Efficient Longitudinal Population Survival Survey Sampling for the Measurement and Verification of Lighting Retrofit Projects

Herman Carstens[a,], Xiaohua Xia[a], Sarma Yadavalli[b], Arvind Rajan[c]

[a]*Centre for New Energy Systems (CNES), Department of Electrical, Electronic, and Computer Engineering, University of Pretoria, South Africa*
[b]*Department of Industrial and Systems Engineering, University of Pretoria, South Africa*
[c]*Advanced Engineering Platform & Department of Electrical and Computer Systems Engineering, School of Engineering, Monash University Malaysia, Bandar Sunway 47500, Malaysia*

Abstract

A method is presented for reducing the required sample sizes for reporting energy savings with predetermined statistical accuracy in lighting retrofit measurement and verification projects, where the population of retrofitted luminaires is to be tracked over time. The method uses a Dynamic Generalised Linear Model with Bayesian forecasting to account for past survey sample sizes and survey results and forecast future population decay, while quantifying estimation uncertainty. A genetic algorithm is used to optimise multi-year sampling plans, and distributions are convolved using a new method of moments technique using the Mellin transform instead of a Monte Carlo simulation. Two cases studies are investigated: single population designs, and stratified population designs, where different kinds of lights are replaced in the same retrofit study. Results show significant cost reductions and increased statistical efficiency when using the proposed Bayesian framework.

Keywords: Bayesian, Population Survival, Retrofit, Sampling Design, Measurement and Verification, Mellin Transform, Method of Moments.

1. Introduction

Measurement and Verification (M&V) is the process by which savings from energy projects are independently and reliably quantified [1]. M&V is usually contractually required and is conducted by an independent third-party which audits the savings claimed by the Energy Services Company (ESCo) or contractor. However, measuring and verifying savings can be prohibitively expensive for some projects if savings are to be reported with the statistical accuracy required by most programmes. This is especially true for household Energy Efficiency (EE) projects [2]. To make M&V more cost-effective and increase the number and value of EE and Demand Side Management (DSM) projects, it is therefore necessary to devise monitoring and verification methods that adhere to the reporting accuracy requirements, but do so at low cost: so-called efficient designs [3].

The cost of an M&V project is related to the uncertainty in the data, and the required statistical accuracy for reporting. There are three kinds of uncertainty that need to be mitigated for an M&V savings estimate to be accurate. These are measurement uncertainty [4, 5], sampling uncertainty, and modelling uncertainty [6–8]. Of these, sampling uncertainty is often the dominant component [9, 10]. Sampling uncertainty arises when the whole population of ECMs or facilities are not monitored. For example, when 100 000 Incandescent Lamp fixtures are retrofitted with Compact Fluorescent Lamps (CFLs) in a residential mass rollout 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. Often, the savings have to be tracked over a number of years. One study found that the cost of electricity saved decreases by 70% for studies where the monitoring period is increased from 1 to 3.9 years [11]. Therefore two factors need to be accounted for when reporting savings. The first is the amount of energy used by the average unit installed by the project. The second is the useful life of the unit, which determines the *persistence* of the savings. Over time the size of the original population shrinks according to its population survival curve, and so should the reported annual savings. Longitudinal M&V should monitor at least two values: the energy use of the equipment in a given year, and the proportion of original equipment that has survived to that year. In this study, we will focus on the latter: population survival. Although energy use can be measured with an energy meter, such meters cannot be used to measure population survival. Population survival should be determined through survey sampling, and the uncertainties associated with such sampling should be included in the overall M&V reporting uncertainty figure. In this paper we propose an efficient method that accounts for past sample sizes and population decay, quantifies uncertainty accurately, and provides a means of predicting future decay with reliably quantifiable uncertainty. This enables sampling planning and efficient survey designs.

This study is structured as follows. In Section 2, a brief review of the relevant literature is given. DGLMs are identified as an appropriate tool for longitudinal uncertainty quantification

Email address: hermancarstens@gmail.com (Herman Carstens)

of population survival in M&V, although they have not been employed in M&V before. DGLMs allow for the application of previously developed [12] autoregressive logistic population decay models to real-world data sets. As a preliminary step in Case Study 2, it was found that this parametrisation fits a wide range of CFL population decay data accurately. The sampling and optimisation model, assumptions, and lamp population survival theory is described in Section 3. It is shown how information such as past sample sizes and survey results can be incorporated into forecasts of future population decay in a mathematically rigorous way. A case study using a single CFL population and real-world data is presented in Section 4. A limitation in the use of Monte Carlo (MC) simulation for constraint determination in a Genetic Algorithm (GA) is illustrated, and an alternative analytical method is applied successfully. A novel penalty soft constraint function is also developed to increase the quality of the GA results. Section 5 expands on the previous case study to include different strata (lamp types) in the same study - which, up to now, has only been possible under certain strict assumptions. Finally conclusions are drawn and recommendations made.

2. Literature Review

2.1. Persistence

In a 1991 M&V guideline by the Oak Ridge National Laboratory, 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" [13]. In a 2015 article [14] and a 2015 technical brief by the Lawrence Berkeley National Laboratory [15] similar comments were made. This study addresses questions around characterising persistence curves and using them for efficient M&V study designs.

We consider technical persistence, or equipment lifetimes, only. 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 [16], and studies should also account for human- and market-related factors [17, 18], although the Uniform Methods Project (UMP) recommends that such factors are not taken into account for residential monitoring programmes [17]. For a foundational introduction to persistence study design, see Vine [19], and for updated treatments, see Hoffman et al., Skumatz, and the UMP [15, 16, 20]. One engineering rather than statistical approach is to use technical degradation factors popular in the United States [15, 21]. These are the lifetimes of the measures relative to the original equipment installed [22] - a single number, rather than curve characterisation or longitudinal studies as implemented below.

Since primary persistence research is expensive, most secondary sources are often used [16, 20]. Primary research studies usually do not track populations, but try to provide a median measure life estimate [23, 24]. Two notable exceptions are the Polish Efficient Lighting Project (PELP) [25] and the Lighting Research Centre at Renssellaer Polytechnic Institute's Specifier Report on CFLs [26]. 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, 9, 27, 28] as it is a well-studied technology with relatively simple principles.

Regarding the curve shapes, The United Nations Framework Convention on Climate Change (UNFCCC)'s Clean Development Mechanism (CDM) recommends a linear decay curve [29]. Logistic decay curves similar to those used in survival analysis have also been introduced [25, 30, 31] and later mproved upon [12] 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 [32], and will have wider applicability. More examples of linear and non-linear survival curve assumptions and study esults for EE appliance models are listed by Young [33].

Persistence monitoring requirements range from 3.9 years [11] to 10 years for CDM lighting projects [34]. Pennsylvania and Texas require 15 years [15]. We will consider 10 and 12 year studies, as this reflects both regulatory requirements nd realistic CFL lifetimes.

2.2. Methods

Two methods are directly applicable to the problem at hand: 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. [35–38], and for an application to EE and DSM persistence studies, [24]. 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 [39]. 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 methods exist in energy monitoring [40]. 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 [41, 42], building on McCoullagh and Nelder's GLM work [43]. Triantafyllopoulos [44] 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 [45–47] and hierarchical models [48]. These models work with parametric distributions that do not describe the complexities of the energy savings calculations discussed in Section 4, but more research in this area is warranted. We have opted for a model similar on West, Harrison, and Migon's advertising awareness study [41], which uses a DGLM with Bayesian forecasting to model binomial survey response data. 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 [49]. However, this method goes beyond current M&V literature by then combining this beta-distributed population survival estimate with normally distributed energy data through a method of moments and the Johnson distribution. It also allows for stratified sampling designs with such distributions, which has also not been recorded in M&V literature, to our knowledge.

3. Modelling

3.1. General remarks

The purpose of the DGLM in this study 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 optimisation routine in Section 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 we will state the modelling assumptions, and give more detailed problem background before presenting the mathematical model. This background includes the characterisation of a CFL population survival model, as well as explaining the necessity and role of Bayesian statistics in the solution.

The model assumptions are as follows:

1. Energy savings, and not population survival, is the key performance indicator of the Energy Efficiency (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 Compact Fluorescent Lamps (CFLs).
4. Only the CFLs installed originally by the project are considered part of the population, although efficient maintenance approaches for such projects have been developed [50–53]
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 to doing so.
8. Survival data can be described as a time series of population proportions (fractions).

3.1.1. Dynamic model for population decay

Survival data are time-dependent and have an autoregressive or dynamic relationship [54]. 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

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

where Φ_k is the proportion of the population surviving at step k, and β and γ are model parameters. Their meanings become apparent when (1) is written in its standard form as a logistic equation:

$$\Phi_t = \frac{1}{\gamma + e^{\beta t - L}}. \tag{2}$$

The slope of the logistic curve is determined by β, and the tarting 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 [19]. The model is sensitive to fixing this value, and so it is best left as a variable. The L term falls away for the utoregressive model in (1) as it shifts the curve left or right, determining the median life of the population. To visualise this logistic curve, data from the PELP [25] has been plotted with a best fit line of (2) in Figure 1. The PELP was a large-scale CFL retrofit study undertaken in the late 1990s where over a million uminaires were installed and tracked over many years. It is a reliable data set and was adopted for use in the South African case [30], but represents only one instance of such data. For Section 5 on stratified sampling, other data sets will be used.

If the points on the logistic curve were known with certainty, redicting 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. Confidence intervals around current and future estimates should, therefore, be determined, but re dependent on the confidence with which previous estimates were determined. This in turn is dependent upon the sizes of

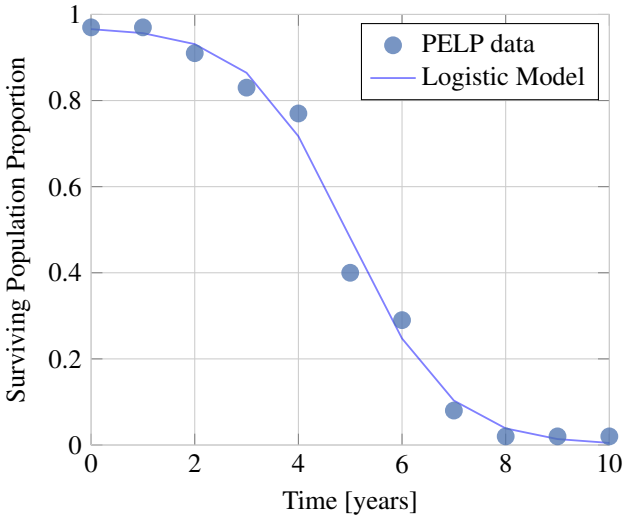


Figure 1: PELP data and best fit line of (2), as reported by (49).

previous samples. Furthermore, the population proportion estimate $\hat{\Phi}_t$ at t is dependent on the regression model and the sample size n_t at t. A larger sample size at t should have 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 this, Bayesian forecasting techniques need to be added.

3.1.2. Bayesian Forecasting

Bayesianism is a branch of statistics that is particularly suited to uncertainty quantification and data analysis. Bayesian theory is both wide and deep. Many useful introductions to the subject have been written [55, 56], and this study will only focus on one aspect that is particularly suitable to the longitudinal survival survey sampling problem: conjugate priors. Bayes' theorem states that

$$Posterior \propto Prior \times Likelihood. \tag{3}$$

That is, the result of the Bayesian calculation (the posterior probability distribution) is the product of a distribution called the prior, and a distribution called the likelihood. The likelihood distribution arises from the data. For survival data, it would be a binomial distribution defined by the observed population proportion, and the sample size. The prior represents all other knowledge the practitioner may have of the system. Although it could be subjective, for this case the prior for a particular year is simply the regression forecast derived from previous samples and forecast according to the DGLM and the population decay model (1). It is generally expressed as $g(\Phi)$. The likelihood function as $P(D = \Phi_{observed}|\Phi)$ reflects the probability of data D taking a value $\Phi_{observed}$, given the parameter Φ. The output of the Bayes equation is then probability density function x, given that $Y_t = y$ has been observed, where Y_t is the number of successes in n_t trials (as will be used below):

$$P(\Phi|D) \propto P(D|\Phi)g(\Phi). \tag{4}$$

In most cases, one cannot convolve (multiply) the prior and likelihood distributions to produce a posterior distribution in a standard closed form that can be described analytically, such as a Normal, Binomial, or Weibull distribution. This step is then performed numerically in packages such as PyMC3 [57], using Markov Chain Monte Carlo (MCMC) techniques. MCMC is the most popular and flexible way of solving Bayesian models. However, in special cases it is possible to solve a model analytically. Although these cases are rare, they have the advantage of accuracy and speed. Because uncertainty propagates over multiple years for longitudinal sampling as in this case, it is preferable to work with precisely defined distributions rather than Monte Carlo (MC) datasets when possible. Regarding speed, modern numeric Bayesian solvers can also solve models in seconds or minutes on standard desktop computers. However, when optimisation needs to be done as in this case, solving such a model thousands of times over becomes impractical. Fortunately, longitudinal population survival sampling models can be solved analytically because 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 our 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. Also, let r_t and s_t be the first and second moments defining the Beta distribution. Then:

$$\mathbf{D}_t = \{\mathbf{D}_{t-1}, Y_t\}, \tag{5}$$

and

$$(\hat{\Phi}_t|\mathbf{D}_t) \sim \text{Beta}(r_t + Y_t, s_t + n_t - Y_t), \tag{6}$$

given that the conjugate prior is Beta:

$$(\hat{\Phi}_t|\mathbf{D}_{t-1}) \sim \text{Beta}[r_t, s_t], \tag{7}$$

and the likelihood function is binomial:

$$(Y_t|\hat{\Phi}_t) \propto \begin{pmatrix} n_t \\ Y_t \end{pmatrix} \hat{\Phi}_t^{Y_t}(1 - \hat{\Phi}_t)^{n_t - Y_t}. \tag{8}$$

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.

3.2. Dynamic Generalised Linear Model

The model proposed below is very similar to a 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 1.

This model was derived from West, Harrison, and Migon's model for Television Viewer Ratings [41, 42]. Their model estimated and forecast television viewer ratings based on sampling a population of viewers and asking them if they were aware of a certain product that was advertised that week. By correlating this awareness to the number of advertisements shown (the perturbation in the system), their model 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 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 (1) can be described by the parameters:

$$\boldsymbol{\theta}_t' = (\beta_t, \gamma_t, \Phi_t). \tag{9}$$

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

$$\hat{\Phi}_t = \mathbf{F}\boldsymbol{\theta}_t \tag{10}$$

where the regression vector is

$$\mathbf{F} = (0, 0, 1). \tag{11}$$

The state vector evolves according to

$$\boldsymbol{\theta}_t = \mathbf{G}_t\boldsymbol{\theta}_{t-1}, \tag{12}$$

however, in this special case

$$\boldsymbol{\theta}_t = \mathbf{g}_t(\boldsymbol{\theta}_{t-1}). \tag{13}$$

Letting $'$ denote a transponent, so that column vectors can be written as row vectors, $\mathbf{g_t(z)}$ and $\mathbf{G_t(z)}$ are defined as follows. Using (1),

$$\mathbf{g}_t(\mathbf{z})' = (\beta, \gamma, \beta\gamma\Phi^2 - \beta\gamma + \Phi) \tag{14}$$

and

$$\mathbf{G}_t(\mathbf{z})' = \begin{bmatrix} 1 & 0 & \gamma\Phi^2 - \Phi \\ 0 & 1 & \gamma\Phi^2 \\ 0 & 0 & 2\beta\gamma\Phi - \beta + 1 \end{bmatrix}. \tag{15}$$

3.2.1. Information Step

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

$$(\boldsymbol{\theta}_{t-1}|\mathbf{D}_{t-1}) \sim [\mathbf{m}_{t-1}, \mathbf{C}_{t-1}], \tag{16}$$

where $\boldsymbol{m}_{t-1}$ is the mean vector estimate of $\boldsymbol{\theta}$, taken from the previous time step's m_t, defined in (39). $\mathbf{C}$ is the covariance matrix. The current state given past data is defined as:

$$(\boldsymbol{\theta}_t|\mathbf{D}_{t-1}) \sim [\mathbf{a}_t, \mathbf{R}_t] \tag{17}$$

with the mean vector of the prior as

$$\mathbf{a}_t = \mathbf{g}_t(\mathbf{m}_{t-1}), \tag{18}$$

as in (13). The variance matrix of the prior is defined as

$$\mathbf{R}_t = \mathbf{G}_t\left(\mathbf{C}_{t-1}\right)\mathbf{G}_t' + \mathbf{W}_t, \tag{19}$$

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

$$\mathbf{W}_t = \mathbf{G}_t\mathbf{U}_t\mathbf{G}_t', \tag{20}$$

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

$$\mathbf{U}_t = 0.03\mathbf{C}_{t-1}, \tag{21}$$

and

$$\mathbf{G}_t = \mathbf{G}_t(\boldsymbol{\theta}_t)_{\boldsymbol{\theta}_t=\mathbf{m}_{t-1}}. \tag{22}$$

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

$$\begin{pmatrix} \mu_t \\ \boldsymbol{\theta}_t \end{pmatrix} \middle| \mathbf{D}_{t-1} \sim \left[\begin{pmatrix} f_t \\ \mathbf{a}_t \end{pmatrix}, \begin{pmatrix} q_t & \mathbf{F}_t'\mathbf{R}_t \\ \mathbf{R}_t\mathbf{F}_t & \mathbf{R}_t \end{pmatrix}\right], \tag{23}$$

where the forecast mean of the information step is

$$f_t = \mathbf{F}_t'\mathbf{a}_t, \tag{24}$$

in accordance with (10), and the forecast variance of the information step is

$$q_t = \mathbf{F}_t'\mathbf{R}_t\mathbf{F}_t. \tag{25}$$

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

$$(\mu_t|\mathbf{D}_{t-1}) \sim \text{Beta}[r_t, s_t], \tag{26}$$

which implies that

$$f_t = \text{E}[\mu_t|\mathbf{D}_{t-1}] = \frac{r_t}{r_t + s_t}, \tag{27}$$

and

$$q_t = \text{V}[\mu_t|\mathbf{D}_{t-1}] = \frac{f_t(1-f_t)}{r_t + s_t + 1}. \tag{28}$$

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 [42] invert (27) and (28) so that:

$$r_t = f_t\left[\frac{f_t(1-f_t)}{q_t} - 1\right] \tag{29}$$

$$s_t = (1-f_t)\left[\frac{f_t(1-f_t)}{q_t} - 1\right]. \tag{30}$$

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

$$f_t = \psi(r_t) - \psi(s_t) \tag{31}$$

and

$$q_t = \dot{\psi}(r_t) + \dot{\psi}(s_t) \tag{32}$$

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. We use the formulae

$$r_t = \left(\frac{1 - f_t}{q_t} - \frac{1}{f_t}\right) f_t^2 \tag{33}$$

$$s_t = \frac{\left(q_t + f_t^2 - f_t\right)(f_t - 1)}{q_t} \tag{34}$$

This then is the prior used for the Bayesian analysis.

3.2.2. Updating Step

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

$$(\hat{\Phi}_t|\mathbf{D}_t) \sim [f_t^*, q_t^*], \tag{35}$$

with the update step forecast mean and variances defined as

$$f_t^* = \mathrm{E}[\hat{\Phi}_t|\mathbf{D}_t] = \frac{r_t + Y_t}{r_t + s_t + n_t}, \tag{36}$$

and

$$q_t^* = \mathrm{VAR}[\hat{\Phi}_t|\mathbf{D}_{t-1}] = \frac{f_t^*(1 - f_t^*)}{r_t + s_t + n_t + 1}. \tag{37}$$

The posterior moments can then be updated as follows:

$$(\boldsymbol{\theta}_t|\mathbf{D}_t) \sim [\mathbf{m}_t, \mathbf{C}_t], \tag{38}$$

where

$$\mathbf{m}_t = \mathbf{a}_t + \mathbf{R}_t\mathbf{F}_t\left(\frac{f_t^* - f_t}{q_t}\right) \tag{39}$$

and

$$\mathbf{C}_t = \mathbf{R}_t - \mathbf{R}_t\mathbf{F}_t\mathbf{F}_t'\mathbf{R}_t\left(\frac{1 - q_t^*}{q_t^2}\right). \tag{40}$$

This method is called Linear Bayesian Estimation.

3.2.3. Forecasting Step

For forecasting to year k given the data at t,

$$(\boldsymbol{\theta}_{t+k}|\mathbf{D}_t) \sim [\mathbf{a}_t(k), \mathbf{R}_t(k)] \tag{41}$$

with $\mathbf{a}_t(k)$ as in (18). $\mathbf{R}_t$ is calculated as follows:

$$\mathbf{R}_t(k) = \mathbf{G}_t(k)\mathbf{R}_t(k-1)\mathbf{G}_t'(k) + \mathbf{W}_t(k) \tag{42}$$

where $\mathbf{W}$ is the evolution variance matrix, defined as

$$\mathbf{W}_t(k) = \mathbf{G}_t(k)(\mathbf{C}_t + \mathbf{U}_t)\mathbf{G}_t'(k), \tag{43}$$

Similarly, $f_t(k)$ and $q_t(k)$ are calculated according to (24) and (25). The parameters of the posterior (forecast) distribution can be calculated through (26), (29), and (30).

3.2.4. Confidence/Credible Interval Estimation

Confidence intervals may also be calculated using the posterior Beta distribution. Strictly speaking, the frequentist confidence interval does not confer a degree of belief or probability as is popularly supposed. Rather, it is the product of a process that produces a quasi-randomly bounded interval containing the true parameter a given percentage of the time [55, 58, 59]. However, the Bayesian credible interval does provide what is conventionally sought, and is produced by the DGLM.

As in (7), the distribution on μ_t is beta. However, in (7) $\mathbf{D}_t$ 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 (29) and (30),

$$r_t^* = f_t^*\left[\frac{f_t^*(1 - f_t^*)}{q_t^*} - 1\right] \tag{44}$$

$$s_t^* = (1 - f_t^*)\left[\frac{f_t^*(1 - f_t^*)}{q_t^*} - 1\right]. \tag{45}$$

Because the probability distributions are asymmetric, an equal-tailed confidence interval will not capture the true nature of the system. Consider a 90% confidence interval on a probability distribution function. In the equal-tailed case, the bottom 5% and top 5% are cut off, and the middle 90% represents the most likely values. However, if the distribution is skewed, it may be the case that the lower confidence limit (LCL) is more or less likely than the upper confidence limit (UCL). In such a case the equal-tailed interval will contain 90% of the values, but not the most likely 90%. We therefore opt to use the Highest Density Interval (HDI). This is sometimes called the Highest Density Region or Highest Probability Density (HPD) interval. For unimodal distributions this is defined as the interval containing 90% of the most likely values, such that the upper and lower bounds of this interval are equally likely. For a skewed distribution this interval will also be skewed.

However, the HDI on the population proportion estimate is not of specific interest for this study. Rather, 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 4.2.

3.3. Optimisation

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 optimisation model. In this section we will formulate such a model.

Two approaches may be followed. The simpler approach would be to forecast one step ahead to $t + 1$ by considering past data, and then determining the sample size n_{t+1} needed to meet the reporting requirements for confidence and precision at that time point. However, this is not an optimal approach. A more comprehensive method would be to consider all future reporting points simultaneously, and to produce a multi-year sampling plan, denoted $\mathbf{n}$. In this sampling plan, sample sizes at different points in time could be traded off against one another for

cost, while still adhering to the reporting requirements. Below we will present and then explain the mathematical optimisation model:

3.3.1. Notation

Let:

d_t	$\begin{cases} 1 & \text{if } n_t > 0 \\ 0 & \text{otherwise} \end{cases}$.
n_t	Decision Variable. Sample size at time t.
$n_{benchmark}$	Non-DGLM solution at time t
v	Cost for initiating sampling at any given time point, in Rand/sampling study
w	Cost per sample in Rand/sample
τ	Present time, where $\tau \in \{1, 2, ..., N\}$
N	Last year of study
e_t	Precision of reported population proportion at time t, where $e_t \in [0, 1]$
$r(\mathbf{n})$	Penalty constraint function
ε	Given precision limit, where $\varepsilon \in [0, 1]$
χ	Number of sampling points where $e_t > \varepsilon$
$\mathbf{M}$	Required reporting points (years), where $\mathbf{M} \subset \{\tau + 1, \tau + 2, ..., N\}$
$\hat{\Phi}_t$	Estimated population proportion surviving at t
LCL_t	Lower Confidence Limit of Highest Density Interval at t

3.3.2. Mathematical formulation

From the notation above, the fitness function can be defined as

$$\min \sum_{t=\tau}^{N} d_t v + n_t w + r(\mathbf{n}), \tag{46}$$

where

$$r(\mathbf{n}) = \sum_{i \in \mathbf{M}} \left(10^5 w(e_t - \varepsilon) + 10^7 + 5wn_{benchmark,\, i}\right) \ \forall t \in \chi \tag{47}$$

and

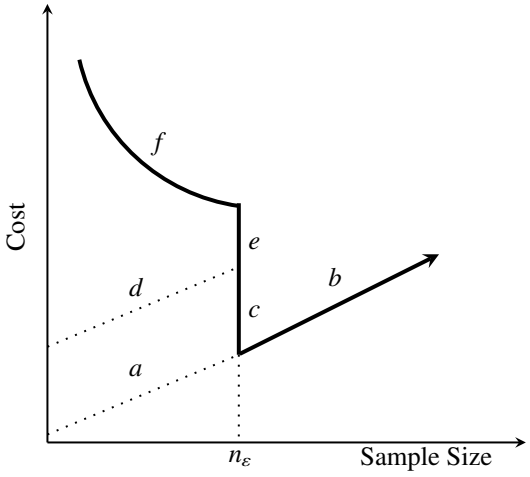
$$e_t = \frac{\hat{\Phi}_t - LCL_t}{\hat{\Phi}_t}. \tag{48}$$

3.3.3. Model Explanation

The proposed fitness function evaluates a particular sampling regime $\mathbf{n}$ based on a number of factors. The first two have real financial significance: the cost w per sample, and a cost v of initiating sampling for a given time point. Thus, if $w = 10$ and $v = 1\,000$, 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 [60, 61]. Levy and Lemeshow do not consider sampling initiation costs [62], while Hansen describes a more thorough approach where costs may vary per stratum [63]. Barnett [60] adds a term for travelling required between samples.

The single constraint on the GA is that the specified statistical confidence and precision levels be adhered to for every reporting point. This summarises the DGLM method of Section 3.2, but is specified as a penalty rather than a hard constraint for reasons given in Section 3.3.4. Consider Figure 2. If there were no constraint, the cost would increase with wn along line ab, and the GA would optimise 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 population of solutions) violate the constraint (placing it on d), the algorithm would tend to optimise *away* from the constraint boundary in the wrong direction. Mutation could transport an individual to b, but it is inefficient to rely solely on this mechanism. Therefore line fe is needed to direct solutions *towards* the constraint rather than away from it.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 optimise 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 therefore built in to ensure that constraint violation is always costly, where $n_{m,\, benchmark,\, i}$ is defined. This step is represented by line ce. We can be sure that it is at or smaller than $n_{benchmark}$: the solution, had no DGLM been used (on which see Section 4.4). Line e compensates for this uncertainty

Figure 2: Genetic Algorithm constraint function $r(\mathbf{n})$ in (47), where n_ε represents the threshold sample size.



The parameter values for the DGLM were set as indicated in Table 2.

3.3.4. Genetic Algorithm

Because only whole-number solutions for sample sizes are valid, and due to the complex nature of the DGLM, this is a discontinuously constrained integer non-linear problem (INLP). As such, the gradient-based methods used for standard optimisation are unsuitable, and a heuristic method is needed.

Table 2: Case Study 1 Model Parameters. $\sim N()$ 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

Because the sampling design solutions are defined as a sequence of numbers, the GA is particularly suited to the problem at hand [64]. The details of the GA are beyond the scope of this paper, but the method can be summarised as follows. Each solution $\mathbf{n}$ 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 optimisation). 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 (46).

3.4. Risk-Conscious Sampling Design

The optimisation described above is for the lowest cost sampling plan. Under this strategy, it is assumed that the outcomes of future surveys fall exactly 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, and the UMP Chapter 2 recommends 10%-30% oversampling [65]. 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. Since it would be simple to enlarge the sample size by urveying additional units, this scenario will not be detrimental to the M&V study. 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 optimisation routine. Nevertheless, the robustness of the current year's required sample size is still of concern. Fortunately, the Bayesian approach is well-suited to risk-conscious sampling design.

From (26), the $\alpha\%$ confidence interval given $\mathbf{D}_{t-1}$ is known. From this, $\hat{\Phi}_{t,\alpha}$ can be determined: population proportion at the upper and lower bounds of the $\alpha\%$ HDI. The DGLM can then be used to simulate the resultant confidence intervals for different sample sizes which all have the $\hat{\Phi}_{t,\alpha}$ survey outcome. In such a manner, the most economical α-robust sample size n_t can be determined. That is, the sample size at which it is $\alpha\%$ certain that the post-sampling confidence interval, given unexpected results, will still have a savings estimate that meets the M&V reporting accuracy requirements. Note that $n_{t,\alpha}$ will probably oversample in the present year since such an outcome is unlikely. However, this sample size will be traded off against future sample sizes, which will then become smaller.

Although this study focusses on optimal sampling solutions, a robust step-ahead algorithm was programmed. Preliminary results show that the robust sample size is not radically different from the optimal sample size. In many cases the method described above does not alter the optimal sampling plan. However, this would depend on the number of previous sampling points, and the sizes of those samples.

4. Case Study 1: Homogeneous Population

4.1. Data

The first case study considered is for a single unstratified population of Compact Fluorescent Lamps which were installed in a lighting retrofit project. The survival data used for this case study is the PELP dataset [25] discussed above. The parameters for (2) were determined to be $[\gamma, \beta, L] = [1.030, 1.056, 5.233]$, so that

$$\hat{\Phi}_{t,sim} = \frac{1}{1.030 + e^{1.056\times t - 5.233}}. \tag{49}$$

A realistic way to use this data would be to generate data points according to $\mathbf{D}_{sim,\,t} \sim Binomial(n = n_t, p = \hat{\Phi}_{t,\,sim})$. 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. We therefore opt to use the PELP data points. Smaller sample sizes will still have larger variances in the DGLM, the way real data would.

For this study, we assume that data has been collected for the first three years, and suppose 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 $\mathbf{M} = \{3, 5, 4, 6\}$. The contract stipulates that reports should have a 90% confidence and 10% precision on the estimate, as per CDM [66] 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]. We therefore consider the 10% precision bound 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. If we suppose 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:

$$E_{saved,\,t} = n_{retrofitted} HOU(0.99P_b - P_r)\hat{\Phi}_t, \qquad (50)$$

Where HOU, is the distribution on daily hours of use and P is the distribution on the power drawn by the old (P_b) and new (P_r) units. The number of lamps retrofitted is $n_{retrofitted}$. We do not consider interactive heating and cooling effects or the in-service rate [17].

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 [26], others 1.75% [67], while others report in the order of ±0.5% in actual operating conditions of fluorescent lamps and pre-retrofitted fixtures generally [68]. We note that 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 [17]. A 2.5% error was selected.

The values used for these distributions are summarised in Table 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 [69] 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. This accords with the CDM assumption [29]. The data are distributed as

$$\text{HOU study estimates} = 2+ \sim Exponential(0.893). \qquad (51)$$

(Another meta-study disaggregated lamp HOUs, and found lower numbers [70]. For information on CFL HOUs disaggregated per installation location, see [71].)

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, we will assume a 4% uncertainty on the HOU. We also assume that the HOU stay the same between the baseline and the reporting period. Should snapback or rebound [17] be proven, two different HOU terms could be defined for the baseline and reporting periods.

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 so numeric Monte Carlo convolution would be the standard way of calculating (50). However, this has two related disadvantages, similar to those discussed in Section 3.1.2. 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 outliers due to noise, are evaluated to be the fittest individuals because they seem to adhere to the constraints, when this is just an artifact 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 sampling plan would not conform to the accuracy requirements. Although these 'false positives' happen rarely, they mislead the algorithm by incorrectly altering the ranking of good solutions. One way to counter this phenomenon would be to increase the number of MC trials to 10^7 or higher. However, it is very expensive to convolve such large datasets for each individual in each generation, and the GA approach then becomes impractical.

The solution to this problem is to calculate the E_{saved} in (50) analytically. This is usually thought to be very difficult or impossible. However, recent work by Kuang and Rajan et. al. [72, 73] have produced a method by which the moments of the posterior of such a convolution may be expressed in terms of the scale and shape parameters of the constituent distributions. This allows for exact expressions of the first four 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 (MTMC) [74]. 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. A Johnson S_B (bounded) distribution [75] 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 [76] to quality [77]. For more information on uncertainty evaluation through moment-based distribution fitting, see Rajan et al. [78]. The Johnson distribution is fitted using Hill's algorithm [79], implemented in Matlab/Octave [80] and then linked to the DGLM in Python.

4.3. Specification of Initial Estimates for DGLM Optimisation

The initial conditions for $\boldsymbol{\theta}'$ in (9) 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.

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 improve the GA effectiveness. The first is that the initial 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, since the GA was found to converge on good sampling patterns, but not to optimize those patterns to the precision limits.

The Python library `DEAP` (Distributed Evolutionary Algorithms in Python) [81] 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 3.

A tournament size of ten supplies a severe selection pressure, and homogenizes 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

Table 3: Genetic Algorithm 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

4.4. Benchmark

The proposed method needs to be benchmarked against a realistic alternative method for solving the problem at hand. We note that Goldberg [9] and the UMP [82] 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. Since we do not assume normality this is not applicable, but otherwise this method is very similar to using Bayesian priors.

First, a note on confidence intervals. Various confidence intervals for binomial proportion sampling have been suggested. Of these, Jeffreys sampling has been shown to be the most accurate and least conservative [83], although less accurate methods are often used in M&V [82]. 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, we therefore define the benchmark 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 and Bayesian Forecasting.

To determine this sample size, $\hat{\Phi}_t$ 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. But like would not be compared to like in this case. A fairer comparison would be to consider the smallest instantaneous sample for the population proportion $\hat{\Phi}_{\mathbf{M},\,DGLM}$ estimated by the DGLM, and find the optimal sample size for that proportion using the GA.

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. A plot of the minimum and average population fitness vs. the

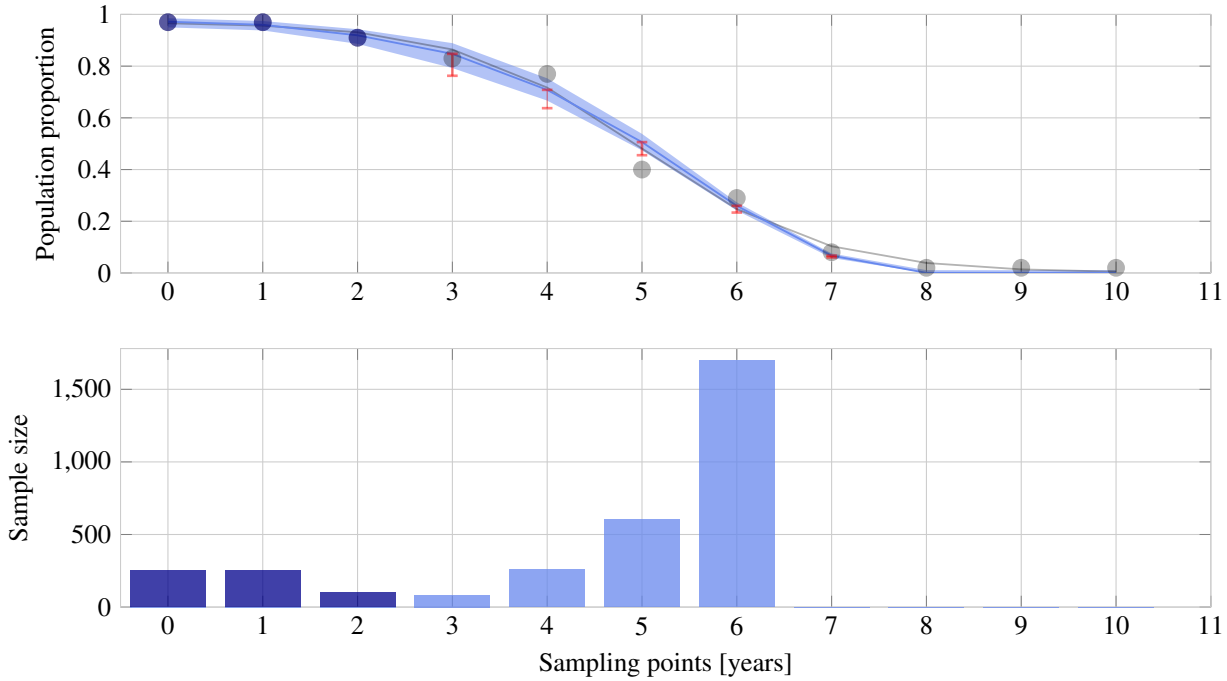


Figure 3: Population proportion inferred from results and sample size data from DGLM for Case Study 1, with forecasting to future years, and optimal sample sizes. The grey lines and points are the PELP data and fits, dark blue past sampling results, and light blue future sample sizes. The 90% confidence area is shaded, and the 10% precision limits indicated by the red error bars.

generations of the GA (not shown due to space constraints) exhibits the classic concave-up shape. The average population fitness decreases rapidly in the first 10 generations, and then approaches a minimum asymptotically in the next 20.

We assume 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

$$\mathbf{n} = [83, 260, 601, 1696, 0, 0, 0], \tag{52}$$

at a cost of R30 400. The benchmark for this realisation is

$$\mathbf{n}_{benchmark} = [153, 303, 707, 2101, 0, 0, 0], \tag{53}$$

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 priors (r_t and s_t) decrease during the forecasting step, as forecasting without data decreases the contribution of prior information as uncertainty increases. 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 3 and Figure 4 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 subsequent to $t-k$. This allows for the updating of past stimates, should that be necessary. These are not shown.

A number of features of the results plotted in Figure 3 warrant attention. It is clear that the actual population proportion result from the survey sample may be different from the true population proportion. We have tried to illustrate this 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 also is 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 3 and Figure 4 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 (50), it is expected that the algorithm constrains the population proportion estimate to less than the 90/10 bound, to meet the overall

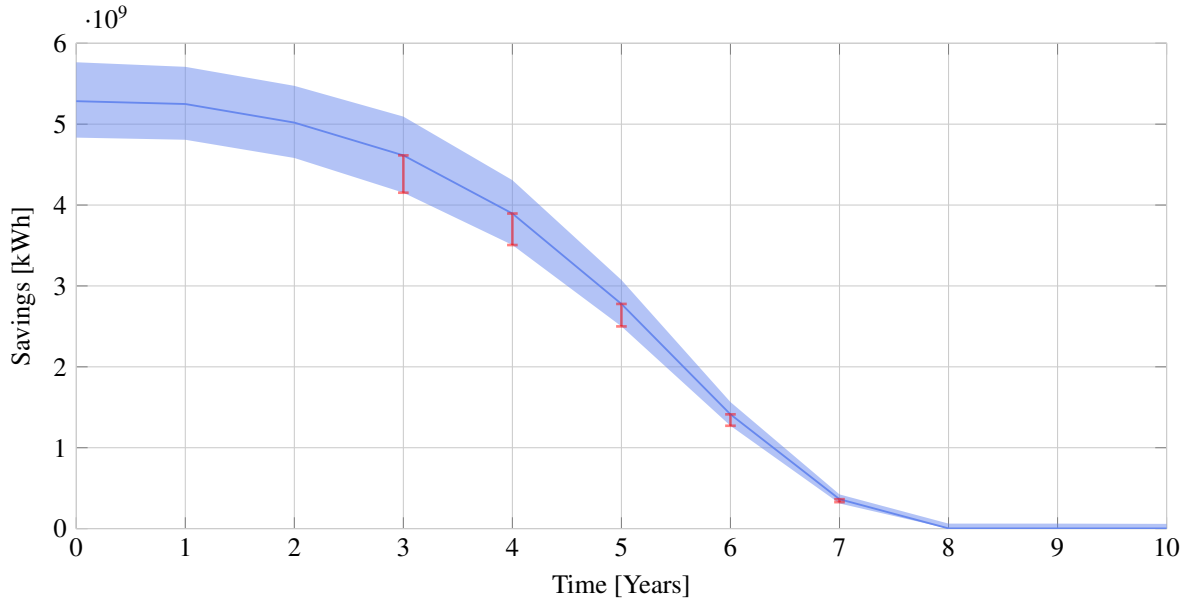


Figure 4: Savings inferred by DGLM using (50) for Case Study 1.

savings 90/10 bound. The overall uncertainty is plotted in Figure 4. 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 uncertainty reporting requirement becomes more stringent as the savings decrease. This leads to a situation where monitoring smaller savings require greater resources [9]. 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.

5. Case Study 2: Stratified Population

The second case study considered is for a heterogeneous population: one 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 undersampled 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 well-studied. For example, Barnett [60] lists formulae for different cases, and the UMP Chapter 11 discussed cost-optimal Pearson allocation for M&V in some detail [82]. 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 (50). Therefore we will not use optimum allocation formulae, but allow the GA to find an efficient allocation.

To generate different realistic population survival curves, the data published in the Lighting Research Center (LRC's) Speci-

Table 4: Case Study 2 Model Parameters. Only those parameters that differ from Table 2 are shown. $\sim N()$ 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	$\sim N(3.11, 0.15)$
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

Table 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

fier Report on CFLs [26] 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 proportion of the population surviving after one, two, or three years is reported. To convert the LRC data to a suitable format, least-squares logistic curves were fitted to the data. The data sets which fitted (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 datasets were selected. The binomial uncertainty as discussed in Section 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 so more preliminary data points could be collected.

The total energy saving is the sum of the three populations described by (50):

$$E_{saved,\ total} = \sum_{i=1}^{3} E_{saved,i} \tag{54}$$

For a three-stratum, twelve-year monitoring project, the sampling solution $\mathbf{n}$ is a 3×12 vector. The GA fitness function also needed to be adapted to account for the expanded sampling term and the three separate E_{saved} distributions, and the moment equations and Johnson distribution developed for (50) was adapted. The first four moments of these three distributions are summed using the standard moment summation formulae. 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 the first case study. 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 4.

5.1. Results and Discussion

Much of the discussion in Section 4.5 is also relevant to this case study. The benchmark 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 5. It is graphically illustrated in Figure 5, and the savings curve is shown in Figure 6. The DGLM method therefore presents a significant advantage over simple regression methods, even if these methods still use the Mellin Transform 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. 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 6), the uncertainty bounds are 'pinched' as the DGLM accounts for the increased certainty.

Also note that no samples of stratum 2 are taken after year 5 (the last past-sampling year). In the simple random sampling case (Section 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 6 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. Conclusion

Dynamic Generalised Linear Models 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 Genetic Algorithm and Mellin Transform Moment Calculator 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 17-40%, depending on stratification and how the study costs are calculated.

Acknowledgement

We wish to thank Jiangfeng Zhang who contributed to an early version of this work. The financial assistance of the Na-

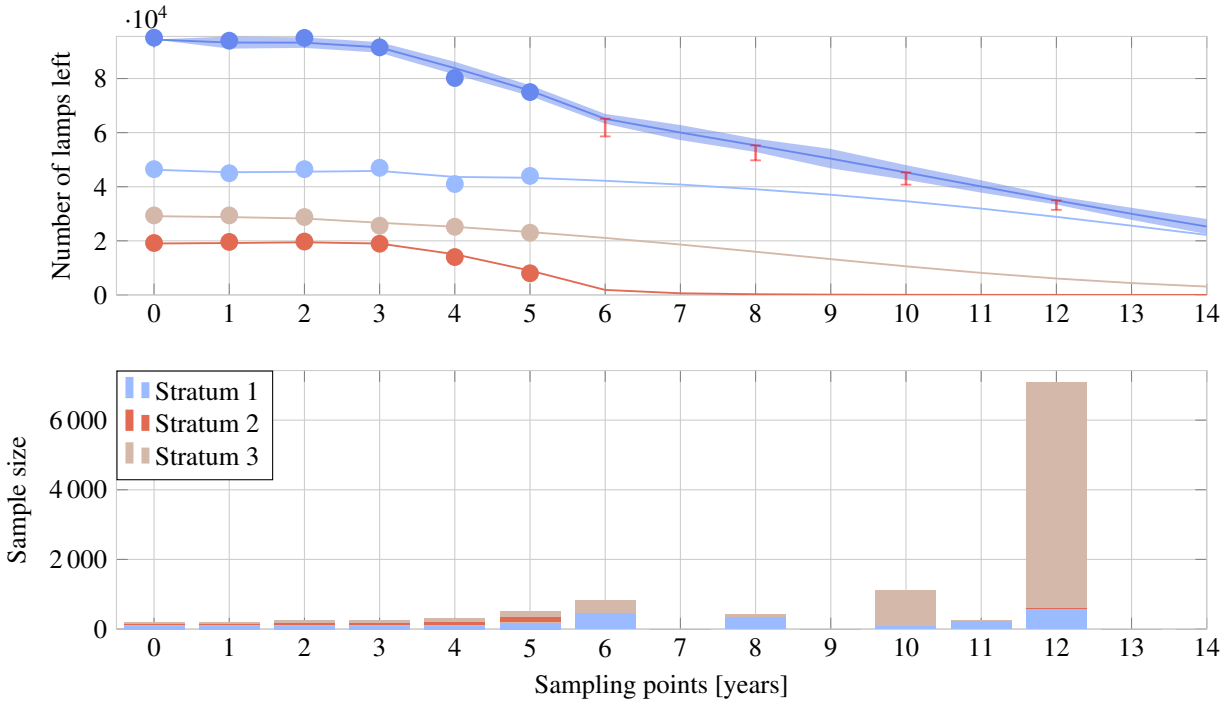


Figure 5: Population proportion inferred from results and sample size data from DGLM for Case Study 2, with forecasting to future years, and optimal sample sizes. Sampling is done up to and including year 5.

ional Hub for the Postgraduate Programme in Energy Efficiency and Demand Side Management towards this research is also acknowledged.

References

[1] Efficiency Valuation Organization, International Performance Measurement and Verification Protocol Vol. 1 (January 2012).

[2] I. Shishlov, V. Belassen, Review of monitoring uncertainty requirements in the CDM, Tech. Rep. Working Paper No. 2014-16, CDC Climat Research (October 2014).

[3] X. Xia, J. Zhang, Mathematical description for the measurement and verification of energy efficiency improvement, Applied Energy 111 (2013) 247–256.

[4] H. Carstens, X. Xia, S. Yadavalli, Measurement uncertainty in energy monitoring: Present state of the art, submitted for Review on September 6, 2016 (August 2016).

[5] H. Carstens, X. Xia, S. Yadavalli, Low-cost energy meter calibration method for measurement and verification, Applied Energy 188. doi:10.1016/j.apenergy.2016.12.028.

[6] American Society of Heating, Refrigeration and Air-Conditioning Engineers, Inc., Guideline 14-2014, Measurement of Energy, Demand, and Water Savings (December 2014).

[7] Z. Olinga, A cost effective approach to handle measurement and verification sampling and modelling uncertainties, Master's thesis, University of Pretoria (2016).

[8] X. Xia, J. Zhang (Eds.), Measurement and Verification Practices: Demystifying M&V Through South African Case Studies, Media in Africa, 2012.

[9] M. L. Goldberg, Reasonable doubts: Monitoring and verification for performance contracting, in: ACEEE Summer Study on Energy Efficiency in Buildings, Vol. 4, American Council for an Energy Efficient Economy, Pacific Grove, California, 1996, pp. 133–143.

[10] H. Carstens, X. Xia, S. Yadavalli, Measurement uncertainty and risk in measurement and verification projects, in: International Energy Programme Evaluation Conference, Long Beach, California, 2015.

[11] M. S. Khawaja, J. Stewart, Long-run savings and cost-effectiveness of home energy report programs, Tech. rep., Cadmus (Winter 2015).

[12] H. Carstens, X. Xia, X. Ye, Improvements to longitudinal Clean Development Mechanism sampling designs for lighting retrofit projects, Applied Energy (126) (2014) 256–265. doi:10.1016/j.apenergy.2014.03.049.

[13] B. Bronfman, H. Michaels, G. Fitzpatrick, S. Nadel, E. Hicks, J. Peters, E. Hirst, J. Reed, M. Hoffman, W. Saxonis, A. Schoh, K. Keating, D. Violette, Handbook of evaluation of utility dsm programs, Tech. Rep. ORNL/CON-336, Oak Ridge National Laboratory (December 1991). doi:10.2172/10120182.

[14] D. M. Grueneich, The next level of energy efficiency: The five challenges ahead, The Electricity Journal 28 (7) (2015) 44–56. doi:10.1016/j.tej.2015.07.001.

[15] I. M. Hoffman, S. R. Schiller, A. Todd, M. A. Billingsley, C. A. Goldman, L. C. Schwartz, Energy savings lifetimes and persistence: Practices, issues and data, Technical Brief LBNL-179191, Lawrence Berkeley National Laboratory (May 2015). doi:10.2172/1236443.

[16] L. A. Skumatz, M. S. Khawaja, 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, Tech. rep., SERA (November 2009).

[17] S. Dimetrosky, The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures, National Renewable Energy Laboratory, 2013, Ch. 6: Residential Lighting Evaluation Protocol.

[18] N. Nord, S. F. Sjøthun, Success factors of energy efficiency measures in buildings in norway, Energy and Buildings 76 (2014) 476–487. doi:10.1016/j.enbuild.2014.03.010.

[19] E. L. Vine, Persistence of energy savings: What do we know and how can it be ensured?, Energy 17 (11) (1992) 1073–1084. doi:10.1016/0360-5442(92)90024-T.

[20] D. Violette, The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures, National Renewable Energy Laboratory, 2013, Ch. 13: Assessing Persistence and Other Evaluation Issues in Cross-Cutting Protocols.

[21] J. Proctor, T. Downey, Summary report of persistence studies: Assess-

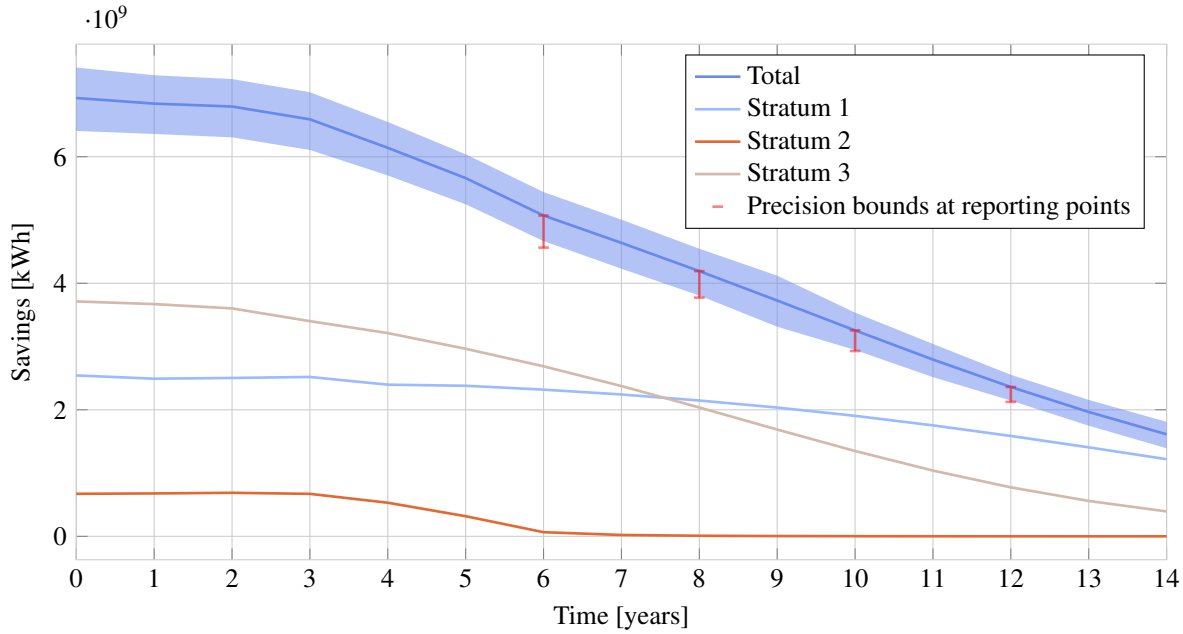


Figure 6: Savings inferred by DGLM using (50) for Case Study 2.

ments of technical degradation factors, Tech. rep., Proctor Engineering (February 1999).

[22] State and Local Energy Efficiency Action Network, Energy Efficiency Program Impact Evaluation Guide, US Department of Energy, DOE/EE-0829 (December 2012).

[23] L. A. Skumatz, D. Whitson, D. Thomas, K. Geraghty, B. Dunford, K. Lorberau, C. Breckinridge, Measure life study II, Tech. Rep. SRC 7851-R1, Bonneville Power Administration (July 1994).

[24] Nexus Market Research, Inc., RLW Analytics, Inc., Residential lighting measure life study, Tech. rep., New England Residential Lighting Program (2008).

[25] Navigant Consulting, Evaluation of the IFC/GEF Poland Efficient Lighting Project CFL Subsidy Program, Tech. Rep. 1, Netherlands Energy Efficient Lighting B.V., International Finance Corporation / Global Environment Facility (1999).

[26] Lighting Research Center, Screwbase compact fluorescent lamp products, Specifier Report 7 (1), Rensellaer Polytechnic Institute (June 1999).

[27] National Renewable Energy Laboratory, Uniform Methods Project. URL http://energy.gov/eere/about-us/ump-home

[28] D. Violette, R. Brakken, A. Shon, 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.

[29] United Nations Framework Convention for Climate Change, Approved Small-Scale Methodology AMS II.J, Demand-Side Activities for Efficient Lighting Technologies.

[30] M. Botha-Moorlach, 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, Tech. rep., Eskom (March 2009).

[31] X. Ye, X. Xia, J. Zhang, Optimal sampling plan for clean development mechanism lighting projects with lamp population decay, Applied Energy 136 (2014) 1184–1192. doi:10.1016/j.apenergy.2014.07.056.

[32] P. O'Connor, A. Kleyner, Practical Reliability Engineering, Wiley, 2011. doi:10.1002/9781119961260.

[33] D. Young, When do energy-efficient appliances generate energy savings? Some evidence from Canada, Energy Policy 36 (1) (2008) 34–46. doi:10.1016/j.enpol.2007.09.011.

[34] United Nations Framework Convention for Climate Change, Clean Development Mechanism Methodology Booklet (November 2015).

[35] T. Clark, M. Bradburn, S. Love, D. Altman, Survival analysis part I: basic concepts and first analyses, British Journal of Cancer 89 (2) (2003) 232. doi:10.1038/sj.bjc.6601118.

[36] M. Bradburn, T. Clark, S. Love, D. Altman, Survival analysis part II: Multivariate data analysis–an introduction to concepts and methods, British journal of cancer 89 (3) (2003) 431. doi:10.1038/sj.bjc.6601119.

[37] M. Bradburn, T. Clark, S. Love, D. Altman, Survival analysis part III: Multivariate data analysis–choosing a model and assessing its adequacy and fit, British Journal of Cancer 89 (4) (2003) 605. doi:10.1038/sj.bjc.6601120.

[38] T. Clark, M. Bradburn, S. Love, D. Altman, Survival analysis part IV: further concepts and methods in survival analysis, British Journal of Cancer 89 (5) (2003) 781. doi:10.1038/sj.bjc.6601117.

[39] F. Tekle, F. Tan, M. Berger, Maximin d-optimal designs for binary longitudinal responses, Computational Statistics & Data Analysis 52 (12) (2008) 5253–5262. doi:10.1016/j.csda.2008.04.037.

[40] Y. Zhang, Z. O'Neill, B. Dong, G. Augenbroe, Comparisons of inverse modeling approaches for predicting building energy performance, Building and E 86 (2015) 177–190. doi:10.1016/j.buildenv.2014.12.023.

[41] M. West, P. J. Harrison, H. S. Migon, Dynamic generalized linear models and Bayesian forecasting, Journal of the American Statistical Association 80 (389) (1985) 73–83. doi:10.1080/01621459.1985.10477131.

[42] J. Harrison, M. West, Bayesian Forecasting & Dynamic Models, Springer, 1999. doi:10.1007/b98971.

[43] P. McCullagh, J. A. Nelder, Generalized linear models, Vol. 37, CRC press, 1989.

[44] K. Triantafyllopoulos, Inference of dynamic generalized linear models: On-line computation and appraisal, International Statistical Review 77 (3) (2009) 430–450. doi:10.1111/j.1751-5823.2009.00087.x. URL http://www.jstor.org.uplib.idm.oclc.org/stable/27919767

[45] D. Gamerman, M. West, An application of dynamic survival models in unemployment studies, The Statistician (1987) 269–274doi:10.2307/2348523.

[46] D. Gamerman, Dynamic Bayesian models for survival data, Applied Statistics (1991) 63–79doi:10.2307/2347905.

[47] E. R Brown, J. G Ibrahim, A Bayesian semiparametric joint hierarchical model for longitudinal and survival data, Biometrics 59 (2) (2003) 221–228. URL http://www.jstor.org.uplib.idm.oclc.org/stable/3695499

[48] D. Gamerman, H. S. Migon, Dynamic hierarchical models, Journal of

the Royal Statistical Society. Series B (Methodological) (1993) 629–642. URL http://www.jstor.org.uplib.idm.oclc.org/stable/2345875

[49] D. Violette, Impact evaluation accuracy and the incorporation of prior information, in: Energy Program Evaluation Conference, Chicago, 1991, pp. 86–92.

[50] B. Wang, X. Xia, J. Zhang, A multi-objective optimization model for the life-cycle cost analysis and retrofitting planning of buildings, Energy and Buildings 77 (2014) 227–235. doi:10.1016/j.enbuild.2014.03.025.

[51] Z. Wu, B. Wang, X. Xia, Large-scale building energy efficiency retrofit: Concept, model and control, Energy 109 (2016) 456–465. doi:10.1016/j.energy.2016.04.124.

[52] X. Ye, X. Xia, L. Zhang, B. Zhu, Optimal maintenance planning for sustainable energy efficiency lighting retrofit projects by a control system approach, Control Engineering Practice 37 (2015) 1–10.

[53] X. Ye, X. Xia, Optimal metering plan for measurement and verification on a lighting case study, Energy 95 (2016) 580–592.

[54] H. Carstens, X. Xia, X. Ye, J. Zhang, Characterising compact fluorescent lamp population decay, in: IEEE Africon Conference, Pointe-Aux-Piments, Mauritius, 2013. doi:10.1109/AFRCON.2013.6757715.

[55] J. Kruschke, Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan, 2nd Edition, Academic Press, 2015.

[56] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Rubin, Bayesian Data Analysis, Vol. 2, Taylor & Francis, 2014.

[57] J. Salvatier, T. V. Wiecki, C. Fonnesbeck, Probabilistic programming in Python using PyMC3, PeerJ Computer Science 2. doi:10.7717/peerj-cs.55.

[58] 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 236 (767) (1937) 333–380.

[59] R. Kacker, A. Jones, On use of Bayesian statistics to make the guide to the expression of uncertainty in measurement consistent, Metrologia 40 (5) (2003) 235. doi:10.1088/0026-1394/40/5/305.

[60] V. Barnett, Sample Survey: Principles and Methods, Arnold, 2002.

[61] A. K. Gupta, D. G. Kabe, Theory of sample surveys, World Scientific, 2011.

[62] P. S. Levy, S. Lemeshow, Sampling of populations: methods and applications, John Wiley & Sons, 2013.

[63] M. H. Hansen, W. N. Hurwitz, W. G. Madow, Sample Survey Methods and Theory, Vol. 1, John Wiley & Sons, 1953.

[64] M. Mitchell, An introduction to genetic algorithms, MIT press, 1998.

[65] D. Gowans, The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures, National Renewable Energy Laboratory, 2013, Ch. 2: Commercial and Industrial Lighting Evaluation Protocol.

[66] United Nations Framework Convention for Climate Change, Approved Small-Scale Methodology AMS II.C, Demand-Side Activities for Specific Technologies.

[67] L. Guan, T. Berrill, R. J. Brown, Measurement of actual efficacy of compact fluorescent lamps (CFLs), Energy and Buildings 86 (2015) 601–607. doi:10.1016/j.enbuild.2014.10.068.

[68] A. H. Lee, Verification of electrical energy savings for lighting retrofits using short-and long-term monitoring, Energy conversion and management 41 (18) (2000) 1999–2008. doi:10.1016/S0196-8904(00)00037-6.

[69] E. Vine, D. Fielding, An evaluation of residential CFL hours-of-use methodologies and estimates: Recommendations for evaluators and program managers, Energy and buildings 38 (12) (2006) 1388–1394. doi:10.1016/j.enbuild.2005.09.008.

[70] W. Gifford, M. Goldberg, P. Tanimoto, D. Celnicker, M. Poplawski, Residential lighting end-use consumption study: Estimation framework and initial estimates, Tech. Rep. PNNL-22182, United States Department of Energy (2012). doi:10.2172/1162372.

[71] C. Jump, J. J. Hirsch, J. Peters, 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, American Council for an Energy Efficient Economy, Pacific Grove, California, 2008, pp. 138–149.

[72] Y. C. Kuang, A. Rajan, M. P.-L. Ooi, T. C. Ong, Standard uncertainty evaluation of multivariate polynomial, Measurement 58 (2014) 483–494. doi:10.1016/j.measurement.2014.09.022.

[73] A. Rajan, M. P.-L. Ooi, Y. C. Kuang, S. N. Demidenko, Analytical standard uncertainty evaluation using Mellin transform, Access, IEEE 3 (2015) 209–222. doi:10.1109/ACCESS.2015.2415592.

[74] Mellin transform-based moment calculator for multivariate polynomials, accessed 20 December 2016. URL http://tc32.ieee-ims.org/content/mellin-transform-based-moment-calculator-multivariate-polynomia

[75] N. L. Johnson, Systems of frequency curves generated by methods of translation, Biometrika 36 (1/2) (1949) 149–176.

[76] J.-G. Simonato, The performance of Johnson distributions for computing value at risk and expected shortfall, SSRN (1706409). doi:10.2139/ssrn.1706409.

[77] N. R. Farnum, Using Johnson curves to describe non-normal process data, Quality Engineering 9 (2) (1996) 329–336. doi:10.1080/08982119608919049.

[78] A. Rajan, Y. C. Kuang, M. P.-L. Ooi, S. N. Demidenko, Benchmark test distributions for expanded uncertainty evaluation algorithms, IEEE Transactions on Instrumentation and Measurement 65 (5) (2016) 1022–1034. doi:10.1109/TIM.2015.2507418.

[79] I. Hill, R. Hill, R. Holder, Algorithm AS 99: Fitting Johnson curves by moments, Journal of the Royal Statistical Society. Series C (Applied statistics) 25 (2) (1976) 180–189. doi:10.2307/2346692.

[80] D. Jones, Johnson Curve Toolbox for Matlab: analysis of non-normal data using the Johnson family of distributions, Tech. rep., College of Marine Science, University of South Florida (2014). URL http://www.marine.usf.edu/user/djones/jctm/jctm.html

[81] F.-A. Fortin, F.-M. De Rainville, M.-A. Gardner, M. Parizeau, C. Gagné, DEAP: Evolutionary algorithms made easy, Journal of Machine Learning Research 13 (2012) 2171–2175.

[82] M. S. Khawaja, J. Rushton, J. Keeling, The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures, National Renewable Energy Laboratory, 2013, Ch. 11: Sample Design Cross-Cutting Protocols.

[83] L. Brown, T. Cai, A. DasGupta, Interval estimation for a binomial proportion, Statistical Science 16 (2) (2001) 101–133. URL http://www.jstor.org/stable/2676784