

ON THE USE OF AGGREGATE SURVEY DATA FOR ESTIMATING REGIONAL MAJOR DEPRESSIVE DISORDER PREVALENCE

DOMINGO MORALES 

UNIVERSITY MIGUEL HERNÁNDEZ DE ELCHE

JOSCHA KRAUSE AND JAN PABLO BURGARD

TRIER UNIVERSITY

Major depression is a severe mental disorder that is associated with strongly increased mortality. The quantification of its prevalence on regional levels represents an important indicator for public health reporting. In addition to that, it marks a crucial basis for further explorative studies regarding environmental determinants of the condition. However, assessing the distribution of major depression in the population is challenging. The topic is highly sensitive, and national statistical institutions rarely have administrative records on this matter. Published prevalence figures as well as available auxiliary data are typically derived from survey estimates. These are often subject to high uncertainty due to large sampling variances and do not allow for sound regional analysis. We propose a new area-level Poisson mixed model that accounts for measurement errors in auxiliary data to close this gap. We derive the empirical best predictor under the model and present a parametric bootstrap estimator for the mean squared error. A method of moments algorithm for consistent model parameter estimation is developed. Simulation experiments are conducted to show the effectiveness of the approach. The methodology is applied to estimate the major depression prevalence in Germany on regional levels crossed by sex and age groups.

Key words: empirical best prediction, generalized linear mixed model, method of moments, parametric bootstrap, small area estimation.

1. Introduction

Major depressive disorder (MDD) is a very prevalent and severe mental disorder that is associated with significantly shortened life expectancy (Walker et al., 2015; Gilman et al., 2017). Empirical studies found that it may cause up to 50% increase in mortality risk relative to non-depressed individuals (Cuijpers et al., 2014). The effect is evident in both short-term and long-term survival rates, due to unnatural causes of death like suicide on the one hand (Zivin et al., 2015), and chronic conditions on the other hand (Glymour et al., 2010; Kozela et al., 2016). See Gilman et al. (2017) for a comprehensive overview. Moreover, empirical studies suggest that both the prevalence and the negative impact of depressive disorders varies geographically, as, for instance, reported by Haan et al. (2019) and Brignone et al. (2020). Against this background, assessing MDD on regional levels marks an important indicator for public health reporting. It allows policy-makers to plan comprehensive health care programs. Moreover, it may facilitate further explorative studies regarding the impact of living conditions on depression, which may improve our understanding of these disorders.

However, quantifying the regional prevalence of MDD is methodologically challenging. The topic is highly sensitive and national statistical institutions rarely have administrative records on

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s11336-021-09808-8>.

Correspondence should be made to Domingo Morales, Operations Research Center, University Miguel Hernández de Elche, Elche, Spain. Email: d.morales@umh.es

this matter. Published figures are typically direct estimates that are produced based on survey data. In Germany, there are two major surveys that collect data on public health: *Sozio-oekonomisches Panel* (SOEP; Wagner et al., 2007) and *Gesundheit in Deutschland aktuell* (GEDA; Lange et al., 2015). The SOEP is a panel survey that is collected annually via computer-assisted personal interviews and questionnaires. In 2011, it contained 20828 individuals of age 18 or older. GEDA, on the other hand, is a telephone survey that is collected in irregular intervals. In 2010, it covered 22050 individuals of age 18 or older. Note that in 2011, the survey was not conducted. The amount of resources used to collect these samples is substantial. However, despite the considerable effort, the obtained sample data still lack in geographic detail. Since the German population is more than 80 million, the achieved sample sizes in both surveys are insufficient to provide reliable estimates on regional levels. The direct survey-based estimates suffer from large sampling variances and do not allow for the analysis of local MDD patterns. Thus, alternative statistical approaches based on auxiliary data are required to estimate the MDD prevalence.

Small area estimation (SAE) solves this problem by combining auxiliary data from multiple regions within a suitable regression model. Prevalence estimates are then obtained via model prediction. See Pfeiffermann (2013), Rao and Molina (2015), Sugawara and Kubokawa (2020), or Morales et al. (2021) for a comprehensive overview on SAE methods. Statistical modeling plays an essential role in SAE. Models can be designed to describe either unit-level data, or records that are aggregated by territories or population subgroups, which is called area-level data. In the first case, the fitted model is used to predict the values of the target variable in the unsampled part of the population. These predictions are subsequently used to construct estimators of regional prevalence and other population parameters of interest. However, it has the disadvantage of requiring information external to the survey, like the population means of the auxiliary variables or even a census file. In the second case, aggregated data modeling allows borrowing information from other domains, auxiliary variables from external data sources, as well as correlation structures. This allows the researcher to introduce model-based prevalence predictors that are more precise than direct estimators. The area-level approach has fewer restrictions to incorporate auxiliary variables and can be used without the need to use confidential individual data. This makes it more flexible and applicable in the field of psychometrics. Against this background, we propose to use an area-level approach to estimate the regional MDD prevalence.

In general, SAE allows for an efficiency advantage over the previously mentioned direct MDD prevalence estimates when the regression model contains sufficient predictive power. This is the case when (i) the chosen model fits the distribution characteristics of the target variable, and (ii) the auxiliary data have explanatory power. Regarding the first aspect, since prevalence figures may be viewed as proportions or counts, binomial, negative binomial, and Poisson mixed models are well-studied options. They are special cases of the generalized linear mixed model (GLMM), as, for instance, discussed by Breslow and Clayton (1993) and applied to SAE problems by Jiang (2003), Ghosh and Maiti (2004), Sugawara et al. (2017), Hobza et al. (2020) and Faltys et al. (2020). The binomial logit approach for the estimation of regional proportions has been used by Molina et al. (2007), Ghosh et al. (2009), Chen and Lahiri (2012), Erciulescu and Fuller (2013), López-Vizcaíno et al. (2013), López-Vizcaíno et al. (2015), Militino et al. (2015), Chambers et al. (2016), Hobza and Morales (2016), Liu and Lahiri (2017), as well as Hobza et al. (2018). The Poisson or negative binomial mixed models were applied to estimate small area counts or proportions by Chambers et al. (2014), Dreassi et al. (2014), and Boubeta et al. (2016; 2017), among others.

Given these references, there are seemingly suitable SAE models for the estimation of the regional MDD prevalence in Germany. However, given the second aspect mentioned above, our auxiliary data situation prevents us from applying them. Using an area-level approach on SOEP and GEDA data requires the calculation of direct estimates for auxiliary variables to improve the MDD prevalence estimates. These records are subject to sampling errors, since the values

have been estimated from survey samples. To ensure that SAE produces reliable results under these circumstances, the model must explicitly account for measurement errors in both the target variable and auxiliary variable records. This issue was first addressed under linear mixed models (LMMs) by Ybarra and Lohr (2008). They proposed an area-level model, where the continuous dependent variable is a direct estimator of a domain characteristic, and the covariates are direct estimators calculated from a different survey. They introduced a generalization of the Fay-Herriot model to covariates measured with error. As the auxiliary variables are direct estimators, their variance matrix is estimated with a design-based estimator applied to the corresponding unit-level data. The considered area-level model contains measurement errors with known variance matrices, but without specifying its distribution. The authors derived predictors of domain parameters and proposed estimators of the corresponding mean squared errors.

The literature about SAE methodology based on measurement error mixed model contains further interesting contributions. Ghosh and Sinha (2007), Torabi et al. (2009), as well as Singh (2011) introduced empirical Bayes predictors of finite population means under measurement error models. Torabi (2013) presented an application of data cloning to the estimation of the parameters of a measurement error nested error regression model. Singh (2011) applied the simulation extrapolation (SIMEX), ordinary corrected scores and Monte Carlo corrected scores methods of bias correction in the Fay-Herriot model, and investigated the performance of the bias-corrected estimators. Marchetti et al. (2015) presented a Big Data application of a measurement error Fay-Herriot model to estimate poverty indicators. Burgard et al. (2020) and Burgard et al. (2021) considered structural versions of the univariate and multivariate Fay-Herriot models. Bell et al. (2019) compared functional, structural, and naïve area-level models in the SAE setup. Concerning the Bayesian approach, Ghosh et al. (2006) introduced hierarchical Bayes predictors of finite population parameters under structural measurement error models. Arima et al. (2015) introduced a measurement error hierarchical model and derived Bayes predictors. Arima et al. (2017) proposed multivariate Fay-Herriot Bayesian predictors of small area means. Arima and Polettini (2019) considered a Bayesian unit-level small area model with misclassified covariates.

The above-cited studies proposed extensions to SAE models based on continuous target variables to account for measurement errors. However, recall that MDD prevalence estimation requires proportions or counts as target variable, which have to be modeled with GLMMs. Unfortunately, there currently no suitable measurement error extensions for GLMMs that could be applied to the SOEP and GEDA. Accordingly, new statistical methodology has to be developed to close this gap. This can be done in two ways, depending on whether frequentist or Bayesian procedures developed for LMMs are adapted to GLMMs. This paper follows the frequentist approach and leaves open the study of Bayesian statistical methodologies.

We propose a new area-level Poisson mixed model that is capable of accounting for measurement errors in covariates. With this feature, it is a GLMM for SAE that allows for robust predictions from uncertain auxiliary data that is measured with error. A detailed model description is provided and empirical best predictors (EBP) based on the model are derived. We further state a parametric bootstrap estimator of the mean squared error (MSE) of prediction. A consistent method of moments algorithm for model parameter estimation is presented. Further, it is demonstrated how the variances of the model parameter estimates can be approximated. Simulation experiments are conducted in order to establish the effectiveness of the approach. The new methodology is subsequently applied for the estimation of the regional MDD prevalence in Germany within the year 2011. For the target variable, we use area-level survey records from the SOEP. For the auxiliary data, we consider direct area-level estimates obtained from GEDA.

The remainder of the paper is organized as follows. Section 2 introduces the problem of interest and describes the data employed in the estimation of regional prevalence. Section 3 presents the new model, derives EBPs, and proposes a parametric bootstrap estimator of the MSEs. Section 4 states the method of moments (MM) for model parameter estimation and elaborates

on variance approximation. Section 5 contains the simulation experiments. Section 6 presents the application of the model-based statistical methodology to the data from the German surveys SOEP and GEDA. Section 7 closes with some conclusive remarks. The paper is supported by a supplemental material file containing three appendices. Appendix A gives the components of the Newton–Raphson algorithm that calculates the MM estimators of the model parameters. Appendix B provides developments that establish consistency of the MM estimators. Appendix C presents additional simulation experiments.

2. The Problem of Interest and the Data

The problem that motivates the research on new statistical methodology is the regional prevalence estimation of MDD in Germany. For regions with big sample size, we can carry out the estimation by means of direct estimators. However, if domain sample sizes are not big enough then direct estimators are not precise. This problem can be addressed by reinforcing the direct information provided by the target variable with the indirect information provided by correlated auxiliary variables or data from other domains. In what follows, we describe the survey data employed in the estimation of regional prevalence.

The base population are the citizens of Germany that are 18 years or older from the year 2011. The domains are defined as cross combinations of (i) the seven Nielsen regions of the country, (ii) binary sex, as well as (iii) the age groups I: 18-44, II: 45-64, III: 65+. There are $D = 42$ domains, and we are interested in estimating the domain prevalences

$$\bar{Y}_d = \frac{1}{N_d} \sum_{j \in U_d} y_{dj}, \quad y_{dj} = \begin{cases} 1, & j \text{ reports to have a MDD} \\ 0, & \text{else} \end{cases}, \quad d = 1, \dots, D,$$

where N_d is the domain size and i denotes an individual in domain U_d . In accordance with the developments in Sect. 3, two data sources are required to implement the model.

For the target variable, we use the 2011 wave of the German survey SOEP with 20828 individuals (Wagner et al., 2007). Please note that the SOEP originally covers citizens that are 16 years or older. However, we have removed individuals of age 16 and 17 from the sample in order to obtain a coherent population for both target variable and auxiliary data. The domain-specific sample sizes range from 278 to 840. See Goebel et al. (2008) for further details on the survey design, interview mode, and data processing in SOEP. The survey mainly covers socioeconomic topics, but also assesses medical conditions and other health-related information. This includes several mental disorders. Against this background, we define the sample counts $y_d, d = 1, \dots, D$, as the number of sampled individuals per domain that reported being diagnosed with MDD in SOEP. The sample counts range from 5 to 84.

For the auxiliary data, we use the 2010 sample of the German health survey GEDA (Lange et al., 2015). Note that in 2011, the survey was not conducted. The sample contains citizens of age 18 and older from the German population that are interviewed via computer-assisted telephone interviews in a representative random sample. The data set contains a broad range of health-related information of participants on the person level. For further information on the survey as well as its sampling design and response rates, see Robert Koch Institute (2013). The GEDA sample size for 2010 is 22050, which is larger than the SOEP sample, which allows using estimates of the first survey as auxiliary data for the second one. Further, we note that GEDA and SOEP are independent surveys. That is, there is no sampling error covariance between the target variable stemming from the SOEP and the covariates stemming from GEDA. If the target variable and the covariates stem from the same survey, the error covariance structure has to be taken into account

additionally. This is not covered by the presented methodology and would lead to a new topic for research. In order to select a statistical model for explaining the behavior of the target variable MDD, we have to select a set of covariates from GEDA that explain the variation of the sample counts of individuals having MDD. The selection is based on methodological considerations on the one hand, and on related literature concerning MDD on the other hand. All in all, we consider the subsequent covariates:

- *Domain size*: Number of individuals in domain U_d (scaled)
- *Sex*: Binary sex (male, female) that domain U_d is associated with
- *Depr. treatment*: Share of people that have been treated for depression in U_d
- *SES score*: Average score of the socioeconomic status in U_d

The variable *Domain size* is obtained from administrative records of the German census 2011. It is included as a more flexible substitute for an intercept in order to stabilize the fitting behavior. The equations of the MM algorithm are based on differences between theoretical moments under a given parametrization and observed sample moments. In the presence of measurement errors, an objective function based on these differences is much more flat relative to situations where values are measured correctly. Including the domain size as fixed effect reduces the variability of the exponent in the exponential function and leads to a numerically more stable estimation process. However, we have scaled the respective values in order to avoid computational infinities arising from the usage of the exponential function in the statistic software R. We let $N_d := |U_d|$ and define the value of *Domain size* for domain U_d as $x_d = N_d/\delta$, where δ is a scalar that is constant over all domains. We choose $\delta = (\sum_{d=1}^D N_d)/(\sum_{d=1}^D n_d)$ in order to obtain values of the same magnitude as the domain sample sizes of the SOEP for 2011.

The binary auxiliary variable *Sex* is included because the domains are constructed as cross combinations of age, region, and sex. It is well known from the literature that females tend to have a higher probability to be diagnosed with major depression than males. See Cyranowski et al. (2000) and Albert (2015) for further details. Accordingly, we expect female domains to show higher depression proportions in the sample than male domains. The variable *Depr. treatment* is an obvious choice given its content. A high share of people that are treated for depression is likely to be associated with an increased number of depressed individuals. Furthermore, the sample counts and *Depr. treatment* showed a reasonably positive correlation with 0.3. And finally, the variable *SES score* is a combined metric that is based on education, income, and professional position. See Robert Koch Institute (2013) for more specific descriptions. Empirical studies found to that the socioeconomic status is negatively associated with depressive disorders. Thus, we expect high socioeconomic status to be associated with lower depression proportions. However, note that there is debate in the literature with regards to which aspect of the status actually drives this relation. For further details, we refer to Jo et al. (2011) and Freeman et al. (2016).

3. Model

3.1. Formulation

This section introduces the area-level measurement error Poisson mixed (MEPM) model. Let U denote a finite population that is segmented into D domains indexed by $d = 1, \dots, D$. We assume a situation where (i) survey data for a variable of interest y , and (ii) either administrative records or survey data regarding auxiliary variables $\mathbf{x} = \{x_1, \dots, x_p\}$ are available. The area-level MEPM model is composed of three hierarchical stages. In the first stage a model, called *sampling* model, gives the distribution of the sample counts y_d . Let p_d be the domain probability of an event of interest and let $v_d = n_d$ be the domain sample size (or a size parameter in other counting

setups). The sampling model indicates that the distribution of the target variable y_d is

$$y_d \sim \text{Pois}(v_d p_d), \quad d = 1, \dots, D.$$

That is to say, the target variable y_d gives the observed domain counts in the sample for the event of interest. Against the background of Sect. 2, y_d refers to the number of individuals in the sample that have a MDD in domain d . In the second stage, the logarithm of the mean parameter $\mu_d = v_d p_d$ (the natural parameter in Poisson regression models) is assumed to vary linearly with the p area-level auxiliary variables contained in \mathbf{x} . In the second stage, we consider a set of random effects $\{v_d : d = 1, \dots, D\}$ that are independent and identically distributed (i.i.d.) according to $v_d \sim N(0, 1)$. In matrix notation, we have

$$\mathbf{v} = \underset{1 \leq d \leq D}{\text{col}} (v_d) \sim N_D(\mathbf{0}, \mathbf{I}_D), \quad f_v(\mathbf{v}) = (2\pi)^{-D/2} \exp \left\{ -\frac{1}{2} \mathbf{v}' \mathbf{v} \right\},$$

where $f_v(\mathbf{v})$ gives the probability density function of \mathbf{v} . The *linking* model is specified as

$$\log \mu_d = \log v_d + \log p_d = \log v_d + \tilde{\mathbf{x}}_{0,d} \boldsymbol{\beta}_0 + \tilde{\mathbf{x}}_{1,d} \boldsymbol{\beta}_1 + \phi v_d, \quad d = 1, \dots, D,$$

where $\tilde{\mathbf{x}}_{0,d}$ and $\tilde{\mathbf{x}}_{1,d}$ are a $1 \times p_1$ and $1 \times p_2$ row vectors, respectively, containing the true aggregated values of $p = p_1 + p_2$ auxiliary variables for domain d . The terms

$$\boldsymbol{\beta}_0 = \underset{1 \leq k \leq p_1}{\text{col}} (\beta_k), \quad \boldsymbol{\beta}_1 = \underset{p_1+1 \leq k \leq p}{\text{col}} (\beta_k),$$

are the corresponding column vectors of regression parameters. By defining $\tilde{\mathbf{x}}_d = (\tilde{\mathbf{x}}_{0,d}, \tilde{\mathbf{x}}_{1,d})$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}'_0, \boldsymbol{\beta}'_1)'$, the linking model can be written in the simpler form.

$$\log \mu_d = \log v_d + \log p_d = \log v_d + \tilde{\mathbf{x}}_d \boldsymbol{\beta} + \phi v_d, \quad d = 1, \dots, D.$$

We assume that $\tilde{\mathbf{x}}_{0,d}$ is known and equal to $\mathbf{x}_{0,d}$, $d = 1, \dots, D$. With respect to practice, $\mathbf{x}_{0,d}$ may correspond to administrative records that are taken from a register. Further, we assume that the $\tilde{\mathbf{x}}_{1,d}$'s are unknown random vectors with predictors $\mathbf{x}_{1,d}$'s calculated from independent data sources. These data sources might be data sets obtained from surveys with larger sample sizes than the survey where the dependent variables y_d 's are recorded. In the third stage, we consider the *measurement error* model

$$\tilde{\mathbf{x}}'_{1,d} = \mathbf{x}'_{1,d} + \mathbf{u}_d, \quad \mathbf{u}_d \sim N_{p_1}(\mathbf{0}, \boldsymbol{\Sigma}_d), \quad d = 1, \dots, D, \quad (1)$$

where $\mathbf{x}_{1,d}$ is a row vector containing unbiased predictors of the components of $\tilde{\mathbf{x}}_{1,d}$, \mathbf{u}_d is a column vector of random measurement errors, $\boldsymbol{\Sigma}_d$ is the known covariance matrix of \mathbf{u}_d , and \mathbf{u}_d and $\mathbf{x}_{1,d}$, $d = 1, \dots, D$, are independent. In practice, $\boldsymbol{\Sigma}_d$ is typically unknown and subsequently estimated from the same unit-level survey data as $\mathbf{x}_{1,d}$. We finally assume that $\mathbf{x}_{1,d}$, \mathbf{u}_d , v_d , $d = 1, \dots, D$, are independent, but we only introduce inference procedures conditionally on $\mathbf{x}_d = (\mathbf{x}_{0,d}, \mathbf{x}_{1,d})$, $d = 1, \dots, D$. In matrix notation, we have

$$\mathbf{y} = \underset{1 \leq d \leq D}{\text{col}} (y_d), \quad \mathbf{X} = \underset{1 \leq d \leq D}{\text{col}} (\mathbf{x}_d), \quad \mathbf{u} = \underset{1 \leq d \leq D}{\text{col}} (\mathbf{u}_d), \quad \mathbf{u}_d = \underset{1 \leq k \leq p_1}{\text{col}} (u_{dk}).$$

Based on these definitions, the probability density function of the measurement errors is

$$f_u(\mathbf{u}) = (2\pi)^{-Dp_1/2} \left(\prod_{d=1}^D |\boldsymbol{\Sigma}_d| \right)^{-1/2} \exp \left\{ -\frac{1}{2} \sum_{d=1}^D \mathbf{u}'_d \boldsymbol{\Sigma}_d^{-1} \mathbf{u}_d \right\}.$$

The area-level MEPM model can be expressed as a single model in the form

$$y_d | v_d, \mathbf{u}_d, \mathbf{x}_d \sim \text{Pois}(v_d p_d), \quad \log \mu_d = \log v_d + \log p_d = \log v_d + \mathbf{x}_d \boldsymbol{\beta} + \mathbf{u}'_d \boldsymbol{\beta}_1 + \phi v_d, \quad (2)$$

$p_d = \exp \{ \mathbf{x}_d \boldsymbol{\beta} + \mathbf{u}'_d \boldsymbol{\beta}_1 + \phi v_d \}$, $d = 1, \dots, D$. From the model assumptions, it holds that

$$P(y_d | \mathbf{v}, \mathbf{u}, \mathbf{X}) = P(y_d | v_d, \mathbf{u}_d, \mathbf{x}_d) = \frac{1}{y_d!} \exp\{-v_d p_d\} v_d^{y_d} p_d^{y_d},$$

$$P(\mathbf{y} | \mathbf{v}, \mathbf{u}, \mathbf{X}) = \prod_{d=1}^D P(y_d | v_d, \mathbf{u}_d, \mathbf{x}_d),$$

$$P(\mathbf{y} | \mathbf{X}) = \int_{R^D} \int_{R^{Dp_1}} P(\mathbf{y} | v_d, \mathbf{u}_d, \mathbf{x}_d) f_v(\mathbf{v}) f_u(\mathbf{u}) d\mathbf{v} d\mathbf{u} = \int_{R^{D(p_1+1)}} \psi(\mathbf{y}, \mathbf{v}, \mathbf{u}) d\mathbf{v} d\mathbf{u},$$

where

$$\begin{aligned} \psi(\mathbf{y}, \mathbf{v}, \mathbf{u}) &= (2\pi)^{-\frac{D}{2}} \exp \left\{ \frac{-\mathbf{v}'\mathbf{v}}{2} \right\} (2\pi)^{-Dp_1/2} \left(\prod_{d=1}^D |\boldsymbol{\Sigma}_d| \right)^{-1/2} \exp \left\{ -\frac{1}{2} \sum_{d=1}^D \mathbf{u}'_d \boldsymbol{\Sigma}_d^{-1} \mathbf{u}_d \right\} \\ &\cdot \prod_{d=1}^D \frac{\exp\{-v_d p_d\} v_d^{y_d} \exp \{ y_d (\mathbf{x}_d \boldsymbol{\beta} + \mathbf{u}'_d \boldsymbol{\beta}_1 + \phi v_d) \}}{y_d!} \\ &= (2\pi)^{-\frac{D}{2}} \exp \left\{ \frac{-\mathbf{v}'\mathbf{v}}{2} \right\} (2\pi)^{-Dp_1/2} \left(\prod_{d=1}^D |\boldsymbol{\Sigma}_d| \right)^{-1/2} \exp \left\{ -\frac{1}{2} \sum_{d=1}^D \mathbf{u}'_d \boldsymbol{\Sigma}_d^{-1} \mathbf{u}_d \right\} \\ &\cdot \left(\prod_{d=1}^D y_d! \right)^{-1} \exp \left\{ \sum_{d=1}^D \{ -v_d \exp \{ \mathbf{x}_d \boldsymbol{\beta} + \mathbf{u}'_d \boldsymbol{\beta}_1 + \phi v_d \} + y_d \log v_d \} \right\} \\ &\cdot \exp \left\{ \sum_{k=1}^p \left(\sum_{d=1}^D y_d x_{dk} \right) \beta_k + \sum_{k=1}^{p_1} \left(\sum_{d=1}^D y_d u_{dk} \right) \beta_k + \phi \sum_{d=1}^D y_d v_d \right\}. \end{aligned}$$

3.2. Empirical Best Prediction

This section derives the EBPs of the domain probability p_d and the mean parameter $\mu_d = v_d p_d$. We proceed according to the following basic strategy. First, the best predictors (BPs) are calculated under the model under the preliminary assumption that the model parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}', \phi)'$ are known. Afterward, $\boldsymbol{\theta}$ is replaced by a consistent estimator $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}', \hat{\phi})'$ to obtain the EBP. We show how to perform consistent model parameter estimation in Sect. 4. We start with the EBP of the domain probability p_d . Recall that the conditional distribution of \mathbf{y} , given \mathbf{v} , \mathbf{u} and \mathbf{X} , is $P(\mathbf{y} | \mathbf{v}, \mathbf{u}, \mathbf{X}) = \prod_{d=1}^D P(y_d | v_d, \mathbf{u}_d, \mathbf{x}_d)$, where

$$P(y_d | v_d, \mathbf{u}_d, \mathbf{x}_d) = \frac{1}{y_d!} \exp\{-v_d p_d\} v_d^{y_d} p_d^{y_d}$$

$$= \frac{v_d^{y_d}}{y_d!} \exp \left\{ y_d (\mathbf{x}_d \boldsymbol{\beta} + \mathbf{u}'_d \boldsymbol{\beta}_1 + \phi v_d) - v_d \exp \{ \mathbf{x}_d \boldsymbol{\beta} + \mathbf{u}'_d \boldsymbol{\beta}_1 + \phi v_d \} \right\}.$$

The probability density functions of v_d and \mathbf{u}_d are

$$f(v_d) = (2\pi)^{-1/2} \exp \left\{ -\frac{1}{2} v_d^2 \right\}, \quad f(\mathbf{u}_d) = (2\pi)^{-p_1/2} |\boldsymbol{\Sigma}_d|^{-1/2} \exp \left\{ -\frac{1}{2} \mathbf{u}'_d \boldsymbol{\Sigma}_d^{-1} \mathbf{u}_d \right\}.$$

The BP of p_d is its conditional expectation under the model, that is, $\hat{p}_d(\boldsymbol{\theta}) = E_\theta[p_d | \mathbf{y}, \mathbf{X}]$. In this case, we have that $E_\theta[p_d | \mathbf{y}, \mathbf{X}] = E_\theta[p_d | y_d, \mathbf{x}_d]$ and

$$\begin{aligned} E_\theta[p_d | y_d, \mathbf{x}_d] &= \frac{\int_{R^{p_1+1}} \exp \{ \mathbf{x}_d \boldsymbol{\beta} + \mathbf{u}'_d \boldsymbol{\beta}_1 + \phi v_d \} P(y_d | v_d, \mathbf{u}_d, \mathbf{x}_d) f(v_d) f(\mathbf{u}_d) dv_d d\mathbf{u}_d}{\int_{R^{p_1+1}} P(y_d | v_d, \mathbf{u}_d, \mathbf{x}_d) f(v_d) f(\mathbf{u}_d) dv_d d\mathbf{u}_d} \\ &= \frac{N_d(y_d, \boldsymbol{\theta})}{D_d(y_d, \boldsymbol{\theta})}. \end{aligned}$$

The numerator and denominator of the fraction are given by

$$\begin{aligned} N_d(y_d, \boldsymbol{\theta}) &= \int_{R^{p_1+1}} \exp \left\{ (y_d + 1) (\mathbf{x}_d \boldsymbol{\beta} + \mathbf{u}'_d \boldsymbol{\beta}_1 + \phi v_d) - v_d \exp \{ \mathbf{x}_d \boldsymbol{\beta} + \mathbf{u}'_d \boldsymbol{\beta}_1 + \phi v_d \} \right\} \\ &\quad \times f(v_d) f(\mathbf{u}_d) dv_d d\mathbf{u}_d, \\ D_d(y_d, \boldsymbol{\theta}) &= \int_{R^{p_1+1}} \exp \left\{ y_d (\mathbf{x}_d \boldsymbol{\beta} + \mathbf{u}'_d \boldsymbol{\beta}_1 + \phi v_d) - v_d \exp \{ \mathbf{x}_d \boldsymbol{\beta} + \mathbf{u}'_d \boldsymbol{\beta}_1 + \phi v_d \} \right\} \\ &\quad \times f(v_d) f(\mathbf{u}_d) dv_d d\mathbf{u}_d. \end{aligned}$$

In principle, the EBP is then calculated according to $\hat{p}_d = N_d(y_d, \hat{\boldsymbol{\theta}}) / D_d(y_d, \hat{\boldsymbol{\theta}})$. However, both integrals contained in $E_\theta[p_d | y_d, \mathbf{x}_d]$ do not have a closed-form solution. As a consequence, the EBP cannot be calculated exactly, but has to be approximated instead. This can be done by using Monte Carlo integration. That is to say, we choose a number of $2L$ symmetric locations at which numerator and denominator are evaluated within the support of the multivariate probability density of both random effects and measurement errors. Afterward, we average the obtained functional values and obtain an approximation to the corresponding integrals. This is summarized in the subsequent procedure.

1. Estimate $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}', \hat{\phi})$.
2. Generate $v_d^{(\ell)}$ i.i.d. $N(0, 1)$, $\mathbf{u}_d^{(\ell)}$ i.i.d. $N_{p_1}(\mathbf{0}, \boldsymbol{\Sigma}_d)$, $v_d^{(L+\ell)} = -v_d^{(\ell)}$, $\mathbf{u}_d^{(L+\ell)} = -\mathbf{u}_d^{(\ell)}$, $d = 1, \dots, D$, $\ell = 1, \dots, L$.
3. Calculate $\hat{p}_d(\hat{\boldsymbol{\theta}}) = \hat{N}_d / \hat{D}_d$, where

$$\begin{aligned} \hat{N}_d &= \frac{1}{2L} \sum_{\ell=1}^{2L} \exp \left\{ (y_d + 1) (\mathbf{x}_d \hat{\boldsymbol{\beta}} + \mathbf{u}_d^{(\ell)'} \hat{\boldsymbol{\beta}} + \hat{\phi} v_d^{(\ell)}) - v_d \exp \{ \mathbf{x}_d \hat{\boldsymbol{\beta}} + \mathbf{u}_d^{(\ell)'} \hat{\boldsymbol{\beta}} + \hat{\phi} v_d^{(\ell)} \} \right\}, \\ \hat{D}_d &= \frac{1}{2L} \sum_{\ell=1}^{2L} \exp \left\{ y_d (\mathbf{x}_d \hat{\boldsymbol{\beta}} + \mathbf{u}_d^{(\ell)'} \hat{\boldsymbol{\beta}} + \hat{\phi} v_d^{(\ell)}) - v_d \exp \{ \mathbf{x}_d \hat{\boldsymbol{\beta}} + \mathbf{u}_d^{(\ell)'} \hat{\boldsymbol{\beta}} + \hat{\phi} v_d^{(\ell)} \} \right\}. \end{aligned}$$

It can be shown that for $L \rightarrow \infty$, it holds that $\hat{N}_d/\hat{D}_d \xrightarrow{a.s.} N_d(y_d, \hat{\theta})/D_d(y_d, \hat{\theta})$ (Calfisch, 1998). That is to say, for L chosen sufficiently large, the upper algorithm can approximate the EBP up to arbitrary order. Based on these developments, we can conclude that the EBP of the mean parameter $\mu_d = v_d p_d$ is $\hat{\mu}_d(\hat{\theta}) = v_d \hat{p}_d(\hat{\theta})$. Note that there may be applications where the researcher wants to avoid Monte Carlo integration, for instance when p_1 is large and heavy computational burden shall be avoided. For such cases, we state synthetic predictors as an alternative that can be calculated very efficiently. The synthetic predictor of p_d is $\tilde{p}_d^{syn} = \exp\{\mathbf{x}_d \hat{\boldsymbol{\beta}}\}$. Likewise, the synthetic predictor of $\mu_d = v_d p_d$ is $\tilde{\mu}_d^{syn} = v_d \exp\{\mathbf{x}_d \hat{\boldsymbol{\beta}}\}$. These predictors have acceptable accuracy when both measurement errors and random effects are negligible in the distribution of the target variable.

3.3. Mean Squared Error Estimation

This section presents a parametric bootstrap algorithm that estimates the mean squared error (MSE) of the EBP $\hat{\mu}_d = \hat{\mu}_d(\hat{\theta})$, which is characterized by

$$MSE(\hat{\mu}_d) = E[(\hat{\mu}_d - \mu_d)^2], \quad d = 1, \dots, D.$$

Let B be the number of bootstrap replicates. The procedure is given as follows.

1. Obtain estimates $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}, \hat{\phi})'$ based on the observations $(y_1, \mathbf{x}_1), \dots, (y_D, \mathbf{x}_D)$
2. For $b = 1, \dots, B$, do
 - (a) Generate $v_d^{(b)} \stackrel{i.i.d.}{\sim} N(0, 1)$, $\mathbf{u}_d^{(b)} \stackrel{i.i.d.}{\sim} N_p(\mathbf{0}, \boldsymbol{\Sigma}_d)$, $d = 1, \dots, D$
 - (b) Calculate $\mu_d^{(b)} = v_d p_d^{(b)}$, $p_d^{(b)} = \exp\{\mathbf{x}_d \hat{\boldsymbol{\beta}} + \mathbf{u}_d^{(b)'} \hat{\boldsymbol{\beta}} + \hat{\phi} v_d^{(b)}\}$, and draw $y_d^{(b)} \sim \text{Pois}(\mu_d^{(b)})$, $d = 1, \dots, D$
 - (c) Obtain estimates $\hat{\boldsymbol{\theta}}^{(b)} = (\hat{\boldsymbol{\beta}}^{(b)}, \hat{\phi}^{(b)})$ based on the observations $(y_1^{(b)}, \mathbf{x}_1), \dots, (y_D^{(b)}, \mathbf{x}_D)$
 - (d) Calculate the EBP $\hat{\mu}_d^{(b)} = \hat{\mu}_d(\hat{\boldsymbol{\theta}}^{(b)})$, $d = 1, \dots, D$
3. Output: $mse^*(\hat{\mu}_d) = \frac{1}{B} \sum_{b=1}^B (\hat{\mu}_d^{(b)} - \mu_d^{(b)})^2$

The MSE estimator based on this naive parametric bootstrap approach has generally $O(D^{-1})$ bias, which is not always acceptable in practice. More refinement can be achieved by double bootstrapping, as proposed Hall and Maiti (2006). We have not implemented the double bootstrap approach because it is computationally very intensive and has a great impact on simulation times. Nevertheless, in some real data setups with small D and high and accurate computing capacity, the double bootstrap approach is recommendable.

4. Model Parameter Estimation

Conditioned to \mathbf{X} , the model likelihood of the MEPM model, $P(\mathbf{y}|\mathbf{X})$, is an integral on $R^{D(p_1+1)}$. This fact is a drawback for estimating the model parameters by the maximum likelihood (ML) method. Therefore, we propose applying the MM method, which does not require approximating integrals. The MM method gives consistent estimators and it is computationally more efficient than the ML method. This section presents a Newton–Raphson algorithm to fit the area-level MEPM model via the MM approach. Note that computational details of the algorithm are located in Appendix A of the supplemental material. Further, Appendix B establishes the consistency of the MM estimators.

4.1. Method of Moments

A natural set of equations for the MM approach is

$$0 = f_k(\boldsymbol{\theta}) = M_k(\boldsymbol{\theta}) - \hat{M}_k = \sum_{d=1}^D E_{\theta}[y_d]x_{dk} - \sum_{d=1}^D y_d x_{dk}, \quad k = 1, \dots, p, \quad (3)$$

$$0 = f_{p+1}(\boldsymbol{\theta}) = M_{p+1}(\boldsymbol{\theta}) - \hat{M}_{p+1} = \sum_{d=1}^D E_{\theta}[y_d^2] - \sum_{d=1}^D y_d^2, \quad (4)$$

where the model-based expectations $E_{\theta}[y_d]$ and $E_{\theta}[y_d^2]$ are given in Appendix A. For solving the system of nonlinear MM equations, we use a Newton–Raphson algorithm. Let $i = 0, 1, 2, \dots$ denote the index of iterations. Further, let $\mathbf{f}(\boldsymbol{\theta}^{(i)})$ be the set of MM equations based on the model parameter vector in the i -th iteration. Likewise, let $\mathbf{H}(\boldsymbol{\theta}^{(i)})$ denote the corresponding Jacobian matrix. The Newton–Raphson algorithm is performed as follows:

1. Set the initial values $i = 0$ and $\boldsymbol{\theta}^{(0)} = (\boldsymbol{\beta}^{(0)}, \phi^{(0)})$.
2. Repeat the following steps till convergence
 - (a) Update $\boldsymbol{\theta}^{(i)}$ by using the equation

$$\boldsymbol{\theta}^{(i+1)} = \boldsymbol{\theta}^{(i)} - \mathbf{H}^{-1}(\boldsymbol{\theta}^{(i)})\mathbf{f}(\boldsymbol{\theta}^{(i)}),$$

- (b) Update the iteration index $i \leftarrow i + 1$.

3. Output: $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}^{(i+1)}$.

Appendix A demonstrates how to compute the components of the Jacobian matrix. Note that the MM estimator is consistent in probability, i.e., $\hat{\boldsymbol{\theta}} \xrightarrow{P} \boldsymbol{\theta}$ as $D \rightarrow \infty$. We proof this result in Appendix B based on the developments of Jiang (1998).

4.2. Variance Estimation and Significance Testing

Hereafter, we show how to estimate the variance of the model parameter estimates obtained from the MM method. This step is important with respect to practice when significance tests for covariates that are measured with error shall be conducted. Let us define

$$\mathbf{M}(\boldsymbol{\theta}) = \underset{1 \leq k \leq p+1}{\text{col}} (M_k(\boldsymbol{\theta})), \quad \hat{\mathbf{M}} = \underset{1 \leq k \leq p+1}{\text{col}} (\hat{M}_k),$$

where M_k and \hat{M}_k are defined as in Sect. 4.1. The asymptotic variance of the MM estimators can be approximated by a Taylor expansion of $\mathbf{M}(\boldsymbol{\theta})$ around $\boldsymbol{\theta}$. This is to say,

$$\hat{\mathbf{M}} = \mathbf{M}(\hat{\boldsymbol{\theta}}) \approx \mathbf{M}(\boldsymbol{\theta}) + \mathbf{H}(\boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}), \quad \hat{\boldsymbol{\theta}} - \boldsymbol{\theta} \approx \mathbf{H}^{-1}(\boldsymbol{\theta})(\hat{\mathbf{M}} - \mathbf{M}(\boldsymbol{\theta})).$$

Under some regularity conditions, it holds that

$$\text{var}(\hat{\boldsymbol{\theta}}) = E[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})'] \approx \mathbf{H}^{-1}(\boldsymbol{\theta})\text{var}(\hat{\mathbf{M}})\mathbf{H}^{-1}(\boldsymbol{\theta}).$$

An estimator of $\text{var}(\hat{\theta})$ is

$$\widehat{\text{var}}(\hat{\theta}) = \mathbf{H}^{-1}(\hat{\theta})\widehat{\text{var}}(\hat{\mathbf{M}})\mathbf{H}^{-1}(\hat{\theta}),$$

where $\hat{\theta} = \theta^{(r+1)}$ is taken from the output of the MM algorithm and $\widehat{\text{var}}(\hat{\mathbf{M}})$ is an estimator of the covariance matrix of $\hat{\mathbf{M}}$. An algorithm to estimate $\text{var}(\hat{\theta})$ is

1. Fit the model to the sample and calculate $\hat{\theta}$.
2. Generate bootstrap samples $\{y_d^{(b)} : d = 1, \dots, D\}$, $b = 1, \dots, B$, from the fitted model.
3. Calculate $\hat{\mathbf{M}}^{(b)}$, $b = 1, \dots, B$, and

$$\bar{\mathbf{M}} = \frac{1}{B} \sum_{b=1}^B \hat{\mathbf{M}}^{(b)}, \quad \widehat{\text{var}}_B(\hat{\mathbf{M}}) = \frac{1}{B} \sum_{b=1}^B (\hat{\mathbf{M}}^{(b)} - \bar{\mathbf{M}})(\hat{\mathbf{M}}^{(b)} - \bar{\mathbf{M}})'$$

4. Output: $\widehat{\text{var}}_A(\hat{\theta}) = \mathbf{H}^{-1}(\hat{\theta})\widehat{\text{var}}_B(\hat{\mathbf{M}})\mathbf{H}^{-1}(\hat{\theta})$.

Alternatively, steps 3 and 4 can be replaced by

3. Fit the model to the bootstrap samples and calculate $\hat{\theta}^{(b)}$, $b = 1, \dots, B$, $\bar{\theta} = \frac{1}{B} \sum_{b=1}^B \hat{\theta}^{(b)}$.
4. Output: $\widehat{\text{var}}_B(\hat{\theta}) = \frac{1}{B} \sum_{b=1}^B (\hat{\theta}^{(b)} - \bar{\theta})(\hat{\theta}^{(b)} - \bar{\theta})'$.

From this procedure, for a given significance level $\alpha \in (0, 1)$ and the corresponding t -value, a model parameter estimate $\hat{\theta}_k$ is considered to be significantly different from zero if

$$\hat{\theta}_k \notin \left(\hat{\theta}_k - t_{D-p, 1-\frac{\alpha}{2}} \sqrt{\widehat{\text{var}}(\hat{\theta}_k)}, \hat{\theta}_k + t_{D-p, 1-\frac{\alpha}{2}} \sqrt{\widehat{\text{var}}(\hat{\theta}_k)} \right).$$

5. Simulation Experiments

This section presents three simulation experiments that are implemented in order to demonstrate the effectiveness of the methodology. We investigate the performance of (i) model parameter estimation, (ii) mean parameter prediction, and (iii) MSE estimation. Note that further simulation results are located in Appendix C of the supplemental material.

5.1. Set Up

A Monte Carlo simulation with $I = 500$, $i = 1, \dots, I$, iterations is conducted. We generate a population of D domains, where D varies over scenarios. For $d = 1, \dots, D$, we define

$$\mu_d = v_d p_d, \quad y_d \sim \text{Pois}(\mu_d), \quad p_d = \exp\{\beta_0 + \mathbf{x}_{1,d}\boldsymbol{\beta}_1 + \mathbf{x}_{2,d}\boldsymbol{\beta}_2 + \mathbf{u}'_{1,d}\boldsymbol{\beta}_1 + \mathbf{u}'_{2,d}\boldsymbol{\beta}_2 + \phi v_d\},$$

where $v_d = 300$, $\beta_0 = -4$, $\boldsymbol{\beta}_1 = (0.5, 0.5)'$, $\boldsymbol{\beta}_2 = -\boldsymbol{\beta}_1$, and $\phi = 0.3$, as well as $v_d \sim N(0, 1)$. Accordingly, we have an intercept and four covariates that are measured with error. Note that the random effects are generated in every iteration individually. The unbiased covariate predictors are drawn from uniform distributions according to $x_{dj} \sim U(1.0, 1.4)$, $j = 1, \dots, 4$, and held fixed

TABLE 1.
Overview of simulation scenarios

Scenario 1	Scenario 2	Scenario 3	Scenario 4
$D = 50$	$D = 75$	$D = 100$	$D = 125$

over all Monte Carlo iterations. The covariate measurement errors $\mathbf{u}_d = (\mathbf{u}'_{1,d}, \mathbf{u}'_{2,d})'$ are drawn in each iteration individually according to

$$\mathbf{u}_d \sim N_4(\mathbf{0}, \Sigma_d), \quad \Sigma_d = \begin{pmatrix} \sigma_{1,d}^2 & \sigma_{12,d} & \sigma_{13,d} & \sigma_{14,d} \\ \sigma_{21,d} & \sigma_{2,d}^2 & \sigma_{23,d} & \sigma_{24,d} \\ \sigma_{31,d} & \sigma_{32,d} & \sigma_{3,d}^2 & \sigma_{34,d} \\ \sigma_{41,d} & \sigma_{42,d} & \sigma_{34,d} & \sigma_{4,d}^2 \end{pmatrix}, \quad d = 1, \dots, D,$$

where $\sigma_{j,d}^2 \sim U(0.05, 0.15)$, $\sigma_{jk,d} = \rho_{jk}\sigma_{j,d}^2\sigma_{k,d}^2$, $\rho_{jk} = 0.5$ for $j = 1$ and $k = 2, 3, 4$, as well as $\rho_{jk} = -0.3$ for $j = 2, 3$ and $k = 3, 4$, $j \neq k$. Overall, we consider four simulation scenarios arising from the four different values for D . The scenarios are defined as follows.

The simulation is conducted with the statistic software package R. The objective is to estimate the mean parameter $\mu_d, d = 1, \dots, D$. We compare the following fitting methods:

- **PM**: maximum likelihood Laplace method for the Poisson mixed model, with the R function `glmer`.
- **MEPM**: methods of moments for the measurement error Poisson mixed model, with own programming.

5.2. Results of Model Parameter Estimation

This section studies the behavior of the MM algorithm in Sect. 4.1. We consider MSE and bias as performance measures. For parameters $\theta \in \{\beta_0, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \phi\}$, they are given by

$$MSE(\theta) = \frac{1}{I} \sum_{i=1}^I (\hat{\theta}^{(i)} - \theta)^2, \quad BIAS(\theta) = \frac{1}{I} \sum_{i=1}^I (\hat{\theta}^{(i)} - \theta).$$

Note that since $\boldsymbol{\beta}_1 = (\beta_1, \beta_2)'$ with $\beta_1 = \beta_2$, and $\boldsymbol{\beta}_2 = (\beta_3, \beta_4)'$ with $\beta_3 = \beta_4$, we average the performance measures for each of for the sake of a compact presentation. Table 2 presents the MSE results, and Table 3 presents the bias results. We start with the regression parameters $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}'_1, \boldsymbol{\beta}'_2)'$. It can be seen that the MEPM model outperforms the PM model in every scenario and for all elements of $\boldsymbol{\beta}$. The MSE of the estimates obtained from the MM approach are always smaller than those obtained from the maximum likelihood Laplace method under the PM model. The additional information obtained from the anticipation of the covariate measurement errors allows for efficiency gains in these settings. The effect is particular evident when looking at the intercept β_0 . Here, the standard PM model cannot clearly identify the contribution of β_0 to the functional description of the target variable, which causes excessive variation in the respective estimates. The MEPM model, on the other hand, shows good performance on that regard. For the bias of the regression parameter estimates, we have mixed results. This is due to the fact that measurement error distributions are symmetric around zero. There is no clear tendency which

TABLE 2.
Mean squared error of model parameter estimation

Scenario	Method	$MSE(\beta_0)$	$MSE(\beta_1)$	$MSE(\beta_2)$	$MSE(\phi)$
1	PM	34.5890	0.4450	0.4625	0.0148
1	MEPM	1.3244	0.3223	0.3202	0.0276
2	PM	33.7888	0.3095	0.2914	0.0069
2	MEPM	1.2405	0.2270	0.2410	0.0173
3	PM	34.4403	0.2551	0.2311	0.0048
3	MEPM	1.3900	0.2506	0.2292	0.0051
4	PM	32.9951	0.1715	0.1805	0.0034
4	MEPM	0.9233	0.1531	0.1623	0.0101

TABLE 3.
Bias of model parameter estimation

Scenario	Method	$BIAS(\beta_0)$	$BIAS(\beta_1)$	$BIAS(\beta_2)$	$BIAS(\phi)$
1	PM	5.7220	0.0116	-0.0199	-0.0511
1	MEPM	0.0710	-0.0271	-0.0039	-0.1055
2	PM	5.6690	-0.0044	0.0202	-0.0191
2	MEPM	0.0587	-0.0420	0.0197	-0.0737
3	PM	5.7485	-0.0152	-0.0033	-0.0123
3	MEPM	0.0389	-0.0164	-0.0037	-0.0147
4	PM	5.6555	0.0122	0.0081	-0.0066
4	MEPM	-0.0416	0.0004	0.0182	-0.0513

of the two approaches obtains better results. However, when looking at the overall efficiency (in terms of the MSE), the MEPM model is better.

We now consider the standard deviation parameter estimates. Here, the PM model obtains more efficient results. By looking at the bias, we see that the MEPM approach underestimates the true random effect standard deviation. This is likely due to the additional anticipation of the covariate measurement errors. The method seems to attribute some of the random effect variance to the uncertainty stemming from the auxiliary data, which causes the underestimation tendency and the overall MSE inefficiency. However, we will see in the next section that the predictions obtained under the MEPM model are superior to those under the PM model regardless of this effect.

5.3. Results of Mean Parameter Prediction

This section studies the behavior of the EBP under the area-level MEPM model presented in Sect. 3.2 against the EBP derived by Boubeta et al. (2016) under the area-level Poisson mixed model. We consider relative mean squared error (RMSE), relative root mean squared error (RRMSE), absolute bias (ABIAS), as well as relative absolute bias (RABIAS) as performance measures. They are given as follows:

$$RMSE_d = \left(\frac{1}{I} \sum_{i=1}^I (\mu_d^{(i)} - \hat{\mu}_d^{(i)})^2 \right)^{1/2}, \quad RRMSE_d = \frac{RMSE_d}{\bar{\mu}_d}, \quad \bar{\mu}_d = \frac{1}{I} \sum_{i=1}^I \mu_d^{(i)},$$

TABLE 4.
Performance of mean parameter prediction

Scenario	Method	$RMSE$	$RRMSE$	$ABIAS$	$RABIAS$	$PoEP$
1	PM Model	1.6575	28.6726	1.2896	22.5440	0.00
1	MEPM Model	1.5813	27.3569	1.2323	21.5249	100.00
2	PM Model	1.6146	27.6157	1.2553	21.4650	0.00
2	MEPM Model	1.5684	26.8232	1.2281	21.0016	100.00
3	PM Model	1.5806	27.0935	1.2251	20.9966	4.00
3	MEPM Model	1.5501	26.5669	1.2076	20.6966	96.00
4	PM Model	1.5474	26.7783	1.1975	20.7312	8.00
4	MEPM Model	1.5214	26.3288	1.1879	20.5686	92.00

$$ABIAS_d = \frac{1}{I} \sum_{i=1}^I |\mu_d^{(i)} - \hat{\mu}_d^{(i)}|, \quad RABIAS_d = \frac{ABIAS_d}{\bar{\mu}_d}.$$

We further define the proportion of efficient predictions (PoEP)

$$PoEP = \frac{1}{D} \sum_{d=1}^D \mathbb{1}_{(RMSE_d < RMSE_d^*)} \cdot 100\%,$$

as well as the subsequent aggregated measures

$$RMSE = \frac{1}{D} \sum_{d=1}^D RMSE_d, \quad ABIAS = \frac{1}{D} \sum_{d=1}^D ABIAS_d,$$

$$RRMSE = \frac{1}{D} \sum_{d=1}^D RRMSE_d \cdot 100, \quad RABIAS = \frac{1}{D} \sum_{d=1}^D RABIAS_d \cdot 100\%,$$

to allow for a more compact presentation. Here, the term $\mathbb{1}_{(RMSE_d < RMSE_d^*)}$ yields the value 1 when the respective method has achieved a smaller RMSE than the alternative, and 0 else. Table 4 presents the results of mean parameter prediction.

We see that the EBP under the MEPM model outperforms the predictor under the PM model in all scenarios and for all considered performance measures. Further, by looking at the measure $PoEP$, we see that the MEPM predictions are more efficient than the PM predictions in 100% of the domains for Scenario 1 and 2, and in over 90% for Scenario 3 and 4. The anticipation of the covariate measurement errors leads to a significant improvement of the domain count predictions, which is in line with the theoretical developments of Sect. 3. This is further visualized in Fig. 1. It contains boxplots of the measure $RRMSE_d$ over all domains per scenario. The results of the PM model are displayed in blue, while the results of the MEPM model are depicted in red. It becomes evident that both the boxes and the median of the $RRMSE_d$ -values are visibly lower under the MEPM model than under the PM model. The strongest efficiency gains are obtained under the scenario with a smaller number of domains. This is important with respect to the empirical application in Sect. 6, where the number of domains is small as well.

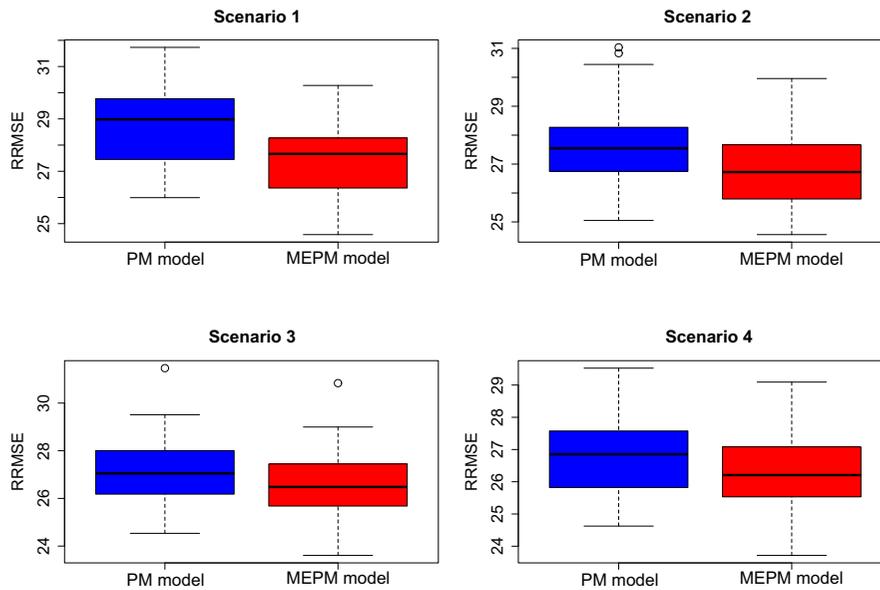


FIGURE 1.
RRMSE of mean parameter prediction

5.4. Results of Mean Squared Error Estimation

This section studies the performance of the parametric bootstrap algorithm for MSE estimation in Sect. 3.3. For this, we use the Monte Carlo MSE obtained measured in the simulation experiment as approximation to the real MSE. That is to say, we define

$$MSE_d = \frac{1}{I} \sum_{i=1}^I (\mu_d^{(i)} - \hat{\mu}_d^{(i)})^2, \quad mse_d^* = \frac{1}{I} \sum_{i=1}^I mse_d^{*(i)},$$

where $mse_d^{*(i)}$ is an MSE estimate obtained for the d -th domain in the i -th Monte Carlo iteration. Further, we consider absolute bias and relative absolute bias as performance measures, which are given by

$$ABIAS_d = \frac{1}{I} \sum_{i=1}^I |mse_d^{*(i)} - MSE_d|, \quad RABIAS_d = \frac{ABIAS_d}{MSE_d} \cdot 100\%.$$

As in the previous two simulation experiments, we use aggregated performance measures

$$mse^* = \frac{1}{D} \sum_{d=1}^D mse_d^*, \quad ABIAS = \frac{1}{D} \sum_{d=1}^D ABIAS_d, \quad RABIAS = \frac{1}{D} \sum_{d=1}^D RABIAS_d$$

for a compact presentation. We choose $B = 300$ for the number of bootstrap replicates used for MSE estimation. Table 5 presents the corresponding simulation results and further visualized in Fig. 2. It can be seen that the parametric bootstrap yields plausible results. There is a slight

TABLE 5.
Simulation results for MSE estimation

Scenario	MSE	mse^*	$ABIAS$	$RABIAS$
1	2.4674	2.2443	0.2643	10.2388
2	2.4741	2.3078	0.2726	10.6629
3	2.4156	2.3800	0.2575	10.4314
4	2.3876	2.2987	0.2312	9.6469

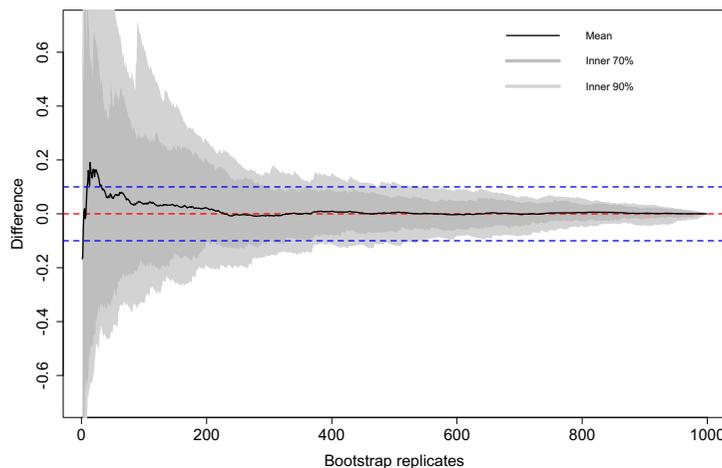


FIGURE 2.
Convergence behavior of MSE estimators per domain

tendency for underestimation for scenarios where the number of domains is small. However, this tendency decreases when increasing the number of domains, although at a rather slow rate. This is because increasing D also increases the number of unknown MSE values with to be estimated under new domain-specific measurement error distributions. Furthermore, recall the comment on the MSE estimator's bias at the end of Sect. 3.3. Depending on the application, researcher may consider using double bootstrap approaches in order to improve MSE estimates. However, with a relative absolute bias of roughly 10%, the MSE estimates are relatively accurate given the fact that a GLMM with measurement errors is used for prediction.

Figure 2 studies the convergence behavior of the parametric bootstrap for MSE estimation. It shows the distribution of the difference $mse_d^*(B) - mse_d^*(1000)$, $d = 1, \dots, D$, when the parametric bootstrap estimator is calculated based on $B \in \{1, \dots, 1000\}$ replicates.

The light gray band displays the inner 90% of differences over all domains, while the gray band marks the inner 70%. The mean is plotted by the black line. The blue dashed lines mark 10% differences, and the red dashed line displays 0% difference. We see that the mean stabilizes at about $B = 250$ iterations. Accordingly, with the $B = 300$ we employed in the simulation, the results are generally stable and allow for an evaluation of the overall MSE level in practice. However, in case regional analysis shall be conducted where MSE estimates are compared between domains, a higher number of iterations is required. This is due to the additional uncertainty resulting from the covariate measurement errors. The inner 70% of MSE estimates are within the 10% difference range for $B = 400$, and the inner 90% only for $B = 600$ or higher. That is to say, if regional

analysis between domains in terms of MSE estimates shall be conducted, we recommend at least $B = 600$.

6. Regional Prevalence Estimation of MDD

This section presents the application of the MEPM model in Sect. 3 to the data and inferential problem described in Sect. 2. Let us first recall that Sect. 3.1 distinguishes between covariates that are measured without error $\tilde{\mathbf{x}}_{0,d}$, and covariates that are measured with error $\tilde{\mathbf{x}}_{1,d}$. In our application, the variables $\{Domain\ size, Sex\}$ are measured without error. The first is retrieved from administrative records, and the second is a dummy variable. Their values can be included directly in the MEPM model. For the remaining covariates $\{Depr.\ treatment, SES\ score\}$, we use the GEDA data to obtain unbiased predictors $\mathbf{x}_{1,d}$ of the true covariate values that are statistically related to the sample counts y_d . For this, we produce Horvitz–Thompson-type direct estimates for $\tilde{\mathbf{x}}_{1,d}$ using the sample observations and survey weights of GEDA. Against the background of Sect. 3, we note that the Horvitz–Thompson estimator is unbiased and asymptotically normal distributed (Hájek, 1960; Berger, 1998; Chen and Rao, 2007). Accordingly, it satisfies the distribution assumptions with regards to the errors stated in Eq. (1). It further allows us to calculate the domain error covariance matrices Σ_d , $d = 1, \dots, D$, which are required in order to account for the measurement errors according to the developments in Sect. 3. For a given covariate x_{dk} in domain U_d that is measured with error, the error variance $\text{var}(u_{dk})$, $k = 1, \dots, p_1$, is equivalent to the variance of the respective Horvitz–Thompson estimator $\text{var}(x_{dk})$. The error covariances are obtained from (Wood, 2008)

$$\text{cov}(u_{dj}, u_{dk}) = \text{cov}(x_{dj}, x_{dk}) = \frac{1}{2} (\text{var}(x_{dj}) + \text{var}(x_{dk}) - \text{var}(x_{dj} - x_{dk})),$$

for $d = 1, \dots, D$ and $j, k = 1, \dots, p_1$ as well as $j \neq k$. Accordingly, for the final area-level MEPM model it holds that $y_d | v_d \sim \text{Pois}(\mu_d)$, where

$$\log \mu_d = \log v_d + \tilde{\mathbf{x}}_{0,d} \boldsymbol{\beta}_0 + \mathbf{x}_{1,d} \boldsymbol{\beta}_1 + \mathbf{u}'_d \boldsymbol{\beta}_1 + \phi v_d, \quad d = 1, \dots, D.$$

We perform variance estimation and significance testing for the model parameters as described in Sect. 4.2. In addition to that, we estimate the MSE of the resulting EBPs according to the parametric bootstrap algorithm presented in Sect. 3.3. For both operations, we choose $B = 1000$ for the number of bootstrap replicates.

Similarly to Sect. 5, we distinguish between the results of (i) model parameter estimation, (ii) domain prevalence estimation, and (iii) uncertainty estimation. We start with model parameter estimation. The results obtained from fitting the MEPM model as described in the last subsection are summarized in Table 6. It displays the estimated values, the standard deviations, the p values with respect to the hypothesis tests $H_0 : \theta_k = 0$, $k = 1, \dots, p + 1$, as well as 90% confidence intervals. Note that the last parameter *RE Std.-Dev.* refers to the random effect standard deviation ϕ . We see that estimated values of the regression parameters are plausible. The covariate *Depr. treatment* is positively linked to the target variable, as could be expected from its definition. The covariates *Sex - male* and *SES score* are negatively associated with the sample counts. This is in line with the literature concerning depressive disorders, as discussed in Sect. 2. The only noticeable variable is *Domain size*, whose coefficient is not significantly different from zero. However, we included it as fixed effect not for explanatory power, but for stability as substitute for an intercept. Thus, we did not expect the covariate to contribute relevantly to the variation of the counts.

For the remaining model parameters, it becomes evident that they are significantly different from zero on at least a 90% confidence level. We note that testing significance in the presence

TABLE 6.
Results of model parameter estimation

Parameter	Estimate	Std.-Dev.	<i>p</i> value	90% CI
<i>Domain size</i>	−0.00061	0.00134	0.64804	(−0.00281, 0.00159)
<i>Sex - male</i>	−0.46819	0.18321	0.01060	(−0.76957, −0.16681)
<i>Depr. treatment</i>	1.68745	0.90707	0.06284	(0.19532, 3.17958)
<i>SES score</i>	−0.25319	0.04167	0.00000	(−0.32173, −0.18464)
<i>RE Std.-Dev.</i>	0.76266	0.09162	–	(0.61195, 0.91337)

of measurement errors is particularly hard. On the one hand, the errors may draw the regression parameter estimates toward zero. On the other hand, anticipating covariate uncertainty tends to increase the variance of regression parameter estimation. Furthermore, the random effects are likely to be masked by the additional variation stemming from measurement errors, as both model components assume normal distributions with expectation zero. Nevertheless, we find all model parameters (except for the sample size) to be significant in the presented application. Not only the regression parameters, but also that the random effect standard deviation is significantly different from zero. This implies that the random effects are clearly statistically visible in the distribution of the target variable after accounting for the measurement errors.

We continue with domain prevalence estimation. The MDD sample prevalence observed in the SOEP for 2011 is 6.0%, where the sex-specific prevalences are 4.2% (male) and 7.8% (female). The MDD prevalence for the German population as estimated by the MEPM model is 6% with a 90% confidence interval of (5.7%, 6.3%). This is in line with the results of reference studies, for instance by Busch et al. (2013). They used the German health interview and examination survey for adults from 2008 to 2011 to estimate the 12-month prevalence of diagnosed MDD. The authors reported a total prevalence of 6% for adult Germans. In general, a methodologically sound comparison of the model predictions to external sources is difficult. Many empirical studies do not report MDD prevalence in particular, but depression in a broader sense that also includes lighter forms, such as depressive mood or states of low spirit. On that note, the public health reporting system of Germany stated the combined 12-month prevalence of depression and depressive moods for 2012 at 8% (Gesundheitsberichtserstattung des Bundes 2020). However, with regards to Busch et al. (2013), they explicitly distinguished between diagnosed MDD and lighter forms. Accordingly, the overall level of our prevalence estimates is plausible.

The male prevalence estimated by the MEPM model is 4.2% with a 90% confidence interval of (3.6%, 4.8%). Likewise, the female prevalence is estimated at 7.7% with an interval of (7.1%, 8.3%). The prevalence for each sex as well as the individual age groups is displayed in Table 7. It contains the respective estimate of the MEPM model parameter (Estimate), a corresponding 90% confidence interval (MEPM 90-CI), the prevalence observed in the sample (Sample), as well as a sample-based 90% confidence interval (Sample 90-CI). We see that the highest MDD prevalence is found in middle-aged females, that is, from 45 to 64 years old. We further see that in the groups of elderly, the MDD frequency overall decreases. These tendencies has also been found by Busch et al. (2013) as well as Robert Koch Institute (2013). Hence, we can conclude that the predicted evolution over the age groups is reasonable. Furthermore, we see that the confidence intervals obtained by MEPM are much smaller than their sample-based counterparts.

The observed sample prevalence and the estimated prevalence over all domains are compared in Fig. 3. We see that the sample proportions vary randomly around the predicted proportions of the model. This is in line with the distribution assumptions of the MEPM model. It indicates that there is no systematic error in the prediction behavior of the proposed method. We further see that

TABLE 7.
Estimated sex- and age-specific MDD prevalence

Age group	Sex	Estimate	MEPM 90-CI	Sample	Sample 90-CI
18 - 44	Combined	5.0%	(4.5%, 5.5%)	4.9%	(3.4%, 6.5%)
18 - 44	Male	3.4%	(2.8%, 4.1%)	3.4%	(2.0%, 4.7%)
18 - 44	Female	6.4%	(5.5%, 7.2%)	6.3%	(4.6%, 8.0%)
45 - 64	Combined	8.2%	(7.8%, 8.6%)	8.4%	(6.6%, 10.3%)
45 - 64	Male	5.8%	(5.3%, 6.3%)	6.1%	(4.4%, 7.7%)
45 - 64	Female	10.3%	(9.7%, 11.0%)	10.5%	(8.5%, 12.6%)
65+	Combined	4.2%	(3.6%, 4.7%)	4.0%	(2.4%, 5.6%)
65+	Male	2.7%	(2.1%, 3.4%)	2.5%	(1.2%, 3.8%)
65+	Female	5.5%	(4.6%, 6.3%)	5.3%	(3.5%, 7.2%)

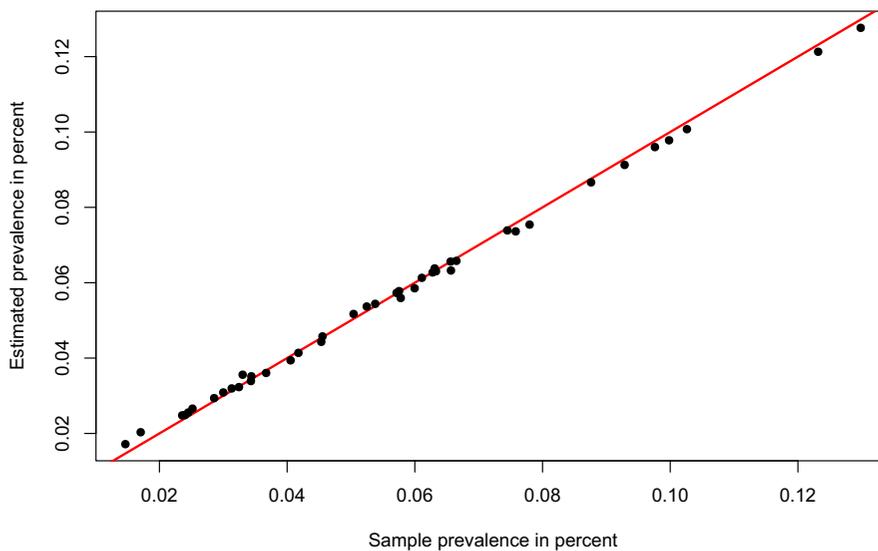


FIGURE 3.
Estimated domain prevalence versus sample domain prevalence

there are no larger deviations, which implies that the MM algorithm has achieved a good fitting performance based on the covariates.

Let us now investigate the spatial distribution of MDD in Germany. The corresponding estimates are presented in Fig. 4. It displays a heat map of the MDD prevalence over the seven Nielsen regions of Germany. Males and females are jointly considered in this graph. We see that the highest prevalence can be found in Bavaria. This is in line with the results of Melchior et al. (2014), who found that the highest depression prevalence in Germany for 2014 is located in Bavaria. However, please note that their findings are based on insurance claims of the statutory health insurance company BKK. Therefore, these results are not directly comparable to our survey-based estimates due to several selectivity issues. See Burgard et al. (2019) for further insights on this problem. With respect to the remaining regions, our results display low prevalences in Baden-Württemberg and in Northwest.

We finally address uncertainty estimation. For this, we compare the results of the parametric bootstrap estimator for MSE estimation in Sect. 3.3 with the uncertainty of the sample prevalence.

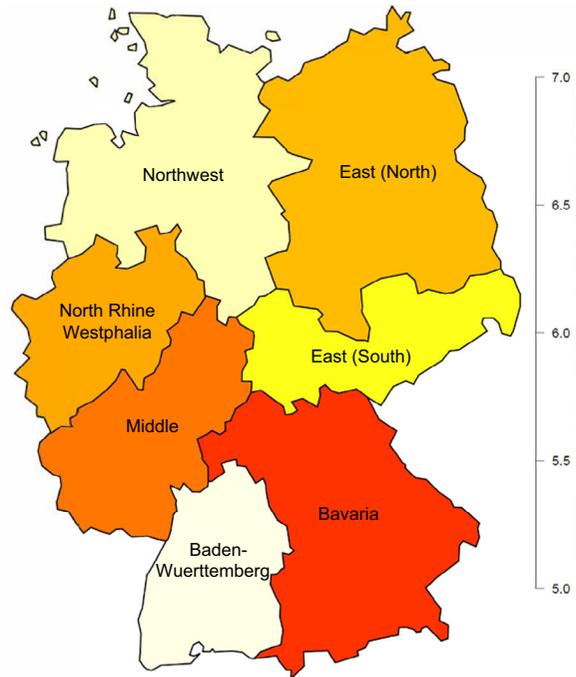


FIGURE 4.
Estimated regional prevalence

The results are summarized in Fig. 5. On the left-hand side, the estimated RMSEs of the EBP is plotted against the standard deviations of the sample proportions. The red line indicates the equality of both uncertainty measures. We see that the dots are all located under the line. This implies that the standard deviations of the sample proportions are always larger than the RMSEs of the EBP. On the right-hand side, the relative efficiency improvement

$$RE_d = 100 \frac{SD(\tilde{p}_d) - rmse^*(\hat{p}_d)}{SD(\tilde{p}_d)}, \quad d = 1, \dots, 42,$$

of the MEPM EBP over the simple sample proportion estimate is depicted. Here, $rmse^*(\hat{p}_d)$ denotes the estimated RMSE based on the parametric bootstrap sample in domain d , and $SD(\tilde{p}_d)$ is the standard deviation of the sample proportion. We see that the efficiency gains range between 14% and 61%. This is also in line with the results presented in Table 7, where the sample-based confidence intervals were visibly larger than those obtained with the MEPM approach. This underlines the efficiency advantage of the method over a standard sample-based estimation.

7. Conclusion

This paper proposed a novel model-based approach to estimate MDD prevalence on regional levels using uncertain data. The newly introduced method is an area-level Poisson mixed model that accounts for measurement errors in auxiliary information. With this feature, it is the first GLMM in SAE that allows for erroneous covariate records. The EBP under the model was derived and a parametric bootstrap algorithm for MSE estimation was presented. We further stated a MM

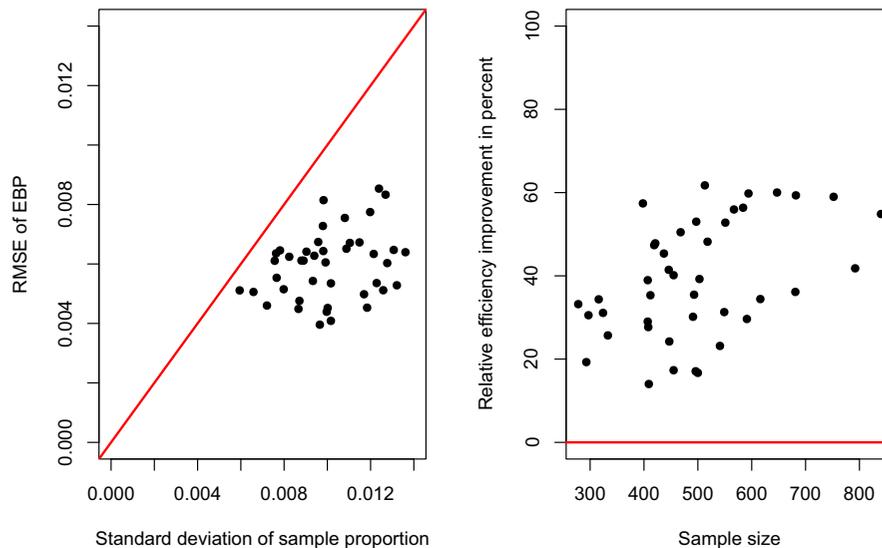


FIGURE 5.
Results of uncertainty estimation

fitting algorithm for model parameter estimation. The effectiveness of the methodology was established in several simulation experiments.

The approach was applied to obtained regional MDD prevalence estimation in Germany. The application presents a prevalence map over the seven Nielsen regions of Germany, where men and women are considered together. The map shows the highest prevalence in Bavaria and Mitte, which is in line with previous findings in the scientific literature. With respect to the rest of the regions, Baden-Württemberg and the North-West have a low prevalence and the East (South), East (North), and North Rhine-Westphalia are in an intermediate position. This type of disaggregated information allows a better understanding of the spatial distribution of health problems.

The proposed methodology enjoys attractive features from both a theoretical and practical perspective. We start with the theoretical aspects. The derived EBP is an empirical version of the BP, which marks the domain proportion predictor with minimum MSE, in the class of unbiased predictors, under the statistical setting our model considers. Furthermore, the presented MM approach is consistent in model parameter estimation, which is indeed a desirable property for any estimator. With regards to practice, the MEPM model is specified on aggregated data, which is much more likely to be available due to less restrictive privacy regulations. This holds in particular for data on sensitive topics, such as MDD. Further, the approach shows robustness in settings that are quite common in empirical research. We rarely have a census or register data on mental disorders. Accordingly, the vast majority of area-level records in this research field are in fact survey-based estimates and, thus, subject to errors.

A drawback of the method is that it can only be fitted with exclusively area-level data if the respective variances and covariances of the auxiliary variable predictors are known. Otherwise, these terms have to be estimated from the respective survey data as well. This would then require the use of unit-level (person-level) data in addition to the area-level records. However, even if this is necessary, the MEPM model is still more practical than any other model-based unit-level approach. In order to quantify the EBP of a domain quantity under a GLMM, it is typically required to have covariate observations for every person within the target population (Hobza et al. 2018). That is to say, we would need to have unit-level data not only from the sampled individuals, but

also from the non-sampled individuals. Such information is rarely available in practice, which makes the MEPM model an attractive alternative for these applications.

In addition to these aspects, the general concept of accounting for measurement errors as presented has a much broader range of applications. Beside the prediction of domain parameters, the methodology may also be used to analyze the outcomes of clinical studies or field experiments. Here, researchers might be interested in the structural relation between several variables, but need to account for measurement interference due to external circumstances. In that case, a corresponding measurement error approach can be used to fit parametric regression models that quantify these relations under the inherent uncertainty.

A related but different problem from error covariates is the incorrect model specification. This is a relevant topic for model-based SAE. Jiang et al. (2011) introduced the OBPs under Fay-Herriot models and under nested error regression models. Chen et al. (2015) extended the OBP approach to area-level Poisson mixed models. These authors show that OBPs are robust with respect to model misspecifications. The extension of the OBP approach to the measurement error Fay-Herriot model of Ybarra and Lohr (2008), or to the measurement error GLMMs, is thus desirable for applications to real data. However, it is not a straightforward task and it will deserve future research.

Acknowledgments

This research is supported by the Spanish grant PGC2018-096840-B-I00 and by the research project “RIFOSS - Research Innovation for Official and Survey Statistics” funded by the German Federal Statistical Office.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

Declarations

Data availability The census data on domain sizes used in this study are freely available at the website of the German Federal Statistical Office. The target variable sample data from the survey SOEP are available on demand at Deutsches Institut für Wirtschaftsforschung (German Institute for Economic Research). The auxiliary variable sample data from the survey GEDA are available on demand at the Robert-Koch Institute, Germany. The sample data are protected under national privacy regulations. The authors of this article have no legal permission to share the data with any third party.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

- Albert, P. R. (2015). Why is depression more prevalent in women? *Journal of Psychiatry & Neuroscience*, 40(4), 219–221.
- Arima, S., Bell, W. R., Datta, G. S., Franco, C., & Liseo, B. (2017). Multivariate Fay–Herriot Bayesian estimation of small area means under functional measurement error. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 180(4), 1191–1209.

- Arima, S., & Poletini, S. (2019). A unit level small area model with misclassified covariates. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, *182*(4), 1439–1462.
- Arima, S., Datta, G. S., & Liseo, B. (2015). Bayesian estimators for small area models when auxiliary information is measured with error. *Scandinavian Journal of Statistics*, *42*(1), 518–529.
- Bell, W. R., Chung, H. C., Datta, G. S., & Franco, C. (2019). Measurement error in small area estimation: Functional versus naive models. *Survey Methodology*, *45*(1), 61–80.
- Berger, Y. G. (1998). Rate of convergence to normal distribution for the Horvitz–Thompson estimator. *Journal of Statistical Planning and Inference*, *67*(2), 209–226.
- Boubeta, M., Lombardía, M. J., & Morales, D. (2016). Empirical best prediction under area-level Poisson mixed models. *TEST*, *25*, 548–569.
- Boubeta, M., Lombardía, M. J., & Morales, D. (2017). Poisson mixed models for studying the poverty in small areas. *Computational Statistics and Data Analysis*, *107*, 32–47.
- Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, *88*(421), 9–25.
- Brignone, E., George, D. R., Sinoway, L., Katz, C., Sauder, C., Murray, A., Gladden, R., & Kraschnewski, J. L. (2020). Trends in the diagnosis of diseases of despair in the United States, 2009–2018: A retrospective cohort study. *BMJ Open*, *10*, e037679.
- Burgard, J., Esteban, M., Morales, D., & Pérez, A. (2020). A Fay–Herriot model when auxiliary variables are measured with error. *TEST*, *29*(1), 166–195.
- Burgard, J., Esteban, M., Morales, D., & Pérez, A. (2021). Small area estimation under a measurement error bivariate Fay–Herriot model. *Statistical Methods and Applications*, *30*, 79–108.
- Burgard, J. P., Krause, J., & Münnich, R. (2019). Adjusting selection bias in German health insurance records for regional prevalence estimation. *Population Health Metrics*, *17*(13), 1–13.
- Busch, M. A., Maske, U. E., Ryl, L., Schlack, R., & Hapke, U. (2013). Prevalence of depressive symptoms and diagnosed depression among adults in Germany—results of the German health interview and examination for adults (DEGS1). *Bundesgesundheitsblatt*, *56*, 733–739.
- Calfisch, R. E. (1998). Monte Carlo and quasi-Monte Carlo methods. *Acta Numerica*, *7*, 1–49.
- Chambers, R., Dreassi, E., & Salvati, N. (2014). Disease mapping via negative binomial regression m-quantiles. *Statistics in Medicine*, *33*, 4805–4824.
- Chambers, R., Salvati, N., & Tzavidis, N. (2016). Semiparametric small area estimation for binary outcomes with application to unemployment estimation for local authorities in the UK. *Journal of the Royal Statistical Association. Series A (Statistics in Society)*, *179*(2), 453–479.
- Chen, J., & Rao, J. N. K. (2007). Asymptotic normality under two-phase sampling designs. *Statistica Sinica*, *17*(3), 1047–1064.
- Chen, S., Jiang, J., & Nguyen, T. (2015). Observed best prediction for small area counts. *Journal of Survey Statistics and Methodology*, *3*, 135–161.
- Chen, S., & Lahiri, P. (2012). Inferences on small area proportions. *Journal of the Indian Society of Agricultural Statistics*, *66*, 121–124.
- Cuijpers, P., Vogelzangs, N., Twisk, J., Kleiboer, A., Li, J., & Penninx, B. W. (2014). Comprehensive meta-analysis of excess mortality in depression in the general community versus patients with specific illnesses. *American Journal of Psychiatry*, *171*(4), 453–462.
- Cyranowski, J. M., Frank, E., Young, E., & Shear, K. (2000). Adolescent onset on the gender difference in lifetime rates of major depression—A theoretical model. *Archives of General Psychiatry*, *57*(1), 21–27.
- Dreassi, E., Ranalli, M. G., & Salvati, N. (2014). Semiparametric m-quantile regression for count data. *Statistical Methods in Medical Research*, *23*, 591–610.
- Erciulescu, A. L., & Fuller, W. A. (2013). Small area prediction of the mean of a binomial random variable. *Joint Statistical Meeting 2013 Proceedings - Survey Research Methods Section, Session 37 Small-Area Estimation: Theory and Applications*, pp. 855–863. Downloadable from http://www.asasrms.org/Proceedings/y2013/files/307932_79947.pdf
- Faltys, O., Hobza, T., & Morales, D. (2020). Small area estimation under area-level generalized linear mixed models. *Communications in Statistics - Simulation and Computation*. <https://doi.org/10.1080/03610918.2020.1836216>
- Freeman, A., Tyrovolas, S., Koyanagi, A., Chatterji, S., Leonardi, M., Ayuso-Mateos, J. L., Tobiasz-Adamczyk, B., Koskinen, S., Rummel-Kluge, C., & Haro, J. M. (2016). The role of socio-economic status in depression: Results from the COURAGE (aging survey in Europe). *BMC Public Health*, *16*(1098), 1–8.
- Gesundheitsberichtserstattung des Bundes (2020). Durch einen Arzt oder Psychotherapeuthen diagnostizierte Depression oder depressive Verstimmung in den letzten 12 Monaten. <http://www.gbe-bund.de>.
- Ghosh, M., Kim, D., Sinha, K., Maiti, T., Katzoff, M., & Parsons, V. L. (2009). Hierarchical and empirical Bayes small domain estimation and proportion of persons without health insurance for minority subpopulations. *Survey Methodology*, *35*, 53–66.
- Ghosh, M., & Maiti, T. (2004). Small-area estimation based on natural exponential family quadratic variance function models and survey weights. *Biometrika*, *91*, 95–112.
- Ghosh, M., & Sinha, K. (2007). Empirical Bayes estimation in finite population sampling under functional measurement error models. *Journal of Statistical Planning and Inference*, *137*, 2759–2773.
- Ghosh, M., Sinha, K., & Kim, D. (2006). Empirical and hierarchical Bayesian estimation in finite population sampling under structural measurement error models. *Scandinavian Journal of Statistics*, *33*, 591–608.

- Gilman, S. E., Sucha, E., Kingsbury, M., Horton, N. J., Murphy, J. M., & Colman, I. (2017). Depression and mortality in a longitudinal study: 1952–2011. *Canadian Medical Association Journal*, *189*(42), E1304–E1310.
- Glymour, M. M., Maselko, J., Gilman, S. E., Patton, K. K., & Avendaño, M. (2010). Depressive symptoms predict incident stroke independently of memory impairments. *Neurology*, *75*(23), 2063–2070.
- Goebel, J., Krause, P., Pischner, R., Sieber, I., & Wagner, G. G. (2008). *Daten- und datenbankstruktur der längsschnittstudie sozio-oekonomisches panel (soep)*. Online. SOEPpapers on Multidisciplinary Panel Data Research 89, DIW Berlin, The German Socio-Economic Panel (SOEP).
- Haan, P., Hammerschid, A., Lindner, R., & Schmieder, J. (2019). Todesfälle durch suizid, alkohol und drogen sinken deutlich bei Männern und Frauen in Ost- und Westdeutschland. *DIW Wochenbericht*, *86*(7–8), 99–105.
- Hájek, J. (1960). Limiting distributions in simple random sampling from a finite population. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, *5*, 361–374.
- Hall, P., & Maiti, T. (2006). On parametric bootstrap methods for small area estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, *68*, 221–238.
- Hobza, T., Marhuenda, Y., & Morales, D. (2020). Small area estimation of additive parameters under unit-level generalized linear mixed models. *SORT*, *44*(1), 3–38.
- Hobza, T., & Morales, D. (2016). Empirical best prediction under unit-level logit mixed models. *Journal of Official Statistics*, *32*(3), 661–692.
- Hobza, T., Morales, D., & Santamaría, L. (2018). Small area estimation of poverty proportions under unit-level temporal binomial-logit mixed models. *TEST*, *27*, 270–294.
- Jiang, J. (1998). Consistent estimators in generalized linear mixed models. *Journal of the American Statistical Association*, *93*(442), 720–729.
- Jiang, J. (2003). Empirical best prediction for small-area inference based on generalized linear mixed models. *Journal of Statistical Planning and Inference*, *111*(1–2), 117–127.
- Jiang, J., Nguyen, T., & Rao, S. (2011). Best predictive small area estimation. *Journal of the American Statistical Association*, *106*(494), 732–745.
- Jo, S.-J., Yim, H. W., Bang, M. H., Lee, M. O., Jun, T.-Y., Choi, J.-S., Lee, M.-S., & Park, Y.-M. (2011). The association between economic status and depressive symptoms: An individual and community level approach. *Psychiatry Investigation*, *8*(3), 194–200.
- Kozela, M., Bobak, M., Besala, A., Micek, A., Kubinova, R., Malyutina, S., Denisova, D., Richards, M., Pikhart, H., Peasy, A., Marmot, M., & Pajak, A. (2016). The association of depressive symptoms with cardiovascular and all-cause mortality in central and eastern Europe: Prospective results of the HAPIEE study. *European Journal of Preventive Cardiology*, *23*(17), 1839–1847.
- Lange, C., Jentsch, F., Allen, J., Hoebel, J., Kratz, A. L., von der Lippe, E., Müters, S., Schmich, P., Thelen, J., Wetzstein, M., Fuchs, J., & Ziese, T. (2015). Data resource profile: German health update (GEDA)—the health interview survey for adults in Germany. *International Journal of Epidemiology*, *44*(2), 442–450.
- Liu, B., & Lahiri, P. (2017). Adaptive hierarchical Bayes estimation of small area proportions. *Calcutta Statistical Association Bulletin*, *69*(2), 150–164.
- López-Vizcaíno, Lombardía, & Morales, D. (2013). Multinomial-based small area estimation of labour force indicators. *Statistical Modelling*, *13*(2), 153–178.
- López-Vizcaíno, Lombardía, & Morales, D. (2015). Small area estimation of labour force indicators under a multinomial model with correlated time and area effects. *Journal of the Royal Statistical Association. Series A (Statistics in Society)*, *178*(3), 535–565.
- Marchetti, S., Giusti, C., Pratesi, M., Salvati, N., Giannotti, F., Pedreschi, D., Rinzivillo, S., Pappalardo, L., & Gabrielli, L. (2015). Small area model-based estimators using big data sources. *Journal of Official Statistics*, *31*(2), 263–281.
- Melchior, H., Schulz, H., Walker, J., Ganninger, M., & Hürter, M. (2014). Unterschiede in der Prävalenz und der Versorgung depressiver Erkrankungen. In F. Knieps & H. Pfaff (Eds.), *Gesundheit in Regionen (BKK Gesundheitsreport 2014)* (pp. 87–92). Medizinisch Wissenschaftliche Verlagsgesellschaft.
- Militino, A. F., Ugarte, M. D., & Goicoa, T. (2015). Deriving small area estimates from information technology business surveys. *Journal of the Royal Statistical Association. Series A (Statistics in Society)*, *178*(4), 1051–1067.
- Molina, I., Saei, A., & Lombardía, M. J. (2007). Small area estimates of labour force participation under a multinomial logit mixed model. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, *170*(4), 975–1000.
- Morales, D., Esteban, M. D., Pérez, A., & Hobza, T. (2021). *A course on small area estimation and mixed models*. Springer.
- Pfeffermann, D. (2013). New important developments in small area estimation. *Statistical Science*, *28*(1), 40–68.
- Rao, J. N. K., & Molina, I. (2015). *Small area estimation*. Wiley Series in Survey Methodology (2nd ed.). Wiley.
- Robert Koch Institute. (2013). *German health update 2010 (GEDA 2010). Public use file third version*. <https://doi.org/10.7797/27-200910-1-1-3>
- Singh, T. (2011). *Efficient small area estimation in the presence of measurement error in covariates*. Ph.D. Thesis, Texas A&M University.
- Sugasawa, S., & Kubokawa, T. (2020). Small area estimation with mixed models: A review. *Japanese Journal of Statistics and Data Science*, *3*, 693–720.
- Sugasawa, S., Kubokawa, T., & Ogasawara, K. (2017). Empirical uncertain Bayes methods in area-level models. *Scandinavian Journal of Statistics*, *44*, 684–706.
- Torabi, M. (2013). Likelihood inference in generalized linear mixed measurement error models. *Computational Statistics and Data Analysis*, *57*(1), 549–557.
- Torabi, M., Datta, G. S., & Rao, J. N. K. (2009). Empirical Bayes estimation of small area means under a nested error linear regression model with measurement errors in the covariates. *Scandinavian Journal of Statistics*, *36*(2), 355–369.

- Wagner, G. G., Frick, J. R., & Schupp, J. (2007). *The German socio-economic panel study (SOEP): Scope, evolution and enhancements*. Online. SOEPpapers on Multidisciplinary Panel Data Research 1, DIW Berlin, The German Socio-Economic Panel (SOEP).
- Walker, E. R., McGree, R. E., & Druss, B. G. (2015). Mortality in mental disorders and global disease burden implications: A systematic review and meta-analysis. *JAMA Psychiatry*, 72(4), 334–341.
- Wood, J. (2008). On the covariance between related Horvitz–Thompson estimators. *Journal of Official Statistics*, 24(1), 53–78.
- Ybarra, L. M. R., & Lohr, S. L. (2008). Small area estimation when auxiliary information is measured with error. *Biometrika*, 95, 919–931.
- Zivin, K., Yosef, M., Miller, E. M., Valenstein, M., Duffy, S., Kales, H. C., Vijan, S., & Kim, H. M. (2015). Associations between depression and all-cause and cause-specific risk of death: A retrospective cohort study in the veterans health administration. *Journal of Psychosomatic Research*, 78(4), 324–331.

Manuscript Received: 19 NOV 2020

Final Version Received: 16 AUG 2021

Accepted: 20 AUG 2021

Published Online Date: 6 SEP 2021