac{n_{retrofitted} b_{1a} m_2 (m_3 - 1)(lb_1 + ub_1)}{(b_{1a} + b_{1b})} \tag{7.5}$$

$$E[y_2] = \frac{n_{retrofitted}^2 b_{1a} (1 + b_{1a}) (lb_1 + ub_1)^2 \left(m_2^2 + s_2^2\right) \left(1 - 2m_3 + m_3^2 + s_3^2\right)}{(b_{1a} + b_{1b})(1 + b_{1a} + b_{1b})} \tag{7.6}$$

$$E[y_3] = \frac{n_{retrofitted}^3 b_{1a} 1 \left(2 + 3b_{1a} + b_{1a}^2\right) (m_3 - 1)(lb_1 + ub_1)^3 \left(m_2^3 + 3m_2 s_2^2\right) \left(1 - 2m_3 + m_3^2 + 3s_3^2\right)}{(b_{1a} + b_{1b})(1 + b_{1a} + b_{1b})(2 + b_{1a} + b_{1b})} \tag{7.7}$$

$$E[y_4] = \frac{AB}{(b_{1a} + b_{1b})(1 + b_{1a} + b_{1b})(2 + b_{1a} + b_{1b})(3 + b_{1a} + b_{1b})} \tag{7.8}$$

where

$$A = n_{retrofitted}^4 b_{1a} \left(6 + 11b_{1a} + 6b_{1a}^2 + b_{1a}^3\right) (lb_1 + ub_1)^4 \left(m_2^4 + 6m_2^2 s_2^2 + 3s_2^4\right) \tag{7.9}$$

and

$$B = 1 - 4m_3^3 + m_3^4 + 6s_3^2 + 3s_3^4 + 6m_3^2 \left(1 + s_3^2\right) - 4m_3 \left(1 + 3s_3^2\right). \tag{7.10}$$

These are used with (6.59)-(6.62) as inputs to the Johnson distribution as in (6.63), which will describe the overall probability distribution on the savings estimate for a specific point in time.

The fitness function (6.44) is modified to include the survey cost term. Let v be the survey initiation cost ($v = 1000$), and w_s the cost per survey sample ($w_s = 10$). Also let $d_t = 1$ for years in which surveying is done, and $d_t = 0$ otherwise. Then the fitness equation is modified to

$$\min \sum_{t=1}^{N} n_{m,\,t} w_m + \sum_{t=1}^{N} n_{s,\,t} w_s + d_t v + r(\mathbf{n}). \tag{7.11}$$

The penalty function is also modified accordingly:

$$r(\mathbf{n}) = \sum_{t \in \mathbf{M}} \left(10^5 (w_s + w_m)(e_t - \varepsilon) + 10^8 + 5(w_m n_{m,\,benchmark,\,t} + w_s n_{s,\,benchmark,\,t} \right) \forall\, t \in \chi. \tag{7.12}$$

In this case, the relative cost of surveying and metering play a significant role in determining an optimal solution, since the GA will trade these sources of uncertainty off against one another. The parameters used for this GA are the same as those listed in Table 5.1. Since these costs are project-specific, the result from any single study is not normative but may illuminate the characteristics of the method and the kinds of results that can be expected. Two cases will be considered below. The first is a simple random sampling case: monitoring a single population of retrofitted lamps over multiple years. The survey and cross-sectional metering sample sizes are then optimized simultaneously to minimise cost while still adhering to the required reporting precision levels. In the second case, the study is expanded so that three distinct sub-populations of lamps are monitored over multiple years to achieve the same objective. This is a combined stratified sampling design.

7.3 CASE STUDY 1: SIMPLE RANDOM SAMPLING DESIGN

The first case considers a single population of retrofitted lamps tracked over a number of years. The lamp population is assumed to decay according to the PELP data points [28, 169].

It is assumed that three years' data has been collected ($\mathbf{D}_{0-2}$), and that reporting is to be done annually for $\mathbf{M} = \{4,5,6,7\}$. In the project, 100 000 CFLs of 11W each replace their 60W incandescent counterparts, and the savings need to be determined. Past meter samples were $\mathbf{n}_{m,\,0-2} = [68,68,68]$, and past survey samples were $\mathbf{n}_{s,\,0-2} = [250,250,100]$.

7.3.1 Benchmark

The combined benchmark is calculated using a GA with the combination of the survey sampling and energy metering uncertainty determined as in (7.1), where the uncertainty in E_r in (7.3) is calculated according to the standard sampling formula of (4.10) combined with the meter measurement error. As in Chapter 6 the survey sampling benchmark was selected as a Jeffreys interval on the proportion [311].

The benchmark is, therefore, an optimal sampling plan in which prior data are not taken into account through the Bayesian method.

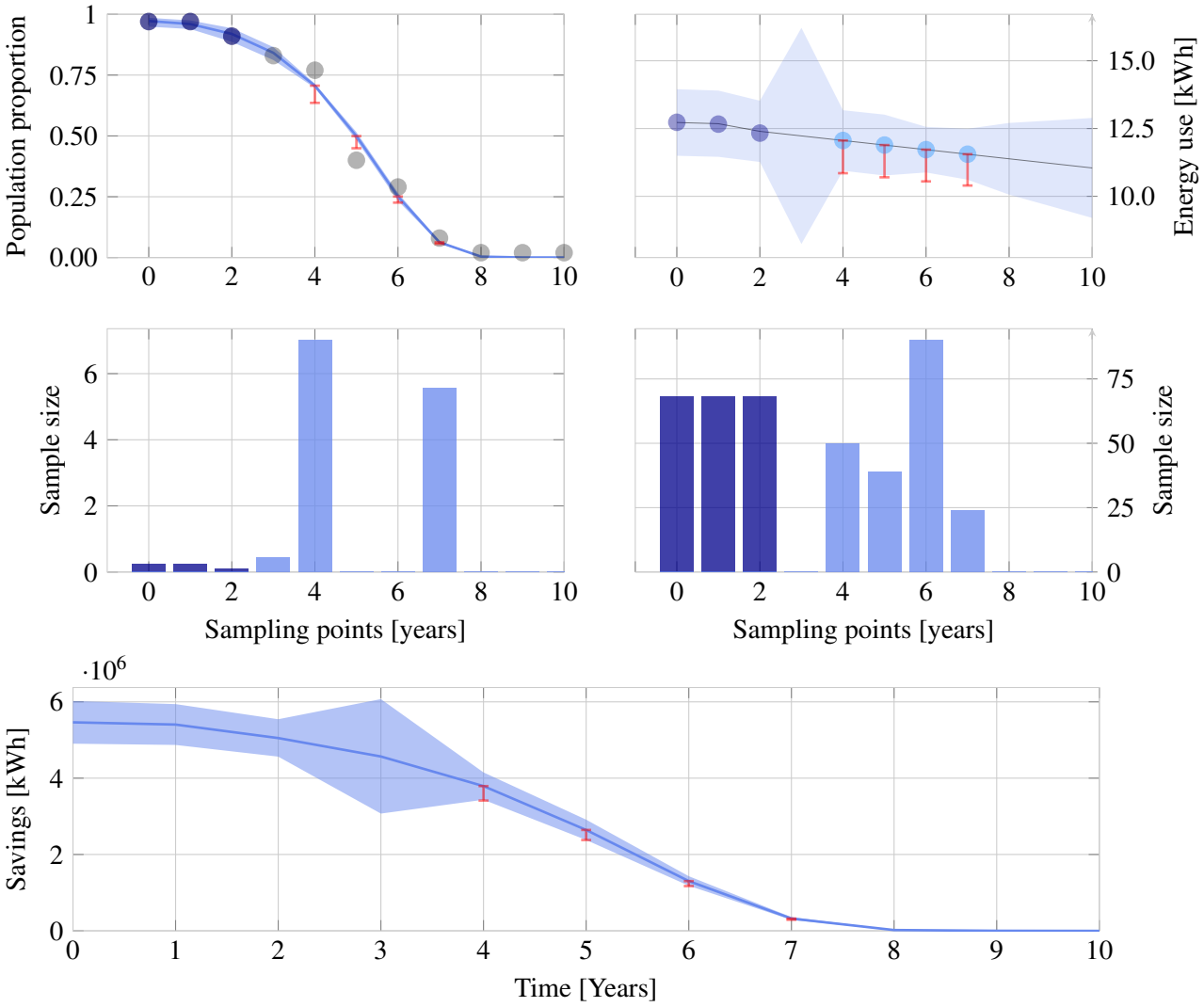


Figure 7.2. Plot of combined survey sampling (top left) and metering (top right) for a single population (Case Study 2), with the combined savings estimate over time at the bottom. Dark blue indicates past samples, and light blue indicates planned future samples.

7.3.2 Results and discussion

An efficient sampling plan is listed in Table 7.1, and has a cost of R772 240. A benchmark sampling plan is listed in Table 7.2, and has a cost of R1 128 940. The Bayesian method therefore achieves a saving of 40.13% for these cases.

The results for this scenario are shown in Figure 7.2. The top four graphs show the individual metering

Table 7.1. Combined sampling plan for Case Study 2. Years beyond seven are not shown since no reporting was required, and no samples were taken.

Years	**3**	**4**	**5**	**6**	**7**
Survey	3448	7008	0	0	5568
Meters	0	50	39	90	24

Table 7.2. Benchmark of the combined sampling plan for Case Study 2. Years beyond seven are not shown since no reporting was required, and no samples were taken.

Years	**3**	**4**	**5**	**6**	**7**
Survey	0	1189	3730	12842	7633
Meters	84	74	92	94	0

and survey sampling plans and results, with the bottom graph combining these results into an overall savings estimate.

No reporting was deliberately specified for $t = 3$, to force the algorithm to forecast for that year. The increase in uncertainty is evident.

As would be expected, the algorithm favours oversampling on the survey side to compensate for the metering cost. Under present assumptions, three hundred survey samples can be taken for the cost of a single meter. However, metering cannot be completely neglected. Furthermore, the additional information contained in a sample decreases with the square root of the sample size. This means that to double the amount of information available from a sample of size n, a sample of size $4n$ $(2\sqrt{n})$ will be needed. The principle of diminishing returns, therefore, applies to large survey sample sizes traded off against small metering samples. Although an additional meter may be more expensive, its relative contribution to uncertainty reduction is greater than the additional three hundred survey samples would be.

The DLM-DGLM shows a clear advantage over existing methods. Smaller sample sizes than existing sampling methods such as (4.10) are needed.

7.4 CASE STUDY 2: COMBINED STRATIFIED SAMPLING DESIGN

To demonstrate the scalability of the method, a stratified sampling design is considered. As before, both survey sampling and meter placement are considered simultaneously over a number of years. However, instead of considering a project with a single population, a project with three different sub-populations is considered. In stratum one, 50 000 incandescent lamps of 60W each, burning for 3.11 hours per day, are replaced by 11W CFLs. In stratum two, 20 000 incandescent lamps of 60W each, that burn for two hours per day, are replaced by 11W CFLs. In stratum three, 30 000 incandescent lamps of 100W each, that burn for 4.11 hours per day are replaced by 14W CFLs. To provide realistic population survival curves, three curves from the LRC data on CFLs are used [28,170]. Curves with short, medium, and long lives were selected. Data points $\mathbf{D}$ were then randomly generated as $\mathbf{D}_{sim,\,t} \sim Binomial[n = n_{s,\,t}, p = \Phi_{t,\,sim}]$, so that large sample size results have less random scatter than small sample size results. It was assumed that meter placement and surveying costs were constant across the strata, although this could easily be changed if there were a reason to do so. The method is unaltered from the simple random sampling case, except for minor changes in the fitness function to sum all three strata regarding cost and uncertainty.

Five years of sampling are assumed to have been conducted in the past. Meter sample sizes were $\mathbf{n}_{m,\,0-4} = [50,50,40,30,20,10]$ for each stratum. Survey sampling was conducted based on the decay rates of the individual populations. For stratum one, the sample sizes were $\mathbf{n}_{s,\,0-4} = [100,100,100,100,200]$. For strata two and three, the sample sizes were $\mathbf{n}_{s,\,0-4} = [50,50,75,75,100,150]$. The reason that the sample sizes increase for the survey sampling is that it is critical to identify the point at which the population curve changes from the plateau to the transition phase. Small sample sizes during these years add disproportionate noise which leads to inaccurate forecasts.

Expansion of the MTMC equations to three strata proceed in the same manner as (6.67)-(6.70).

7.4.1 Benchmark

Wherever possible, stratified sampling designs are preferable to simple random sampling designs, because the intra-stratum variance is homogenised, leading to smaller sample sizes [40]. Stratified

Table 7.3. Stratified survey sampling plans for Case Study 3. Benchmark (top), Efficient (bottom)

Years	6	7	8	9	10	11	12
Stratum 1	886	45	923	132	440	0	849
Stratum 2	945	60	872	0	284	0	447
Stratum 3	783	11	189	23	363	0	183
Stratum 1	780	0	291	0	553	0	848
Stratum 2	692	0	238	0	141	0	100
Stratum 3	403	0	799	0	259	0	97

designs should, therefore, be benchmarked against other stratified designs. The most efficient stratified sampling design for normally distributed strata with unequal variances is the 'Neyman allocation'. If different costs are incurred for different strata, the cost-weighted Neyman allocation should be used. These methods cannot capture the complexities of the case at hand, however. To provide a robust benchmark, we expand the method described in Section 7.3.1 to the stratified case. In effect, a GA is used to devise a stratified sampling design with all the complexity of the proposed method, except for the Bayesian forecasting and dynamic model components.

7.4.2 Results and discussion

One efficient sampling result is shown in Table 7.3 and Table 7.4 at a cost of R1 417 010. The benchmark is R1 918 350, representing a 26.55% saving. It is evident that the algorithm favours placing meters and doing surveys in strata where many lamps are left, as these have the highest contribution to overall energy use. In other respects, the result is similar to the simple random sampling case. The survey component is oversampled to offset the high cost of metering.

The result shows the scalability of the method to multiple strata, as well as the advantage of doing so. By stratifying the population, smaller sample sizes are needed. The Neyman allocation method recommended by M&V guidelines [40] is efficient and accurate, provided that only simple stratified designs be attempted. However, the method proposed in this paper is more flexible and allows for intricate, real-world stratified designs needed for most M&V projects.

Table 7.4. Stratified meter sampling plans for Case Study 3. Benchmark (top), Efficient (bottom)

Years	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Stratum 1	49	0	108	0	146	0	159
Stratum 2	11	0	0	0	0	0	0
Stratum 3	26	0	61	0	22	0	26
Stratum 1	40	0	62	0	116	0	109
Stratum 2	0	0	0	0	0	0	0
Stratum 3	35	0	41	0	30	18	0

7.5 CONCLUSION

The DLM in combination with a DGLM can be used to model metering and surveying simultaneously, and is shown to reduce overall M&V project costs by almost 40% for the simple random sampling case, while still adhering to the 90/10 reporting uncertainty requirement. This figure depends on the cost profile of the specific project, however. The method is then expanded to a stratified sampling case with three metered and surveyed sub-populations, for which sampling and metering costs are reduced by 26.6%.

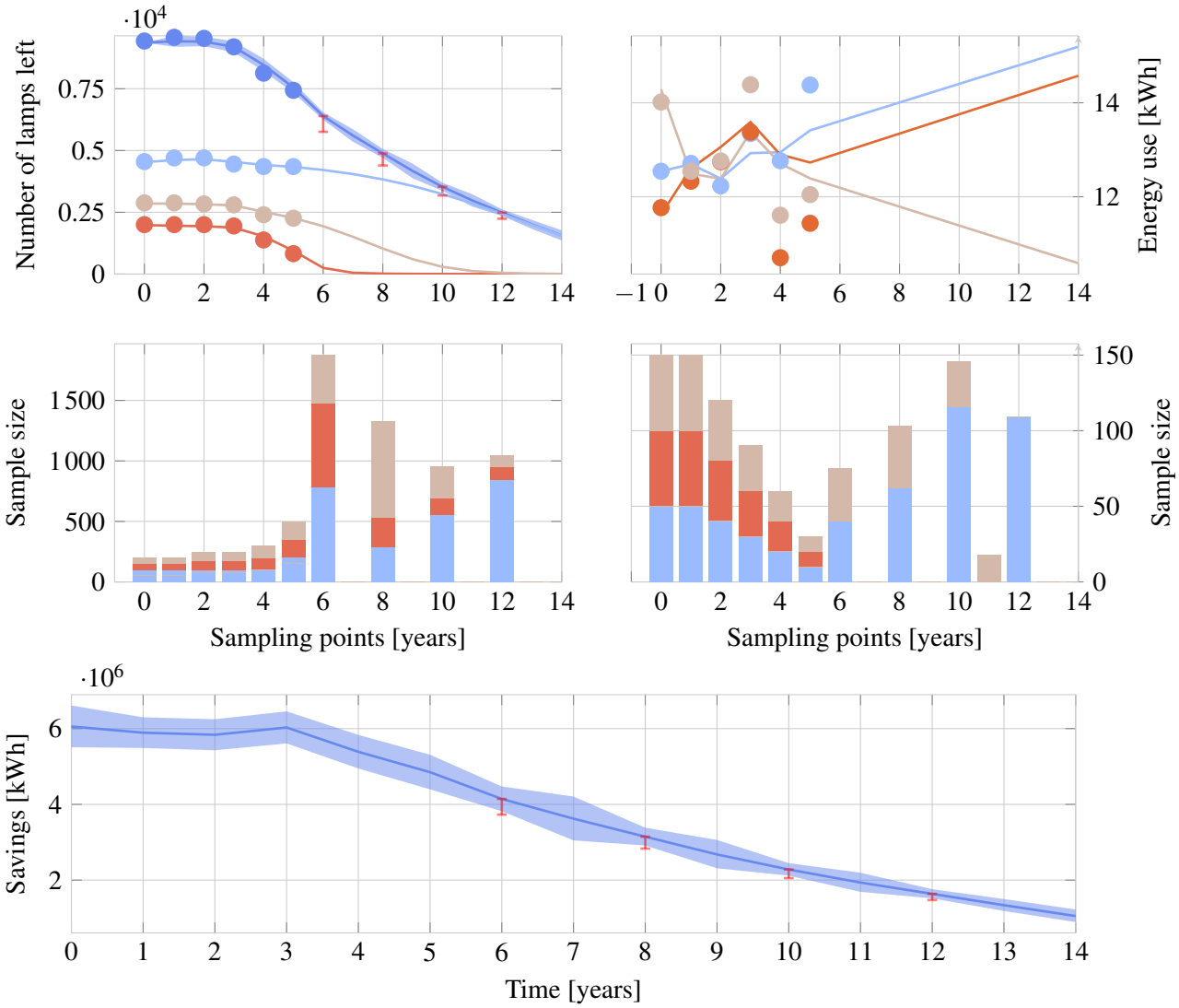


Figure 7.3. Efficient combined sampling plan using the DLM and DGLM for one random model realisation (Case Study 3). Stratum 1 is in blue, stratum 2 in red, and stratum 3 in brown. Combined values are shown in blue. The 10% error bars are shown in red, and the 90% confidence intervals in light blue.

CHAPTER 8 CONCLUSION

The aim of this research is to illustrate the application of Bayesian statistics to energy M&V. This is done by addressing the three main uncertainty drivers in M&V: measurement, sampling and modelling. The Bayesian paradigm as explained in Chapter 3 describes uncertainty as a state of knowledge and treats parameters of interest as random variables with probability density functions. It also allows for the transparent use of prior knowledge when evaluating measurement results. Although such priors could be subjective, this thesis uses results from the SIMEX algorithm for mismeasurement correction, and prior sampling results for future sampling planning.

8.1 MEASUREMENT UNCERTAINTY

It is shown that energy metering uncertainty makes a relatively small contribution to overall M&V uncertainty for cases where sampling is done (Section 4.2), although practitioners should be careful to dismiss it out-of-hand, as shown in Figure 5.9. Rather, careful M&V has the potential to reduce metering uncertainty costs by allocating financial resources to its mitigation only when warranted. Chapter 4 proposes one method of doing so, by using low-cost calibrators and more advanced mathematical calibration techniques, rather than costly advanced meters and simpler mathematics.

For most M&V sampling projects, the conclusion is that more information can be gained by installing a large number lower accuracy meters, rather than a small number of high-accuracy meters.

It is also found that the errors-in-variables effect may be significant in M&V, as discussed in Section 2.3.2. The SIMEX procedure with Bayesian refinement was used to mitigate the effect on energy measurement itself, removing attenuation bias and improving parameter estimation. However, in M&V

the errors-in-variables effect will usually be a factor in the measurement of independent variables such as temperature and occupancy.

8.2 SAMPLING UNCERTAINTY

The second aspect addressed in this thesis is sampling uncertainty. Current methods based on frequentist assumptions do have particular problems that could be solved by Bayesian methods (See Chapter 3). Previously proposed methods also have certain disadvantages (cf. Section 2.2.2.3) which are improved upon through the DLM and Bayesian forecasting proposed in Chapter 5. During the optimization phase, it was shown that Monte Carlo simulation is not reliable inside a heuristic such as the GA (Section 6.4.2). The MTMC method provides a more computationally efficient and stable alternative.

The DLM and DGLM in Chapters 5 and 6 were found to have superior uncertainty quantification capabilities when compared to previous methods, and also address the modelling uncertainty aspect of M&V for the cases to which they were applied. When these were combined with a GA, M&V monitoring savings in the order of 17-66% were achieved compared to the benchmarks. However, the robustness of such plans is of concern. Efficient sampling plans are based on the forecasts being perfectly accurate. When this is not the case, an efficient plan may have inadequate statistical precision.

8.3 RECOMMENDATIONS FOR M&V PRACTICE

1. M&V studies designed using the standard sampling formulae are usually underpowered, casting some doubt on the validity of results with small sample sizes. A greater focus on robustness will go some way in solving this problem.

2. The Mellin Transform Moment Calculation method deserves greater use in M&V, as does the Johnson distribution. The implementation of these methods in open-source data analysis software such as Python has the potential to increase the quality of M&V results significantly.

3. The use of power meters measuring fundamental quantities or using the IEEE 1459 definition of power is recommended for M&V.

4. Reporting precision (such as 90/10) is often required for annual savings. However, if this requirement is changed to the precision of the savings to date for longitudinal projects, monitoring costs would be decreased significantly.

8.4 RECOMMENDATIONS FOR FURTHER RESEARCH

1. This thesis applies Bayesian statistics to a few problems in M&V, but many more remain. Probably the most significant opportunity is for the application of hierarchical modelling to complex M&V projects, as alluded to in Section 3.5.1 and done in Booth *et al.* [274].

2. Low-cost metering remains a promising field for M&V. Metering is becoming cheaper and smarter, especially in developed countries. However, to implement it at scale for M&V in developing countries presents a significant opportunity. In this regard virtual instrumentation as in Section 2.2.3.6 shows promise and should be investigated further.

3. Further research on M&V sampling planning should focus on efficiency in the context of robustness.

4. It would be useful to develop methods for meter cross-calibration in a smart grid.

5. Survival Analysis is a powerful tool for persistence research. Much of this work has already been done in other fields, and applying it to M&V should be simple and effective.

6. The DLM and DGLM used in this thesis are relatively simple. They can be expanded to include periodicity or seasonality such as daily or weekly load profiles, thereby increasing the resolution at which time-series modelling can be done. They can also be expanded to include covariates such as temperature and occupancy. Fully-fledged DLMs have the potential to be powerful M&V tools

REFERENCES

[1] Efficiency Valuation Organization, *International Performance Measurement and Verification Protocol Vol. 1*, January 2012.

[2] *Taxation Laws Amendment Act No. 25 of 2015*. South African Government Gazette no. 39588, 2015.

[3] *SANS 50010:2011: Measurement and Verification of Energy Savings*, Standards Division, South African Bureau of Standards Std., Rev. 1.

[4] United Nations Framework Convention for Climate Change, *Clean Development Mechanism Methodology Booklet*, November 2015.

[5] M. S. Khawaja and J. Stewart, "Long-run savings and cost-effectiveness of home energy report programs," Cadmus, Tech. Rep., Winter 2015.

[6] R. Baumgartner, *The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures*. National Renewable Energy Laboratory, January 2012 - March 2013 2013, ch. 12: Survey Design and Implementation Cross-Cutting Protocols for Estimating Gross Savings.

[7] A. Conant, H. Liu, J. Proctor, and B. Wilcox, "Sources of error in home energy use calculations: Evaluation in real-world laboratory homes," in *Proceedings of the International Energy Program Evaluation Conference*. Long Beach, California: International Energy Programme Evaluation Conference, August 2015.

[8] Q. Wang, G. Augenbroe, J.-H. Kim, and L. Gu, "Meta-modeling of occupancy variables and analysis of their impact on energy outcomes of office buildings," *Applied Energy*, vol. 174, pp. 166–180, 2016.

[9] S. Lanzisera, S. Dawson-Haggerty, H. I. Cheung, J. Taneja, D. Culler, and R. Brown, "Methods for detailed energy data collection of miscellaneous and electronic loads in a commercial office building," Lawrence Berkeley National Laboratory, Berkeley, California, Tech. Rep. LBNL-6384E, 2014.

[10] A. Kamilaris, B. Kalluri, S. Kondepudi, and T. K. Wai, "A literature survey on measuring energy usage for miscellaneous electric loads in offices and commercial buildings," *Renewable and Sustainable Energy Reviews*, vol. 34, pp. 536–550, 2014.

[11] R. Ward, R. Choudhary, Y. Heo, and A. Rysanek, "Exploring the impact of different parameterisations of occupant-related internal loads in building energy simulation," *Energy and Buildings*, vol. 123, pp. 92–105, 2016.

[12] R. D. Wilkinson, *Large-Scale Inverse Problems and Quantification of Uncertainty*. Chichester, UK: John Wiley & Sons Ltd., 2010, ch. Bayesian Calibration of Expensive Multivariate Computer Systems.

[13] D. Coakley, P. Raftery, and M. Keane, "A review of methods to match building energy simulation models to measured data," *Renewable and Sustainable Energy Reviews*, vol. 37, pp. 123–141, 2014.

[14] M. C. Kennedy and A. O'Hagan, "Bayesian calibration of computer models," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 63, no. Part 3, pp. 425–646, 2001.

[15] A. Baker, "Sample size selection in energy efficiency research and evaluation - the use and abuse of the coefficient of variation," in *International Energy Program Evaluation Conference*. Rome, Italy: Research Into Action, June 2012.

[16] Y. Dodge, Ed., *The Oxford Dictionary of Statistical Terms*. Oxford, 2010.

[17] American Society of Heating, Refrigeration and Air-Conditioning Engineers, Inc., *Guideline 14-2014, Measurement of Energy, Demand, and Water Savings*, December 2014.

[18] K. Birch, "Measurement good practice guide no. 36: Estimated uncertainties in testing," British Measurement and Testing Association, Teddington, Middlesex, United Kingdom, Tech. Rep., March 2003.

[19] J. Kruschke, *Doing Bayesian Data Analysis: a Tutorial with R, JAGS, and Stan*, 2nd ed. Academic Press, 2015.

[20] J. Neyman, "Outline of a theory of statistical estimation based on the classical theory of probability," *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, vol. 236, no. 767, pp. 333–380, 1937.

[21] R. Kacker and A. Jones, "On use of Bayesian statistics to make the Guide to the expression of Uncertainty in Measurement consistent," *Metrologia*, vol. 40, no. 5, p. 235, 2003.

[22] *IEC 62053-21: Electricity metering equipment (a.c.) Ű Particular requirements Ű Part 21: Static meters for active energy (classes 1 and 2)*, International Electrotechnical Commission Std.

[23] *IEC 62053-22 Electricity metering equipment (a.c.) - Particular requirements -Part 22: Static meters for active energy (classes 0,2 S and 0,5 S)*, International Electrotechnical Commission Std.

[24] *IEC 62053-23:2003 Electricity metering equipment (a.c.) - Particular requirements - Part 23: Static meters for reactive energy (classes 2 and 3)*, International Electrotechnical Commission Std.

[25] *IEC 60044-8: Instrument transformers Part 8: Electronic current transformers*, International Electrotechnical Commission Std.

[26] H. Carstens, X. Xia, and S. Yadavalli, "Measurement uncertainty in energy monitoring: Present state of the art," August 2016, in Press.

[27] ——, "Low-cost energy meter calibration method for measurement and verification," *Applied Energy*, vol. 188, pp. 563–575, 2017.

[28] ——, "Efficient longitudinal population survival survey sampling for the measurement and verification of building retrofit projects," *Energy and Buildings*, vol. 150, pp. 163–176.

[29] ——, "Efficient metering and surveying sampling designs in longitudinal measurement and verification for lighting retrofit," *Energy and Buildings*, vol. 154, pp. 430–447.

[30] *Uniform Methods Project*, National Renewable Energy Laboratory. [Online]. Available: http://energy.gov/eere/about-us/ump-home

[31] American Society of Heating, Refrigeration and Air-Conditioning Engineers, Inc., *Guideline 14-2002, Measurement of Energy and Demand Savings*, June 2002.

[32] J. Haberl, C. Culp, and D. Claridge, "Ashrae's guideline 14-2002 for measurement of energy and demand savings: How to determine what was really saved by a retrofit," in *Proceedings of the Fifth International Conference for Enhanced Building Operations*, no. ESL-IC-05-10-50. Energy Systems Laboratory, October 2005.

[33] T. Reddy and D. Claridge, "Uncertainty of "measured" energy savings from statistical baseline models," *HVAC&R Research*, vol. 6, no. 1, pp. 3–20, 2000.

[34] *Guidelines for Verifying Savings from Commissioning Existing Buildings*, California Commissioning Collaborative, 2012.

[35] J. Granderson, S. Touzani, C. Custodio, M. D. Sohn, D. Jump, and S. Fernandes, "Accuracy of automated measurement and verification (M&V) techniques for energy savings in commercial buildings," *Applied Energy*, vol. 173, pp. 296–308, 2016.

[36] Phillip, "Review of prior commercial building energy efficiency retrofit evaluation: A report to Snohomish public utility district," Lawrence Berkeley National Laboratory, Tech. Rep. LBNL-6874E, April 2015.

[37] D. Gowans, *The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures*. National Renewable Energy Laboratory, April 2013 2013, ch. 2: Commercial and Industrial Lighting Evaluation Protocol.

[38] S. Dimetrosky, *The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures*. National Renewable Energy Laboratory, April 2013, ch. 6: Residential Lighting Evaluation Protocol.

[39] D. Mort, *The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures*. National Renewable Energy Laboratory, April 2013, ch. 9: Metering Cross-Cutting Protocols.

[40] M. S. Khawaja, J. Rushton, and J. Keeling, *The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures*. National Renewable Energy Laboratory, January 2012 - March 2013 2013, ch. 11: Sample Design Cross-Cutting Protocols.

[41] D. Violette, *The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures*. National Renewable Energy Laboratory, January 2012 - March 2013 2013, ch. 13: Assessing Persistence and Other Evaluation Issues in Cross-Cutting Protocols.

[42] *International Performance Measurement and Verification Protocol: Statistics and Uncertainty for IPMVP*, Efficiency Valuation Organization, June 2014.

[43] State and Local Energy Efficiency Action Network, *Energy Efficiency Program Impact Evaluation Guide*, US Department of Energy, December 2012, DOE/EE-0829.

[44] *M&V Guidelines: Measurement and Verification for Federal Energy Projects*, 2nd ed., U.S. Department of Energy Office of Energy Efficiency and Renewable Energy,

Federal Energy Management Program (FEMP), September 2000. [Online]. Available: http://www1.eere.energy.gov/femp/pdfs/mv_guidelines.pdf

[45] E. Vine, G. Kats, J. Sathaye, and H. Joshi, "International greenhouse gas trading programs: a discussion of measurement and accounting issues," *Energy Policy*, vol. 31, no. 3, pp. 211–224, 2003.

[46] R. Sonnenblick and J. Eto, "A framework for improving the cost-effectiveness of DSM program evaluations," Lawrence Berkeley Laboratory, Berkeley, California, Tech. Rep. LBL37158, 1995.

[47] I. Shishlov and V. Belassen, "Review of monitoring uncertainty requirements in the CDM," CDC Climat Research, Tech. Rep. Working Paper No. 2014-16, October 2014.

[48] A. Michaelowa, D. Hayashi, and M. Marr, "Challenges for energy efficiency improvement under the CDM - the case of energy-efficient lighting," *Energy Efficiency*, vol. 2, no. 4, pp. 353–367, 2009.

[49] *Approved Small-Scale Methodology AMS II.C, Demand-Side Activities for Specific Technologies*, United Nations Framework Convention for Climate Change.

[50] *Approved Small-Scale Methodology AMS II.J, Demand-Side Activities for Efficient Lighting Technologies*, United Nations Framework Convention for Climate Change.

[51] X. Ye, X. Xia, and J. Zhang, "Optimal sampling plan for clean development mechanism lighting projects with lamp population decay," *Applied Energy*, vol. 136, pp. 1184–1192, 2014.

[52] X. Ye and X. Xia, "Optimal metering plan for measurement and verification on a lighting case study," *Energy*, vol. 95, pp. 580–592, 2016.

[53] H. Carstens, X. Xia, and X. Ye, "Improvements to longitudinal Clean Development Mechanism sampling designs for lighting retrofit projects," *Applied Energy*, no. 126, pp. 256–265, May 2014.

[54] *Concept note: Uncertainty of measurements in large-scale CDM methodologies*, no. CDM-EB73-AA-A04. UNFCCC CDM, May 2013.

[55] D. Violette and P. Rathbun, *The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures*. National Renewable Energy Laboratory, 2014, ch. 23 Estimating Net Savings: Common Practices.

[56] X. Xia and J. Zhang, "Mathematical description for the measurement and verification of energy efficiency improvement," *Applied Energy*, vol. 111, pp. 247–256, 2013.

[57] F. Crenna, G. B. Rossi, and L. Bovio, "Probabilistic measurement evaluation for the implementation of the measuring instrument directive," *Measurement*, vol. 42, no. 10, pp. 1522–1531, 2009.

[58] L. Pendrill and H. Källgren, "Exhaust gas analysers and optimised sampling, uncertainties and costs," *Accreditation and Quality Assurance*, vol. 11, no. 10, pp. 496–505, 2006.

[59] L. Pendrill, "Optimised measurement uncertainty and decision-making when sampling by variables or by attribute," *Measurement*, vol. 39, pp. 829–840, 2006.

[60] T. Fearn, S. A. Fisher, M. Thompson, and S. L. Ellison, "A decision theory approach to fitness for purpose in analytical measurement," *Analyst*, vol. 127, no. 6, pp. 818–824, 2002.

[61] A. Rysanek and R. Choudhary, "Optimum building energy retrofits under technical and economic uncertainty," *Energy and Buildings*, vol. 57, pp. 324–337, 2013.

[62] E. Mills, H. Friedman, T. Powell, N. Bourassa, D. Claridge, T. Haasl, and M. A. Piette, "The cost-effectiveness of commercial-buildings commissioning," Lawrence Berkeley National Laboratory, Tech. Rep. LBNL-56637 (Rev.), December 2004.

[63] J. Friege and E. Chappin, "Modelling decisions on energy-efficient renovations: A review," *Renewable and Sustainable Energy Reviews*, vol. 39, pp. 196–208, 2014.

[64] E. Mills, S. Kromer, G. Weiss, and P. A. Mathew, "From volatility to value: analysing and managing financial and performance risk in energy savings projects," *Energy Policy*, vol. 34, no. 2, pp. 188–199, 2006.

[65] J. L. Mathieu, D. S. Callaway, and S. Kiliccote, "Variability in automated responses of commercial buildings and industrial facilities to dynamic electricity prices," *Energy and Buildings*, vol. 43, no. 12, pp. 3322–3330, 2011.

[66] P. Lee, P. Lam, and W. Lee, "Risks in Energy Performance Contracting (EPC) projects," *Energy and Buildings*, vol. 92, pp. 116–127, 2015.

[67] D. Violette, "Impact evaluation accuracy and the incorporation of prior information," in *Energy Program Evaluation Conference*, Chicago, 1991, pp. 86–92.

[68] D. Violette, R. Brakken, A. Shon, and J. Greer, "Statistically adjusted engineering (sae) estimates: What can the evaluation analyst do about the engineering side of the analysis?" in *International Program Evaluation Conference*, 1993.

[69] M. L. Goldberg, "Reasonable doubts: Monitoring and verification for performance contracting," in *ACEEE Summer Study on Energy Efficiency in Buildings*, vol. 4. Pacific Grove, California: American Council for an Energy Efficient Economy, 1996, pp. 133–143.

[70] L. A. Irwin, "A high accuracy standard for electricity meters," in *2010 IEEE Power and Energy Society Transmission and Distribution Conference and Exposition*, 2010, pp. 1–3.

[71] ——, "White paper: A high accuracy standard for electricity meters," Schneider Electric, Tech. Rep., 2011.

[72] H. Carstens, X. Xia, and S. Yadavalli, "Measurement uncertainty and risk in measurement and verification projects," in *International Energy Programme Evaluation Conference*, Long Beach, California, August 2015.

[73] M. Riddle and R. T. Muehleisen, "A guide to Bayesian calibration of building energy models," in *Building Simulation Conference*, 2014.

[74] Y. Heo, D. J. Graziano, L. Guzowski, and R. T. Muehleisen, "Evaluation of calibration efficacy under different levels of uncertainty," *Journal of Building Performance Simulation*, vol. 8, no. 3, pp. 135–144, 2015.

[75] Q. Wang, B. D. Lee, G. Augenbroe, and C. J. Paredis, "An application of normative decision theory to the valuation of energy efficiency investments under uncertainty," *Automation in Construction*, vol. 73, pp. 78–87, 2017.

[76] R. Kammerud, K. L. Gillespie, and M. Hydeman, "Economic uncertainties in chilled water system design," *ASHRAE Transactions*, vol. 105, no. SE-99-16-3, pp. 1075–1085, 1999.

[77] Y. Heo, "Bayesian calibration of building energy models for energy retrofit decision-making under uncertainty," Ph.D. dissertation, Georgia Institute of Technology, December 2011.

[78] J. Jackson, "Promoting energy efficiency investments with risk management decision tools," *Energy Policy*, vol. 38, pp. 3865 – 3873, 2010.

[79] S. Burhenne, O. Tsvetkova, D. Jacob, G. P. Henze, and A. Wagner, "Uncertainty quantification for combined building performance and cost-benefit analyses," *Building and Environment*, vol. 62, pp. 143–154, 2013.

[80] P. Lee, P. Lam, F. W. Yik, and E. H. Chan, "Probabilistic risk assessment of the energy saving shortfall in energy performance contracting projects–a case study," *Energy and Buildings*, vol. 66, pp. 353–363, 2013.

[81] Q. Deng, X. Jiang, Q. Cui, and L. Zhang, "Strategic design of cost savings guarantee in energy performance contracting under uncertainty," *Applied Energy*, vol. 139, pp. 68–80, 2015.

[82] Q. Deng, X. Jiang, L. Zhang, and Q. Cui, "Making optimal investment decisions for energy service companies under uncertainty. A case study," *Energy*, vol. 88, pp. 234–243, 2015.

[83] T. Reddy, "Literature review on calibration of building energy simulation programs: Uses, problems, procedures, uncertainty, and tools," *ASHRAE Transactions*, vol. 112, no. Part 1, pp. 226–240, 2006.

[84] J. Sun and T. A. Reddy, "Calibration of building energy simulation programs using the analytic optimization approach (rp-1051)," *HVAC&R Research*, vol. 12, no. 1, pp. 177–196, 2006.

[85] T. A. Reddy, I. Maor, and C. Panjapornpon, "Calibrating detailed building energy simulation programs with measured data – Part I: General methodology (RP-1051)," *HVAC&R Research*, vol. 13, no. 2, pp. 221–241, 2007.

[86] ——, "Calibrating detailed building energy simulation programs with measured data – part II: Application to three case study office buildings," *HVAC&R Research*, vol. 13, no. 2, pp. 243–265, 2007.

[87] B. D. Lee, Y. Sun, G. Augenbroe, and C. J. Paredis, "Toward better prediction of building performance: a workbench to analyze uncertainty in building simulation," in *BS2013: 13th Conference of the International Building Performance Simulation Association*, Chambery, France, August 2013.

[88] Y. Sun, Y. Heo, M. Tan, H. Xie, C. J. Wu, and G. Augenbroe, "Uncertainty quantification of microclimate variables in building energy models," *Journal of Building Performance Simulation*, vol. 7, no. 1, pp. 17–32, 2014.

[89] W. Tian, "A review of sensitivity analysis methods in building energy analysis," *Renewable and Sustainable Energy Reviews*, vol. 20, pp. 411–419, 2013.

[90] J. Sanyal, J. New, R. E. Edwards, and L. Parker, "Calibrating building energy models using supercomputer trained machine learning agents," *Concurrency and Computation: Practice and Experience*, vol. 26, no. 13, pp. 2122–2133, 2014.

[91] Y. Heo, R. Choudhary, and G. Augenbroe, "Calibration of building energy models for retrofit analysis under uncertainty," *Energy and Buildings*, vol. 47, pp. 550–560, 2012.

[92] Y. Heo and V. M. Zavala, "Gaussian process modeling for measurement and verification of building energy savings," *Energy and Buildings*, vol. 53, pp. 7–18, 2012.

[93] Y. Heo, G. Augenbroe, and R. Choudhary, "Quantitative risk management for energy retrofit projects," *Journal of Building Performance Simulation*, vol. 6, no. 4, pp. 257–268, 2013.

[94] Y. Heo, G. Augenbroe, D. Graziano, R. T. Muehleisen, and L. Guzowski, "Scalable methodology for large scale building energy improvement: Relevance of calibration in model-based retrofit analysis," *Building and Environment*, vol. 87, pp. 342–350, 2015.

[95] Q. Li, G. Augenbroe, and J. Brown, "Assessment of linear emulators in lightweight Bayesian calibration of dynamic building energy models for parameter estimation and performance prediction," *Energy and Buildings*, vol. 124, pp. 194–202, 2016.

[96] W. Tian, S. Yang, Z. Li, S. Wei, W. Pan, and Y. Li, "Identifying informative energy data in Bayesian calibration of building energy models," *Energy and Buildings*, vol. 119, pp. 363–376, 2016.

[97] Z. Olinga, "A cost effective approach to handle measurement and verification sampling and modelling uncertainties," Master's thesis, University of Pretoria, 2016.

[98] X. Ye, X. Xia, L. Zhang, and B. Zhu, "Optimal maintenance planning for sustainable energy efficiency lighting retrofit projects by a control system approach," *Control Engineering Practice*, vol. 37, pp. 1–10, 2015.

[99] X. Ye, "Optimal measurement and verification plan on lighting," Ph.D. dissertation, University of Pretoria, March 2015.

[100] R. Cox, *The Algebra of Probable Inference*. Baltimore.: Johns Hopkins Press, 1961.

[101] S. De Wit and G. Augenbroe, "Analysis of uncertainty in building design evaluations and its implications," *Energy and Buildings*, vol. 34, no. 9, pp. 951–958, 2002.

[102] Y. Sun, L. G. b, J. Wu, and G. Augenbroe, "Exploring hvac system sizing under uncertainty," *Energy and Buildings*, vol. 81, pp. 243–252, 2014.

[103] P. Lee, P. Lam, W. Lee, and E. Chan, "Analysis of an air-cooled chiller replacement project using a probabilistic approach for energy performance contracts," *Applied Energy*, vol. 171, pp. 415–428, 2016.

[104] V. Corrado and H. E. Mechri, "Uncertainty and sensitivity analysis for building energy rating," *Journal of Building Physics*, vol. 33, no. 2, pp. 125–156, 2009.

[105] W. Tian and R. Choudhary, "A probabilistic energy model for non-domestic building sectors applied to analysis of school buildings in greater london," *Energy and Buildings*, vol. 54, pp. 1–11, 2012.

[106] A. Booth and R. Choudhary, "Decision making under uncertainty in the retrofit analysis of the UK housing stock: Implications for the Green Deal," *Energy and Buildings*, vol. 64, pp. 292–308, 2013.

[107] Z. O'Neill and B. Eisenhower, "Leveraging the analysis of parametric uncertainty for building energy model calibration," in *Building simulation*, vol. 6, no. 4. Springer, 2013, pp. 365 – 377.

[108] M. Manfren, N. Aste, and R. Moshksar, "Calibration and uncertainty analysis for computer models – a meta-model based approach for integrated building energy simulation," *Applied Energy*, vol. 103, pp. 627–641, 2013.

[109] Y. Sun, "Closing the building energy performance gap by improving our predictions," Ph.D. dissertation, Georgia Institute of Technology, August 2014.

[110] A. C. Harding and D. W. Nutter, "Measurement and verification of industrial equipment: Sampling interval and data logger considerations," *Energy Engineering*, vol. 113, no. 6, pp. 7–33, 2016.

[111] H. Carstens, "Improvements to longitudinal Clean Development Mechanism sampling designs for lighting retrofit projects," Master's thesis, University of Pretoria, 2014.

[112] M. W. Ahmad, M. Mourshed, D. Mundow, M. Sisinni, and Y. Rezgui, "Building energy metering and environmental monitoring–a state-of-the-art review and directions for future research," *Energy and Buildings*, vol. 120, pp. 85–102, 2016.

[113] American Society for Heating, Refrigeration, and Air Conditioning Engineers, "Engineering analysis of experimental data," American Society for Heating, Refrigeration, and Air Conditioning Engineers, Tech. Rep., 1986.

[114] *ANSI C12.20 American National Standard for Electricity Meters - 0.2 and 0.5 Accuracy Classes*, National Electrical Manufacturers Association Std.

[115] M. Abbas, L. M. Al-Hadhrami, and S. A. Vaqar, "Perspectives in electric energy metrology," in *2015 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*. IEEE, 2015, pp. 1243–1248.

[116] L. G. Polese, S. Frank, M. Sheppy, C. Lobato, E. Rader, J. Smith, and N. Long, "Monitoring and characterization of miscellaneous electrical loads in a large retail environment," National Renewable Energy Laboratory, Tech. Rep. NREL/TP-5500-60668, February 2014.

[117] M. C. Lorek, F. Chraim, K. S. Pister, and S. Lanzisera, "COTS-based stick-on electricity meters for building submetering," *IEEE Sensors Journal*, vol. 14, no. 10, pp. 3482–3489, 2014.

[118] Personal Correspondence with primary author, July 2015, low Cost metering manufacturer.

[119] "5th CEER Benchmarking Report on the Quality of Electricity Supply 2011," Council for European Energy Regulators, Tech. Rep., 2011.

[120] *CENELEC Harmonisation Document HD 472 S1:1988*, European Committee for Electrotechnical Standardization Std.

[121] ANSI, *C84.1-1982: American National Standard Standard for Electric Power System and Equipment - Voltage Ratings (60Hz).*, American National Standards Institute Std.

[122] D. Gallo, C. Landi, N. Pasquino, and N. Polese, "A new methodological approach to quality assurance of energy meters under nonsinusoidal conditions," *IEEE Transactions on Instrumentation and Measurement*, vol. 56, no. 5, pp. 1694–1702, 2007.

[123] A. Cataliotti, V. Cosentino, and S. Nuccio, "The measurement of reactive energy in polluted distribution power systems: An analysis of the performance of commercial static meters," *IEEE Transactions on Power Delivery*, vol. 23, no. 3, pp. 1296–1301, 2008.

[124] A. Cataliotti, V. Cosentino, A. Lipari, and S. Nuccio, "Metrological characterization and operating principle identification of static meters for reactive energy: an experimental approach under nonsinusoidal test conditions," *IEEE Transactions on Instrumentation and Measurement*, vol. 58, no. 5, pp. 1427–1435, 2009.

[125] D. Gallo, C. Landi, and M. Luiso, "Power meter verification issue: Reactive power measurement in non sinusoidal conditions," in *2015 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*. IEEE, 2015, pp. 1255–1260.

[126] P. V. Barbaro, A. Cataliotti, V. Cosentino, and S. Nuccio, "Behaviour of reactive energy meters in polluted power systems," in *XVIII IMEKO World Congress, Metrology for a Sustainable Development, Rio de Janeiro, Brazil*, vol. 172, no. 2, 2006.

[127] A. J. Berrisford, "Smart meters should be smarter," in *2012 IEEE Power and Energy Society General Meeting*, 2012, pp. 1–6.

[128] A. Cataliotti, V. Cosentino, D. Di Cara, A. Lipari, and S. Nuccio, "A DAQ-based sampling wattmeter for IEEE Std. 1459-2010 powers measurements. uncertainty evaluation in nonsinusoidal conditions," *Measurement*, vol. 61, pp. 27–38, 2015.

[129] A. Ferrero, "Measuring electric power quality: Problems and perspectives," *Measurement*, vol. 41, no. 2, pp. 121–129, 2008.

[130] A. Cataliotti, V. Cosentino, and S. Nuccio, "Static meters for the reactive energy in the presence of harmonics: an experimental metrological characterization," *IEEE Transactions on Instrumentation and Measurement*, vol. 58, no. 8, pp. 2574–2579, 2009.

[131] C. Budeanu, "Reactive and fictitious powers," *Romanian National Institute*, no. 2, p. , 1927.

[132] S. Fryze, "Active, reactive and apparent powers in circuits with distorted voltages and currents," *Elektrotechnische Zeitschrift*, vol. 53, no. 25-27, pp. 596–599, 1932.

[133] A. Cataliotti, V. Cosentino, A. Lipari, and S. Nuccio, "On the calibration of reactive energy meters under non sinusoidal conditions," *XIX IMECO Kongres Fundamental and Applied Metrology, Lisbon*, pp. 719–723, 2009.

[134] *IEEE Std. 1459-2010: IEEE standard definitions for the measurement of electric power quantities under sinusoidal, Nonsinusoidal, Balanced or Unbalanced Conditions*, IEEE Std.

[135] R. Arseneau and E. So, "Calibration services in support of smart grid applications," in *2012 IEEE Power and Energy Society General Meeting*. IEEE, 2012, pp. 1–4.

[136] P. V. Barbaro, A. Cataliotti, V. Cosentino, and S. Nuccio, "A novel approach based on nonactive power for the identification of disturbing loads in power systems," *IEEE Transactions on Power Delivery*, vol. 22, no. 3, pp. 1782–1789, 2007.

[137] E. O'Driscoll and G. E. O'Donnell, "Industrial power and energy metering–a state-of-the-art review," *Journal of Cleaner Production*, vol. 41, pp. 53–64, 2013.

[138] S. Nuccio and C. Spataro, "A Monte Carlo method for the auto-evaluation of the uncertainties in analog-to-digital conversion-based measurements," *COMPEL: The International Journal for Computation and Mathematics in Electrical and Electronic Engineering*, vol. 23, no. 1, pp. 148–158, 2004.

[139] C. Spataro, "ADC based measurements: a common basis for the uncertainty estimation," in *17th Symposium IMEKO TC 4, 3rd Symposium IMEKO TC 19 and 15th IWADC Workshop:*

Instrumentation for the ICT Era, Kosice, Slovakia, September 2010.

[140] M. J. Korczynski and A. Hetman, "A calculation of uncertainties in virtual instrument," in *Instrumentation and Measurement Technology Conference, 2005. IMTC 2005. Proceedings of the IEEE*, vol. 3. IEEE, 2005, pp. 1697–1701.

[141] A. Ferrero, R. Gamba, and S. Salicone, "A method based on random-fuzzy variables for online estimation of the measurement uncertainty of dsp-based instruments," *IEEE Transactions on Instrumentation and Measurement*, vol. 53, no. 5, pp. 1362–1369, 2004.

[142] Y. C. Kuang, A. Rajan, M. P.-L. Ooi, and T. C. Ong, "Standard uncertainty evaluation of multivariate polynomial," *Measurement*, vol. 58, pp. 483–494, 2014.

[143] A. Rajan, M. P.-L. Ooi, Y. C. Kuang, and S. N. Demidenko, "Analytical standard uncertainty evaluation using Mellin transform," *Access, IEEE*, vol. 3, pp. 209–222, 2015.

[144] A. Rajan, Y. C. Kuang, M. P.-L. Ooi, and S. N. Demidenko, "Benchmark test distributions for expanded uncertainty evaluation algorithms," *IEEE Transactions on Instrumentation and Measurement*, vol. 65, no. 5, pp. 1022–1034, 2016.

[145] ——, "Moment-based measurement uncertainty evaluation for reliability analysis in design optimization," in *Instrumentation and Measurement Technology Conference Proceedings (I2MTC), 2016 IEEE International*. IEEE, 2016, pp. 1–6.

[146] A. Cataliotti, V. Cosentino, D. Di Cara, A. Lipari, S. Nuccio, and C. Spataro, "Uncertainty evaluation in power measurements with commercial data acquisition boards," in *International Workshop on ADC Modelling, Testing and Data Converter Analysis and Design and IEEE 2011 ADC Forum*, 2011.

[147] *Directive 2004/22/EC of the European Parlaiment and of the council of 31 March 2004 on measuring instruments*, 2004.

[148] A. D. Femine, D. Gallo, C. Landi, and M. Luiso, "Advanced instrument for field calibration of electrical energy meters," *IEEE Transactions on Instrumentation and Measurement*, vol. 58, no. 3, pp. 618–625, 2009.

[149] D. Amicone, A. Bernieri, L. Ferrigno, and M. Laracca, "A smart add-on device for the remote calibration of electrical energy meters," in *IEEE Instrumentation and Measurement Technology Conference, 2009. I2MTC'09*. IEEE, 2009, pp. 1599–1604.

[150] N. Calamaro, V. Elkonin, Y. Beck, and D. Shmilovitz, "Electric energy metering developments at the smart grid: technology, accuracy, standardization, and verification," Society for Electrical and Electronics Engineers in Israel, Tech. Rep., 2013. [Online]. Available: www.seeei.org/el2013/article/1030-art.pdf

[151] B. D'Apice, D. Gallo, C. Landi, and N. Rignano, "Distributed laboratory for metrological confirmation of power quality instruments," in *Proceedings of the 2005 IEEE International Conference on Virtual Environments, Human-Computer Interfaces and Measurement Systems, 2005. VECIMS 2005.*, 2005, pp. 13 – 15.

[152] Y. Sun and G. Augenbroe, "Urban heat island effect on energy application studies of office buildings," *Energy and Buildings*, vol. 77, pp. 171–179, 2014.

[153] W. J. Fisk, D. Faulkner, and D. P. Sullivan, "Accuracy of co2 sensors," Lawrence Berkeley National Laboratory, Tech. Rep. LBNL-1095E, 2008.

[154] Y. Agarwal, B. Balaji, R. Gupta, J. Lyles, M. Wei, and T. Weng, "Occupancy-driven energy management for smart building automation," in *Proceedings of the 2nd ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Building*. ACM, 2010, pp. 1–6.

[155] D. Yan, W. O'Brien, T. Hong, X. Feng, H. B. Gunay, F. Tahmasebi, and A. Mahdavi, "Occupant behavior modeling for building performance simulation: current state and future challenges," *Energy and Buildings*, vol. 107, pp. 264–278, 2015.

[156] T. Hong, S. Taylor-Lange, S. DŠ'Oca, D. Yan, and S. P. Corgnati, "Advances in research and

applications of energy-related occupant behavior in buildings," *Energy and Buildings*, no. 116, pp. 694–102, 2015.

[157] TUV NEL, "Flare gas measurement using ultrasonic transit time: Good practice guide," National Measurement System, Tech. Rep.

[158] A. Johnson and et al., "Measurement challenges and metrology for monitoring CO2 emissions from smokestacks - workshop summary," National Institute for Standards and Measurement, NIST Special Publication 1201, December 2015.

[159] C. A. Gueymard and D. R. Myers, "Evaluation of conventional and high-performance routine solar radiation measurements for improved solar resource, climatological trends, and radiative modeling," *Solar Energy*, vol. 83, no. 2, pp. 171–185, 2009.

[160] B. Bronfman, H. Michaels, G. Fitzpatrick, S. Nadel, E. Hicks, J. Peters, E. Hirst, J. Reed, M. Hoffman, W. Saxonis, A. Schoh, K. Keating, and D. Violette, "Handbook of evaluation of utility DSM programs," Oak Ridge National Laboratory, Tech. Rep. ORNL/CON-336, December 1991.

[161] D. M. Grueneich, "The next level of energy efficiency: The five challenges ahead," *The Electricity Journal*, vol. 28, no. 7, pp. 44–56, 2015.

[162] I. M. Hoffman, S. R. Schiller, A. Todd, M. A. Billingsley, C. A. Goldman, and L. C. Schwartz, "Energy savings lifetimes and persistence: Practices, issues and data," Lawrence Berkeley National Laboratory, Technical Brief LBNL-179191, May 2015.

[163] L. A. Skumatz, M. S. Khawaja, and J. Colby, "Lessons learned and next steps in energy efficiency measurement and attribution: Energy savings, net to gross, non-energy benefits, and persistence of energy efficiency behavior," SERA, Tech. Rep., November 2009.

[164] N. Nord and S. F. Sjøthun, "Success factors of energy efficiency measures in buildings in Norway," *Energy and Buildings*, vol. 76, pp. 476–487, 2014.

[165] E. L. Vine, "Persistence of energy savings: What do we know and how can it be ensured?" *Energy*, vol. 17, no. 11, pp. 1073–1084, 1992.

[166] J. Proctor and T. Downey, "Summary report of persistence studies: Assessments of technical degradation factors," Proctor Engineering, Tech. Rep., February 1999.

[167] L. A. Skumatz, D. Whitson, D. Thomas, K. Geraghty, B. Dunford, K. Lorberau, and C. Breckinridge, "Measure life study II," Bonneville Power Administration, Tech. Rep. SRC 7851-R1, July 1994.

[168] Nexus Market Research, Inc. and RLW Analytics, Inc., "Residential lighting measure life study," New England Residential Lighting Program, Tech. Rep., 2008.

[169] Navigant Consulting, "Evaluation of the IFC/GEF Poland Efficient Lighting Project CFL Subsidy Program," Netherlands Energy Efficient Lighting B.V., International Finance Corporation / Global Environment Facility, Tech. Rep. 1, 1999.

[170] Lighting Research Center, "Screwbase compact fluorescent lamp products," Rensellaer Polytechnic Institute, Specifier Report 7 (1), June 1999.

[171] M. Botha-Moorlach and G. Mckuur, "A report on the factors that influence the demand and energy savings for Compact Fluorescent Lamp door-to-door rollouts in South Africa," Eskom, Tech. Rep., March 2009.

[172] P. O'Connor and A. Kleyner, *Practical Reliability Engineering*. Wiley, 2011.

[173] D. Young, "When do energy-efficient appliances generate energy savings? Some evidence from Canada," *Energy Policy*, vol. 36, no. 1, pp. 34–46, 2008.

[174] D. C. Montgomery, *Design and Analysis of Experiments*, 7th ed. Wiley New York, 1997.

[175] "Regression for M&V: Reference guide," Bonneville Power Administration, Tech. Rep., May 2012.

[176] T. Walter, P. N. Price, and M. D. Sohn, "Uncertainty estimation improves energy measurement and verification procedures," *Applied Energy*, vol. 130, pp. 230–236, 2014.

[177] J. L. Mathieu, P. N. Price, K. Sila, and M. A. Piette, "Quantifying changes in building electricity use, with application to demand response," *IEEE Transactions on Smart Grid*, vol. 41, no. 4, pp. 374–381, 2009.

[178] J. A. Shonder and P. Im, "Bayesian analysis of savings from retrofit projects," *ASHRAE Transactions*, vol. 118, p. 367, 2012.

[179] J. Granderson, S. Touzani, S. Fernandes, and C. Taylor, "Application of automated measurement and verification to utility energy efficiency program data," *Energy and Buildings*, vol. 142, pp. 191–199, 2017.

[180] Y. Zhang, Z. O'Neill, B. Dong, and G. Augenbroe, "Comparisons of inverse modeling approaches for predicting building energy performance," *Building and E*, vol. 86, pp. 177–190, 2015.

[181] International Standard Organization, *Guide 98–3 (2008) Uncertainty of measurement Part 3: Guide to the expression of uncertainty in measurement (GUM: 1995)*, International Standard Organization Std.

[182] United Kingdom Accreditation Service, *UKAS M3003: The Expression of Uncertainty and Confidence in Measurement*, United Kingdom Accreditation Service Std., Rev. 3, November 2012.

[183] European Accreditation, *EA-4/02 M: 2013: Evaluation of the Uncertainty of Measurement In Calibration*, European Accreditation Std.

[184] International Standards Organization, *ISO 17025: General Requirements for the Competence of Testing and Calibration Laboratories*, Std., 2005.

[185] Joint Committee for Guides in Metrology, *OIML G1-101 Evaluation of measurement data - Supplement 1 to the "Guide to the expression of uncertainty in measurement" - Propagation of distributions using a Monte Carlo method*, International Organization of Legal Metrology, Paris, France, 2008.

[186] S. Burhenne, D. Jacob, and G. P. Henze, "Sampling based on sobol' sequences for monte carlo techniques applied to building simulations," in *Proceedings of the International Conference on Building Simulation*, 2011, pp. 1816–1823.

[187] I. Lira and D. Grientschnig, "Bayesian assessment of uncertainty in metrology: a tutorial," *Metrologia*, vol. 47, pp. R1–R14, 2010.

[188] Mellin transform-based moment calculator for multivariate polynomials. IEEE TC-32 - Fault Tolerant Measurement Systems. Accessed 20 December 2016. [Online]. Available: http://tc32.ieee-ims.org/content/mellin-transform-based-moment-calculator-multivariate-polynomials

[189] N. L. Johnson, "Systems of frequency curves generated by methods of translation," *Biometrika*, vol. 36, no. 1/2, pp. 149–176, 1949.

[190] J.-G. Simonato, "The performance of Johnson distributions for computing value at risk and expected shortfall," *SSRN*, no. 1706409, 2011.

[191] N. R. Farnum, "Using Johnson curves to describe non-normal process data," *Quality Engineering*, vol. 9, no. 2, pp. 329–336, 1996.

[192] I. Lira, "The gum revision: the Bayesian view toward the expression of measurement uncertainty," *European Journal of Physics*, vol. 37, no. 2, p. 025803, 2016.

[193] W. Estler, "Measurement as inference: Fundamental ideas," *CIRP Annals - Manufacturing Technology*, vol. 48, no. 2, pp. 611–631, 1999.

[194] G. B. Rossi, "A probabilistic model for measurement processes," *Measurement*, vol. 34, no. 2, pp. 85–99, 2003.

[195] M. G. Cox, G. B. Rossi, P. M. Harris, and A. Forbes, "A probabilistic approach to the analysis of measurement processes," *Metrologia*, vol. 45, no. 5, p. 493, 2008.

[196] R. J. Carroll, D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu, *Measurement Error in Nonlinear Models: a Modern Perspective*. CRC press, 2006.

[197] P. Gustafson, *Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments*. CRC Press, 2003.

[198] P. Eguía, E. Granada, J. Alonso, E. Arce, and A. Saavedra, "Weather datasets generated using kriging techniques to calibrate building thermal simulations with TRNSYS," *Journal of Building Engineering*, vol. 7, pp. 78–91, 2016.

[199] M. J. Ree and T. R. Carretta, "The role of measurement error in familiar statistics," *Organizational Research Methods*, vol. 9, no. 1, pp. 99–112, January 2006.

[200] E. T. Jaynes, "Straight Line Fitting - A Bayesian Solution," in *Tenth Annual MAXENT Workshop, University of Wyoming*, W. T. Grandy and L. Schick, Eds. Kluwer Academic Publishers, Holland, July 1990.

[201] R. J. Carroll and D. Ruppert, *Transformation and Weighting in Regression*. CRC Press, 1988, vol. 30.

[202] R. Ridge, "Errors in variables: a close encounter of the third kind," in *Proceedings of the Energy Evaluation Conference*, Chicago, 1997, pp. 479–487.

[203] N. Rivers, M. Jaccard *et al.*, "Electric utility demand side management in canada," *Energy Journal-Cleveland*, vol. 32, no. 4, p. 93, 2011.

[204] D. M. Violette, B. Provencher, and I. Sulyma, "Assessing bottom-up and top-down approaches for assessing DSM programs and efforts," in *International Program Evaluation Conference*, Rome, 2012.

[205] L. Huntley, "Establishing traceability for a high performance AC/DC transfer standard," *ISA Transactions*, vol. 29, no. 4, pp. 41–47, 1990.

[206] J. Somppi, "A case study in characterizing & disciplining electrical calibrator instrumentation to improve test accuracies & measurement uncertainties," in *NCSL International Workshop and Symposium*, 2007.

[207] R. B. Schumacher, "Systematic measurement errors," *Journal of Quality Technology*, vol. 13, no. 1, pp. 10–24, 1981.

[208] D. Deaver and W. Everett, "How to maintain your confidence," in *Proc. 1993 NCSL Workshop and Symposium*, 1993.

[209] W. Wong, "What TUR do you really need? putting statistical theory into practice," Fluke Corporation, Tech. Rep.

[210] D. Deaver, "Managing calibration confidence in the real world," in *NCSL Workshop and Symposium*, 1995.

[211] ——, "Guardbanding in the world of ISO Guide 25," in *NCSL Workshop and Symposium*, 1998.

[212] A. R. Eagle, "A method for handling errors in testing and measuring," *Industrial Quality Control*, vol. 10, no. 3, pp. 10–15, 1954.

[213] G. B. Rossi and F. Crenna, "A probabilistic approach to measurement-based decisions," *Measurement*, vol. 39, no. 2, pp. 101–119, 2006.

[214] *ISO 14253–1, 1998: Inspection by measurement of workspieces and measuring equipment; Part 1: Decision rules for proving conformance or nonconformance with specifications,*, International Standards Organization Std.

[215] S. D. Phillips, W. T. Estler, M. S. Levenson, and K. R. Eberhardt, "Calculation of measurement uncertainty using prior information," *Journal of Research of the National Institute of Standards*

and Technology, vol. 103, no. 6, pp. 625–632, November-December 1998.

[216] W. A. Fuller, *Measurement Error Models*. Wiley-Interscience, 2006.

[217] C. Carobbi, G. Pellicci, and S. Vieri, "Error modeling of static energy meters," in *XIX IMEKO World Congress, Fundamental and Applied Metrology*, 2009, pp. 820–824.

[218] H. Küchenhoff and R. Carroll, "Segmented regression with errors in predictors: Semi-parametric and parametric methods," *Statistics in Medicine*, vol. 16, no. 2, pp. 169–188, 1997.

[219] M. C. Burkhart, Y. Heo, and V. M. Zavala, "Measurement and verification of building systems under uncertain data: A Gaussian process modeling approach," *Energy and Buildings*, vol. 75, pp. 189–198, 2014.

[220] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis*. Taylor & Francis, 2014, vol. 2.

[221] G. S. Pavlak, A. R. Florita, G. P. Henze, and B. Rajagopalan, "Comparison of traditional and Bayesian calibration techniques for gray-box modeling," *Journal of Architectural Engineering*, vol. 20, no. 2, pp. 04 013 011–2–16, 2013.

[222] A. Kucukelbir, D. Tran, R. Ranganath, A. Gelman, and D. M. Blei, "Automatic differentiation variational inference," *arXiv preprint arXiv:1603.00788*, 2016.

[223] T. Clark, M. Bradburn, S. Love, and D. Altman, "Survival analysis part I: basic concepts and first analyses," *British Journal of Cancer*, vol. 89, no. 2, p. 232, 2003.

[224] M. Bradburn, T. Clark, S. Love, and D. Altman, "Survival analysis part II: Multivariate data analysis–an introduction to concepts and methods," *British journal of cancer*, vol. 89, no. 3, p. 431, 2003.

[225] ——, "Survival analysis part III: Multivariate data analysis–choosing a model and assessing its adequacy and fit," *British Journal of Cancer*, vol. 89, no. 4, p. 605, 2003.

[226] T. Clark, M. Bradburn, S. Love, and D. Altman, "Survival analysis part IV: further concepts and methods in survival analysis," *British Journal of Cancer*, vol. 89, no. 5, p. 781, 2003.

[227] F. Tekle, F. Tan, and M. Berger, "Maximin d-optimal designs for binary longitudinal responses," *Computational Statistics & Data Analysis*, vol. 52, no. 12, pp. 5253–5262, 2008.

[228] N. H. Tehrani, U. T. Khan, and C. Crawford, "Baseline load forecasting using a Bayesian approach," in *2016 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*. IEEE, 2016, pp. 1–4.

[229] M. West, P. J. Harrison, and H. S. Migon, "Dynamic generalized linear models and Bayesian forecasting," *Journal of the American Statistical Association*, vol. 80, no. 389, pp. 73–83, 1985.

[230] J. Harrison and M. West, *Bayesian Forecasting & Dynamic Models*. Springer, 1999.

[231] P. McCullagh and J. A. Nelder, *Generalized Linear Models*. CRC press, 1989, vol. 37.

[232] K. Triantafyllopoulos, "Inference of dynamic generalized linear models: On-line computation and appraisal," *International Statistical Review*, vol. 77, no. 3, pp. 430–450, 2009. [Online]. Available: http://www.jstor.org.uplib.idm.oclc.org/stable/27919767

[233] D. Gamerman and M. West, "An application of dynamic survival models in unemployment studies," *The Statistician*, pp. 269–274, 1987.

[234] D. Gamerman, "Dynamic Bayesian models for survival data," *Applied Statistics*, pp. 63–79, 1991.

[235] E. R Brown and J. G Ibrahim, "A Bayesian semiparametric joint hierarchical model for longitudinal and survival data," *Biometrics*, vol. 59, no. 2, pp. 221–228, 2003. [Online]. Available: http://www.jstor.org.uplib.idm.oclc.org/stable/3695499

[236] M. De Iorio, W. O. Johnson, P. Müller, and G. L. Rosner, "Bayesian nonparametric nonproportional hazards survival modeling," *Biometrics*, vol. 65, no. 3, pp. 762–771, 2009.

[237] D. Gamerman and H. S. Migon, "Dynamic hierarchical models," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 629–642, 1993. [Online]. Available: http://www.jstor.org.uplib.idm.oclc.org/stable/2345875

[238] T. Bayes, R. Price, and J. Canton, *An essay towards solving a problem in the doctrine of chances*. C. Davis, Printer to the Royal Society of London, 1763.

[239] D. V. Lindley, "The philosophy of statistics," *Journal of the Royal Statistical Society: Series D (The Statistician)*, vol. 49, no. 3, pp. 293–337, 2000.

[240] J. K. Kruschke, "Bayesian estimation supersedes the t-test," *Journal of Experimental Psychology: General*, vol. 142, no. 2, pp. 573–603, 2013.

[241] H. Jeffreys, *Theory of Probability*. Oxford: Clarendon, 1961.

[242] D. C. Montgomery and G. C. Runger, *Applied Statistics and Probability for Engineers*. John Wiley & Sons, 2010.

[243] J. O. Berger and M. Delampady, "Testing precise hypotheses," *Statistical Science*, pp. 317–335, 1987.

[244] J. P. Ioannidis, "Why most published research findings are false," *PLoS Med*, vol. 2, no. 8, p. e124, 2005.

[245] K. S. Button, J. P. Ioannidis, C. Mokrysz, B. A. Nosek, J. Flint, E. S. Robinson, and M. R. Munafò, "Power failure: why small sample size undermines the reliability of neuroscience," *Nature Reviews Neuroscience*, vol. 14, no. 5, pp. 365–376, 2013.

[246] D. V. Lindley, "The choice of sample size," *Journal of the Royal Statistical Society: Series D (The Statistician)*, vol. 46, no. 2, pp. 129–138, 1997.

[247] J. Bernardo, "Statistical inference as a decision problem: the choice of sample size," *The Journal of the Royal Statistical Society: Series D (The Statistician)*, vol. 46, no. 2, pp. 151–153, 1997.

[248] S. Senn, *Determining the Sample Size*. Wiley, 2007.

[249] A. Gelman and C. Hennig, "Beyond subjective and objective in statistics," *Journal of the Royal Statistical Society. Series A*, vol. 180, no. Part 4, pp. 1–31, 2017.

[250] B. de Finetti, *Theory of Probability*. Wiley, 1974.

[251] A. Gallo, "A refresher on regression analysis," *Harvard Business Review*, 4 November 2015. [Online]. Available: https://hbr.org/2015/11/a-refresher-on-regression-analysis

[252] R. T. Clemen and T. Reilly, *Making Hard Decisions with Decisions Tools*. South-Western, 2001.

[253] R. L. Winkler, "The assessment of prior distributions in bayesian analysis," *Journal of the American Statistical association*, vol. 62, no. 319, pp. 776–800, 1967.

[254] A. Tversky and D. Kahneman, "Judgment under uncertainty: Heuristics and biases," in *Utility, probability, and human decision making*. Springer, 1975, pp. 141–162.

[255] E. Jaynes, *Probability Theory: The Logic of Science*. Cambridge University Press, 2003.

[256] J. Berger, "The case for objective Bayesian analysis," *Bayesian Analysis*, vol. 1, pp. 385–402, 2006.

[257] A. Gelman, A. Jakulin, M. G. Pittau, and Y.-S. Su, "A weakly informative default prior distribution for logistic and other regression models," *The Annals of Applied Statistics*, pp. 1360–1383, 2008.

[258] W. F. Darnieder, "Bayesian methods for data-dependent priors," Ph.D. dissertation, The Ohio State University, 2011. [Online]. Available: https://etd.ohiolink.edu/rws_etd/document/get/osu1306344172/inline

[259] C. Shannon and W. Weaver, *The Mathematical Theory of Information*. Urbana, Illinois: University of Illinois Press, 1963.

[260] E. Jaynes, "Information theory and statistical mechanics," *Physical Review Letters*, vol. 106, pp. 620–630, 1957.

[261] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equation of state calculations by fast computing machines," *The journal of chemical physics*, vol. 21, no. 6, pp. 1087–1092, 1953.

[262] C. Robert and G. Casella, *Monte Carlo Statistical Methods*. Springer Science & Business Media, 2013.

[263] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Transactions on pattern analysis and machine intelligence*, no. 6, pp. 721–741, 1984.

[264] M. D. Hoffman and A. Gelman, "The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1593–1623, 2014.

[265] M. Plummer *et al.*, "JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling," in *Proceedings of the 3rd international workshop on distributed statistical computing*, vol. 124. Vienna, 2003, p. 125.

[266] B. Carpenter, A. Gelman, M. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. A. Brubaker, J. Guo, P. Li, and A. Riddell, "Stan: A probabilistic programming language," *Journal of Statistical Software*, 2016, (In Press).

[267] J. Salvatier, T. V. Wiecki, and C. Fonnesbeck, "Probabilistic programming in Python using PyMC3," *PeerJ Computer Science*, vol. 2, April 2016.

[268] D. Ruch, J. Kissock, and T. Reddy, "Prediction uncertainty of linear building energy use models with autocorrelated residuals," *Journal of Solar Energy Engineering*, vol. 121, no. 1, pp. 63–68, 1999.

[269] J. K. Kissock, J. S. Haberl, and D. E. Claridge, "Inverse modeling toolkit: Numerical algorithms (rp-1050)," *Transactions-American society of heating refrigerating and air conditioning engineers*, vol. 109, no. 2, pp. 425–434, 2003.

[270] M.-T. Ke, C.-H. Yeh, and C.-J. Su, "Cloud computing platform for real-time measurement and verification of energy performance," *Applied Energy*, vol. 188, pp. 497–507, 2017.

[271] J. VanderPlas, "Frequentism and Bayesianism: a Python-driven primer," *arXiv preprint arXiv:1411.5018*, 2014.

[272] K. L. Lange, R. J. Little, and J. M. Taylor, "Robust statistical modeling using the t distribution," *Journal of the American Statistical Association*, vol. 84, no. 408, pp. 881–896, 1989.

[273] K. Agnew and M. Goldberg, *The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures*. National Renewable Energy Laboratory, January 2012 - March 2013 2013, ch. 8: Whole-Building Retrofit with Consumption Data Analysis Evaluation Protocol.

[274] A. Booth, R. Choudhary, and D. Spiegelhalter, "A hierarchical Bayesian framework for calibrating micro-level models with macro-level data," *Journal of Building Performance Simulation*, vol. 6, no. 4, pp. 293–318, 2013.

[275] P. Gregory, *Bayesian Logical Data Analysis for the Physical Sciences: A Comparative Approach with Mathematica® Support*. Cambridge University Press, 2005.

[276] I. M. Sobol, "Global sensitivity indices for nonlinear mathematical models and their monte carlo estimates," *Mathematics and computers in simulation*, vol. 55, no. 1, pp. 271–280, 2001.

[277] A. Saltelli, "Making best use of model evaluations to compute sensitivity indices," *Computer Physics Communications*, vol. 145, no. 2, pp. 280–297, 2002.

[278] A. Saltelli, P. Annoni, I. Azzini, F. Campolongo, M. Ratto, and S. Tarantola, "Variance based sensitivity analysis of model output. design and estimator for the total sensitivity index," *Computer Physics Communications*, vol. 181, no. 2, pp. 259–270, 2010.

[279] J. Herman and W. Usher, "SALib: An open-source python library for sensitivity analysis," *The Journal of Open Source Software*, vol. 2, no. 9, jan 2017. [Online]. Available: https://doi.org/10.21105%2Fjoss.00097

[280] M. D. Morris, "Factorial sampling plans for preliminary computational experiments," *Technometrics*, vol. 33, no. 2, pp. 161–174, 1991.

[281] F. Campolongo, J. Cariboni, and A. Saltelli, "An effective screening design for sensitivity analysis of large models," *Environmental modelling & software*, vol. 22, no. 10, pp. 1509–1518, 2007.

[282] K. Menberg, Y. Heo, and R. Choudhary, "Sensitivity analysis methods for building energy models: Comparing computational costs and extractable information," *Energy and Buildings*, vol. 133, pp. 433–445, 2016.

[283] S. R. Schiller, "National energy efficiency evaluation, measurement and verification (EM&V) standard: scoping study of issues and implementation requirements," Lawrence Berkeley National Laboratory, Tech. Rep. LBNL-4265E, 2011.

[284] J. Cook and L. Stefanski, "Simulation-extrapolation estimation in parametric measurement error models," *Journal of the American Statistical Association*, vol. 89, pp. 1314–1328, 1994.

[285] S. Van Der Walt, S. C. Colbert, and G. Varoquaux, "The numpy array: a structure for efficient numerical computation," *Computing in Science & Engineering*, vol. 13, no. 2, pp. 22–30, 2011.

[286] E. Jones, T. Oliphant, P. Peterson *et al.*, "SciPy: Open source scientific tools for Python," 2001–, [Online; accessed 2016-07-30]. [Online]. Available: http://www.scipy.org/

[287] J. Nocedal and S. Wright, *Numerical Optimization*. New York: Springer, 2006.

[288] A. Gelman *et al.*, "Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper)," *Bayesian analysis*, vol. 1, no. 3, pp. 515–534, 2006.

[289] United States Energy Information Administration. (2017) Commercial building energy consumption survey (CBECS). United States Department of Energy. [Online]. Available: https://www.eia.gov/consumption/commercial/

[290] R. Kalman, "A new approach to linear filtering and prediction problems," *Journal of Basic Engineering*, vol. 82, pp. 35–45, 1960.

[291] ——, "New methods in Wiener filtering theory," in *Proceedings of the First Symposium of Engineering Applications of Random Function Theory and Probability*, J. Bogdanoff and F. Kozin, Eds. New York: Wiley, 1963.

[292] R. Luus, "Statistical inference of the multiple regression analysis of complex survey data," Ph.D. dissertation, University of Stellenbosch, December 2016.

[293] E. Vine and D. Fielding, "An evaluation of residential CFL hours-of-use methodologies and estimates: Recommendations for evaluators and program managers," *Energy and buildings*, vol. 38, no. 12, pp. 1388–1394, 2006.

[294] F.-A. Fortin, F.-M. De Rainville, M.-A. Gardner, M. Parizeau, and C. Gagné, "DEAP: Evolutionary algorithms made easy," *Journal of Machine Learning Research*, vol. 13, pp. 2171–2175, jul 2012.

[295] K. C. So and E. Wu, "Developing cost-effective inspection sampling plans for energy-efficiency programs at Southern California Esdison," *Interfaces*, vol. 46, no. 6, pp. 522–532, 2016.

[296] M. Moinester and R. Gottfried, "Sample size estimation for correlations with pre-specified confidence interval," *The Quantitative Methods for Psychology*, vol. 10, no. 2, pp. 124–130, 2014.

[297] B. Wang, X. Xia, and J. Zhang, "A multi-objective optimization model for the life-cycle cost analysis and retrofitting planning of buildings," *Energy and Buildings*, vol. 77, pp. 227–235, 2014.

[298] Z. Wu, B. Wang, and X. Xia, "Large-scale building energy efficiency retrofit: Concept, model and control," *Energy*, vol. 109, pp. 456–465, 2016.

[299] H. Carstens, X. Xia, X. Ye, and J. Zhang, "Characterising Compact Fluorescent Lamp population decay," in *IEEE Africon Conference*, Pointe-Aux-Piments, Mauritius, 2013.

[300] X. Xia, "Control problems in building energy retrofit and maintenance planning," *Annual Reviews in Control*, 2017.

[301] V. Barnett, *Sample Survey: Principles and Methods*. Arnold, 2002.

[302] A. K. Gupta and D. G. Kabe, *Theory of Sample Surveys*. World Scientific, 2011.

[303] P. S. Levy and S. Lemeshow, *Sampling of Populations: Methods and Applications*. John Wiley & Sons, 2013.

[304] M. H. Hansen, W. N. Hurwitz, and W. G. Madow, *Sample Survey Methods and Theory*. John Wiley & Sons, 1953, vol. 1.

[305] L. Guan, T. Berrill, and R. J. Brown, "Measurement of actual efficacy of compact fluorescent lamps (CFLs)," *Energy and Buildings*, vol. 86, pp. 601–607, 2015.

[306] A. H. Lee, "Verification of electrical energy savings for lighting retrofits using short-and long-term monitoring," *Energy conversion and management*, vol. 41, no. 18, pp. 1999–2008, 2000.

[307] W. Gifford, M. Goldberg, P. Tanimoto, D. Celnicker, and M. Poplawski, "Residential lighting end-use consumption study: Estimation framework and initial estimates," United States Department of Energy, Tech. Rep. PNNL-22182, 2012.

[308] C. Jump, J. J. Hirsch, J. Peters, and D. Moran, "Welcome to the dark side: The effect of switching on CFL measure life," in *ACEEE Summer Study on Energy Efficiency in Buildings*, vol. 2. Pacific Grove, California: American Council for an Energy Efficient Economy, 2008, pp. 138–149.

[309] I. Hill, R. Hill, and R. Holder, "Algorithm AS 99: Fitting Johnson curves by moments," *Journal of the Royal Statistical Society. Series C (Applied statistics)*, vol. 25, no. 2, pp. 180–189, 1976.

[310] D. Jones, "Johnson Curve Toolbox for Matlab: analysis of non-normal data using the Johnson family of distributions," College of Marine Science, University of South Florida, Tech. Rep., 2014. [Online]. Available: http://www.marine.usf.edu/user/djones/jctm/jctm.html

[311] L. Brown, T. Cai, and A. DasGupta, "Interval estimation for a binomial proportion," *Statistical Science*, vol. 16, no. 2, pp. 101–133, 2001. [Online]. Available: http://www.jstor.org/stable/2676784