Energy and Buildings 154 (2017) 430–447

Contents lists available at ScienceDirect

Energy and Buildings

journal homepage: www.elsevier.com/locate/enbuild

Efficient metering and surveying sampling designs in longitudinal Measurement and Verification for lighting retrofit

Herman Carstens[a,*], Xiaohua Xia[a], Sarma Yadavalli[b]

[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

ARTICLE INFO

Article history:
Received 13 April 2017
Received in revised form 21 July 2017
Accepted 27 August 2017
Available online 31 August 2017

Keywords:
Bayesian
Retrofit
Dynamic Linear Model
Measurement and Verification
Mellin Transform
Persistence
Metering
Survey
Sampling

ABSTRACT

Measurement and Verification (M&V) is often required for energy efficiency or demand side management projects in buildings, to demonstrate that savings were in fact achieved. For projects where sampling has to be done, these costs can be the most significant driver of the overall M&V project cost, especially in multi-year (longitudinal) projects. This study presents a method for calculating efficient combined metering and survey sample designs for longitudinal M&V of retrofit projects. In this paper, a building lighting retrofit case study is considered. A Dynamic Linear Model (DLM) with Bayesian forecasting is used. The Bayesian component of the model determines the sample size-weighted uncertainty bounds on multi-year metering studies, with results from previous years incorporated into the overall calculation to reduce forecast uncertainty. The DLM is compared to previous meter sampling methods, and an investigation into the robustness of efficient sampling plans is also conducted. The Mellin Transform Moment Calculation method is then used to combine the DLM with a Dynamic Generalised Linear Model describing the uncertainty in survey results for the longitudinal monitoring of lamp population decay. A genetic algorithm is employed to optimise the combined sampling design. Besides the reliable uncertainty quantification features of the method, results show a reduction in sampling costs of 40% for simple random sampling, and approximately 26.6% for stratified sampling, as compared to realistic benchmark methods.

1. Background

Energy Measurement and Verification (M&V) is the process by which energy savings from Energy Efficiency or Demand Side Management (EEDSM) projects (most often implemented for buildings) are independently and reliably quantified [1]. For example, 500,000 Compact Fluorescent Lamps (CFLs) may have replaced their incandescent counterparts in a countrywide residential mass roll-out programme. For such a project to be eligible for tax rebates such as the 12L incentive in South Africa [2] or the United Nations Clean Framework Convention for Climate Change (UNFCCC) Development Mechanism (CDM) programme [3], an M&V team would be asked to quantify the savings realised. The output of an M&V report is an estimate of the energy savings achieved by the project. This figure must usually be reported with regulator-specified degree of statistical precision, which in turn determines the level of monitoring required. The statistical precision is stated in terms of an 'expanded uncertainty', such as 90/10. This means that the 90% confidence bounds on the estimated savings should be within 10% of the mean.

Because the energy saving of a project represents the absence of energy use, it cannot be measured directly. Rather, energy measurements are made or samples are taken during the pre- and post-retrofit periods. An energy model is constructed (or 'trained') using pre-retrofit data, and is then used to predict what the energy use during the post-retrofit period *would have been*, had no intervention taken place. The difference between these values and the measured values is the energy saving.

There are three main uncertainty drivers in such an M&V model which need be accounted for to report savings with realistic statistical precision. These are measurement, sampling, and modelling uncertainty. Controlling these uncertainties can be expensive. In longitudinal studies, metering and sampling uncertainties are the main cost drivers. Many meters need to be installed, and multiple inspectors need to visit geographically diverse sites to install meters and inspect the number of surviving retrofit units. The M&V cost due to minimising metering and sampling uncertainty may even affect the retrofit project feasibility. For example, Michaelowa, Hayashi, and Marr [4] document that no lighting retrofit projects were undertaken under the stringent CDM AM0046 [5] require-

[*] Corresponding author.
E-mail address: herman.carstens@up.ac.za (H. Carstens).

http://dx.doi.org/10.1016/j.enbuild.2017.08.080
0378-7788/© 2017 Elsevier B.V. All rights reserved.

ment. Only when the alternative AMS II.C [6] and AMS II.J [7] were adopted, did M&V stringency requirements allow for project feasibility and significant uptake. The same effect is present in other M&V projects. Therefore, a research gap exists for methods that can design statistically and financially efficient M&V plans: plans which achieve the same precision as other plans, but at a lower cost in terms of units sampled and money spent [8]. Such methods would not only increase M&V accuracy, but also project profitability. Bayesian methods have been recommended for such situations where finances and uncertainty interact [9]. Efficient methods should also consider measurement, sampling, and modelling uncertainty simultaneously, and trade them off against each another. The need for efficient M&V designs is especially acute in multi-year (longitudinal) M&V studies. Although they are also costly themselves, longitudinal studies have been found to reduce the reported cost of savings by up to 70%, compared to single-year M&V studies [10]. In such longitudinal studies, information from previous years could be used to reduce current and future uncertainties in the savings estimates or to reduce sample sizes. Although this is a common problem, it does not have a straightforward solution for efficient sampling design. Research addressing these gaps will, therefore, enhance both the theory and practice of M&V.

As in the example above, this paper will focus on multi-year lamp retrofit projects in which incandescent lamps are replaced by Compact Fluorescent Lamps (CFLs). Lamp retrofit projects are popular in M&V as case studies [1,11–13], since the operation of lamps is simple, they are mostly independent of covariates such as outside air temperature, and they are well-studied; not many technologies have such readily available data on persistence as CFLs do, for example. They, therefore, serve as a useful introduction to a method, which can be extended later to include considerations such as covariates or other complicating factors.

Such longitudinal energy monitoring projects have two components or dimensions that need to be considered when calculating total energy use and uncertainty, and therefore when designing such studies. The first is population survival: establishing how many of the originally installed (retrofitted) units are still effective at a given point in time. This entails survey sampling and has been the focus of previous works [14–20]. The second factor is the average annual energy use per unit. For lighting studies, this can be calculated with measured operational hours by lighting loggers and estimated power use of lamps. In M&V jargon this is called the 'retrofit isolation with key parameter measurement' approach [1]. Alternatively, meters may be installed on a sample of the lighting circuits, which is called 'retrofit isolation with all parameter measurement'. Even though metering is cross-sectional (in the spatial dimension), there is still a longitudinal component in multi-year cross-sectional metering designs. Results up to the previous year's sample should in some way inform the current parameter and uncertainty estimates. This calls for a regression model or a Bayesian approach, both of which will be adopted below.

Once such a model has been constructed, survey sampling results and uncertainties should be combined with metering results and uncertainties to calculate the overall energy use (and savings) estimation, and overall reporting uncertainty. This will result in a more realistic uncertainty value being used for efficient study design. The American Society of Heating, Refrigeration, and Air-conditioning Engineers (ASHRAE's) Guideline 14 on Measurement of Energy, Demand, and Water Savings [21] (henceforth referred to as G14) does provide a method for combining the three kinds of uncertainty mentioned above. However, such a holistic view of M&V uncertainty has not been adopted in the design of efficient M&V methods yet (the literature is discussed below). For example, the 90/10 criterion has previously been taken to apply to sampling uncertainty only, and not to the combined estimated savings figure, incorporating sampling, measurement, and modelling uncertainties. The proposed method integrates these uncertainty drivers in an optimizable manner. It also takes past metering and survey results into account when calculating the current energy use values and uncertainties. Incorporating past data in a mathematically sound yet informative manner has been a problem for M&V sampling design. Past samples in a longitudinal project contain information, both in their results and in their sample sizes. Since uncertainty in the parameter estimates decreases with more information, these past samples can be used to decrease uncertainty in the current estimates. The more information is available from past samples, the less information is needed from present and future samples to meet the uncertainty criteria for reporting. This means that smaller sample sizes may be specified for present and future points, if past data can be used. This increases statistical and financial efficiency. However, applying this information from past samples in a mathematically sound and time-sensitive manner is important. If this can be done, the method can then be used to forecast future uncertainties under different sampling regimes. An optimization algorithm can then be employed to select an efficient regime, thereby minimising M&V costs and increasing project feasibility.

A substantial body of literature about general M&V methods exists. A foundational mathematical description [22] has been provided, but most studies focus on regression methods for baseline determination, and not on sampling. For useful surveys of state-of-the-art regression methods, see Zhang et al. [23] and Granderson et al. [24]. Recently, Ke et al. have used Particle Swarm Optimization (PSO) to reduce modelling uncertainty in a regression problem [25] (although the use of PSO rather than matrix inversion for regression requires further motivation). Tehrani et al. have also used recursive Bayesian regression in a novel way for M&V adjusted baseline forecasting [26], and Shonder and Im [27] have also adopted a Bayesian approach.

Standard statistical sampling theory has been applied to M&V by internationally accepted guidelines. The required sample size is usually expressed in the form

$$n = \frac{CV^2 z^2}{p^2} \tag{1}$$

where p is the relative precision and z is the standard score. Therefore, 68 samples are needed for a 90% confidence interval ($z = 1.645$) at 10% precision, when the Coefficient of Variation CV = 0.5 [28]. The CV of a process provides a normalised measure of its standard deviation with respect to its mean. Therefore a process with a standard deviation of 50 and a mean of 100 has the same CV as a process with a standard deviation of two and a mean of four – their relative standard deviations are equal. Besides the G14, the two other leading international M&V guidelines, the International Performance Measurement and Verification Protocol (IPMVP) [1] and the Uniform Methods Project (UMP) [11], both recommend variations on (1), but do not consider longitudinal studies. The G14 [21] provides a method for aggregating results obtained over time based on Reddy and Claridge's seminal work [29], but does not consider varying sample sizes, and does not quantify uncertainty as well as a Bayesian approach would [27,30]. It is well known that uncertainty quantification in standard regression can be a problem for anything but very simple cases, and methods such as bootstrapping and cross-validation are used for more complex cases [31,32]. A Bayesian approach proves to be a flexible and powerful alternative for efficient, exact uncertainty quantification.

2. Motivation

Standard sampling theory for non-longitudinal cases is well established – both for simple random, and stratified cases, and also incorporates cost considerations [33–36]. Luus adopted a frequentist approach to complex sampling problems in her PhD thesis [31], and used bootstrapping to quantify uncertainty. The thesis provided an excellent overview of advanced sampling techniques, but the uncertainty quantification method is computationally very expensive, and not realistic for optimal sampling designs such those investigated below. Not many studies have attempted to devise efficient sampling methods for longitudinal sampling in retrofit projects, and those that do cannot incorporate population survival survey sampling (non-normal sampling) for overall M&V plans as will be done below. The most directly relevant work was done by Ye and his co-authors [18–20,37]. Improvements on Ye et al.'s method were suggested by Carstens et al. [16,38], and an extension considering modelling uncertainty was done by Olinga [39]. Ye et al.'s method reduces 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. Further work on the problem is motivated by the following observations on these previous methods:

- The aggregation of results from multiple years should 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. For example, 68 metering results from year one should not be added to 68 metering results from year two, so that the total sample size is 136. Due to serial correlation (autocorrelation), samples in year two will contain less information than samples in year one.
- The second factor used previously to reduce meter sample sizes is finite population correction. However, FPC only becomes relevant for population sizes below 1000 and is therefore not applicable to the large-scale studies considered.
- The method also assumes that the means of the metering results for all years are stationary. This is realistic assumption, as energy use may increase or decrease due to various factors. The method proposed below does not make this assumption.
- In the previous model, confidence and precision levels are undefined for years in which no sample is taken. The result is that the precision of the model stays constant when no sampling is done. For example, 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 would be more realistic to increase uncertainty for years in which no sampling is done. The method proposed below does this in a mathematically rigorous manner.
- In previous work, low-cost meters with lower accuracies are selected for low-CV populations [20]. However, high-accuracy meters only enhance the overall accuracy in low-CV cases, when process variability plays a smaller role relative to measurement uncertainty [40]. 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 [41]. This is considered in Section 3.2.1. Also, the time resolution of the meter does not refer to how often the meter measures current and voltage, but the time period over which the meter integrates when storing a data point [42]. The measurement interval is shorter than the integration interval. The integration interval can also be set, and is not five minutes as was supposed for a Class 1 meter in previous works.
- Regarding optimization, gradient-descent methods were employed previously. However, the optimization function is an integer non-linear program (INLP) with discontinuities [16]. Heuristic methods will therefore be used to provide more reliable results, as discussed in Section 3.2.2. Last, the earlier method assumes that 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 [14], and will be incorporated in Section 4.

The method proposed in this paper seeks to improve on the areas above by providing a Bayesian approach to the lighting retrofit monitoring problem, which has been suggested for energy monitoring as far back as 1991 [43]. This Bayesian approach extends previous work [14] from only population survival survey sampling to also include metering placement and overall M&V study design – which has not been done before to our knowledge. Bayesian statistics allows for the use of information from prior meter-samples to be incorporated in a mathematically consistent manner. In this framework, the prior probability distributions are combined with the current sampling data, called the 'likelihood'. Together, these form the posterior probability distribution, from which the uncertainty in the posterior estimate can be quantified. Although a Bayesian prior may be chosen subjectively in other cases, it is determined by the underlying mathematics and previous sampling results for our case. Much of this work is based on West and Harrison's *Bayesian Forecasting and Dynamic Models* [44]. Triantafyllopoulos [45] 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 [46–49] and hierarchical models [50], which applies to the sampling problem described in Section 4. More general introductions to Bayesian theory have been written by Kruschke [51] and Gelman [52], and an introduction to Bayesian measurement theory may also be useful to readers unfamiliar with the approach [53].

The paper is structured as follows. Section 3.1 discusses the theory and methodology of longitudinal cross-sectional metering uncertainty quantification, and presents Dynamic Linear Model (DLM) with Bayesian forecasting. A demonstration in a minimal working example is given, and a case study from previous work is analysed to compare differences of approach, and results. A more complete case study is presented in Section 3.3, using an optimization algorithm. An investigation into the execution of efficient sampling plans is also done. This concludes the first part of the paper dealing with metering alone. The second part of the paper combines this metering method with a survey sampling method from previous work, to obtain a combined efficient monitoring plan. A brief introduction of previous work on population survival survey sampling is presented in Section 4.1, so that both of these models can be integrated into comprehensive energy monitoring case studies in Section 4. These case studies consider the simple random sampling case (Section 4.3), as well as the stratified sampling case (Section 4.4). Finally, conclusions are drawn and recommendations are made in Section 5.

3. Methodology

3.1. Uncertainty quantification for cross-sectional metering sampling models

As mentioned above, there are two components to a longitudinal M&V model: population survival survey sampling, and metering. This section focusses on metering. 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 method below describes how the sample size for such a case can be minimized within the reporting constraints.

3.1.1. 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 in the circuit. Seasonality can be built into the model to increase model granularity to monthly or hourly levels [44], but is not considered here. Last, 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 second-order (straight-line) linear model is used, although other linear functions may also be specified.

Last, it is assumed that samples are independent in time. This means that the same facilities cannot be sampled repeatedly in consecutive years. A new random selection of facilities needs to be made in each sampling year. Although this was not done in previous work on this problem [18–20,37] 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) [54]. 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. A possible solution is to use an autocorrelation correction factor. In previous work we used an exponential windowing function [16,38], but a sample size adjustment factor as per G14 [21] 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, which is a difficult and expensive task in itself. As Violette [43] 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 [55].

3.1.2. 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 work the average annual energy use measured by the meters was fixed at the beginning of the study [18,19,37,38]. For models taking population decay into account (cf. Section 4), the population decay rates were fixed at study inception, and not updated as new information became available. Only the uncertainties are updated as the model progresses through time, given the sampling plan $\mathbf{n}_m$. These differences are illustrated in Fig. 1 vs. Fig. 10. In a dynamic modelling framework, new data alter 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 [56,57], applied to time-series analysis rather than control. However, according to West and Harrison, “To say that ‘Bayesian forecasting is Kalman filtering’ is akin to saying that statistical inference is regression” [44]. 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 they may not do so. A more complete explanation of LBE in the context of DLMs is given by West and Harrison [44].

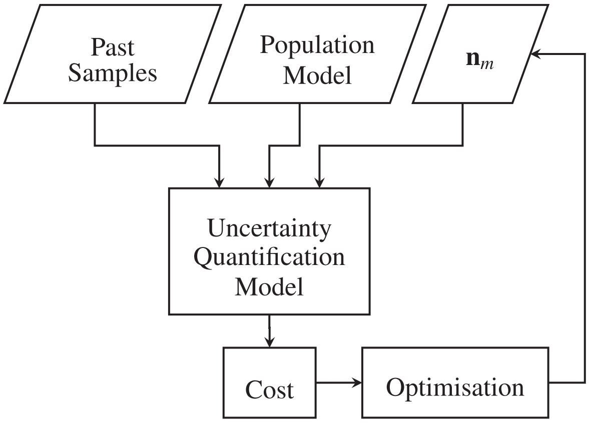

Fig. 1. Flow diagram illustrating existing methods [18,19,37,38], where $\mathbf{n}_m$ denotes the metering plan.

For simple special cases, the DLM estimate at a given point in time would be equal to the Ordinary Least Squares (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). 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 [31,32].

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 m is planned. In the second case, a sample of 20 m is planned. Only their means are used in the regression model. How should the model distinguish between these two plans? For small sample sizes, random draws from a Monte Carlo simulation will not reflect the variance of underlying distribution accurately. 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 is 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.

Similar work on Dynamic Generalised Linear Models (DGLMs) has already been done in the context of lamp population survival surveys [14]. In that case, *Generalised* Linear Models were needed since population proportions are binomially distributed. A parallel in Kalman filtering would be an extended Kalman filter, which has some non-Bayesian elements combined with OLS theory [44]. However, in the case under investigation, normal distributions can be assumed with reasonable confidence, and a DLM is adequate.

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

$$Y_t = \mathbf{F}'\boldsymbol{\theta}_t + \nu, \quad \nu \sim N[0, V] \tag{2}$$

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

$$\boldsymbol{\theta}_t = \mathbf{G}\boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t, \quad \boldsymbol{\omega}_t \sim N[0, \mathrm{W}_t] \tag{3}$$

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. [14]) 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.

Variable definitions

Since we assume 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

$$\hat{E}_{r,t} = \beta_t t + constant \tag{4}$$

The state vector for this system is then

$$\boldsymbol{\theta}_t = [\hat{E}_{r,t}, \beta_t], \tag{5}$$

where the regression vector is

$$\mathbf{F}' = [1, 0], \tag{6}$$

so that (2) is satisfied by yielding $Y_t = \hat{E}_{r,t}$. The evolution matrix is defined as

$$\mathbf{G} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \tag{7}$$

so that (3) 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 all previous data $\mathbf{D}_{t-1}$.

For a linear growth model such as the one under consideration, given that the mean estimate at time t is μ_t, the linear algebra reduces to

$$Y_t = \mu_t + \nu_t \tag{8}$$

$$\mu_t = \mu_{t-1} + \beta_{t-1} + \omega_{t-1} \tag{9}$$

$$\beta_t = \beta_{t-1} + \omega_{t-1}. \tag{10}$$

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 (14), $\mathbf{R}_t$ the prior variance in (15), and $\mathbf{A}_t$ the adaptive vector in (20) (not used explicitly in forecasting). Let the data up to the previous time step be $\mathbf{D}_{t-1}$, and the | sign indicate "given", or "conditional on". In the LBE scheme only the first and second moments are specified. The joint distribution on Y_t and $\boldsymbol{\theta}_t$ is then

$$\begin{pmatrix} Y_t & \\ & | \ \mathbf{D}_{t-1} \\ \boldsymbol{\theta}_t & \end{pmatrix} \sim \left[\begin{pmatrix} f_t \\ \mathbf{a}_t \end{pmatrix}, \begin{pmatrix} Q_t & Q_t\mathbf{A}'_t \\ \mathbf{A}_t Q_t & \mathbf{R}_t \end{pmatrix} \right]. \tag{11}$$

In this study, the equation above describes a normal distribution, although other kinds can also be described this way. Again, West and Harrison [44] provide a full explanation of the DLM and distributions on all parameters. For the purpose of 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_{t+1} is defined as

$$(\hat{E}_{r,\ \ t+1}|\mathbf{D}_t) = f_{t+1} = \mathbf{F}'\mathbf{a}_{t+1}. \tag{12}$$

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

$$\mathbf{a}_t = \mathbf{G}\mathbf{a}_{t-1}. \tag{13}$$

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

$$Q_{t+1} = \mathbf{F}'\mathbf{R}_{t+1}\mathbf{F}. \tag{14}$$

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

$$\mathbf{R}_{t+1} = \mathbf{G}\mathbf{R}_t\mathbf{G}' + \mathbf{W}_t. \tag{15}$$

The evolution variance $\mathbf{W}_t$ can be static, but from previous work [14] we prefer to update it according to

$$\mathbf{W}_t = \mathbf{G}\mathbf{U}_t\mathbf{G}' \tag{16}$$

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

$$\mathbf{U}_t = \delta\mathbf{C}_t. \tag{17}$$

$\mathbf{W}_t$ 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.

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 (12), (13), (15), and (16) 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

$$(\boldsymbol{\theta}_t|\mathbf{D}_t) \sim T[\mathbf{m}_t, \mathbf{C}_t]. \tag{18}$$

These are calculated as follows. Because data are available, rather than using (14), the variance on E_t is updated according to

$$Q_t = \mathbf{F}'\mathbf{R}_t\mathbf{F} + k_t V \tag{19}$$

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 (19) 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 becomes available. It is calculated as

$$\mathbf{A}_t = \mathbf{R}_t\mathbf{F}Q_t^{-1}. \tag{20}$$

The state is updated by

$$\mathbf{m}_t = \mathbf{a}_{t-1} + \mathbf{A}_t e_t \tag{21}$$

where

$$e_t = Y_t - f_t, \tag{22}$$

and

$$\mathbf{C}_t = \mathbf{R}_t - \mathbf{A}_t\mathbf{A}'_t Q_t. \tag{23}$$

3.1.3. DLM demonstration and comparison to previous methods

To demonstrate how (and verify that) the DLM works, a hypothetical case is considered, and is illustrated in Fig. 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 as in (1). We denote a metering sample size at time t by $n_{m,t}$. The demonstration sampling plan (the vector containing the sample sizes for future years) $\mathbf{n}_m$ is

$$\mathbf{n}_m = [68, 68, 68, 68, 200, 68, 0, 0, 0, 68, 0]. \tag{24}$$

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 Fig. 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 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.

3.1.4. Case Study 1: comparison to previous method

In this section, the DLM will be compared against the earlier method [18–20,37,38], using the case study from [38]. 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 earlier 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, the prior on the slope is set to zero.
- The earlier method uses Finite Population Correction (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 1000, or 0.16% of the installed population in the benchmark study [38]. Therefore it does not affect the calculation and may be neglected in the DLM.

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

- The previous approach uses frequentist confidence intervals. As Neyman, who developed these intervals, remarked, these intervals do not really convey a degree of belief. Rather, they are the product of a process that produces an interval which contains the true value a given percentage of the time [58]. Since the bounds are random [59], using such intervals for risk calculation is problematic. The Bayesian credible interval used by the DLM does, however, produce the interval sought for uncertainty quantification. The two intervals do sometimes agree numerically, but their interpretations are different and should not be equated [51,52].
- The improvements to the previous model [38] 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 (17) 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.
- The increase in uncertainty for years in which no sampling is done, cannot be removed without removing a fundamental component of the DLM. It is therefore difficult to compare it to a model in which it is assumed that the uncertainty stays constant over years of non-sampling.

With these caveats in mind, a case study for the previous method [38] is analysed by the DLM, using the optimal sample sizes determined using that method. This case study has become somewhat of a benchmark since all models solving this problem consider it. In this case study based on a real UNFCCC CDM project [60], 607,559 CFLs rated at 20 W were distributed to households in the South African provinces of the Northern Cape, Free State, Gauteng, Limpopo, and Mpumalanga, to replace 100 W ICLs. Crushing certificates for the replaced lamps were obtained to verify that they were indeed replaced. It was assumed that they burn for an average of 4.5 h 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, in order 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

$$\mathbf{n}_m = [68, 68, 28, 16, 8, 8, 6, 6, 4, 4, 2]. \tag{25}$$

Results and discussion

The resulting uncertainty bounds using the earlier method's sampling plan, calculated with the DLM, is plotted in Fig. 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 undersampling 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. For example, increasing the prior on the slope of the regression line to a number above zero results in undersampling for all years. Such changes do not affect DLM models accounting for inter-sample variance as strongly.

When decreasing the effective sample size by using an autocorrelation factor of 0.25 [21], it is found that year six is also undersampled. However, when the exponentially windowed sam-

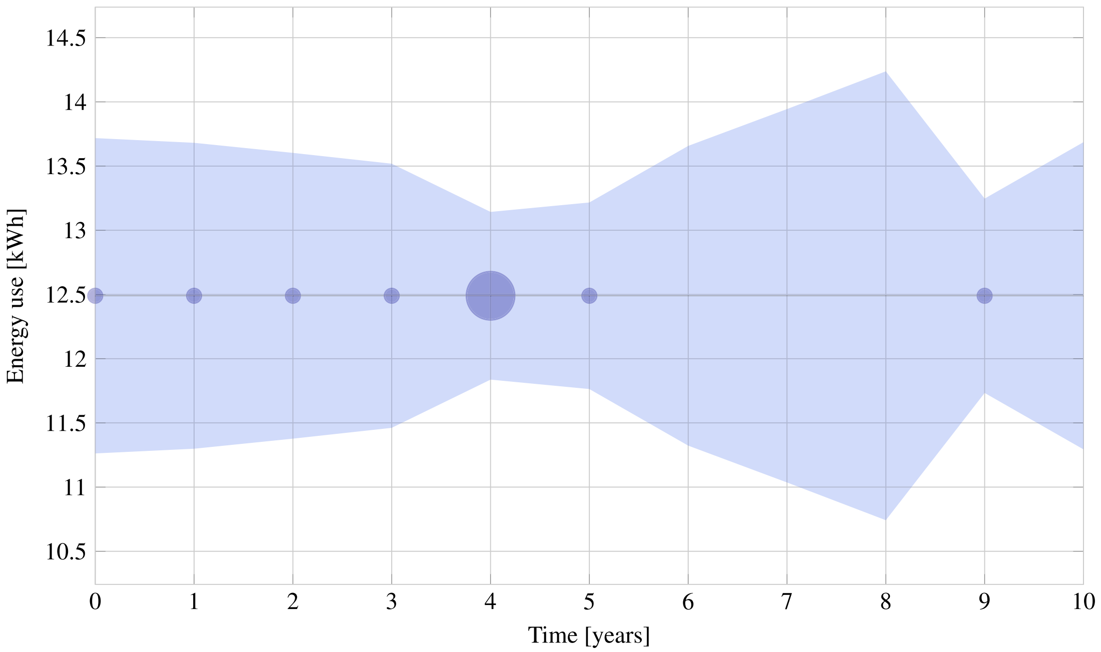

Fig. 2. DLM demonstration where the relative sizes of the markers provide a qualitative indication of sample sizes. The blue shaded area represents the instantaneous 90% credible interval around the estimate. Hypothetical case where all sampling results fall on the mean. Sample sizes are $\mathbf{n}_{m,0-3,9} = 68$, and $n_{m,4} = 200$.

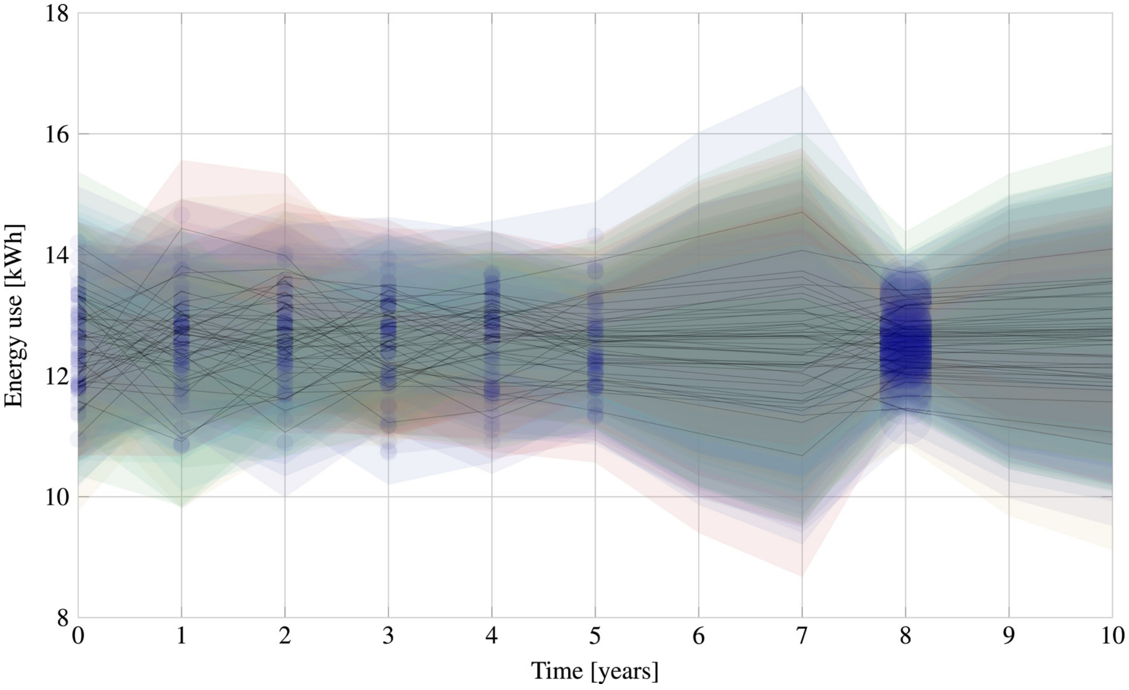

Fig. 3. DLM demonstration reflecting true sampling results. Multiple realisations shown.

pling plan is used, the confidence bounds are much closer to the precision limits for all years.

Although the results indicate that the previous methods do not yield ‘optimal’ or even efficient sampling designs, the improved model with exponential windowing [38] is relatively safe to use, under its assumptions of a stationary mean, etc. Although convenient, these assumptions can be restrictive and unrealistic, however, as discussed in the bulleted points above. To mitigate them, a randomised control trial will have to be designed. This would involve having a treatment group (retrofits installed), and a control group (no retrofits installed), where these two groups are similar in all other relevant aspects. The difference between their energy use would have to be reported with 90/10 accuracy, meaning that the energy use in each group would have to be determined with an accuracy exceeding 90/10, making them much larger. Selecting such groups would be difficult: those who volunteer that their energy use be monitored for ten years, and who do not plan to use energy efficient lighting during that time, may not be representative of the population as a whole: called self-selection bias. The groups may also change over time: young couples may have children, and the children of older couples may move out, for example. People may renovate, disqualifying them and leading to

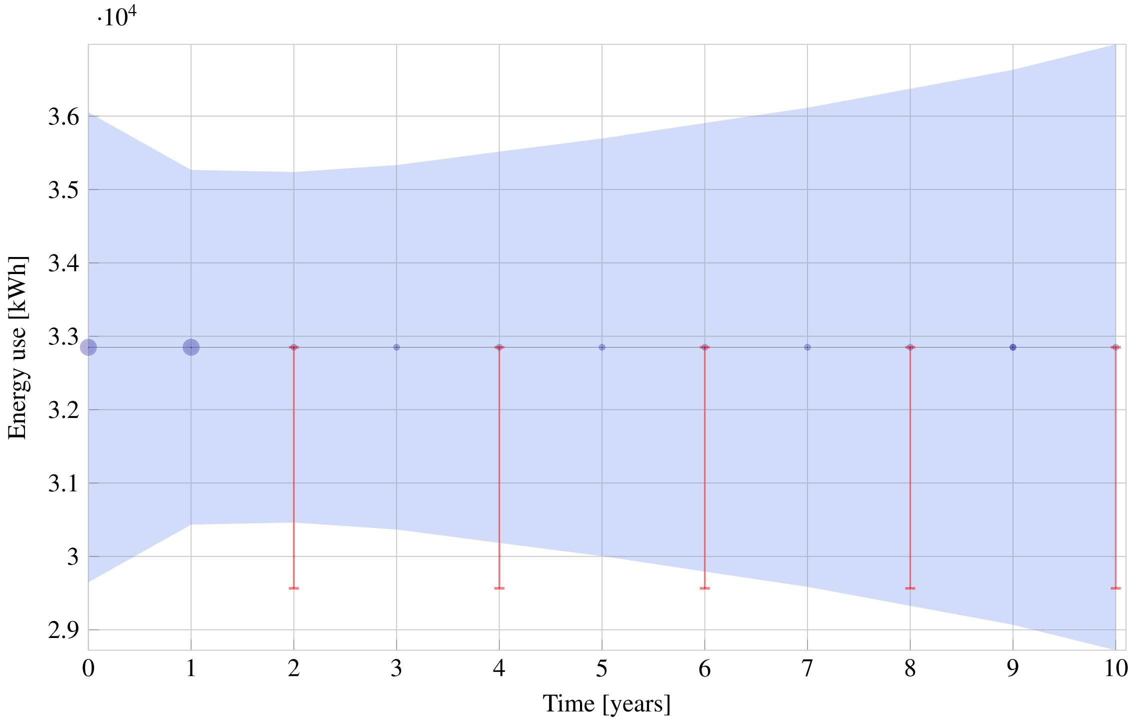

Fig. 4. 90% Confidence bounds (shaded) compared to 10% precision limits (red error bars) for previous sampling plan using earlier method [19,38], analysed by DLM. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

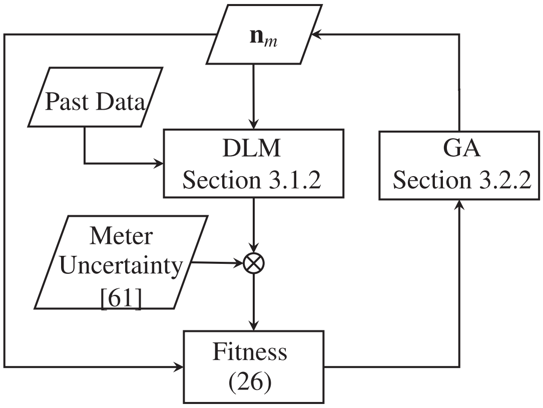

Fig. 5. Flow diagram of cross sectional metering sampling designs as in Section 3.2.

subject dropout. Such phenomena would skew the measurements and indicate spurious trends, and would need to be accounted for in the sampling design in cases were the same facilities are continually metered and taken to represent the whole population. The DLM presents fewer such practical and mathematical difficulties, as will be shown in the next section.

3.2. Efficient cross-sectional metering sampling designs using DLMs

In the previous subsection, the DLM was compared to earlier methods. In this one, the DLM is used in an optimization routine to design an efficient sampling plan, given past data. The flow is illustrated in Fig. 5. The extension of the above methodology to an optimization problem will first be discussed in theory, and a case study will then be presented. Note that the study commences at $t=0$.

We note that 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 so as 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. We therefore refer to these as *naïve* efficient designs, following the convention in mismeasurement studies [62–64]. A robust design with more meters, on the other hand, will therefore 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 3.3.3) and robust efficiency is recommended for future research.

3.2.1. Adding 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 [40] that metering uncertainty makes a small contribution to overall uncertainty for sampling designs with standard variance assumptions. We assume Class 1 m [65] are used with Class 1 Current Transformers (CTs) [66], as these are common for revenue metering. Since no load profiles are assumed for the study, a flat error rate of 3%

Table 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

is assumed. (For plots showing the change in error rate as a function of the rated current of the instruments, see [41]). The 3% figure allows for the combined meter-CT accuracy, as well as for low-cost calibration [61]. However, at this level, it can be shown [40] that the difference made by metering error is so small that the required sample sizes do not change due to the additional uncertainty.

3.2.2. 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 [38], and will in our 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 work [14,38]. The GA was implemented via the `DEAP` Python library [67].

The parameters used to tune the GA for this case will need to be the same as for the optimization in Section 4, and are shown in Table 1. The mutation function was set so that the genes that are selected for mutation are altered by adding a number from the distribution $\sim Normal(-10, 500)$.

Previous cross-sectional efficient metering studies have considered installation, maintenance, and meter removal costs separately for each meter [18,19,37,38]. 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 3.1.1, 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 Section 4, however. From industry experience, we set this rate at R3000 (South African Rand) per meter per sampling point, although it may vary significantly by contract and supplier.

3.2.3. Notation

Let:

χ	Number of sampling points where $e_t \gg> \varepsilon$
$n_{m,benchmark}$	Non-DLM solution at time t
$n_{m,t}$	Decision variable. Sample size at time t $n=\{\tau, \tau+1, \ldots, N\}$
w_m	Cost per meter in Rand/sample
τ	Present time, where $\tau \in \{1, 2, \ldots, N\}$
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 $\mathbf{M} \subset \{\tau+1, \tau+2, \ldots, N-1\}$
$\hat{E}_{r,t}$	Estimate of average annual energy use at t
$LCL_{m,t}$	Lower Confidence Limit at t

3.2.4. Mathematical formulation

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

$$\min \sum_{t=\tau}^{N} n_{m,t} w_m + r(\mathbf{n}_m), \tag{26}$$

where

$$r(\mathbf{n}_m) = \sum_{t\in\mathbf{M}} \left(10^5 w_m(e_t - \varepsilon) + 10^7 + 5 w_m n_{m,benchmark,t}\right) \forall t \in \chi \tag{27}$$

and

$$e_t = \frac{\hat{E}_{r,t} - LCL_{m,t}}{\hat{E}_{r,t}}. \tag{28}$$

3.2.5. Description

The decision variable is the metering sampling plan $\mathbf{n}_m$, the individual elements of which are written as $n_{m,t}$ in (26).

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 (26) 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 [14]. 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 determined so that solutions that do incur a penalty are directed into the feasible region, rather than away from it [14]. Consider Fig. 6. If there were no constraint, the cost would increase with $n_{m,t}w_m$ 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 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

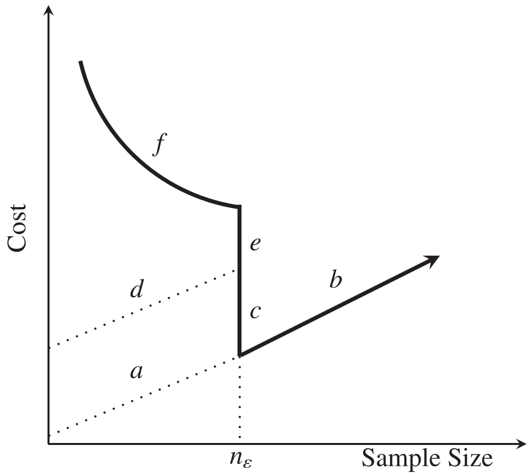

Fig. 6. Genetic Algorithm constraint function $r(\mathbf{n})$ in (27), where n_ε represents the threshold sample size.

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 by (1). This step is represented by line *ce*.

Regarding (28), 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.

3.3. Case Study 2: efficient cross-sectional metering design

Because the method creates the possibility to measure such projects in more realistic ways than before, no adequate data are available, and synthetic data based on industry standards will be combined with real data from similar projects, extending Case Study 1.

We assume that the luminaires are 11 W CFLs that operate for an average of 3.11 h per day [68], or $E = 12.49$ kWh per year. The CV in the sample is set to 0.5; a standard M&V assumption [69]. 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.

We model the true energy use 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, since the meters are installed in only a sample of the population buildings. 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, $\mathbf{M} = \{3, 5, 7, 9\}$.

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 a 99% confidence. Therefore $3\sigma = 12.49/4$, with the prior variance specified as σ^2.

3.3.1. 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 $(1 - R^2)$ [13]. R^2 is the coefficient of determination, which is the square of the Pearson moment correlation coefficient. This is similar to 'ratio-estimation', where the additional information contained in the known ratio or regression line can be used to reduce the sample size. 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 Coefficient of Variation on the Standard Deviation on the Root Mean Square Error [21,24], 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 as an extension of G14 [21] and Reddy and Claridge's work [29].

We therefore benchmark the method against the standard M&V approach of (1). Since metering error has been determined to not affect sample size, it may be neglected.

3.3.2. Case Study 2: results and discussion

The values generated for the first three points are $\mathbf{D}_{0-2} = [12.39, 13.02, 12.71]$, where the sample sizes are $\mathbf{n}_{m,0-3} = [68, 68, 68]$. One efficient sampling for one realisation of results is shown in Fig. 7. The planned sample sizes $\mathbf{n}_m$ are

$$\mathbf{n}_{m,DLM} = [56, 0, 36, 0, 32, 0, 26], \tag{29}$$

while standard sampling theory yields

$$\mathbf{n}_{m,Benchmark} = [68, 0, 68, 0, 68, 0, 68]. \tag{30}$$

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

The red error bars in Fig. 7 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 (27) 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

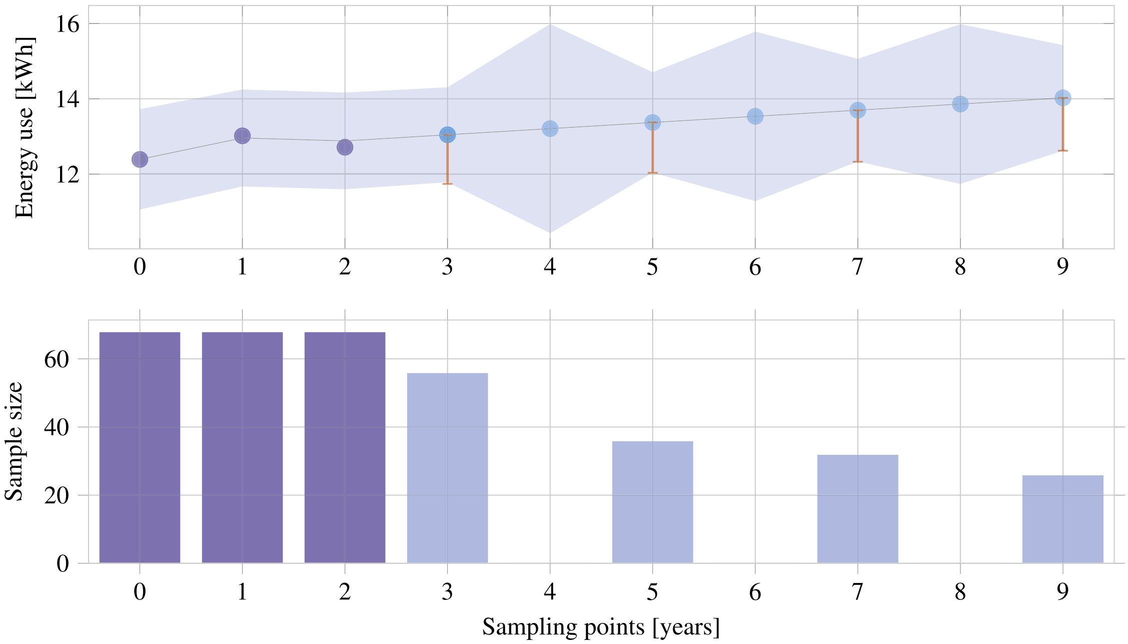

Fig. 7. Efficient sampling plan using the DLM for one random model realisation. (For interpretation of the references to color in the text, the reader is referred to the web version of this article.)

'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 [16,38], but has not always been adopted [18–20,37].

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 Fig. 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 Fig. 8. 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 relationship is illustrated in Fig. 8. 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 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 (as illustrated in Fig. 8, when such sample sizes will yield inadequate precision). The forecasting uncertainty bounds should certainly be considered. (Note that the forecasting uncertainty bounds in Fig. 7 are instantaneous future sample sizes which include results from planned future samples). Nonetheless, the relationship shown in Fig. 8 is true in the sense that if that relationship is correct, the required future sample sizes do follow that curve.

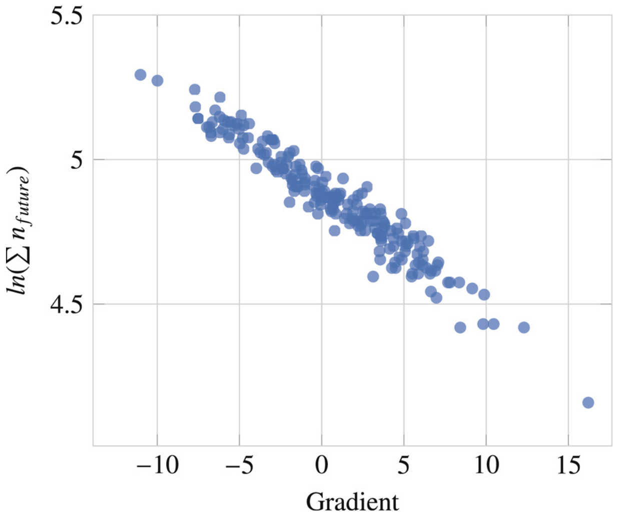

Fig. 8. Natural logarithm of the total number of future samples under efficient sampling plans, as a function of the gradients of the regression lines on past samples, e.g. in Fig. 7.

3.3.3. The reliability of efficient sampling designs

After an efficient sampling plan has been designed, it should be executed. In this section, we investigate the reliability of efficient sampling plans, in terms of compliance to reporting precision requirements. Since the sampling plan needs to be updated every

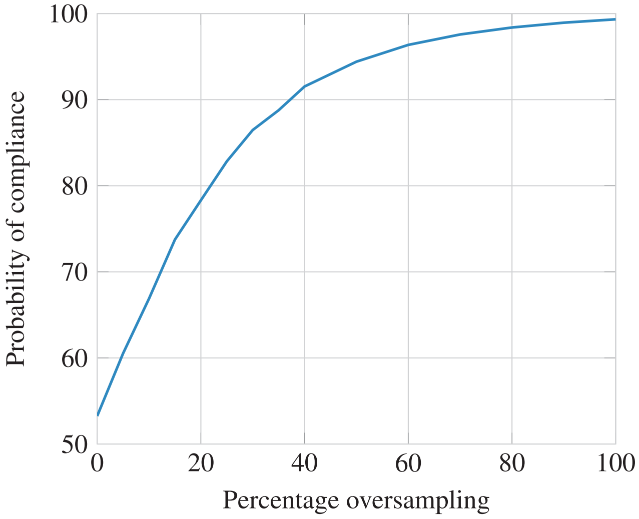

Fig. 9. The effect of oversampling on the probability on non-compliance, as per Section 3.3.3.

time new data becomes available, we investigate only the next time step beyond the sampling plan already devised. We suppose that three years' data are available ($\mathbf{D}_{0-2}$), and that the fourth year is forecast, planned, and executed. We simulate such scenarios and analyse the result. The investigation proceeds as follows:

1. Generate $\mathbf{D}_{0-2}$ from the distribution $\sim Normal\left(12.49, \frac{12.49CV}{\sqrt{\mathbf{n}_{\mathrm{m},\,0-2}}}\right)$.
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 $n_{m,3}$: $D_3 \sim Normal\left(12.49, \frac{12.49CV}{\sqrt{n_{m,\ 3}}}\right)$.
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 3.2, 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 Fig. 9. 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 (1) recommended by the leading guidelines [1,21,70]. When simulating $n = 68$ from a distribution with $CV = 0.5$, one finds that the interval produced includes the true value and satisfies the 90/10 criterion in only 50% of cases. Since this formula is so common in M&V it will be used in this study, but M&V professionals are encouraged to do this simulation and consider the implications on M&V sampling designs.

Fig. 8 also illustrates that efficient designs are sensitive to the apparent gradient inferred from past samples. The gradient illustrated in Fig. 7 is slightly positive, leading to smaller sample sizes than if the gradient were very negative (due to the randomness in the realisations of the sampling points). There is a danger that efficient sample sizes will undersample in cases where energy use seemingly increases dramatically but is only due to randomness in the samples.

We therefore investigate two rudimentary risk mitigation strategies. The first is to oversample by a given percentage. The second is to use the information from the DLM to determine a robust sample size.

In the first approach, we oversample by 0–100% and plot the results in Fig. 9. This relationship depends 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 UMP recommends 10–30% oversampling [71] for other reasons; a useful recommendation for the considerations under discussion as well.

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 Fig. 9.

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 if cost is narrowly defined as metering cost. In a robustly efficient sampling plan, on the hand, the cost of metering is traded off against the cost of non-compliance to uncertainty reporting requirements. Considering non-compliance makes naïve efficient plans costly, 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 cost a naïve efficient plan, however.

The analysis above represents a very simple robust plan, and future work may develop more a complete, robust framework, similar to that of Rysanek and Choudhary [72].

4. Combined longitudinal and cross-sectional sampling

Thus far, the paper has only considered metering sampling, which is the first aspect of a complete longitudinal energy retrofit M&V study design. The second aspect is longitudinal population survival survey sampling. This was investigated in detail in previous work [14]. That work will be summarised below to provide context, after which the population survival survey sampling component will be integrated with the metering sampling component discussed above to give a complete longitudinal M&V design for a building lighting retrofit project.

4.1. Longitudinal population survival survey sampling

The total energy saved by a retrofit project in a given year would be proportional to the number of retrofitted units installed by the project which are still active during that year. The purpose of these surveys is therefore to estimate the proportion of the population surviving at time t, which is denoted $\hat{\Phi}_t$.

The decay of many populations, including lamp populations, can be described by a logistic function [17,73,74]. Such a logistic function has been developed [15,38], and was applied to the problem at hand by using a Dynamic Generalised Linear Model (DGLM)

[14]. A DGLM is similar to the DLM described in Section 3.1.2, but uses a Generalised Linear Model (GLM), because the survey result is distributed according to the beta/binomial distributions. These distributions result from pass/fail Bernoulli trials obtained through telephone interviews or site visits. The beta and binomial distributions form a conjugate prior pair, with the same advantages regarding the speed and accuracy of this solution over numerical solutions as for the DLM of Section 3.1.2.

A difficulty arises when combining this beta-distributed survey result with other parameter distributions. These parameters could be meter results of average annual energy use (as in Section 3.3), or estimates such as annual hours of use and average power draw per luminaire. If all the distributions are normal (as has been the case up to this point), then their convolution can be calculated quickly and accurately using analytical equations. However, parameters are often normally distributed and need to be convolved with the beta population proportion estimate – an operation that is usually only done by Monte Carlo (MC) methods. MC is powerful and versatile, but for accurate uncertainty determination in a GA it can be prohibitively expensive. Also, in a GA with hundreds of individuals over several generations, the GA algorithm identifies ‘outlier distributions’ as the fittest individuals. The apparent fitness of these anomalies is merely the result of a random, favourable realisation of the underlying distributions. This is rare enough in standard MC to be irrelevant, but becomes important in a threshold heuristic optimization program where slight MC noise can mean the difference for an unfit individual to be seen as fit. These outlier distributions seem to adhere to the precision limits at lower cost when they actually violate the constraint. They are then registered falsely by the GA as fit individuals. To mitigate this effect, a recently developed technique called Mellin Transform Moment Calculation (MTMC) [75,76] was used instead of MC. MTMC takes the scale and shape parameters of the input distributions, and describes the moments of the convolved resultant distribution analytically, where this convolution can be any polynomial function. The first four moments (mean, variance, skewness, kurtosis) from the MTMC were then used to fit a Johnson distribution, which is nearly identical to the MC result and can be used to calculate uncertainty bounds cheaply and consistently. For more information on uncertainty evaluation through moment-based distribution fitting, see Rajan et al. [77]. As in Section 3.3, the MTMC uncertainty estimates are applied in a GA to find an optimal multi-year sampling plan.

Finally, because the beta distribution can be asymmetrical, Highest Density Intervals (HDIs) are preferred to standard equal-tailed confidence intervals [14,51].

4.2. Combining survey sampling with metering data

Instead of combining the survey result uncertainty from the previous section with estimates for energy consumption (hours of use and power consumption), it will be combined with more accurate meter sampling results. For this case, metering and survey sample sizes need to traded off against one another to ensure adherence to the overall uncertainty reporting bounds, at low cost. A diagram illustrating the how the various components discussed so far fit into the overall plan is shown in Fig. 10. This is different to previous combined sampling designs (Fig. 1), where only meter sampling was optimised, 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

$$\hat{\mathbf{E}}_{saved} \sim \hat{\Phi}\mathbf{n}\Delta\hat{\mathbf{E}}, \tag{31}$$

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 [21] 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 be their respective standard deviations. Assuming that there is measurement error in the meter of 2.52% as described in Section 3.2.1, the uncertainty distribution on the ratio of the power draws P_b/P_r can be described by the distribution

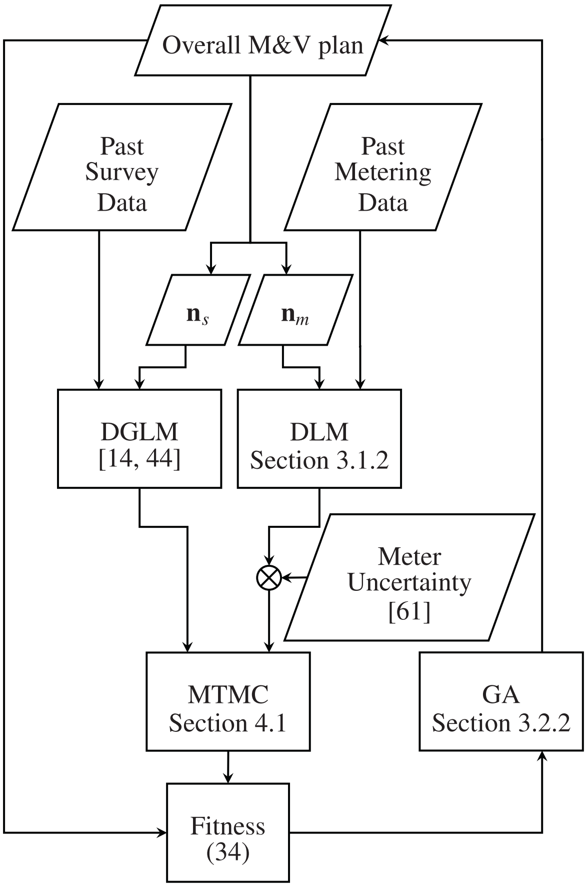

Fig. 10. Flow diagram illustrating proposed method for combining metering and surveying data. The metering plan is denoted $\mathbf{n}_m$, and the sampling plan $\mathbf{n}_s$.

$$P_b/P_r \sim N\left(\frac{P_b}{P_r}, z\frac{P_b}{P_r}\sqrt{\left(\frac{s_r}{P_r}\right)^2 + \left(\frac{s_b}{P_b}\right)^2}\right) \tag{32}$$

as per the ASHRAE’s guideline RA96 [78]. The annual energy saving per luminaire given this ratio can then be expressed as

$$\Delta\hat{\mathbf{E}} \sim \hat{\mathbf{E}}_r(P_b/P_r - 1). \tag{33}$$

As mentioned previously, the MTMC method can then be used to calculate the first four moments of E_{saved} in (31). These are used as inputs to the Johnson distribution, which will describe the overall probability distribution on the savings estimate for a specific point in time.

The fitness function (26) 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

$$\min \sum_{t=1}^{N} n_{m,t} w_m + \sum_{t=1}^{N} n_{s,\ t} w_s + d_t v + r(\mathbf{n}). \tag{34}$$

The penalty function is also modified accordingly:

$$r(\mathbf{n}) = \sum_{t \in \mathbf{M}} \Big(10^5 (w_s + w_m)(e_t - \varepsilon) + 10^8 \quad (35)$$

$$+\, 5(w_m n_{m,benchmark,t} + w_s n_{s,benchmark,t} \Big) \; \forall \;\; t \in \chi. \quad (36)$$

In this case, the relative cost of surveying and metering play a large role in determining an optimal solution, since the GA will trade these sources of uncertainty off against one another when evaluating different solutions. 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 the 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 optimised 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.

In the interest of brevity, these case studies will not be described in as much detail as those above or from previous work [14], from which they are expanded. However, the details will remain the same unless otherwise stated.

4.3. Case Study 3: combined 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 Polish Efficient Lighting Project (PELP) data points [14,17]. This was a large study of over one million lamps, tracked over a number of years. Unlike the metering data, PELP points were used rather than randomly generated points according to the past sample sizes. This is because while a randomly varying upward trend in the metering data is not of concern, an apparently upward trend in a logistic curve such as that of the population decay data, results in an invalid model. The meter data, however, were generated as before.

It is assumed that three years' data has been collected (years 0–2), and that reporting is to be done annually for $\mathbf{M} = \{\mathbf{4, 5, 6, 7}\}$. In the project, 100,000 CFLs of 11 W each replace their 60 W 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 $\mathbf{n}_{s,0-2} = [250, 250, 100]$.

4.3.1. Benchmark

The combined benchmark is calculated using a GA with the combination of the survey sampling and metering uncertainty determined as in (31), where the uncertainty in E_r in (33) is calculated according to (1) combined with the meter measurement error. As in previous work on longitudinal survey sampling, the survey sampling benchmark was selected as a Jeffreys interval on the proportion [79].

The benchmark is therefore an optimal sampling plan in which prior data are not taken into account through the Bayesian method.

4.3.2. Case Study 3: results and discussion

An efficient sampling plan is listed in Table 2, at a cost of R772240. A benchmark sampling plan is listed in Table 3, at a cost of R1,128,940. The Bayesian method therefore achieves a saving of 40.13% for these cases.

The results for this case study are shown in Fig. 11. The top four graphs show the individual metering 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 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 $4n$, (or $2\sqrt{n}$) 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 surveys samples would be.

The method shows a clear advantage over existing methods. Smaller sample sizes than existing sampling methods such as (1) are needed. Not only does the DLM-DGLM method offer a more complete consideration of sampling, but it does so using well-established and mathematically consistent methods.

4.4. Case Study 4: 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. These sub-populations (or strata) can be specified according to technology, application, location, or any criterion that would make one group of units distinct from a second group of units. An example is given by Ye [37] where 263,519 CFLs and 140,777 LEDs were installed as retrofits in 45 provincial hospitals in South Africa. No population decay data was available for that project, and it is assumed that it was imputed with the PELP data discussed above. Two-stratum designs are relatively simple to solve by other means as well, and so to illustrate the scalability of the proposed method, a three stratum case along similar lines is considered below. Although the numbers of units may differ, all other aspects are the same.

In stratum one, 50,000 incandescent lamps of 60 W each, burning for 3.11 h per day, are replaced by 11 W CFLs. In stratum two, 20 000 incandescent lamps of 60 W each, that burn for 2 h per day, are replaced by 11 W CFLs. In stratum three, 30,000 incandescent lamps of 100 W each, that burn for 4.11 h per day are replaced by 14 W CFLs. To provide realistic population survival curves, three curves from the Lighting Research Centre's Specifier Report on CFLs are used [14,74]. The data were transformed from time-to-failure data (i.e. "3.3 years to 20% failures") to lamp survival (i.e. "at 3 years, 18% had failed", for example), since for M&V studies the monitoring interval is fixed. 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 are 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 in terms of cost and uncertainty.

Five years of sampling are assumed to have been conducted in the past. Meter samples 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

Table 2
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 3
Benchmark of 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

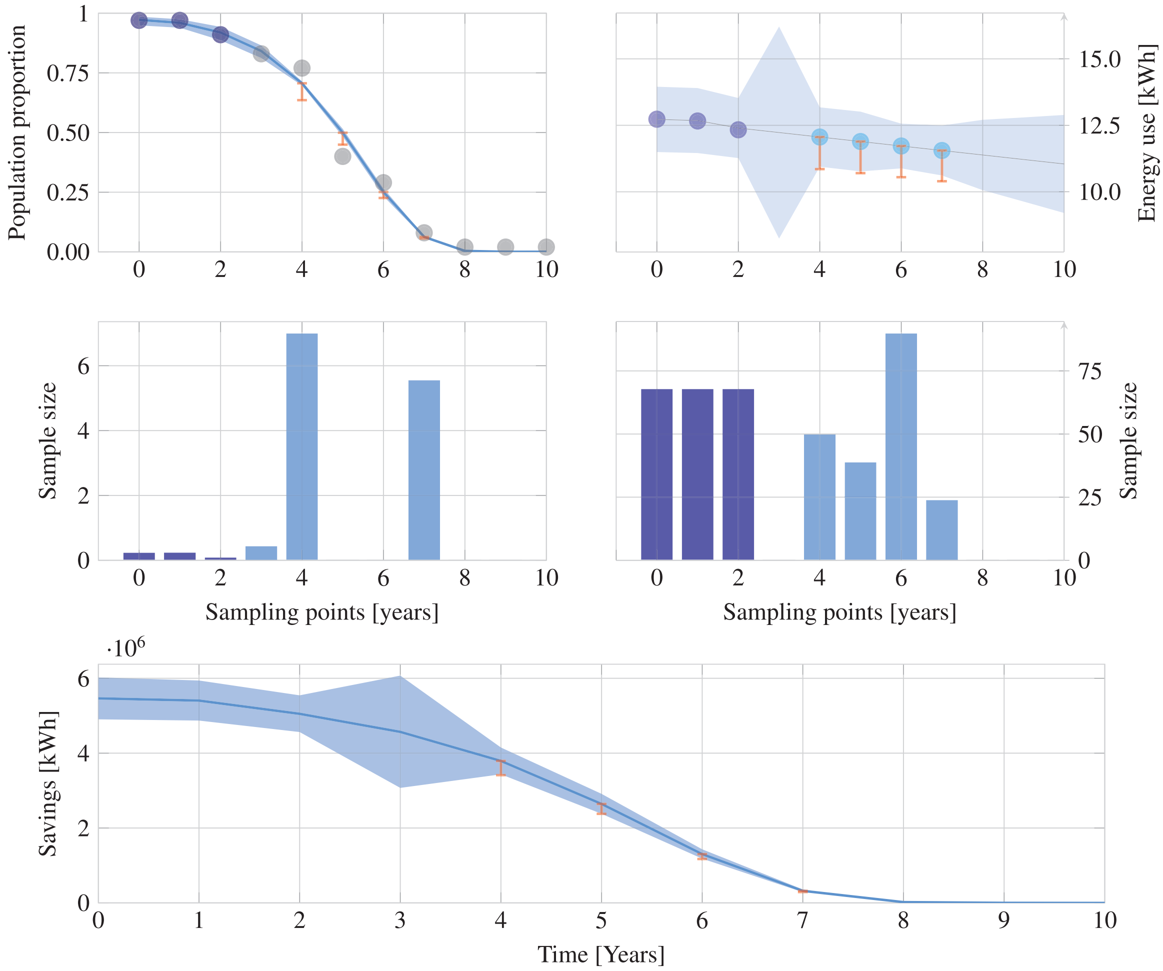

Fig. 11. Plot of combined survey sampling (top left) and metering (top right) for a single population (Case Study 3), with the combined savings estimate over time at the bottom. Dark blue indicates past samples, and light blue indicates planned future samples. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

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.

4.4.1. 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 [70]. Stratified 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 benchmark method described in Section 4.3.1 to the stratified case. In effect, a GA is used to devise a stratified sampling design with all the complexity of proposed method, except for the Bayesian forecasting and dynamic model components.

4.4.2. Case Study 4: results and discussion

One efficient sampling result is shown in Table 4 and Table 5 and has a cost of R1,417,010. The benchmark is R1,918,350, rep-

Table 4
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

Table 5
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

resenting 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 noisiness of efficient samples is also shown in the top right subplot of Fig. 12. Bear in mind that the *y*-axis is only between 12 and 14. Nevertheless, the random variation in relatively small samples (50–150 m) does result in spurious trends in the regression lines.

The overall energy savings curve (bottom subplot of Fig. 12), is a smooth line in this case. However, if the different strata had significantly different population survival characteristics, a cascade-profile would be observed.

Both metering and survey sample sizes seem to increase towards the end of the study. That is because as the savings figure becomes smaller, the relative precision bound (in this case 10%) becomes more stringent. For example, 10% of 4×10^6 (year 6) is much larger than 1.5×10^6 (year 12). This is counter-productive: much of the project's budget is spent on measuring small savings. It would be more sensible to place the precision reporting requirement on the total savings to date figure, rather than the reported annual savings figure. It is possible that some jurisdictions have implemented such an approach, although we are not aware of such cases.

The Neyman allocation method recommended by M&V guidelines [70] is efficient and accurate, provided that only simple stratified designs are attempted, without considering other factors such as different sources of uncertainty. However, the method proposed in this paper is more flexible, and allows for complex, real-world stratified designs needed for most M&V projects. The method finds smaller sample sizes and distributes them intelligently over time so that uncertainty constraints are adhered to, while reducing costs.

5. Conclusion and recommendations

A Dynamic Linear Model (DLM) with Bayesian forecasting is shown to provide superior uncertainty quantification and sample designs compared to standard and previously proposed methods, and does so under more realistic conditions. 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 one half of possible scenarios, given the assumptions in the case study. This is due to the lack of statistical power in the standard M&V sampling formula.

The DLM, in combination with a Dynamic Generalised Linear Model (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%.

DLMs are recommended as a useful alternative to standard linear regression for M&V, should reliable uncertainty quantification be required. At the moment the model works with annual energy data, as this is the frequency at which reporting is done most often. However, DLMs can be extended to finer resolutions by taking seasonality and periodicity into account. The addition of covariates such as temperature may then become necessary. It is possible to add these. This further extension recommended for future research.

Because DLMs are updated on-line, and regression need not be redone every time new data becomes available, it holds promise for M&V 2.0, where M&V big data need to be processed continuously to give real-time feedback. DLMs provide the option to quantify uncertainty while maintaining a relatively low computational overhead, and its application in this domain should be investigated further.

Further research into robust M&V sampling decision frameworks is also recommended, since the disadvantages of the low statistical power of the standard sampling formula proposed by most M&V guidelines are well illustrated in this study.

M&V guidelines recommend the normal-distribution approach to binomial sampling (useful for population survival survey analysis). This approach has been shown to be inaccurate [79], and should be amended to more accurate methods.

From a regulation perspective, it is recommended that statistical accuracy constraints on reported savings pertain to the total savings

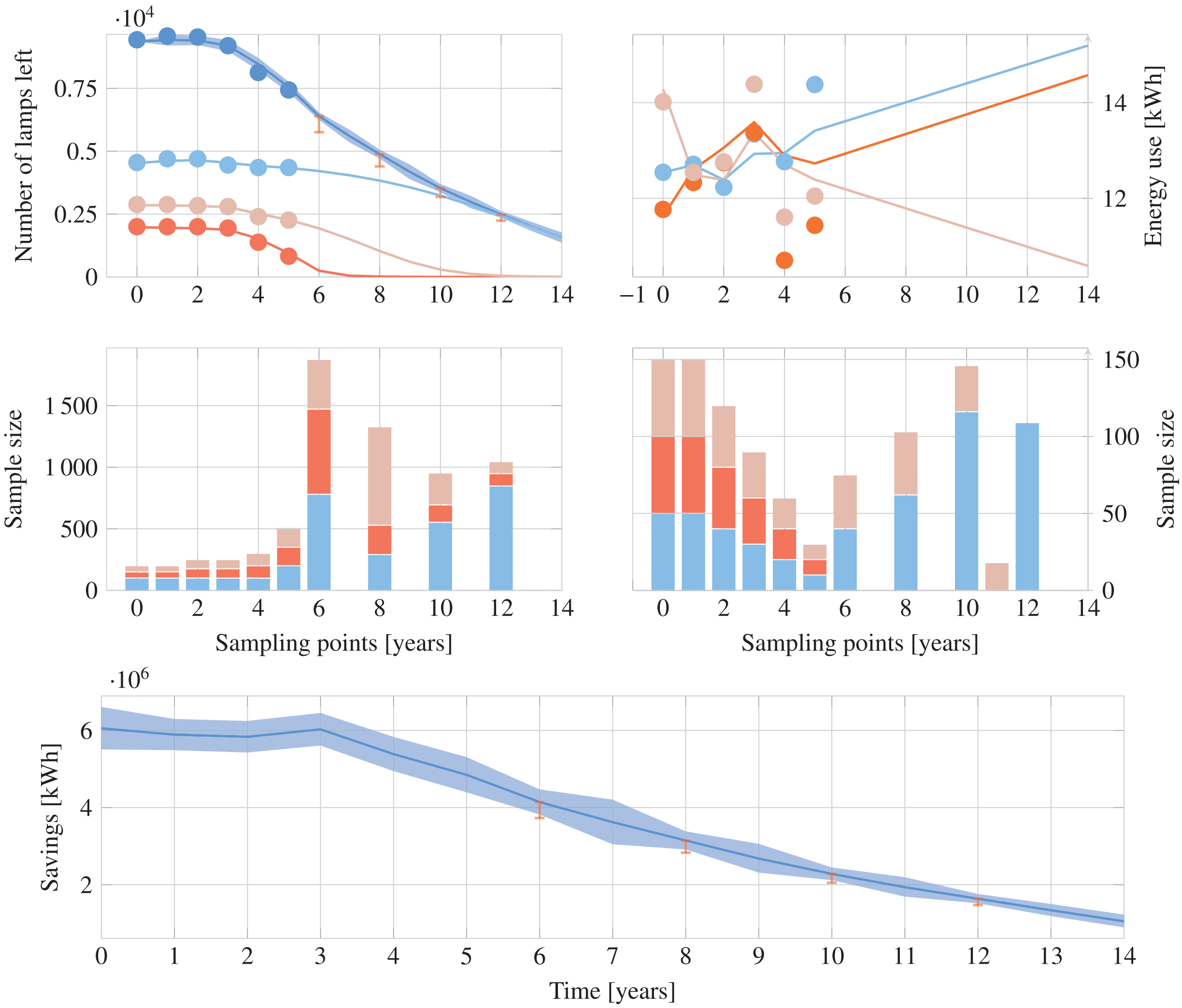

Fig. 12. Efficient combined sampling plan using the DLM and DGLM for one random model realisation (Case Study 4). Stratum 1 is in blue, stratum 2 in red, and stratum 3 in brown. Combined values are shown in blue. The 10% error bars are shown in red, and the 90% confidence intervals in light blue. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

to date, and not to the annual reported savings. This would make a large difference to M&V budgets, increasing project feasibility without sacrificing overall project objectives.

Acknowledgement

The financial assistance of the National Hub for the Postgraduate Programme in Energy Efficiency and Demand Side Management towards this research is hereby acknowledged.

References

[1] Efficiency Valuation Organization, International Performance Measurement and Verification Protocol, vol. 1, 2012.
[2] Taxation Laws Amendment Act No. 25 of 2015, South African Government Gazette no. 39588, 2015.
[3] United Nations Framework Convention for Climate Change, Clean Development Mechanism Methodology Booklet, 2015.
[4] A. Michaelowa, D. Hayashi, M. Marr, Challenges for energy efficiency improvement under the CDM – the case of energy-efficient lighting, Energy Effic. 2 (4) (2009) 353–367.
[5] CDM Executive Board, Approved Baseline and Monitoring Methodology AM0046, Tech. Rep., UNFCCC, 2007.
[6] United Nations Framework Convention for Climate Change, Approved Small-Scale Methodology AMS II.C, Demand-Side Activities for Specific Technologies.
[7] United Nations Framework Convention for Climate Change, Approved Small-Scale Methodology AMS II.J, Demand-Side Activities for Efficient Lighting Technologies.
[8] Y. Dodge (Ed.), The Oxford Dictionary of Statistical Terms, Oxford, 2010.
[9] J.L. Mathieu, D.S. Callaway, S. Kiliccote, Variability in automated responses of commercial buildings and industrial facilities to dynamic electricity prices, Energy Build. 43 (12) (2011) 3322–3330.
[10] M.S. Khawaja, J. Stewart, Long-run savings and cost-effectiveness of home energy report programs, Tech. Rep., Cadmus (Winter), 2015.
[11] National Renewable Energy Laboratory, Uniform Methods Project. http://energy.gov/eere/about-us/ump-home.
[12] 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? International Program Evaluation Conference (1993).
[13] M.L. Goldberg, Reasonable doubts: monitoring and verification for performance contracting ACEEE Summer Study on Energy Efficiency in Buildings, vol. 4, American Council for an Energy Efficient Economy, Pacific Grove, CA, 1996, pp. 133–143.
[14] H. Carstens, X. Xia, S. Yadavalli, A. Rajan, Efficient longitudinal population survival survey sampling for the measurement and verification of building retrofit projects, Energy Build. 150 (2017) 163–176.
[15] H. Carstens, X. Xia, X. Ye, J. Zhang, Characterising compact fluorescent lamp population decay, in: IEEE Africon Conference Pointe-Aux-Piments, Mauritius, 2013, http://dx.doi.org/10.1109/AFRCON.2013.6757715.
[16] H. Carstens, Improvements to Longitudinal Clean Development Mechanism Sampling Designs for Lighting Retrofit Projects, Master's Thesis, University of Pretoria, 2014.
[17] 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.

[18] X. Ye, X. Xia, J. Zhang, Optimal sampling plan for clean development mechanism lighting projects with lamp population decay, Appl. Energy 136 (2014) 1184–1192, http://dx.doi.org/10.1016/j.apenergy.2014.07.056.
[19] X. Ye, X. Xia, L. Zhang, B. Zhu, Optimal maintenance planning for sustainable energy efficiency lighting retrofit projects by a control system approach, Control Eng. Pract. 37 (2015) 1–10.
[20] X. Ye, Optimal Measurement and Verification Plan on Lighting, Ph.D. Thesis, University of Pretoria, 2015.
[21] American Society of Heating, Refrigeration and Air-Conditioning Engineers, Inc, Guideline 14-2014, Measurement of Energy, Demand, and Water Savings, 2014.
[22] X. Xia, J. Zhang, Mathematical description for the measurement and verification of energy efficiency improvement, Appl. Energy 111 (2013) 247–256.
[23] Y. Zhang, Z. O'Neill, B. Dong, G. Augenbroe, Comparisons of inverse modeling approaches for predicting building energy performance, Build. E 86 (2015) 177–190, http://dx.doi.org/10.1016/j.buildenv.2014.12.023.
[24] J. Granderson, S. Touzani, C. Custodio, M.D. Sohn, D. Jump, S. Fernandes, Accuracy of automated measurement and verification (M&V) techniques for energy savings in commercial buildings, Appl. Energy 173 (2016) 296–308, http://dx.doi.org/10.1016/j.apenergy.2016.04.049.
[25] M.-T. Ke, C.-H. Yeh, C.-J. Su, Cloud computing platform for real-time measurement and verification of energy performance, Appl. Energy 188 (2017) 497–507, http://dx.doi.org/10.1016/j.apenergy.2016.12.034.
[26] N.H. Tehrani, U.T. Khan, 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, http://dx.doi.org/10.1109/CCECE.2016.7726749.
[27] J.A. Shonder, P. Im, Bayesian analysis of savings from retrofit projects, ASHRAE Trans. 118 (2012) 367.
[28] Efficiency Valuation Organization, International Performance Measurement and Verification Protocol: Statistics and Uncertainty for IPMVP, 2014.
[29] T. Reddy, D. Claridge, Uncertainty of "measured" energy savings from statistical baseline models, HVAC&R Res. 6 (1) (2000) 3–20.
[30] Y. Heo, Bayesian Calibration of Building Energy Models for Energy Retrofit Decision-Making Under Uncertainty, Ph.D. Thesis, Georgia Institute of Technology, 2011.
[31] R. Luus, Statistical Inference of the Multiple Regression Analysis of Complex Survey Data, Ph.D. Thesis, University of Stellenbosch, 2016.
[32] T. Walter, P.N. Price, M.D. Sohn, Uncertainty estimation improves energy measurement and verification procedures, Appl. Energy 130 (2014) 230–236, http://dx.doi.org/10.1016/j.apenergy.2014.05.030.
[33] V. Barnett, Sample Survey: Principles and Methods, Arnold, 2002.
[34] A.K. Gupta, D.G. Kabe, Theory of Sample Surveys, World Scientific, 2011.
[35] P.S. Levy, S. Lemeshow, Sampling of Populations: Methods and Applications, John Wiley & Sons, 2013.
[36] M.H. Hansen, W.N. Hurwitz, W.G. Madow, Sample Survey Methods and Theory, vol. 1, John Wiley & Sons, 1953.
[37] X. Ye, X. Xia, Optimal metering plan for measurement and verification on a lighting case study, Energy 95 (2016) 580–592.
[38] H. Carstens, X. Xia, X. Ye, Improvements to longitudinal Clean Development Mechanism sampling designs for lighting retrofit projects, Appl. Energy 126 (2014) 256–265, http://dx.doi.org/10.1016/j.apenergy.2014.03.049.
[39] Z. Olinga, A Cost Effective Approach to Handle Measurement and Verification Sampling and Modelling Uncertainties, Master's Thesis, University of Pretoria, 2016.
[40] H. Carstens, X. Xia, S. Yadavalli, Measurement uncertainty and risk in measurement and verification projects, in: International Energy Programme Evaluation Conference, Long Beach, CA, 2015.
[41] 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.
[42] A.C. Harding, D.W. Nutter, Measurement and verification of industrial equipment: sampling interval and data logger considerations, Energy Eng. 113 (6) (2016) 7–33, http://dx.doi.org/10.1080/01998595.2016.11772066.
[43] D. Violette, Impact evaluation accuracy and the incorporation of prior information, in: Energy Program Evaluation Conference, Chicago, 1991, pp. 86–92.
[44] J. Harrison, M. West, Bayesian Forecasting & Dynamic Models, Springer, 1999, http://dx.doi.org/10.1007/b98971.
[45] K. Triantafyllopoulos, Inference of dynamic generalized linear models: on-line computation and appraisal, Int. Stat. Rev. 77 (3) (2009) 430–450, http://dx.doi.org/10.1111/j.1751-5823.2009.00087.x.
[46] D. Gamerman, M. West, An application of dynamic survival models in unemployment studies, Statistician (1987) 269–274, http://dx.doi.org/10.2307/2348523.
[47] D. Gamerman, Dynamic Bayesian models for survival data, Appl. Stat. (1991) 63–79, http://dx.doi.org/10.2307/2347905.
[48] E.R. Brown, J.G. Ibrahim, A Bayesian semiparametric joint hierarchical model for longitudinal and survival data, Biometrics 59 (2) (2003) 221–228 http://www.jstor.org.uplib.idm.oclc.org/stable/3695499.
[49] M. De Iorio, W.O. Johnson, P. Müller, G.L. Rosner, Bayesian nonparametric nonproportional hazards survival modeling, Biometrics 3 (2009) 762–771 http://www.jstor.org.uplib.idm.oclc.org/stable/27919767.
[50] D. Gamerman, H.S. Migon, Dynamic hierarchical models, J. R. Stat. Soc. Ser. B (Methodol.) (1993) 629–642 http://www.jstor.org.uplib.idm.oclc.org/stable/2345875.
[51] J. Kruschke, Doing Bayesian Data Analysis: A Tutorial With R, JAGS, and Stan, 2nd Ed., Academic Press, 2015.
[52] A. Gelman, J.B. Carlin, H.S. Stern, D.B. Rubin, Bayesian Data Analysis, vol. 2, Taylor & Francis, 2014.
[53] M.G. Cox, G.B. Rossi, P.M. Harris, A. Forbes, A probabilistic approach to the analysis of measurement processes, Metrologia 45 (5) (2008) 493 http://stacks.iop.org/Met/45/493.
[54] United States Energy Information Administration, Commercial Building Energy Consumption Survey (CBECS), 2017 https://www.eia.gov/consumption/commercial/.
[55] K. Agnew, M. Goldberg, The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures, National Renewable Energy Laboratory, 2013, Ch 8: Whole-Building Retrofit with Consumption Data Analysis Evaluation Protocol.
[56] R. Kalman, A new approach to linear filtering and prediction problems, J. Basic Eng. 82 (1960) 35–45, http://dx.doi.org/10.1115/1.3662552.
[57] R. Kalman, New methods in Wiener filtering theory, in: J. Bogdanoff, F. Kozin (Eds.), Proceedings of the First Symposium of Engineering Applications of Random Function Theory and Probability, Wiley, New York, 1963.
[58] J. Neyman, Outline of a theory of statistical estimation based on the classical theory of probability, Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Sci. 236 (767) (1937) 333–380 http://www.jstor.org/stable/91337.
[59] D.C. Montgomery, G.C. Runger, Applied Statistics and Probability for Engineers, 4th Ed., John Wiley & Sons, New York, 2007.
[60] UNFCCC, Project Design Document Form: Gauteng, Free State, Mpumalanga, Limpopo, & Northern Cape CFL Replacement Project (1) in South Africa, Tech. Rep., Version 06, 2012.
[61] H. Carstens, X. Xia, S. Yadavalli, Low-cost energy meter calibration method for measurement and verification, Appl. Energy 188 (2017) 563–575, http://dx.doi.org/10.1016/j.apenergy.2016.12.028.
[62] P. Gustafson, Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments, CRC Press, 2003.
[63] W.A. Fuller, Measurement Error Models, Wiley-Interscience, 2006.
[64] R.J. Carroll, D. Ruppert, L.A. Stefanski, C.M. Crainiceanu, Measurement Error in Nonlinear Models: A Modern Perspective, CRC Press, 2006.
[65] IEC 62053-21: Electricity metering equipment (a.c.) – particular requirements – part 21: Static meters for active energy (classes 1 and 2).
[66] IEC 60044-8: Instrument transformers part 8: Electronic current transformers.
[67] F.-A. Fortin, F.-M. De Rainville, M.-A. Gardner, M. Parizeau, C. Gagné, DEAP: evolutionary algorithms made easy, J. Mach. Learn. Res. 13 (2012) 2171–2175.
[68] E. Vine, D. Fielding, An evaluation of residential CFL hours-of-use methodologies and estimates: recommendations for evaluators and program managers, Energy Build. 38 (12) (2006) 1388–1394, http://dx.doi.org/10.1016/j.enbuild.2005.09.008.
[69] 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, Research Into Action, Rome, Italy, 2012.
[70] 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.
[71] 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.
[72] A. Rysanek, R. Choudhary, Optimum building energy retrofits under technical and economic uncertainty, Energy Build. 57 (2013) 324–337, http://dx.doi.org/10.1016/j.enbuild.2012.10.027.
[73] 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, 2009.
[74] Lighting Research Center, Screwbase compact fluorescent lamp products, Specif. Rep. 7 (1) (1999), Rensellaer Polytechnic Institute (June).
[75] Y.C. Kuang, A. Rajan, M.P.-L. Ooi, T.C. Ong, Standard uncertainty evaluation of multivariate polynomial, Measurement 58 (2014) 483–494, http://dx.doi.org/10.1016/j.measurement.2014.09.022.
[76] 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, http://dx.doi.org/10.1109/ACCESS.2015.2415592.
[77] A. Rajan, Y.C. Kuang, M.P.-L. Ooi, S.N. Demidenko, Benchmark test distributions for expanded uncertainty evaluation algorithms, IEEE Trans. Instrum. Meas. 65 (5) (2016) 1022–1034, http://dx.doi.org/10.1109/TIM.2015.2507418.
[78] American Society for Heating, Refrigeration, and Air Conditioning Engineers, Engineering Analysis of Experimental Data, Tech. Rep., American Society for Heating, Refrigeration, and Air Conditioning Engineers, 1986.
[79] L. Brown, T. Cai, A. DasGupta, Interval estimation for a binomial proportion, Stat. Sci. 16 (2) (2001) 101–133 http://www.jstor.org/stable/2676784.