Research Article

Applied Stochastic Models in Business and Industry

Received 23 September 2016, Revised 30 October 2016, Accepted 30 October 2016 Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/asmb.2217

Panel-based stratified cluster sampling and analysis for photovoltaic outdoor measurements

Andrey Pepelyshev[a], Evgenii Sovetkin[b] and Ansgar Steland[b*†]

We study a stratified multisite cluster-sampling panel time series approach in order to analyse and evaluate the quality and reliability of produced items, motivated by the problem to sample and analyse multisite outdoor measurements from photovoltaic systems. The specific stratified sampling in spatial clusters reduces sampling costs and allows for heterogeneity as well as for the analysis of spatial correlations due to defects and damages that tend to occur in clusters. The analysis is based on weighted least squares using data-dependent weights. We show that this does not affect consistency and asymptotic normality of the least squares estimator under the proposed sampling design under general conditions. The estimation of the relevant variance–covariance matrices is discussed in detail for various models including nested designs and random effects. The strata corresponding to damages or manufacturers are modelled via a quality feature by means of a threshold approach. The analysis of outdoor electroluminescence images shows that spatial correlations and local clusters may arise in such photovoltaic data. Further, relevant statistics such as the mean pixel intensity cannot be assumed to follow a Gaussian law. We investigate the proposed inferential tools in detail by simulations in order to assess the influence of spatial cluster correlations and serial correlations on the test's size and power. © 2016 The Authors. Applied Stochastic Models in Business and Industry published by John Wiley & Sons, Ltd.

Keywords: Big data; data science; energy; panel design; photovoltaics; stratification; time series analysis; weighted regression

1. Introduction

Photovoltaics (PV) contributes substantially to the power supply in many developed countries. PV systems may consist of hundreds and thousands of PV modules, which are exposed to heavy operating conditions over many years. This exposure may result in degrading performance and damages, as well as defects and defaults of modules. Collecting on-site outdoor measurements in order to detect such defects and assess their temporal development and impact, as well as to evaluate the overall quality and economic value of such systems, is a challenging task and requires careful designs for sampling and analysis.

In this article, we propose a panel-based methodology for the medium-term and long-term evaluation and analysis of a heterogeneous set of PV systems. There are several problem-specific issues that have to be taken into account and motivated the specification of the proposed approach. In practice, PV systems often consist of PV modules from different module types and even different manufacturers. Further, various other site-specific factors may affect response variables related to quality and reliability issues, for example, the exposition to salt or snow, the type of the electrical connectors or the direct current (DC) to alternating current (AC) converter, also called solar inverter.

Another issue is that quality measurements taken at PV modules may be affected by spatial correlations. There are several phenomena, which may result in such dependencies. For example, it may happen that the modules are installed in the same order as they are produced, such that correlations from the production line are propagated. Another source of correlation is the fact that the PV modules are electrically connected in strings and mechanically mounted in rows (typically one to six), on racks. In this way, electrical, as well as mechanical, issues may cluster. For example, hot spots visible in infrared irradiance images result from low current solar cells (e.g. due to a damage) connected in a string with good cells,

[a] *School of Mathematics, Cardiff University, Senghennydd Road, Cardiff, Wales CF24 4AG, UK*
[b] *Institute of Statistics, RWTH Aachen University, Wüllnerstr. 3, D-52056 Aachen, Germany*
**Correspondence to: Ansgar Steland, Institute of Statistics, RWTH Aachen University, Wüllnerstr. 3, D-52056 Aachen, Germany.*
[†] *E-mail: steland@stochastik.rwth-aachen.de*

such that the low current cell becomes a resistor. The passage of the current through the resistor results in resistive heating, that is, power conversion from electrical energy to thermal energy and power losses. This may result in complete cell defaults. Defects such as those hot spots may increase the electrical load of neighbouring cells and PV modules, so the issue is propagated and thus may occur in cluster. In the present work, we do not try to identify possible causes for such correlations. Instead, we develop a methodology for sampling and statistical analysis, which allows for such correlations.

Those considerations, as well as the fact that collecting measurements from PV modules randomly spread over a relatively large area, generate much higher sampling costs than taking measurements from modules that are located side by side. This motivates to take measurements in spatial clusters where the locations of the clusters are randomly selected.

We consider a stratified random effects linear model with correlated error terms and propose a weighted least squares estimation approach, because ignoring stratification and applying the usual regression estimator as implemented in standard statistical packages would lead to biased and thus invalid inference. Simulations not reported here show that the weighted estimator $\widehat{\theta}_n$ is robust with respect to moderate departures from the planned stratification sampling scheme. Correlations, due to cluster sampling, have to be taken into account and are modelled by a random effect. However, we propose to employ a nonparametric estimator of the variance–covariance matrix, which does not require certain structured dependencies, such that the approach goes even beyond the random effects model and allows for arbitrary unstructured variance–covariance models for the clusters.

Clearly, the analysis of the temporal development of the quality requires a panel design, which may result in serially correlated data. We model the temporal development by linear time series models. As strata usually represent important characteristics that may change over time, for example, certain focus defects, we further elaborate a model that relates the strata proportions to an underlying quality feature. We also discuss how to set up confidence intervals for the defect rates, taking into account the complex sampling design.

The statistical properties of the resulting approach are extensively investigated by Monte Carlo simulations under realistic model specifications to allow statisticians and engineers involved in such a study to balance sampling costs and statistical estimation accuracy as well as statistical power. In particular, the simulations aim to clarify the impact of several model parameters of interest such as the degree of within-cluster correlations.

The organization of the rest of the paper is as follows. In Section 2, we briefly review some basic facts of stratified sampling and provide details of the proposed stratified cluster sampling panel approach. Section 3 presents the proposed approach to model the data and to make valid inference, focusing on cross-sectional stratified cluster sampling for a set of heterogeneous PV systems. The extension to a panel time series design is discussed in Section 5. Section 6 deals with a temporal model for defect rates and discusses how one may construct asymptotically valid confidence intervals for the canonical estimators. By analysing real outdoor measurements from electroluminescence imaging, we demonstrate, firstly, that the spatial correlations and local clusters arise in such PV data and, secondly, that the normality assumptions imposed by classical linear mixed models are violated (Section 7), providing empirical justification for our approach. The results of extensive Monte Carlo simulations are presented in Section 8.

2. The general sampling design

The general sampling approach, which combines the concepts of stratified sampling, random cluster sampling and a panel design, can certainly be applied to a large number of field studies. Nevertheless, we shall explain and elaborate the approach in terms of the PV industrial application that motivated the work, namely, to design a methodology for the evaluation and analysis of large PV systems.

The aim was to elaborate a solution that allows to answer different important questions ranging from the estimation of defect rates to the testing of site-specific effects (such as the module type) based on measurements taken at several time points. Simultaneously, the sampling costs have to be taken into account and balanced with the statistical properties, which generally motivates the choices of the sample sizes.

Initially, at time zero, we propose to draw, at each site, a large random sample of size $n = 400$; samples at later stages may consist of 200 modules. The initial sample may be stratified by the module type (or manufacturer), as well as by a small number of important *focus defects* or predamages, such as an increased number of microcracks in the electroluminescence (EL) image supplied by the manufacturer or made after having installed the PV system. The reason to conduct stratified sampling is that for those focus defects a minimal sample size, say 40, should be guaranteed, in order to allow certain strata-wise statistical analyses, even when the strata correspond to rare events. Suppose, for simplicity, that there are $k = 4$ strata and stratum 0 represents *good* modules, whereas the other strata with probabilities between 0.05 and 0.15 represent defects. The proposed stratified sampling is always conducted in multiples of the strata sample sizes $(m_1, \dots, m_4) = (80, 40, 40, 40)$, respectively, according to the proportions $2 : 1 : 1 : 1$. Let w_i denote the true proportion of stratum i in the population. Then the distribution function (d.f.) of a randomly drawn measurement X is given

by $F(x) = \sum_{i=1}^{k} w_i F_i(x)$, where F_i is the d.f. of the ith stratum. If stratified sampling is applied, one replaces the w_i by $m_i/(m_1 + \cdots + m_k)$. Given estimates $\widehat{F}_i(x)$ for the ith stratum, one may easily estimate the population's d.f. $F(x)$.

It may happen that the classification of the modules with respect to the strata is unknown at the beginning of the study and thus has to be determined after having drawn the initial random sample of PV modules, for example, by appropriate measurements (e.g. based on expert audits, power output measurements and EL or infrared imaging) and by applying clustering techniques. In this case, one may proceed as follows: we observe for each quality feature X a bivariate sample $(X_1, \delta_1), \ldots, (X_n, \delta_n)$, where $\delta_i \in \{1, \ldots, k\}$ indicates the observed strata of the ith observation. Let us assume that $E(\delta_i)$ coincides with the true class of X_i for all i. Then the unknown proportion w_i of each strata can be easily estimated by $\widehat{w}_i = \widetilde{n}_i/n$ with $\widetilde{n}_i = \sum_{j=1}^{n} \mathbf{1}(\delta_j = i)$, for $i = 1, \ldots, k$. Further, one may estimate $F(x)$ by $\widehat{F}(x) = \sum_{i=1}^{k} \widehat{w}_i \widehat{F}_i(x)$. In general, those strata proportions (defect rates) do not coincide with the required strata sample sizes. As discussed in [1], one then may continue sampling until all strata have the required sample sizes.

Consider now a multisite study where a, $a \in \mathbb{N}$, PV systems (sites) are under investigation, such that the pooled PV modules represent a random sample of the underlying overall population of PV modules. By weighting the aforementioned estimates with the proportion of PV modules, $n_v/(n_1 + \cdots + n_a)$, where n_v denotes the number of PV modules installed at site v, $v = 1, \ldots, a$, one may easily generalize the aforementioned estimators.

The initial (stratified) large random sample allows assessment of the PV systems by current-voltage curves (IV curves) and EL imaging, in order to evaluate the initial quality and economic value of the systems in terms of power output, power losses and damages, defects (such as microcracks or cell breaks) and predicted future returns, right from the beginning. To assess the temporal development, one has to conduct a longitudinal study resulting in time series observations. Because, in practice, monitoring all initially selected PV modules over a longer time span is infeasible, we propose to select a panel. From the initial sample at time $t = 0$ one selects, say, 50 *anchor* modules by random, forming the core of the *panel* for the longitudinal study. Under the proposed stratification scheme, one may simply select half of the modules in each stratum by random. At each time instant, each of those 50 modules and four direct neighbours are measured. This means the sampling is conducted in randomly selected clusters (batches) consisting of $b = 5$ modules, thus resulting in a sample of size 200 at each time point. Consequently, we are given a panel time series where the panel of PV modules is observed over the time span of interest.

It is important to select the anchor modules spatially at random, within the strata defined by the stratification variable (module type and manufacturer), in order to ensure that the whole area of the PV site is covered. Taking that sample as a random sample of possibly correlated batches has two additional advantages: the engineer can easily detect clusters of defects and may perhaps infer their causes. In addition, those clusters can be used to estimate local spatial correlations. Usually, there is an additional effect present in such a cluster. For example, for PV outdoor measurements, there is a row effect as the modules are typically installed in two rows, and according to expert knowledge, measurement from the upper and lower row may differ.

We assume that the panel is remeasured at later time instants, for example, on an annual basis. In what follows, we assume that the study is planned for a small number of time points, such that serial correlations may be present. Then we are given only a short (vector) time series, such that methods from time series analysis cannot be applied. Therefore, we shall rely on a multiple testing procedure to test for interesting effects, especially main effects associated with a site or a certain defect. According to our Monte Carlo investigations, the proposed multiple testing approach is not severely affected by serial correlations, which justifies the approach: ignoring serial correlations provides a good approximation for the settings studied in our simulations.

3. Stochastic model and inference

In order to analyse data collected at, say, a PV sites according to the aforementioned sampling design, we consider a stratified linear model with random effects; for a general exposition on linear models, we refer to [2]. This framework allows us to take into account fixed site-specific effects represented by regression vectors $\mathbf{x}_v$ for site $v \in \{1, \ldots, a\}$. For an empirical study designed to investigate influential factors, one may select the sites with respect to factors of interest, such as the type of the converter, the type of grounding, the geographical location or the exposure to external factors, such as salt or snow. If the engineer is interested to keep things simple, we propose to consider binary effects (coded by $+1$ and -1) and to consider a (nearly) orthogonal design to obtain statistically sound estimates. Orthogonal design vectors allow for uncorrelated estimation of the associated regression coefficients (main site effects), whereas a nearly orthogonal design aims at approximating orthogonality when it cannot be achieved exactly; for an exposition, we refer to [3]. Table I illustrates, as an example, a nearly orthogonal design for the case of $a = 5$ PV sites and four binary factors.

Let $Y_{\ell j}^{(v)}$ be the jth observation of stratum ℓ at a fixed site v and let $\mathbf{Y}^{(v)} = (Y_{\ell j}^{(v)} : j = 1, \ldots, m_\ell, \ell = 1, \ldots, k)$. We may and shall assume that all observations are sorted in such a way that the first m_1 entries correspond to the first strata, the

Table I. A design for five sites and four factors with two levels.

Number of site	Location	Place	Solar inverter	Grounding
1	+1	+1	+1	+1
2	+1	−1	−1	−1
3	−1	+1	−1	−1
4	−1	−1	+1	−1
5	−1	−1	−1	+1

next m_2 to strata two and so forth. To analyse the data in the presence of regressors as introduced earlier, we assume the linear model

$$Y_{\ell j}^{(v)} = \mu + \boldsymbol{\beta}'\mathbf{x}_v + \boldsymbol{\zeta}'\mathbf{z}_{v\ell j} + \widetilde{\epsilon}_{v\ell j}, \tag{3.1}$$

where

$$\widetilde{\epsilon}_{v\ell j} = \epsilon_{B,\lfloor (j-1)/b \rfloor + 1} + \epsilon_{v\ell j}, \tag{3.2}$$

for $\ell = 1, \dots, k, j = 1, \dots, m_\ell$. Here, $\mathbf{x}_v$ is the, say, p-dimensional regression vector with the experimental conditions of site v (extensions are discussed next), $\mathbf{z}_{v\ell j}$ is a q-dimensional vector of additional explanatory variables, $p, q \in \mathbb{N}$, μ is the grand mean, $\boldsymbol{\beta}$ and $\boldsymbol{\zeta}$ are the regression coefficients and $\epsilon_{\ell vj}$ are mean zero and independently distributed error terms representing the measurement error when observing the jth measurement of ℓth stratum. ϵ_{Br} represents the mean zero random effect of the rth cluster (batch), $r = 1, \dots, m_\ell/b$, assumed to be i.i.d. with common variance σ^2 and independent from the measurement errors $\epsilon_{v\ell j}$. At this point, we confine ourselves to that model and postpone discussion of several extensions to the subsequent sections. Notice that the design matrix of model (3.1) takes the form $\mathbf{X}^{(v)} = (\mathbf{D}^{(v)} | \mathbf{Z}^{(v)})$, where $\mathbf{Z}^{(v)}$ is the matrix with rows $\mathbf{z}'_{v\ell j}$, $\mathbf{D}^{(v)} = \mathbf{1}_{n_v} \otimes \mathbf{x}'_v$ and $\otimes$ is the Kronecker product. The upper left block of the block matrix $\mathbf{X}^{(v)\prime}\mathbf{X}^{(v)}$ is given by $\mathbf{D}^{(v)\prime}\mathbf{D}^{(v)}$, which is a diagonal matrix for an orthogonal design.

The aforementioned model can be used to analyse data from one site or data from all sites. Let us first discuss the former case. Fix $v \in \{1, \dots, a\}$ and put $\mathbf{e}^{(v)} = (\epsilon_{v\ell j} : j = 1, \dots, m_\ell, \ell = 1, \dots, k)'$. Observe that the first m_1 entries correspond to the first strata and follow the d.f. $F_1^{(v)}$, the next m_2 to strata two and have distribution $F_2^{(v)}$ and so forth. For simplicity, we assume that the size of all batches is equal and each batch contains only PV modules of the same stratum. Then the variance–covariance matrix of $\mathbf{e}^{(v)}$ is given by

$$\boldsymbol{\Sigma}_v = \mathrm{Cov}\,(\mathbf{e}^{(v)}) = \bigoplus_{\ell=1}^{k} \bigoplus_{r=1}^{m_\ell/b} \boldsymbol{\Sigma}_\ell^{(v)}, \tag{3.3}$$

where

$$\boldsymbol{\Sigma}_\ell^{(v)} = \sigma^2 \mathbf{J}_b + \sigma_{v\ell}^2 \mathbf{I}_b \tag{3.4}$$

with $\sigma_{v\ell}^2 = \int x^2 dF_\ell^{(v)}$ being the variance of the measurement error at site v for stratum ℓ. Here, $\mathbf{A} \oplus \mathbf{B}$ stands for the direct sum of two matrices $\mathbf{A}$ and $\mathbf{B}$, that is, the block diagonal matrix with upper left submatrix $\mathbf{A}$ and lower right submatrix $\mathbf{B}$, $\mathbf{J}_b = \mathbf{1}_b\mathbf{1}_b'$ with $\mathbf{1}_b = (1, \dots, 1)' \in \mathbb{R}^b$ and $\mathbf{I}_b$ is the b-dimensional identity matrix. It is easy to see that the within-cluster correlation for stratum ℓ is given by $\rho_{v\ell} = \sigma^2/(\sigma^2 + \sigma_{v\ell}^2)$.

Of course, the structured variance–covariance (3.4) results from the linear random effects model (3.2). More generally, we may also allow for an unstructured variance–covariance matrix. To do so, partition $\mathbf{e}^{(v)} = \left(\mathbf{e}_{11}^{(v)\prime}, \dots, \mathbf{e}_{1,m_1/b}^{(v)\prime}, \dots, \mathbf{e}_{k1}^{(v)\prime}, \dots, \mathbf{e}_{k,m_k/b}^{(v)\prime}\right)'$, where the b-dimensional random vectors $\mathbf{e}_{\ell r}^{(v)}$ consist of the corresponding mean zero error terms $\epsilon_{v\ell j}$ with marginal d.f. F_ℓ. The unstructured (fully unspecified) cluster sampling model now assumes that

$$\mathbf{e}_{\ell r}^{(v)} \overset{\text{i.i.d.}}{\sim} \left(0, \boldsymbol{\Sigma}_\ell^{(v)}\right), \tag{3.5}$$

for unknown variance–covariance matrices $\boldsymbol{\Sigma}_\ell^{(v)}$, $\ell = 1, \dots, k$.

If the aforementioned model is used to analyse the data of the whole study, one puts

$$\mathbf{Y} = (Y^{(1)} \dots, Y^{(a)\prime})', \quad \text{and} \quad \mathbf{e} = (\mathbf{e}^{(1)\prime}, \dots, \mathbf{e}^{(a)\prime})'.$$

Assuming independence across sites, the variance–covariance matrix of the errors is given by

$$\boldsymbol{\Sigma} = \bigoplus_{v=1}^{a} \boldsymbol{\Sigma}_v.$$

The design matrix is $\mathbf{X} = (\mathbf{X}^{(1)\prime}|\cdots|\mathbf{X}^{(a)\prime})'$ with $\mathbf{X}^{(v)}$ as aforementioned. A straightforward calculation shows that the upper left submatrix of $\mathbf{X}'\mathbf{X}$ is a diagonal matrix for an orthogonal design, and the same applies to $\mathbf{X}'\mathbf{L}\mathbf{X}$ for any diagonal matrix $\mathbf{L}$.

The parameter vector of interest is $\boldsymbol{\theta} = (\mu, \boldsymbol{\zeta}', \boldsymbol{\beta}')'$. Following [4], we consider the weighted estimator $\widehat{\boldsymbol{\theta}}_n$ of $\boldsymbol{\theta}$ as a solution of the following minimization problem

$$\sum_{v=1}^{a}\sum_{\ell=1}^{k}\frac{\widehat{w}_\ell^{(v)}}{m_\ell}\sum_{j=1}^{m_\ell}\left(Y_{\ell j}^{(v)} - \mathbf{x}_v'\boldsymbol{\beta} - \mu\right)^2 \to \min_{\boldsymbol{\theta}}.$$

This estimator can be written explicitly as follows

$$\widehat{\boldsymbol{\theta}} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}Y,$$

where $\mathbf{Y} = \left(y_{11}^{(1)}, \ldots, y_{1m_1}^{(1)}, y_{21}^{(2)}, \ldots, y_{k,m_k}^{(a)}\right)'$ and

$$\mathbf{W} = \operatorname{diag}\left(\widehat{w}_1^{(1)}/m_1, \ldots, \widehat{w}_1^{(1)}/m_1, \widehat{w}_2^{(1)}/m_2, \ldots, \widehat{w}_k^{(a)}/m_k\right).$$

Notice that if the weights and therefore $\mathbf{W}$ are deterministic or estimated from an independent sample, then the covariance matrix (given the learning sample to estimate W) of the estimator $\widehat{\boldsymbol{\theta}}_n$ has the usual form

$$\operatorname{Cov}(\widehat{\boldsymbol{\theta}}_n) = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\boldsymbol{\Sigma}\mathbf{W}\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}.$$

If b and k are fixed, $\mathbf{e}$ is a b-dependent series, such that the least squares estimator can be seen to be consistent and asymptotically normal under routine regularity conditions, as $\min_\ell m_\ell \to \infty$.

For general stochastic weights, we have the following sufficient conditions for the validity of the method, which we formulate for a fixed site and thus omit the corresponding index. Suppose that

$$m_\ell/n \to \mu_\ell \in (0, \infty), \tag{3.6}$$

for $\ell = 1, \ldots, k$. Further, let us assume that the weighted least squares estimator satisfies the following regularity assumptions: Allowing for (non-random) regressors that may depend on the strata $\ell \in \{1, \ldots, k\}$ and the repeated measurement $j \in \{1, \ldots, m_\ell\}$, we assume that

$$\frac{1}{n}\sum_{\ell=1}^{k}\sum_{j=1}^{m_\ell} w_\ell \mathbf{x}_{\ell j}\mathbf{x}_{\ell j}' \to \mathbf{C}, \tag{3.7}$$

as $\min_\ell m_\ell \to \infty$, for some regular matrix $\mathbf{C}$, the second moments of the regressors $\mathbf{x}_{\ell j} = (x_{\ell j1}, \ldots, x_{\ell jp})'$ converge strata-wise,

$$\frac{1}{m_\ell}\sum_{j=1}^{m_\ell} x_{\ell jv}^2 \to m_{2v}^{(\ell)}, \tag{3.8}$$

as $m_\ell \to \infty$, for vectors $\mathbf{m}_2^{(\ell)} = \left(m_{21}^{(\ell)}, \ldots, m_{2p}^{(\ell)}\right)'$, $\ell = 1, \ldots, k$, and

$$\frac{1}{\sqrt{n}}\sum_{\ell=1}^{k}\sum_{j=1}^{m_\ell} w_\ell \mathbf{x}_{\ell j}\widetilde{\epsilon}_{\ell j} \xrightarrow{d} \mathcal{N}(0, \mathbf{S}), \tag{3.9}$$

as $\min_\ell m_\ell \to \infty$, for some regular matrix $\mathbf{S}$ that depends on the cluster variance–covariance matrix $\boldsymbol{\Sigma}_\ell$, $\ell = 1, \ldots, k$. If the stochastic weights satisfy

$$\widehat{w}_\ell = w_\ell + O_P\left(1/\sqrt{m_\ell}\right), \qquad \ell = 1, \ldots, k, \tag{3.10}$$

then the weighted least squares estimator with stochastic weights $\widehat{w}_\ell$ has the same asymptotic distribution as the weighted least squares estimator with deterministic weights w_ℓ:

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{d} \mathcal{N}(0, \mathbf{C}^{-1}\mathbf{S}(\mathbf{C}^{-1})'),$$

as $\min_\ell m_\ell \to \infty$. That central limit theorem also yields the weak consistency of the proposed estimator. It is worth mentioning that Assumption (3.10) is very weak. In particular, the weights can be estimated in-sample, that is, from the

same data to be analysed. For the readers convenience, a proof that adopts arguments as detailed in [5, ch. 8.2.1] is sketched in the Appendix.

It remains to discuss how to estimate the unobservable variance–covariance matrix. Recall that the covariance matrix $\mathbf{\Sigma}$ of $\mathbf{Y}$ is a block-diagonal matrix with blocks of size $b \times b$ where each block corresponds to a cluster. There are k different blocks $\mathbf{\Sigma}_\ell^{(v)}$, $\ell = 1, \dots, k$, corresponding to the k different strata. We estimate $\mathbf{\Sigma}_\ell^{(v)}$ nonparametrically as follows. Based on the weighted least squares estimator $\widetilde{\boldsymbol{\theta}}$, we calculate the corresponding residuals. The residuals for the ℓth stratum are collected in a matrix, $\mathbf{E}_\ell^{(v)}$ with k columns and number of rows equal to the number of clusters in this stratum. Then, $\mathbf{\Sigma}_\ell^{(v)}$ is estimated by the sample covariance matrix $\widehat{\mathbf{\Sigma}}_\ell^{(v)}$ corresponding to $\mathbf{E}_\ell^{(v)}$, which is a consistent estimator under fairly general conditions even if the specific structure of dependencies of the linear model approach does not hold. Now, the variance–covariance matrices $\mathbf{\Sigma}^{(v)}$ and $\mathbf{\Sigma}$ are estimated by substituting the $\mathbf{\Sigma}_\ell^{(v)}$ by their estimates in the aforementioned formulas.

4. Comparison with linear mixed models and extensions to nested models

In this section, we compare the aforementioned nonparametric modelling and estimation approach with the classical linear mixed models approach [6] and discuss its extension to nested models as arising, for example, when repeated measurements at each module are available. The linear mixed model is similar as (3.1) but imposes much stronger assumptions. It is given by

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon},$$

where it is assumed that $\mathbf{y}|\boldsymbol{\gamma} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma}, \sigma^2\mathbf{I})$ and $\boldsymbol{\gamma} \sim \mathcal{N}(0, \sigma^2\mathbf{D})$, for some unknown matrix $\mathbf{D}$.

In order to estimate the fixed effects parameter vector, $\boldsymbol{\beta}$, of the model and the distribution parameters, $(\sigma, \mathbf{D})$, of the random effects $\boldsymbol{\gamma}$, one employs maximum likelihood or restricted maximum likelihood approaches. In the latter approach, the optimization problem is split into two parts, namely, estimation of fixed effects parameters and random effects [7].

In order to compare our ansatz with the classical linear mixed model approach, we remark that the model described in Section 3 can be easily extended to several random effects. For the application discussed in this paper, we shall be mostly interested in the so-called nested random effects models, as there is strong evidence that our real data have this structure (Section 7), where this issue is discussed in greater detail and substantiated by data analysis.

The nested structure arises when we not only observe a measurement $Y_{lj}^{(v)}$ for each module but also a measurement $Y_{ljc}^{(v)}$ for each cell $c = 1, \dots, n_{jl}$ inside module $j = 1, \dots, m_l$ in batch $l = 1, \dots, k$ of site (or strata) v, where n_{jl} denotes the number of cells of the corresponding module. Then model (3.1) takes the form

$$Y_{ljc}^{(v)} = \mu + \boldsymbol{\beta}'\mathbf{x}_v + \boldsymbol{\zeta}'\mathbf{z}_{vljc} + \bar{\epsilon}_{vljc}, \qquad c = 1, \dots, n_{jl}, j = 1, \dots, m_l, l = 1, \dots, k,$$

where the error terms $\bar{\epsilon}_{vljc}$ are mean zero with a variance–covariance matrix $\mathbf{\Sigma}$ reflecting the nested random effects structure. A common approach is to assume that, for independent mean zero random variables $\epsilon_{B,vl}$ (random effects for modules inside batch l), $\epsilon_{C,vj}$ (random effects for cells inside module j) and ϵ_{vljc} (measurement error), all with finite variances,

$$\bar{\epsilon}_{vljc} = \epsilon_{B,vl} + \epsilon_{C,vlj} + \epsilon_{vljc},$$

for $c = 1, \dots, n_{jl}, j = 1, \dots, m_l, l = 1, \dots, k$.

Observe that this specification is in agreement with typical assumptions made in the classical linear mixed model, especially equi-correlated errors within a module. Those assumptions, however, can be relaxed easily in our approach.

Observe that, similar as in Section 3, the covariance matrix $\mathbf{\Sigma}$ has a nested block structure; that is, each non-zero block is a block matrix on its own and has the following form

$$\mathbf{\Sigma} = \bigoplus_{v=1}^{n} \bigoplus_{l=1}^{k} \left(\mathbf{I}_{m_l} \otimes \left(\mathbf{\Gamma}_0^{(v)} - \mathbf{\Gamma}^{(v)} \right) + \mathbf{J}_{m_l} \otimes \mathbf{\Gamma}^{(v)} \right), \tag{4.1}$$

where m_l is the number of modules in batch l (possibly different for each batch) and the matrices $\mathbf{\Gamma}_0^{(v)}$ and $\mathbf{\Gamma}^{(v)}$ are square matrices with dimensions equal to the number of cells in a module. Under the aforementioned specification, $\mathbf{\Gamma}_0^{(v)}$ has identical elements on the diagonal as well as identical off-diagonal elements and the matrix $\mathbf{\Gamma}^{(v)}$ has identical elements everywhere.

We may weaken those assumptions for the variance–covariance matrix of the error terms by assuming only unknown correlations between different cells in a module (that are not necessarily equal). In that case, both matrices $\mathbf{\Gamma}_0^{(v)}$ and $\mathbf{\Gamma}^{(v)}$ in (4.1) are unknown and have to be estimated as well.

Estimators for the variance–covariance matrix $\mathbf{\Sigma}$ can be constructed in a similar way as described at the end of Section 3, namely, by averaging sample covariance matrices of the vector of observations corresponding to the cells inside each module.

Contrary to a linear mixed models approach, we do not require a normality assumption of the error terms, which is restrictive for applications and, as shown in Section 7, is not valid for our data, because the method of estimation does not require to specify distributions and the conditions we impose for asymptotic normality of the fixed effects coefficients are nonparametric. Further, in our approach, we may allow weaker assumptions on the variance–covariance structure of the error terms, which allows to take into consideration a more general model for the observations of the linear mixed models.

5. Extended panel time series model and inference

Let us now extend the model of the previous section to the case that the anchor modules and the associated clusters are observed at T time instants thus forming a vector time series. We shall also generalize the model for the cluster effect to the case that a random effect inducing within-cluster correlations is only present with a certain probability.

To allow for serial correlations, we consider an additional additive component that follows a linear time series model. For general expositions on such models and extensions, especially to the class of ARMA (autoregressive-moving-average) models and linear processes, we refer to [8] and [5].

We assume that panels of a sites are observed over time at T time equidistant points, where T is small. At the t-th time point, $t = 1, \dots, T$, measurements are taken according to a stratified cluster sampling approach with four strata, stratified sample sizes $(m_1, \dots, m_4) = m^*(2, 1, 1, 1)$ for some integer m^* and a cluster size b. We formulate the model for a fixed site and therefore omit the site index v. So let us assume that

$$Y^{(v)}_{\ell jt} = \mu + \boldsymbol{\beta}' \mathbf{x}_{\ell jt} + \xi^t_{\ell j} + \epsilon^t_{\ell d_j}, \tag{5.1}$$

where $\xi^t_{\ell j}$ are the strata-related error components to be discussed next and $\epsilon^t_{\ell d_j}$ are cluster-related error terms, $d_j = \lfloor (j-1)/b \rfloor + 1, j = 1, \dots, m_\ell, \ell = 1, \dots, k$ and $t = 1, \dots, T$. Further regressors are omitted to keep the exposition brief but are straightforward to take into account.

The cluster-related random effect is now assumed to follow a mixture distribution

$$\mathcal{L}\left(\epsilon^t_{d_j}\right) = p\mathcal{N}(0, \sigma^2) + (1-p)\delta_0, \tag{5.2}$$

where δ_0 stands for the Dirac measure in 0. This means, with probability p within-cluster correlations are present, which degree is controlled by the parameter σ.

The strata-related error $\xi^t_{\ell j}$ is assumed to be governed by a stationary autoregressive AR(1) model,

$$\xi^t_{\ell j} = \gamma \xi^{t-1}_{\ell j} + c\varepsilon^t_{\ell j}, \qquad t = 1, \dots, T, \tag{5.3}$$

where γ and c are parameters and $\varepsilon^t_{\ell j}$ is a Gaussian white noise process; that is, $\varepsilon^t_{\ell j}$ are i.i.d. $\mathcal{N}(0, 1)$. Extensions to ARMA models,

$$\phi(L)\xi^t_{\ell j} = \psi(L)\varepsilon^t_{\ell j}, \tag{5.4}$$

for lag polynomials $\phi(L)$ and $\psi(L)$, L the lag operator, are straightforward. For simplicity of exposition, we confine ourselves to an AR(1) model. The AR parameter γ controls the dependence between time points and is assumed to satisfy the stationarity condition $\gamma \in (-1, 1)$. The variance of the strata-related errors becomes now a function of both γ and c,

$$\sigma^2_{\ell j} = \mathrm{Var}\ \left(\xi^t_\ell\right) = \frac{c}{\sqrt{1-\gamma^2}}.$$

The model (5.1) combines spatial correlations as well as serial correlations of the measurements. Statistical inference can be conducted as follows: at each time instant t, the available observations are analysed using the test introduced in the previous section. The global null hypothesis that there is no (main) effect at all is rejected, if one test rejects. Obviously, we have to deal with the issue of a multiple testing problem, because T hypotheses are now tested. There are two well-known approaches to handle this, namely, the Bonferroni correction method and the Šidák corrections, where the latter assume independent samples. Because in model (5.1) the time series component is additive, any statistic U_t that depends only on the time t observations treats its value as a constant, which therefore can be absorbed in the intercept. Hence, $U_1, \dots, U_T$ are conditionally independent given $\{Y^t_{\ell jv} : t = 1, \dots, T\}$, and it is easy to verify that the Šidák corrections apply.

6. Modelling and analysing strata proportions (defect rates)

Let us now discuss how to estimate the strata proportions and how to assess the estimation error within the proposed stratified cluster sampling approach. Estimation of the strata proportions is, of course, interesting in its own right, but it is deeply motivated by the fact that for the PV application of interest, the strata usually correspond to relevant defects or damages of solar cells and PV modules, respectively. Even if continuous quality measurements are available such that the approaches of the previous sections are applicable, valid confidence intervals for the defect rates are of interest.

Of course, in the presence of quality measurements, the defect rates are related to the distribution of those measurements. Therefore, let us consider the following threshold model that links the strata to an underlying random variable X. Given strictly ordered thresholds τ_ℓ, $\ell = 0, \dots, k$, with $\tau_0 = -\infty$ and $\tau_k = \infty$, where k is the number of strata, a PV module with quality measurement X, for example, the power output, belongs to stratum ℓ, if

$$\tau_{\ell-1} < X < \tau_\ell.$$

The corresponding proportions are

$$p_\ell = P(\tau_{\ell-1} < X < \tau_\ell),$$

for $\ell = 1, \dots, k$. That threshold model can be linked to the model of the previous section as follows: for given strata proportions at each time instant t, one may calculate the associated thresholds from the d.f. of $Y^{(v)}_{\ell jt}$, which can be easily determined for given model parameters. Having calculated all resulting thresholds $\tau^{(v)}_{\ell t}$, one can determine the right strata for each simulated data set $\{Y^{(v)}_{\ell jt}\}$ by simple classification. After these preparations, for each (simulated) sample of size n_v, we may therefore calculate the relative frequencies of the strata,

$$\widehat{p}_n(t, \ell) = \frac{1}{m_\ell} \sum_{j=1}^{m_\ell} \mathbf{1}\left(\tau^{(v)}_{\ell-1,t} < Y^{(v)}_{\ell jt} < \tau^{(v)}_{\ell t}\right), \qquad \ell = 1, \dots, k, \tag{6.1}$$

to estimate the true strata proportions $p(t, \ell) = E\widehat{p}_n(t, \ell)$.

In order to calculate confidence intervals for the true strata proportions, it is necessary to take into account the dependence structure induced by the stratified cluster sampling approach. The usual formulas based on the binomial distribution are not valid, of course.

We may rely on the following generic asymptotic result (see the Appendix for a derivation): let $\boldsymbol{\xi}_i = (\xi_{i1}, \dots, \xi_{ib})' \in \mathbb{R}^b$ are r i.i.d. random vectors with common variance–covariance matrix $\boldsymbol{\Sigma}$, where each coordinate is distributed according to the Bernoulli distribution with parameter p. Then

$$\sqrt{n}(\hat{p}_n - p) \xrightarrow{d} \mathcal{N}\left(0, 1/b\mathbf{1}_b'\boldsymbol{\Sigma}_\xi\mathbf{1}_b\right), \tag{6.2}$$

where $n = rb$ and b stands for the cluster size. The unknown variance–covariance matrix $\boldsymbol{\Sigma}_\xi$ can be estimated nonparametrically as discussed in Section 3 when replacing the Y-observations by the corresponding indicators appearing in (6.1). Denote the corresponding estimator by $\widehat{\boldsymbol{\Sigma}}_\xi$.

The central limit theorem (6.2) justifies the following asymptotic confidence intervals,

$$\left[\hat{p}_n - z_{1-\alpha/2}\frac{\widetilde{\sigma}_n}{\sqrt{n}}, \hat{p}_n + z_{1-\alpha/2}\frac{\widetilde{\sigma}_n}{\sqrt{n}}\right],$$

where

$$\widetilde{\sigma}_n^2 = 1/b\mathbf{1}_b'\widehat{\boldsymbol{\Sigma}}_n\mathbf{1}_b$$

and $z_{1-\alpha/2}$ denotes the $(1-\alpha/2)$-quantile of the standard normal distribution.

7. Data analysis

We analysed real data collected at a PV system in order to investigate whether the complex structure of (nested) spatial correlations arises in real PV outdoor measurements. Further, we were interested in checking whether or not the strong assumptions, as imposed by the classical linear mixed model approach to analyse such data, hold. The analysed data are field measurements taken at the beginning of the PV-Scan study. From each PV module of a randomly drawn batch of

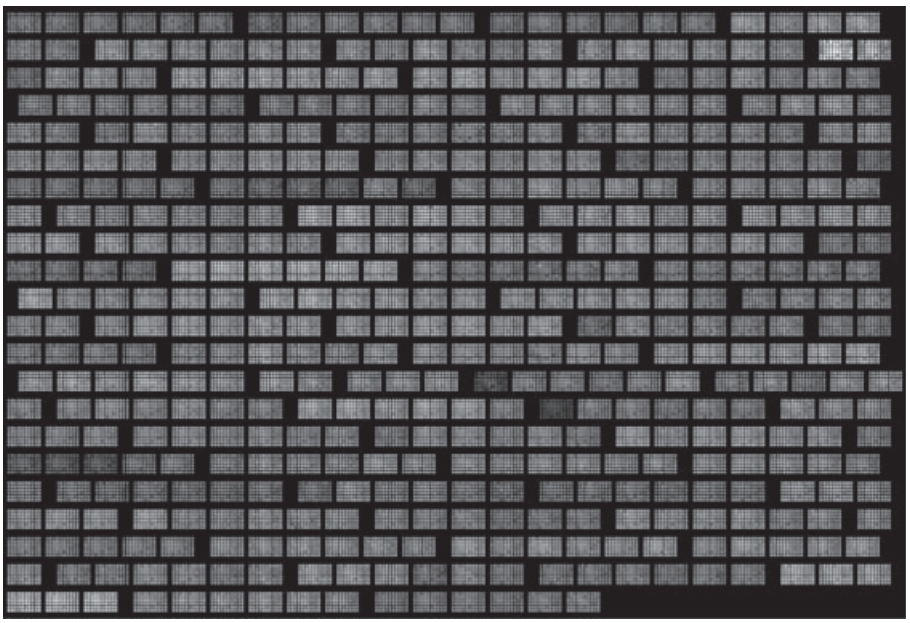

Figure 1. Average cell electroluminescence intensity grouped by batches. Each dot represents the average cell activity. The batch-to-batch variation is noticeably larger than the within-batches and within-module variations.

modules belonging to the same string, an EL image was taken without demounting the module, resulting in more than 2000 images.

Firstly, each EL image was preprocessed to correct for optical distortions of the camera's lens and perspective distortions that arise when the camera's sensor is not parallel to the PV module. A particular issue arising here is the robust estimation of the horizontal and vertical lines determining the boundaries of the module's cells and a cell's areas, which are separated by the grid fingers collecting the current. Here, a specialized algorithm was developed based on robust regression [9]. Then the cells of each PV module were extracted automatically and resized to a common (pixel) resolution. Lastly, several summary statistics were calculated including the average EL intensity of each solar cell, on which we shall focus in what follows.

Figure 1 depicts the average EL intensity of all modules of a PV system. Each PV module corresponds to a subrectangle of 60 dots representing the 60 average cell intensities. The PV modules corresponding to a batch are put side by side, and the batches are plotted side by side as well, from left to right and row by row, without any correspondence to their (physical) spatial location.

It is clearly visible that the batch-to-batch variation of the average EL intensity is substantially larger than that of the within-batches variation. This indicates that a random effects models is appropriate for modelling and analysis.

We analysed the dependence of the average cell intensity on the batch (random), module (random, nested within batch) and the cell (random, nested within module). The model was fitted using the R package lme4 using maximum likelihood under the assumption of Gaussianity. The variance, σ^2, of the random batch effect is estimated by $\widehat{\sigma}_n^2 = 226.10$ with standard error $\widehat{\mathrm{sd}}(\widehat{\sigma}_n^2) = 15.037$ and the variance, σ_M^2, of the random module effect by $\widehat{\sigma_M}^2 = 14.49$ with standard error $\widehat{\mathrm{sd}}(\widehat{\sigma_M}^2) = 3.807$. Hence, confidence intervals at the usual confidence levels do not cover 0, indicating the significance of the effects. Fitting a reduced model without the module effect leads to $\widehat{\sigma}_n^2 = 229.7$ with standard error $\widehat{\mathrm{sd}}(\widehat{\sigma}_n^2) = 15.16$. Testing for the significance by means of an χ^2-test yields a p-value of $2.2 \cdot 10^{-16}$.

However, the question arises whether the data satisfy the assumption of normality. Figure 2 shows a quantile–quantile plot of the model residuals. The departure from normality is obvious and also confirmed when conducting a significance test such as the Shapiro test, shedding doubt on parametric approaches, such as maximum likelihood estimation and inference, for such data.

Contrary to the normal assumption behind the maximum likelihood approach for random effects linear models, the methodology developed in this paper allows for non-normal errors. In addition, for equal batch sizes, we allow for an arbitrary covariance matrix of the measurements taken within a batch, whereas the standard approach assumes a structured covariance matrix with equally correlated measurements within batches and modules.

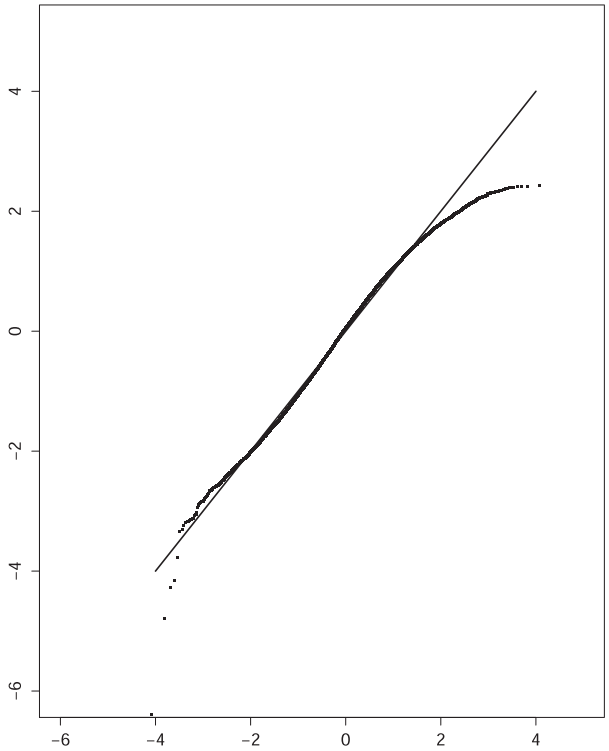

Figure 2. Quantile–quantile plot of the standardized residuals of the full random effects model based on maximum likelihood estimation.

8. Monte Carlo simulations

We conducted several simulation studies in order to investigate the statistical properties of the aforementioned methodology. The aim was to balance the sampling costs and the required statistical properties (power of tests and accuracy of estimators) under distributional assumptions that are realistic for PV outdoor measurements. All Monte Carlo simulations are based on 6000 repetitions except otherwise stated. We used the mixed model (3.1) with fixed regressors $\mathbf{x}_v$ as specified in Table I and true coefficients $\boldsymbol{\beta} = (1, 0.7, 0.4, \beta_4)'$, where β_4 varies to study type I error rates and power under different settings.

8.1. Power of the stratified cluster sampling design

We first investigated the statistical power when testing a main effect in the presence of correlated clusters, for a fixed site. The size of the clusters was selected as $b = 5$, the number of strata was $k = 4$ and the $80 : 40 : 40 : 40$ stratified sampling as discussed earlier was applied. The distributions of the strata were chosen as Gaussian distributions, $\mathcal{N}(a_\ell, \sigma_\ell^2)$, with parameters $a_1 = 0$, $\sigma_1^2 = 4$, $a_2 = 0$, $\sigma_2^2 = 9$, $a_3 = -2$, $\sigma_3^2 = 5$ and $a_4 = -6$, $\sigma_4^2 = 3$ for the four strata. Table II provides the proportions of the strata.

Clearly, the variance of the strata distributions affects the degree of the within-cluster correlation, which is also controlled by the variance σ^2 of the random effect that was assumed to follow an $\mathcal{N}(0, \sigma^2)$ law. Table III shows the resulting within-cluster correlations for all strata for different values of σ^2.

First, 400 observations were generated for each site according to Model (3.1) (with $\boldsymbol{\zeta} = 0$) and used to estimate the w_ℓ. Then 200 observations were simulated under the stratified sampling scheme.

Table II. Proportions of strata.

Number of site	w_1	w_2	w_3	w_4
1	0.7	0.15	0.1	0.05
2	0.7	0.1	0.1	0.1
3	0.7	0.07	0.08	0.15
4	0.65	0.15	0.1	0.1
5	0.6	0.1	0.2	0.1

Table III. Within-cluster correlation for the strata for the model (3.1).

	σ					
Strata	0	0.25	0.5	1	1.5	2
1	0	0.015	0.059	0.200	0.360	0.500
2	0	0.007	0.027	0.100	0.200	0.308
3	0	0.012	0.048	0.167	0.310	0.444
4	0	0.020	0.077	0.250	0.429	0.571

Table IV. Size and power for stratified sampling (uniform across sites).

	σ					
β_4	0	0.25	0.5	1	1.5	2
0	0.066	0.077	0.096	0.187	0.275	0.339
0	0.059	0.060	0.065	0.073	0.074	0.074
0.1	0.248	0.241	0.220	0.185	0.145	0.123
0.2	0.717	0.694	0.632	0.463	0.329	0.250
0.3	0.969	0.958	0.921	0.776	0.584	0.448

First row: size of the test when ignoring cluster correlations.

Table V. Strata sample sizes (non-uniform across sites).

Number of site	m_1	m_2	m_3	m_4
1	100	40	30	30
2	80	40	40	40
3	110	25	25	40
4	90	50	30	30
5	90	30	50	30

Being interested in the type I error rate and statistical power under alternatives when testing a main effect, we fixed $\mu = 200$ and $\boldsymbol{\beta} = (1, 0.7, 0.4, \beta_4)$ and investigated testing the null hypothesis $H_0 : \beta_4 = 0$ against the associated two-sided alternative $H_1 : \beta_4 \neq 0$ at the nominal 5% significance level.

Table IV provides the simulated rejection rates. It can be seen that the test behaves well in terms of the type I error rate even for substantial cluster correlations. The first row in Table IV provides the corresponding results when ignoring the cluster correlations, that is, when assuming independent observations. We can see that the effect of cluster correlations is substantial and leads to a heavily overreacting test when falsely assuming independent data. Indeed, even for small cluster correlations of 0.2, the size of the significance test is unacceptable, such that ignoring such correlations leads to invalid inference. The correlations also affect the statistical power, which decreases as the degree of dependence increases. However, for the chosen sampling approach, the power is still acceptable for correlations corresponding to $\sigma \leqslant 1.5$.

In order to investigate the practical influence of a non-uniform sampling across sites, such samples were generated using the strata sample sizes given in Table V.

The simulated size and power when testing a main effect under non-uniform stratified sampling are shown in Table VI. Again, the first row shows the simulated rejection rates when the cluster correlations are not taken into account. Overall, the results presented in Table VI show that the corresponding rejection rates are quite similar to the findings for uniform stratified sampling.

8.2. Equivalent sample size

A sound and comprehensible measure of the amount of power loss due to cluster correlations is to determine the *equivalent sample size* that leads to the same statistical power when no correlations are present. For that purpose, Table VII provides the rejection rates for different sample sizes determined when no cluster correlations have to be taken into account, in order to allow the engineer to gain insight into that issue. For example, the case of $n = 200$ for $\sigma = 0.5$ (third column in

Table VI. The power of the test for the hypothesis $H_0 : \beta_4 = 0$ (non-uniform stratified sampling).

	σ					
β_4	0	0.25	0.5	1	1.5	2
0	0.054	0.062	0.082	0.168	0.253	0.309
0	0.052	0.057	0.061	0.065	0.066	0.066
0.1	0.253	0.261	0.220	0.170	0.132	0.110
0.2	0.741	0.718	0.634	0.459	0.322	0.232
0.3	0.975	0.959	0.935	0.789	0.593	0.443

First row: results when ignoring cluster correlations.

Table VII. Size and power for different sample sizes, $n = n_v$, when no correlations are present.

	n					
β_4	175	150	125	100	75	50
0	0.055	0.054	0.052	0.052	0.053	0.054
0.1	0.221	0.211	0.189	0.161	0.134	0.118
0.2	0.667	0.598	0.534	0.458	0.356	0.277
0.3	0.936	0.905	0.859	0.763	0.661	0.489

Table VIII. Power of the proposed test for uniform stratified sampling for real error distributions.

	σ_B					
β_4	0	0.25	0.5	1	1.5	2
0	0.042	0.051	0.063	0.057	0.05	0.056
0.1	0.283	0.264	0.212	0.153	0.11	0.075
0.2	0.735	0.74	0.661	0.424	0.287	0.204
0.3	0.97	0.967	0.943	0.73	0.527	0.36

Table IX. Power of the proposed test for non-uniform stratified sampling for real error distributions.

	σ_B					
β_4	0	0.25	0.5	1	1.5	2
0	0.065	0.047	0.051	0.057	0.044	0.054
0.1	0.305	0.275	0.266	0.146	0.117	0.115
0.2	0.79	0.797	0.684	0.468	0.32	0.213
0.3	0.985	0.975	0.959	0.805	0.522	0.413

Table IV) leads to similar power as the case of $n = 150$ for $\sigma = 0$ (second column in Table VII). We also see that the case of $n = 200$ for $\sigma = 1$ is approximately equivalent to the case of $n = 100$ for $\sigma = 0$.

8.3. Benchmarking

It is well known that non-normal measurement errors may affect the size as well as the power of statistical tests. When analysing the power output of PV modules, the distribution may take quite different forms owing to the production process and the way how module classes are determined. Therefore, we benchmarked the proposed approach by simulating the rejection rates when sampling the measurement error from historical data sets available to us. For that purpose, the density was estimated from the historical data by the Parzen–Rosenblatt kernel density estimator with bandwidth selected using the Sheather–Jones method [10]. For alternative approaches to the issue of bandwidth selection, we refer to [11] and the references given there. The random effect modelling the cluster correlations was again assumed to be Gaussian.

The results are shown in Table VIII for uniform stratified sampling and in Table IX for the case of non-uniform stratified sampling. For each stratum, a different real historical data set was used to simulate the measurement error. It can be seen that the rejection rates differ from the corresponding values for Gaussian errors, but the proposed test generally behaves very well for the real-world error distributions.

8.4. Influence of serial correlations

Having in mind the empirical findings discussed earlier, we were mainly interested in the influence of the degree of spatial cluster correlation on the statistical power for different sample sizes when testing a main effect, in the presence of a noticeable serial correlation between time points.

Therefore, we varied the sample size by conducting the stratified sampling scheme with proportions $2:1:1:1$ and a cluster size of $b = 5$. This means, in order to generate data sets of different sample sizes, the number of observations collected at each site, that is, the size of the (stratified) panel, varied between 125 and 1125 at each site. In addition, at time $t = 0$, an initial random sample of size 400 was drawn to estimate the proportions of the strata. For the time series simulations, $T = 8$ time points were chosen.

Recall that in model (5.1), the strata-related error depends on two parameters: γ governs the dependence of the remeasured modules at different time instants and σ_ℓ controls the variance of the strata-related error. This means, each stratum has its own parameter σ_i. In our simulation, we used the fixed values $(\sigma_1, \sigma_2, \sigma_3, \sigma_4) = (4, 9, 5, 3)$, whereas γ was selected from the set of values $\{0, 0.2, 0.7\}$. Here, $\gamma = 0$ represents the case of independent observations over time.

Recall that the parameters σ and p related to the spatial clusters that determine the degree of within-cluster correlation. They were chosen as $p = 0.1$ and $\sigma \in \{0, 1, 2, 3, 4\}$, thus defining four settings I–IV. Table X provides the corresponding values of the coefficient of within-cluster correlation, ρ, in each stratum and for the different values of the model parameter σ.

The power of the significance test for the null hypothesis $H_0^{(v)} : \beta_4 = 0$ against various alternatives was studied, $v = 1, \ldots, a$, employing corrected critical values to take into account that we have to deal with a multiple testing problem. For this set of simulations, each rejection rate was simulated using 3000 repetitions.

Figure 3 provides the power of the considered test at a fixed time instant, for different alternatives for β_4. The AR parameter γ determining the degree of serial correlations was chosen as $\gamma = 0.7$ to study the case of rather strong serial correlations. The within-cluster correlations occur with probability $p = 0.1$, and their size is determined by the choice $\sigma = 4$.

Table X. Correlations ρ.

	$\sigma_1 = 4$	$\sigma_2 = 9$	$\sigma_3 = 5$	$\sigma_4 = 3$	Setting
$\sigma = 0$	0	0	0	0	I
$\sigma = 1$	0.33	0.25	0.31	0.37	II
$\sigma = 2$	0.5	0.4	0.47	0.54	III
$\sigma = 3$	0.6	0.5	0.57	0.63	IV
$\sigma = 4$	0.67	0.57	0.64	0.7	V

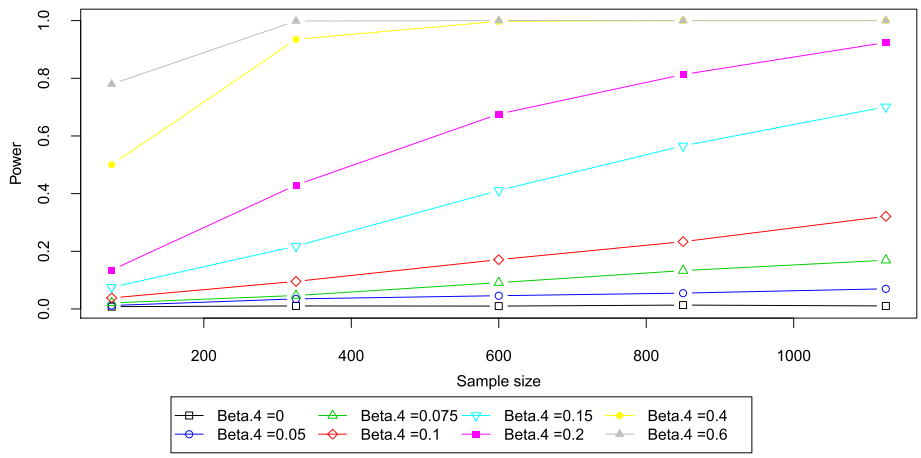

Figure 3. Power of the test for different alternatives under setting V, for fixed $\gamma = 0.7$, $\sigma = 4$.

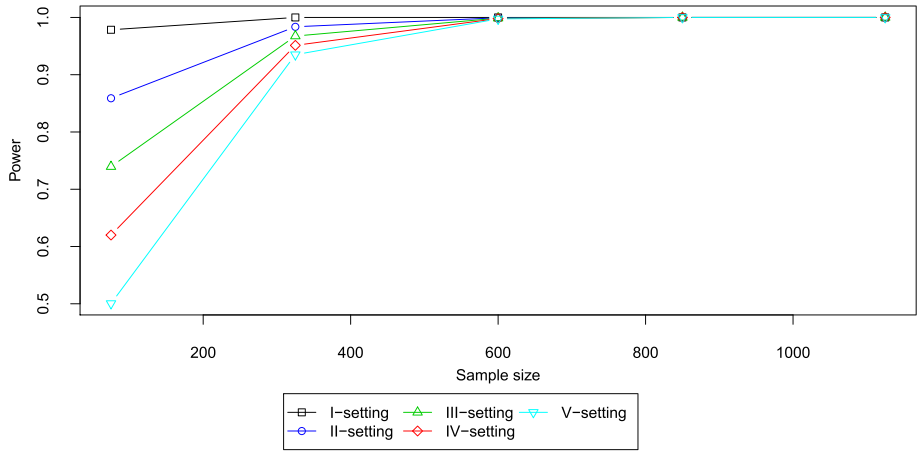

Figure 4. Power of the test under the alternative $\beta_4 = 0.4$ for the parameter settings I–V in Table X leading to within-cluster correlations between 0.25 and 0.7, for fixed $\gamma = 0.7, \beta_4 = 0.15$.

Table XI. Strata proportions over the time.

	Time periods							
First stratum	0.0536	0.0652	0.0727	0.0893	0.1091	0.1091	0.1091	0.1091
Second stratum	0.0893	0.1304	0.1091	0.1429	0.1455	0.1636	0.1636	0.1636
Third stratum	0.0893	0.1522	0.1455	0.1607	0.1636	0.1636	0.1818	0.2
Fourth stratum	0.7679	0.6522	0.6727	0.6071	0.5818	0.5636	0.5455	0.5272

Figure 4 demonstrates the influence of the within-cluster correlation on the statistical power.

A further finding of this set of simulations is that for the case of a short time series as studied here, the serial correlations have only minor importance. This is likely due to the fact that at each time instant a quite large number of *independent* random vectors (consisting of the cluster measurements) are available, and those observations are used for inference. The situation may change, if one picks a small subset of the data, for example, a strata or a couple of clusters, and observes them over time. For such an analysis, however, longer time series as considered here have to be collected.

8.5. Accuracy of sample strata proportions

Lastly, we investigated the accuracy of the proposed confidence intervals for the strata proportions under the proposed stratified cluster sampling approach. We employed the refined panel time series model from the last section in combination with the threshold model for the strata, as explained in Section 6.

Again, we assume that four strata are of interest. The corresponding proportions at each time instant t were fixed according to Table XI. From those probabilities, the associated thresholds were calculated for a fixed set of parameter values of the simulation model. Those thresholds were then used to determine the strata of each simulated observation of simulated random sample.

The fourth stratum can be interpreted as a good lot, whereas the other three strata represent certain damages or defects. For simplicity, let us consider confidence intervals only for the first stratum, which has the lowest probability.

For engineers, it is instructive to look at simulated examples illustrating how the study and the associated confidence intervals could look like under different simulated worlds. Figures 5 and 6 depict confidence intervals for a simulated sample with a relatively large sample size $n = 2500$. The green (dotdash-circle) line corresponds to the true proportion of the strata determined by the simulation model. The red (dash-diamond) line are the bound of the 95% confidence interval, and the black (triangle) line is the estimated proportion, calculated from one simulated sample.

Figures 7–10 depict the corresponding results for fixed simulated samples of sizes 500 (and 200, respectively) and the four settings I–IV as in Table X corresponding to different degrees of the within-cluster correlation. Simulations not reported here in detail have shown that other parameters of the model are relevant when calculating those confidence intervals.

Applied Stochastic Models in Business and Industry

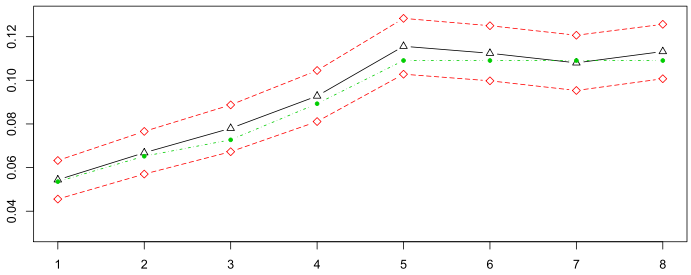

Figure 5. Sample size 2500. Setting I.

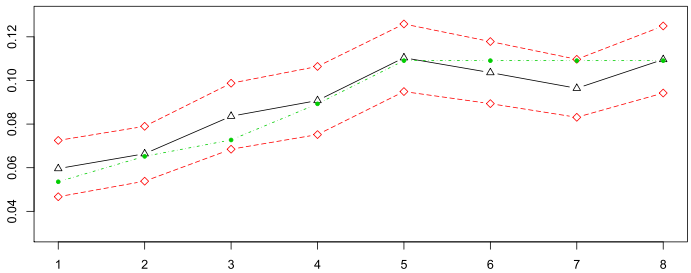

Figure 6. Sample size 2500. Setting V.

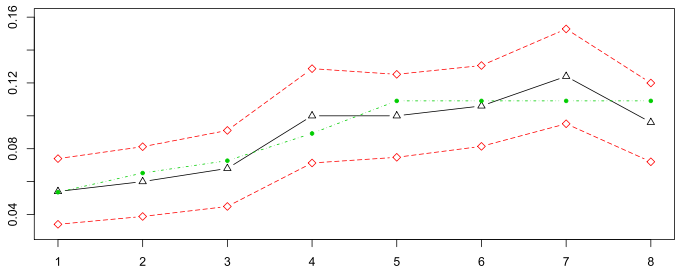

Figure 7. Sample size 500. Setting I.

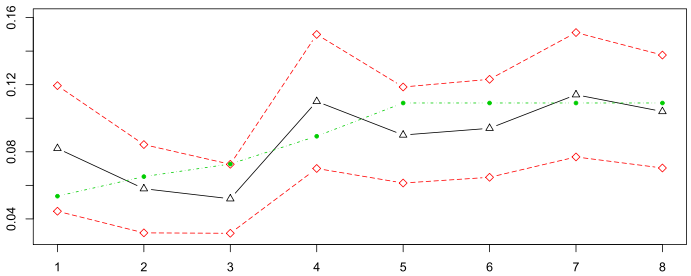

Figure 8. Sample size 500. Setting V.

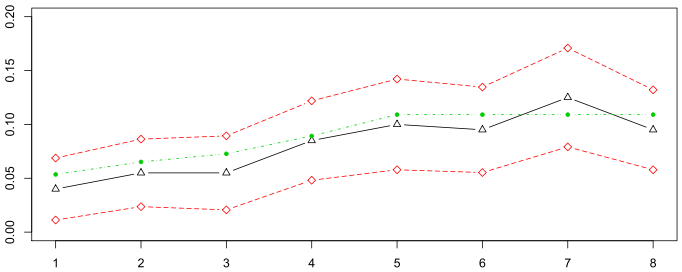

Figure 9. Sample size 200. Setting I.

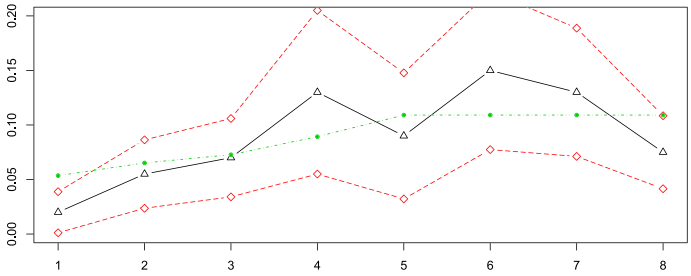

Figure 10. Sample size 200. Setting V.

9. Conclusions

A comprehensive methodology for sampling and the statistical analysis of multisite PV field studies is developed and thoroughly investigated. The proposed approach allows for stratification (e.g. due to predamages or different manufacturers), various forms of correlations, such as spatial correlations, correlations due to random effects or serial correlations, and non-normal measurement errors. In order to reduce the costs of sampling, it is proposed to take measurements at randomly selected clusters of neighbouring modules. To allow for the analysis of temporal developments, a panel design is employed. Estimation is based on weighted least squares, which provides consistent estimates under weak regularity conditions and allows for feasible inference based on hypothesis tests and confidence intervals. We rely on a nonparametric approach to estimate the unknown and possibly unstructured variance–covariance matrix of the errors, thus going beyond the scope of the classical mixed linear model. Several extensions of the basic approach are discussed in greater detail, including nested mixed models and serially correlated errors. The methodology also includes accompanying proposals for the analysis of binary (count) data collected under the proposed sampling design.

The statistical properties are extensively studied by Monte Carlo simulations. Among the investigated properties are the influence of within-cluster correlations, serial correlations due to temporal effects and different strata proportions. Those simulations aim at a thorough understanding of the impact of those various issues, which quite naturally arise in such a study, on the statistical power of hypothesis testing. Here, a major aim is to provide guidance when determining the design and selecting the sample sizes. The results generally support the overall approach and the proposed choices for certain parameters such as the sample sizes of the panel(s). The analysis of a real data set taken at a PV system and leading to massive amount of image data of all involved solar cells illustrates that the hypothesized correlations are present in real data and have to be taken into account.

Appendix A: Derivations

This appendix is devoted to derivations of the asymptotic results on which the proposed statistical assessment is based.

Denote the diagonal elements of $\widehat{\mathbf{W}}$ by $\hat{w}_{\ell j}$ and those of $\mathbf{W}$ by $w_{\ell j}$ and notice that they only depend on ℓ. The number of observations is $n = \sum_{\ell=1}^{k} m_\ell$. The following estimate assumes that the regressors $x_{\ell j}$ are univariate; otherwise, one argues component-wise to obtain a bound for a vector norm. We have

$$\begin{aligned} R_n &= \left| \frac{1}{\sqrt{n}} \sum_{\ell=1}^{k} \sum_{j=1}^{m_\ell} (\hat{w}_{\ell j} - w_{\ell j}) x_{\ell j} \widetilde{\epsilon}_{\ell j} \right| \\ &= \left| \sum_{\ell=1}^{k} \sqrt{\frac{m_\ell}{n}} \sqrt{m_\ell} (\hat{w}_\ell - w_\ell) \frac{1}{m_\ell} \sum_{j=1}^{m_\ell} x_{\ell j} \widetilde{\epsilon}_{\ell j} \right|. \end{aligned}$$

We obtain $R_n = o_P(1)$, as $n \to \infty$, if

$$\frac{1}{m_\ell} \sum_{j=1}^{m_\ell} x_{\ell j} \widetilde{\epsilon}_{\ell j} \xrightarrow{P} 0, \tag{A.1}$$

as $m_\ell \to \infty$, for each $\ell = 1, \ldots, k$. The following proof under condition (3.8) works for arbitrary dependence structures within each cluster as long as the error terms have identical (marginal) distributions. Recall that b is fixed, whereas $m_\ell / b \to \infty$. Observe that

Applied Stochastic Models in Business and Industry

$$\frac{1}{m_\ell}\sum_{j=1}^{m_\ell} x_{\ell j}\widetilde{\epsilon}_{\ell j} = \frac{1}{b}\sum_{k=1}^{b}\frac{1}{m_\ell/b}\sum_{r=1}^{m_\ell/b} x_{\ell,(r-1)b+k}\widetilde{\epsilon}_{\ell,(r-1)b+k}$$
$$=: \frac{1}{b}\sum_{k=1}^{b}\sum_{r=1}^{m_\ell/b} a_{\ell r}e_{\ell r}.$$

Clearly, $e_{\ell r}$, $r = 1, \dots, m_\ell/b$, are i.i.d., because the error terms in each strata share the same marginal distribution. By assumption, the second and first moments of the regression constants $x_{\ell j}$, $j = 1, \dots, m_\ell$, converge. Hence, the array of constants, $a_{\ell r} = x_{\ell,(r-1)b+k}/(m_\ell/b)$, satisfies

$$A_1 = \sum_{r=1}^{m_\ell/b} |a_{\ell r}|$$
$$= \frac{b}{m_\ell}\sum_{r=1}^{m_\ell/b} |x_{\ell,(r-1)b+k}| \leqslant \frac{b}{m_\ell}\sum_{j=1}^{m_\ell} |x_{\ell j}| \to b m_1^{(\ell)},$$

say, such that A_1 is bounded, and

$$A_2 = \sum_{r=1}^{m_\ell/b} |a_{\ell r}|^2$$
$$= \frac{1}{m_\ell}\frac{b^2}{m_\ell}\sum_{r=1}^{m_\ell/b} x^2_{\ell,(r-1)b+k}$$
$$\leqslant \frac{1}{m_\ell}\frac{b^2}{m_\ell}\sum_{j=1}^{m_\ell} x^2_{\ell j} = o(1).$$

Hence, (A.1) follows from the law of large numbers [12, Theorem 5, p. 104]. We have shown that

$$R_n = n^{-1/2}(\widehat{\mathbf{W}} - \mathbf{W})\mathbf{X}'\mathbf{e} = o_P(1).$$

Consequently, we may conclude that $n^{-1/2}\widehat{\mathbf{W}}\mathbf{X}'\mathbf{e}$ inherits its asymptotic normal distribution from $n^{-1/2}\mathbf{W}\mathbf{X}'\mathbf{e}$. Next, observe that

$$\left\|\frac{1}{n}\sum_{\ell=1}^{k}\sum_{j=1}^{m_\ell}(\widehat{w}_\ell - w_\ell)\mathbf{x}_{\ell j}\mathbf{x}'_{\ell j}\right\|_F \leqslant \max_{1\leqslant\ell\leqslant k}|\widehat{w}_\ell - w_\ell|\left\|\frac{1}{n}\sum_{\ell=1}^{k}\sum_{j=1}^{m_\ell}\mathbf{x}_{jl}\mathbf{x}'_{jl}\right\|_F.$$

Notice that the elements of the matrix $\frac{1}{n}\sum_{\ell=1}^{k}\sum_{j=1}^{m_\ell}\mathbf{x}_{jl}\mathbf{x}'_{jl}$ are bounded, because by the Cauchy–Schwarz inequality

$$\left|\frac{1}{n}\sum_{\ell=1}^{k}\sum_{j=1}^{m_\ell} x_{\ell j\nu}x_{\ell j\mu}\right| \leqslant \sqrt{\frac{1}{n}\sum_{\ell=1}^{k}\sum_{j=1}^{m_\ell} x^2_{\ell j\nu}}\sqrt{\frac{1}{n}\sum_{\ell=1}^{k}\sum_{j=1}^{m_\ell} x^2_{\ell j\mu}},$$
$$\to \sqrt{\sum_{\ell=1}^{k}\mu_\ell m^{(\ell)}_{2\nu}\sum_{\ell=1}^{k}\mu_\ell m^{(\ell)}_{2\mu}} < \infty,$$

as $n \to \infty$. The aforementioned facts show that

$$\|n^{-1}\mathbf{X}'\widehat{\mathbf{W}}\mathbf{X} - n^{-1}\mathbf{X}'\mathbf{W}\mathbf{X}\|_F = \|n^{-1}\mathbf{X}'(\widehat{\mathbf{W}} - \mathbf{W})\mathbf{X}\|_F = o_P(1),$$

as $n \to \infty$, such that the triangle inequality leads to

$$n^{-1}\mathbf{X}'\widehat{\mathbf{W}}\mathbf{X} \to \mathbf{C},$$

as $\min m_\ell \to \infty$. Now, the assertions follow from the representation

$$\sqrt{n}(\widehat{\theta}_n - \theta) = (n^{-1}\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}n^{-1/2}\mathbf{W}\mathbf{X}'\mathbf{e}$$

using straightforward arguments such as Slutzky's lemma; for sake of brevity, details are omitted (e.g. [5]).

To derive the limiting distribution of the strata proportion estimators under the sampling design, recall the following version of the multivariate δ method: let U be the open subset in $\mathbb{R}^d$ with $\mathbf{a} \in U$. Let the mapping $f : U \to \mathbb{R}^k$ be such that

(i) $f \in C(U)$,
(ii) partial derivatives exist $\frac{\partial f_i}{\partial x_j}$, for $1 \leqslant i \leqslant d$ and $1 \leqslant j \leqslant k$.

Denote the gradient of the function f_i by Δf_i and put $\Delta_f = (\Delta f_1, \dots \Delta f_l)$. If

$$c_n(\boldsymbol{\eta}_n - \mathbf{a}) \xrightarrow{d} \mathcal{N}(0, \boldsymbol{\Sigma}),$$

as $n \to \infty$, then

$$\mathrm{P}\left(\boldsymbol{\eta}_n \in U\right) \to 1,$$

as $n \to \infty$, and

$$c_n\left(f(\boldsymbol{\eta}_n) - f(\mathbf{a})\right) \mid \boldsymbol{\eta}_n \in U \xrightarrow{d} \mathcal{N}(0, \Delta_f'(\mathbf{a})\boldsymbol{\Sigma}\Delta_f(\mathbf{a})),$$

as $n \to \infty$:

The multivariate central limit theorem yields

$$\sqrt{r}\left(\frac{\xi_1 + \dots \xi_r}{r} - p\mathbf{1}_p\right) \xrightarrow{d} \mathcal{N}(0, \boldsymbol{\Sigma}),$$

as $n \to \infty$. An application of the δ method with $f(x_1, \dots, x_b) = \frac{1}{b}(x_1 + \dots + x_b)$ leads to

$$\sqrt{r}\left(f\left(\frac{\xi_1 + \dots + \xi_r}{r}\right) - f(p, \dots, p)\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{b^2}\mathbf{1}_b'\boldsymbol{\Sigma}\mathbf{1}_b\right),$$

as $n \to \infty$, where $f(p, \dots, p) = p$ and $f\left((\xi_1 + \dots + \xi_r)/r\right) = \frac{1}{rb}\sum_{i=1}^{r}\sum_{j=1}^{b}\xi_{ij} = \hat{p}_n$. Therefore

$$\sqrt{r}(\hat{p}_n - p) \xrightarrow{d} \mathcal{N}\left(0, 1/b^2\mathbf{1}_b'\boldsymbol{\Sigma}\mathbf{1}_b\right),$$

and finally multiplying everything with $\sqrt{b}$, we obtain

$$\sqrt{n}(\hat{p}_n - p) \xrightarrow{d} \mathcal{N}\left(0, 1/b\mathbf{1}_b'\boldsymbol{\Sigma}\mathbf{1}_b\right),$$

as $n \to \infty$, which establishes the assertion.

Acknowledgements

The authors acknowledge financial support of BMWi (Federal Ministry for Economic Affairs and Energy) under grant no. 0325588B, PV-Scan collaborative research project. The methodology of the paper has been discussed extensively with our PV-Scan partners, especially Sunnyside upP, Cologne, TÜV Rheinland Energie und Umwelt, Cologne, and Solarfabrik AG, Freiburg i.B. We thank TÜV Rheinland Energie und Umwelt for providing us real data.

References

1. Steland A. Sampling plans for control-inspection schemes under independent and dependent sampling designs with applications to photovoltaics. In *Frontiers in Statistical Quality Control 11*. Springer: Cham, 2015; 287–317.
2. Rao CR. *Linear Statistical Inference and its Applications* second edition. John Wiley & Sons: New York-London-Sydney, 1973. Wiley Series in Probability and Mathematical Statistics.
3. Dean A, Morris M, Stufken J, Bingham D. *Handbook of Design and Analysis of Experiments*. Chapman and Hall/CRC: Boca Raton, 2015. Handbooks of Modern Statistical Methods.
4. DuMouchel WH, Duncan GJ. Using sample survey weights in multiple regression analyses of stratified samples. *Journal of the American Statistical Association* 1983; **78**(383):535–543.
5. Steland A. *Financial Statistics and Mathematical Finance*. Springer: Chichester, 2012.
6. Faraway JJ. *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*, vol. 124. CRC press: Boca Raton, 2016.
7. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 2015; **67**(1):1–48.
8. Brockwell PJ, Davis RA. *Time Series: Theory and Methods* Second, Springer Series in Statistics. Springer-Verlag: New York, 1991.
9. Sovetkin E, Steland A. On statistical preprocessing of PV field image data using robust regression. In *Advances in Mathematics and Statistical Sciences*, Mastorakis NE, Ding A, Shitikova MV (eds), vol. 40. WSEAS Press.: Dubai, United Arabian Emirates, 2015; 48–51.

Applied Stochastic Models in Business and Industry

10. Sheather SJ, Jones MC. A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society Series B* 1991; **53**(3):683–690.
11. Golyandina N, Pepelyshev A, Steland A. New approaches to nonparametric density estimation and selection of smoothing parameters. *Computational Statistics & Data Analysis* 2012; **56**(7):2206–2218.
12. Chandra TK. *A First Course in Asymptotic Theory of Statistics*. Alpha Science International, 1999.