Towards a Cautious Modelling of Missing Data in Small Area Estimation

Julia Plass JULIA.PLASS@STAT.UNI-MUENCHEN.DE
Aziz Omar AZIZ.OMAR@STAT.UNI-MUENCHEN.DE
Thomas Augustin AUGUSTIN@STAT.UNI-MUENCHEN.DE
Department of Statistics, LMU Munich, Germany (Plass, Omar, Augustin)
Department of Mathematics, Insurance and Applied Statistics, Helwan University Egypt (Omar)

Abstract

In official statistics, the problem of sampling error is rushed to extremes when not only results on sub-population level are required, which is the focus of Small Area Estimation (SAE), but also missing data arise. When the nonresponse is wrongly assumed to occur at random, the situation becomes even more dramatic, since this potentially leads to a substantial bias. Even though there are some treatments jointly considering both problems, they are all reliant upon the guarantee of strong assumptions on the missingness. For that reason, we aim at developing cautious versions of well known estimators from SAE by exploiting the results from a recently suggested likelihood approach, capable of including tenable partial knowledge about the nonresponse behaviour in an adequate way. We generalize the synthetic estimator and propose a cautious version of the so-called LGREG-synthetic estimator in the context of design-based estimators. Then, we elaborate why the approach above does not directly extend to model-based estimators and proceed with some first studies investigating different missingness scenarios. All results are illustrated through the German General Social Survey 2014, also including area-specific auxiliary information from the German Federal Statistical Office's data report.

Keywords: small area estimation; LGREG-synthetic estimator; missing data; partial identification; sensitivity analysis; likelihood; logistic regression; logistic mixed model; German General Social Survey.

1. Introduction

Survey methodology distinguishes between sampling and non-sampling error (cf., e.g., Biemer, 2010). Sampling error occurs when only a subset, but not the whole population can be included in a survey, yet the aim is to generalize the results beyond the units that have been sampled. Sampling error is especially severe if the population is composed of several sub-populations and the samples drawn from these sub-populations are not large enough to permit a satisfying precision on sub-population level. A set of methods has been introduced to tackle such situations and is referred to as *Small Area Estimation* (SAE). The main approach of SAE is to use additional data sources, such as administrative records and census data, as auxiliary data in an attempt to increase the effective sample size (cf., e.g., Münnich et al., 2013; Rao and Molina, 2015).

A common non-sampling error encountered in inference is item-nonresponse. Applying the EM-algorithm and Multiple Imputations are the recent practices (cf., e.g., Little and Rubin, 2014). Both techniques force point-identifiability, i.e. uniqueness of parameters, by requiring the assump-

tion that the missingness is occurring randomly (MAR), i.e. independently of the true underlying value of the variable of interest given covariates. Since the MAR assumption is generally not testable and wrongly imposing it may cause a substantial bias, results have to be treated with caution.

According to the methodology of partial identification in the spirit of Manski (2003), one does not have to insist on strong assumptions to obtain a result at all. Allowing for partially identified parameters enables to incorporate tenable knowledge only. In this way, one receives imprecise – but credible – results, which are refined if additional knowledge about the missingness is available. In this context, there are already several approaches refraining from strong assumptions on the missingness process (cf., e.g., Couso and Dubois, 2014; Denœux, 2014). These cautious procedures also represent a popular field of research of the ISIPTA symposia (cf.,e.g., Cattaneo and Wiencierz, 2012; Schollmeyer and Augustin, 2015; Utkin and Coolen, 2011). Since we may not conjure information about the missingness process or make other strong modelling assumptions (cf., e.g., Couso and Sánchez, 2016; Hüllermeier, 2014), uncertainty due to nonresponse has to be interpreted as lack of knowledge. Thus, approaches, explicitly communicating the associated uncertainty, are indispensable. In the context of official statistics this point was recently stressed by Manski (2015).

Since nonresponse may seriously reduce the already small sample size in SAE jointly considering both issues is especially challenging. As far as we know, already existing approaches dealing with nonresponse in SAE are based on strong assumptions on the missingness process, as MAR or the missing not at random (NMAR) assumption plus strict distributional assumptions. Thus, considering a cautious approach for dealing with nonresponse in SAE represents the core of this paper. To pursue this goal, in Section 2 we start by introducing the notation for the setting considered here followed by an introduction to our application using the German General Social Survey. Afterwards, we give a basic overview about prominent design-based estimators applicable in our situation in Section 3. Two design-based estimators, the classical synthetic estimator and the LGREG-synthetic estimator, are generalized in Section 4. While cautious versions are given for the case of including no missingness assumptions at all, the case of including weak assumptions is considered for both estimators by relying on the cautious likelihood approach developed in Plass et al. (2015). In Section 5 the results are illustrated by means of the application example. In Section 6 we discuss why our approach cannot be directly extended to prominent model-based estimators and then perform a first sensitivity analysis. Section 7 concludes by summarizing the major points and giving some remarks on further research.

2. Setting

Technically, our setting is as follows: Let the population U under study have a total size of N units, and be divided into M non-overlapping domains (areas) U_i, each containing units j, $j = 1, \ldots, N_i$ with N_i as the size of U_i, $i = 1, \ldots, M$. Let Y be a binary variable of interest that is assumed to have a relation with a set of k precisely observed categorical covariates $X_1, \ldots, X_k$ through a certain model. Cross classifying the categorical covariates forms a k-dimension table with a total number of cells v, where the g-th cell – representing the g-th subgroup of the population – contains known joint absolute frequency $X_i^{[g]}$, $g = 1, \ldots, v$, $i = 1, \ldots, M$. To infer about π_i, the probability of a certain category of Y in area i, a sample s of size n is selected, such that a sample s_i of size n_i is selected from area i with $\sum_{i=1}^{M} n_i = n$. Within s_i, sample units j, $j = 1, \ldots, n_i$ ($j \in s_i$) are selected with inclusion probability $1/w_{ij}$, where w_{ij} are the usual sample weights. Sample values of the covariates, denoted by $x_{1ij}, \ldots, x_{kij}$, are assumed to be completely observed, while

of sample values of Y, denoted by y_{ij}, some are missing. Accordingly, s_i is partitioned into $s_{i,obs}$ and $s_{i,mis}$ that refer to sample units with observed and unobserved values of Y, respectively. If we additionally split by g, the samples are denoted by $s_i^{[g]}$, $s_{i,obs}^{[g]}$ and $s_{i,mis}^{[g]}$.

Application example: To illustrate the setting (and later on the results), we rely on the German General Social Survey (GGSS) (GESIS Leibniz Institute for the Social Sciences, 2016). We are interested in the area-specific ratio of people at risk of poverty, where German federal states are the areas completely partitioning the overall domain "Germany" (i.e. $M = 17$)[1]. We construct a binary response variable with values "poor" and "rich" by comparing the collected equivalent income measured on the OECD modified scale with the poverty risk threshold given by 60% of the median net equivalent income, i.e. 986.65€ for year 2014 (DESTATIS, Statistisches Bundesamt, 2016b). The poverty variable shows 454 missing values. As covariates, we use the highest school leaving certificate, which – for ease of presentation – is dichotomized, distinguishing between categories "no Abitur"[2] and "Abitur" only, as well as sex.[3] We base the analysis on the sample with $|s| = 3466$, $|s_{obs}| = 3012$, $|s_{mis}| = 454$. The German Federal Statistical Office' data report (DESTATIS, Statistisches Bundesamt, 2016a) provides area-specific totals $X_i^{[g]}$, $i = 1, \ldots, M$, $g = 1, \ldots, v$, split by the values of the covariates, i.e. the absolute frequencies of the four subgroups "male-no Abitur", "male-Abitur", "female - no Abitur " and "female - Abitur" in area i.

3. Theoretical Background of Design-Based Estimators

SAE techniques result in producing estimators $\hat{\pi}_i$ for area of interest i, $i = 1, \ldots, M$, that are either design-based or model-based.[4] In this paper, we mainly refer to design-based estimators, while we consider model-based ones in Section 6 only. Design-based estimators are either direct estimators that only use data from the targeted area, or indirect estimators that rely on data from other areas as well. This is justified under the assumption of similarity between the areas made to *borrow strength* from other areas.

The Horvitz-Thompson (HT) estimator (Horvitz and Thompson, 1952) $\hat{\pi}_{i,HT} = \frac{1}{N_i} \sum_{j=1}^{n_i} w_{ij} y_{ij}$ for an area i, well known in sampling theory, provides a method to estimate the mean of subpopulation (area) i, thereby accounting for the different sampling probabilities of respondents by sampling weights. The so-called *synthetic estimator* from SAE is a design-based indirect estimator, which is built upon the HT estimator, incorporating not only information from the area of interest, but averaging over all M areas. Thus, the area specific probability π_i is estimated as

$$\hat{\pi}_{i,\text{SYN}} \equiv \hat{\pi}_{\text{SYN}} = \frac{1}{N} \sum_{i=1}^{M} \sum_{j \in s_i} w_{ij} y_{ij} = \frac{1}{N} \sum_{i=1}^{M} N_i \cdot \hat{\pi}_{i,HT} \ , \ \forall i = 1, \ldots, M \ . \tag{1}$$

Since there is no distinction between areas and sample information is included about the response variable only, it merely serves as a basis for further estimators.

1. Although Germany is divided into 16 federal states, the GGSS differentiates between 17 ones, additionally distinguishing between "former East-Berlin" and "former West-Berlin".
2. The "Abitur" is the general qualification for university entrance in Germany.
3. Since there should not be any regional differences with regard to covariate sex, the reason for the inclusion of this covariate rather lies in the interest of illustrating the subgroup specific analysis in a proper way than in an increase of explanatory power in the subject matter context.
4. While properties of design-based estimators (e.g. bias and variance) are evaluated under sampling distribution over all samples with population parameters held fixed, model-based estimators usually condition on the selected sample, and inference regarding them is carried out with respect to the underlying model (cf., e.g., Rao and Molina, 2015).

An estimator that employs sample data as well as area specific auxiliary information on the joint totals $X_{1i}, \ldots, X_{ki}$ is the GREG-synthetic estimator (cf. Särdnal et al., 1992), where we here use its logistic version, the *LGREG-synthetic estimator* (cf. Lehtonen and Veijanen, 1998). Applying the LGREG-synthetic estimator is split into two steps:

First, the regression coefficients β_0, β_1,..., β_k are estimated by means of a standard logistic regression model linking π_{ij}, i.e. the probability for individual j, $j = 1, \ldots, n_i$ in s_i, $i = 1, \ldots, M$, to have the value $y_{ij} = 1$, to the linear predictor containing the individual auxiliary information, here always assuming that all interactions are incorporated.[5] Referring to the application example, we consider two covariates, hence the model includes β_0, β_1, β_2 and an interaction $\beta_{1:2}$, expressing the joint effect of both covariates. According to the aim of borrowing strength, one obtains global regression coefficients. From the estimated global regression coefficients, by applying the response function of a standard logistic regression model, we receive global predictions that only depend on the values of the covariate, but are independent of the area. To stress this, we write $\hat{\pi}^{[g]}$, $g = 1, \ldots, v$, instead of $\hat{\pi}_{ij}$ in our case of categorical covariates. The calculation of these predictions becomes simpler here: Due to the strict monotonicity of the response function, the categorical nature of the covariates and the inclusion of all interactions, a unique relation between the regression coefficients and the predictions can be shown (as, e.g., addressed in Plass et al., 2017). Consequently, we can directly calculate the subgroup specific predictions by

$$\hat{\pi}^{[g]} = \sum_{i=1}^{M} \sum_{j \in s_i^{[g]}} \frac{y_{ij}}{n^{[g]}} \ , \tag{2}$$

with $n^{[g]}$ denoting the cell-count in subgroup g, $g = 1, \ldots, v$.

Second, area-specific information is used: In our setting, the original LGREG-estimator (cf., e.g., Lehtonen and Veijanen, 1998, p.52) for a certain area of interest i can be expressed as

$$\hat{\pi}_{i,LGREG} = \sum_{g=1}^{v} \Big(\overbrace{\sum_{j \in s_{i,g}} w_{ij} y_{ij}}^{\text{HT-part}} + \overbrace{\hat{\pi}^{[g]} \left(X_i^{[g]} - \sum_{j \in s_{i,g}} w_{ij} \right)}^{\text{correction term}} \Big) / N_i \ . \tag{3}$$

It can be understood as the HT estimator corrected by a term accounting for under- and overrepresentation of certain constellations of covariates in the sample, present in case of $X_i^{[g]} > \sum_{j \in s_{i,g}} w_{ij}$ and $X_i^{[g]} < \sum_{j \in s_{i,g}} w_{ij}$, resprectively. The subgroup specific representation in (3) will turn out to be beneficial in context of developing a cautious version (cf. Section 4.2 and 4.3).

4. Cautious Versions of Design-based Estimators under Nonresponse

Since the already established ways of dealing with nonresponse in SAE require strong assumptions, we aim at improving the presented prominent estimators by striving for a proper reflection of the available information on the missingness process. For this purpose, we use the framework of the cautious approach developed for the more general case of coarse[6] categorical data in Plass et al.

5. This is quite natural in this context, since only then the full information about the subgroup specific information, also provided by the auxiliary information in terms of totals, is used.
6. The data problem only distinguishes between fully observed and completely unobserved values, while coarse data additionally include partial observations, e.g. in the sense of grouped data (cf. Heitjan and Rubin, 1991).

(2015) and further extended in Plass et al. (2017) to practically frame the inclusion of auxiliary information. We start by recalling the basic elements of this approach in the following section.

4.1 A Cautious Approach for Dealing with Nonresponse

An observation model $\mathcal{Q}$ is used as a medium to frame the procedure of incorporating auxiliary information on the incompleteness. Restricting to the missing data problem and a binary response variable and considering the problem for subgroup g, $g = 1, \ldots, v$, the model $\mathcal{Q}^{[g]}$ is determined by the set of missingness parameters $q^{[g]}_{na|y}$, i.e. the probability associated with refusing the answer ("na"), given a certain subgroup g and the true value $y \in \{0, 1\}$ of the response variable.[7] In the spirit of partial identification, one can start by incorporating "no" assumptions[8] on $q^{[g]}_{na|y}$, then restricting these missingness parameters successively by certain conceivable conditions. The cautious approach includes this observation model into a classical categorical likelihood problem. For this purpose, a connection between the parameters $\pi^{[g]}$ and $p^{[g]}_{\mathfrak{y}}$ is established via the observation model, where $p^{[g]}_{\mathfrak{y}}$ refers to the observed value $\mathfrak{y} \in \{0, 1, na\}$, thus treating the missing values as a category of its own. The invariance of the likelihood allows to rewrite the log-likelihood in terms of $p^{[g]}_{\mathfrak{y}}$, which can be uniquely maximized in terms of the parameters of interest by relying on the theorem of total probability, receiving

$$\begin{aligned}\ell(\pi^{[g]},\, q^{[g]}_{na|0},\, q^{[g]}_{na|1}) =& n^{[g]}_1\Big(\ln(\pi^{[g]}) + \ln(1 - q^{[g]}_{na|1})\Big) + n^{[g]}_0\Big(\ln(1 - \pi^{[g]}) + \ln(1 - q^{[g]}_{na|0})\Big) \\ &+ n^{[g]}_{na}\Big(\ln(\pi^{[g]} q^{[g]}_{na|1} + (1 - \pi^{[g]}) q^{[g]}_{na|0})\Big)\,, \end{aligned} \tag{4}$$

where $n^{[g]}_1$, $n^{[g]}_0$ and $n^{[g]}_{na}$ refer to the respective observed cell counts within subgroup g, which later on have to be replaced by appropriate sample weights. By maximizing the log-likelihood in (4), we determine the generally set-valued[9] estimators, whose one-dimensional projections can be represented by the lower and upper bounds of intervals, namely $\hat{\underline{\pi}}^{[g]}$, $\hat{\overline{\pi}}^{[g]}$, $\hat{\underline{q}}^{[g]}_{na|0}$, $\hat{\overline{q}}^{[g]}_{na|0}$, $\hat{\underline{q}}^{[g]}_{na|1}$ and $\hat{\overline{q}}^{[g]}_{na|1}$. Thereby, $\hat{\underline{\pi}}^{[g]}$ is attained under $\hat{\overline{q}}^{[g]}_{na|0}$ and $\hat{\underline{q}}^{[g]}_{na|1}$, while $\hat{\overline{\pi}}^{[g]}$ is associated with $\hat{\underline{q}}^{[g]}_{na|0}$ and $\hat{\overline{q}}^{[g]}_{na|1}$.

By considering $q^{[g]}_{na|1} = R \cdot q^{[g]}_{na|0}$, with missing ratio $R \in \mathcal{R} \subseteq \mathbb{R}^+_0$ (also cf. Nordheim (1984)),[10] and $\mathcal{R}$ as the set of missing ratios, assumptions about the missingness can be incorporated. Specific values of R are associated with a particular missingness scenario, thus point-identifying $\pi^{[g]}$. For instance, $R = 1$ represents the missingness scenario under gMAR[11], requiring $q^{[g]}_{na|1} = q^{[g]}_{na|0}$. Partial (weak) assumptions, like incorporating $R \in \mathcal{R}$ into (4), thus refine the result obtained from the log-likelihood optimization without the inclusion of any missingness assumptions. Since it can be shown that $\hat{\underline{\pi}}^{[g],\mathcal{R}}$, $\hat{\overline{q}}^{[g],\mathcal{R}}_{na|0}$ and $\hat{\underline{q}}^{[g],\mathcal{R}}_{na|1}$ as well as $\hat{\overline{\pi}}^{[g],\mathcal{R}}$, $\hat{\underline{q}}^{[g],\mathcal{R}}_{na|0}$ and $\hat{\overline{q}}^{[g],\mathcal{R}}_{na|1}$, i.e. the bounds under the partial assumptions expressed by $\mathcal{R} = [\underline{R}, \overline{R}]$, are achieved under missingness ratios $\underline{R}$ and $\overline{R}$, respectively, one does not have to optimize the log-likelihood for all values in $[\underline{R},\ \overline{R}]$, but optimizing under $\underline{R}$ and $\overline{R}$ is sufficient. While $\mathcal{R} = [0,\ 1]$ corresponds to $q^{[g]}_{na|1} \leq q^{[g]}_{na|0}$, a cautious version of

7. Referring to the framework of analyzing contingency tables, it is natural to drop the reference to individual j.
8. In fact, we confine ourselves to very general assumptions detailed in Plass et al. (2017).
9. The mapping relating $\hat{\pi}^{[g]}$ to $\hat{p}^{[g]}_{\mathfrak{y}}$ is generally not injective.
10. Here we consider a different R than in Plass et al. (2015).
11. Conditioning on subgroup g generalizes the typical MAR assumption.

gMAR is given by $\mathcal{R} = [\max(0,\ 1-\tau),\ 1+\tau], \tau \geq 0$, where the degree of cautiousness is given by the definition of the neighborhood τ (cf. Plass et al., 2017).

4.2 Cautious SAE: Including no Missingness Assumptions

In case of considering $\mathcal{R} = \mathbb{R}_0^+$, i.e. incorporating no assumption on the missingness, the result of the cautious likelihood approach (Plass et al., 2015, p. 251) can be shown to correspond to the one obtained from cautious data completion, plugging in all potential precise sample outcomes compatible with the observations (cf. Augustin et al., 2014, §7.8). Thus, here the lower and upper bound of the synthetic estimator in (1) can be calculated in this case by considering the extreme cases of regarding all missing values as $y_{ij} = 0$, $\forall j \in s_{i,mis},\ i = 1,\dots,M$, or all as $y_{ij} = 1$, $\forall j \in s_{i,mis},\ i = 1,\dots,M$:

$$\underline{\hat{\pi}}_{i,SYN} = \frac{1}{N}\sum_{i=1}^{M}\sum_{j\in s_{i,obs}} w_{ij}y_{ij}\ , \qquad \overline{\hat{\pi}}_{i,SYN} = \frac{1}{N}\sum_{i=1}^{M}\Big(\sum_{j\in s_{i,obs}} w_{ij}y_{ij} + \sum_{j\in s_{i,mis}} w_{ij}\Big)\ . \tag{5}$$

In order to study the bounds $\underline{\hat{\pi}}_{i,LGREG}$ and $\overline{\hat{\pi}}_{i,LGREG}$, it turns out to be beneficial to break the summation over all areas into a term for area i^* [12] of interest and a summation over all other areas $i \neq i^*$. With the regularity condition that sampling weights within area i are equal such that $w_{ij} = w_i, \forall j = 1,\dots,n_i$, and defining $n^{[g]}$ and $n_i^{[g]}$ to be respectively the number of units in s and s_i existing in subgroup g, $g = 1,\dots,v$, $i = 1,\dots,M$, we can rewrite $\hat{\pi}_{i^*,LGREG}$ in (3) as

$$\sum_{g=1}^{v}\left(\Big(\sum_{\substack{i=1\\ i\neq i^*}}^{M}\sum_{j\in s_i^{[g]}}\frac{y_{ij}}{n^{[g]}}\Big)\Big(X_{i^*}^{[g]} - n_{i^*}^{[g]}w_{i^*}\Big) + \sum_{j\in s_{i^*}^{[g]}}\frac{y_{i^*j}}{n^{[g]}}\Big(X_{i^*}^{[g]} - w_{i^*}(n_{i^*}^{[g]} + n^{[g]})\Big)\right)/N_{i^*}\ , \tag{6}$$

$$\text{with}\ \sum_{j\in s_i^{[g]}}\frac{y_{ij}}{n^{[g]}} = \sum_{j\in s_{i,obs}^{[g]}}\frac{y_{ij}}{n^{[g]}} + \sum_{j\in s_{i,mis}^{[g]}}\frac{y_{ij}}{n^{[g]}} \quad \text{and} \quad \sum_{j\in s_{i^*}^{[g]}}\frac{y_{i^*j}}{n^{[g]}} = \sum_{j\in s_{i^*,obs}^{[g]}}\frac{y_{i^*j}}{n^{[g]}} + \sum_{j\in s_{i^*,mis}^{[g]}}\frac{y_{i^*j}}{n^{[g]}}\ ,$$

when missing data are included. The problem consists of finding the values of y_{ij} for the nonrespondents that minimize (maximize) Equation (6). Since Equation (6) is a sum of subgroup specific quantities, optimization for each subgroup g, $g = 1,\dots,v$, separately is sufficient. Provided that $X_{i^*}^{[g]} \geq n_{i^*}^{[g]}w_{i^*}$, we can directly infer that the term referring to the areas $i \neq i^*$ is minimized (maximized) if all the y_{ij}'s, $j \in s_{i,mis}$ are equal to zero (one). Otherwise, the other extreme allocation of zeros and ones is chosen to obtain the minimum (maximum). Analogous considerations can be accomplished in the term associated with area i^*, now based on the condition $X_{i^*}^{[g]} \geq w_{i^*}(n_{i^*}^{[g]} + n^{[g]})$.

4.3 Cautious SAE: First Attempts to Include (Partial) Missingness Assumptions

When partial assumptions in the sense of $R \in [\underline{R}, \overline{R}]$ are tenable, it is useful to express the cautious synthetic estimator and the LGREG-synthetic estimator in terms of $\hat{\pi}^{\mathcal{R}}$, $\hat{q}_{na|0}^{\mathcal{R}}$ and $\hat{q}_{na|1}^{\mathcal{R}}$ obtained by optimizing a log-likelihood as given in (4) under the constraints expressed by R. By again splitting

12. Whenever a differentiation between quantities summing up over all regions and quantities referring to a specific region is needed, we explicitly write i^* for the region under consideration.

$j \in s_i$ into $j \in s_{i,obs}$ and $j \in s_{i,mis}$, the lower bound for the synthetic estimator is received as[13]

$$\underline{\hat{\pi}}^{\mathcal{R}}_{SYN} = \frac{1}{N}\sum_{i=1}^{M}\Big(\sum_{j\in s_{i,obs}} w_{ij}y_{ij} + \underline{\hat{q}}^{\mathcal{R}}_{na|i1}\cdot\underline{\hat{\pi}}^{\mathcal{R}}_{i}\cdot\sum_{j\in s_i} w_{ij}\Big)\,, \tag{7}$$

where $\underline{\hat{q}}^{\mathcal{R}}_{na|i1}\cdot\underline{\hat{\pi}}^{\mathcal{R}}_{i}\cdot\sum_{j\in s_i} w_{ij}$ is the – here smallest – estimated weighted number of nonrespondents with $y_{ij} = 1$, $j \in s_{i,mis}$, under the missingness assumption in focus. Thereby, the included estimators are received by refraining from a subgroup specific consideration, thus regarding $\ell(\pi^{\mathcal{R}},\ q^{\mathcal{R}}_{na|0},\ q^{\mathcal{R}}_{na|1})$ instead of $\ell(\pi^{[g],\mathcal{R}},\ q^{[g],\mathcal{R}}_{na|0},\ q^{[g],\mathcal{R}}_{na|1})$ (cf. (4)). Analogously, $\overline{\hat{\pi}}^{\mathcal{R}}_{SYN}$ is achieved by using $\overline{\hat{q}}^{\mathcal{R}}_{na|i1}$ and $\overline{\hat{\pi}}^{\mathcal{R}}_{i}$ within (7).

To derive the cautious LGREG-synthetic estimator described by $\underline{\hat{\pi}}^{\mathcal{R}}_{i^*,LGREG}$ and $\overline{\hat{\pi}}^{\mathcal{R}}_{i^*,LGREG}$, we base our presentation on the lower bound, while the upper bound is obtained analogously vice versa. Basically, there are two ways to generalize the LGREG-synthetic estimator to a cautious version: One could either consider one overall likelihood or make consistent use of the fact that the LGREG-synthetic estimator is a combination of two estimators, a global one motivated by the idea of "borrowing strength" and another one referring to area i^*. Here, we address the second possibility, while the first one should be studied in further research. For this purpose, we start by maximizing two log-likelihoods, namely $\ell(\pi^{[g],\mathcal{R}},\ q^{[g],\mathcal{R}}_{na|0},\ q^{[g],\mathcal{R}}_{na|1})$ and $\ell(\pi^{[g],\mathcal{R}}_{i^*},\ q^{[g],\mathcal{R}}_{na|i^*0},\ q^{[g],\mathcal{R}}_{na|i^*1})$, under $\underline{R}$ and $\overline{R}$ to derive the respective projections of the generally set-valued estimators. In a next step, we then approach the calculation of $\underline{\hat{\pi}}^{\mathcal{R}}_{i^*,LGREG}$ by including those estimators that minimize

$$\sum_{g=1}^{v}\Bigg(\overbrace{\sum_{j\in s^{[g]}_{i^*,obs}} w_{i^*}y_{i^*j} + \hat{q}^{[g],\mathcal{R}}_{na|i^*1}\hat{\pi}^{[g],\mathcal{R}}_{i^*}\cdot\sum_{j\in s^{[g]}_{i^*}} w_{i^*j}}^{\text{HT-part}} + \overbrace{\hat{\pi}^{[g],\mathcal{R}}(X^{[g]}_{i^*} - n^{[g]}_{i^*}w_{i^*})}^{\text{correction term}}\Bigg)/N_{i^*}\,, \tag{8}$$

which is a version of the classical LGREG-synthetic estimator in Equation (3), where the HT-part is represented in terms of $\hat{\pi}^{[g],\mathcal{R}}_{i^*}$ and $\hat{q}^{[g],\mathcal{R}}_{na|i^*1}$, guaranteeing for the partial assumptions under consideration. Due to the distinct estimation of $\pi^{[g]}$ and $\pi^{[g]}_{i^*}$, we now try to take the associated dependence into account: The interrelation between both estimators may be clearly inferred by considering the representations

$$\hat{\pi}^{[g]}_{i^*} = \Big(\sum_{j\in s^{[g]}_{i^*}} y_{ij}\Big)/n^{[g]}_{i^*} \quad\text{and}\quad \hat{\pi}^{[g]} = \Big(\sum_{\substack{i=1\\ i\neq i^*}}^{M}\sum_{j\in s^{[g]}_{i}} y_{ij} + \sum_{j\in s^{[g]}_{i^*}} y_{ij}\Big)/n^{[g]} \tag{9}$$

(here for ease of representation given without splitting into $s_{i,obs}$ and $s_{i,mis}$), both including respondents from area i^*.[14] Whenever $X^{[g]}_{i^*} > n^{[g]}_{i*}$, we achieve $\underline{\hat{\pi}}^{\mathcal{R}}_{i^*,LGREG}$ if $\underline{\hat{\pi}}^{[g],\mathcal{R}}_{i^*}$, $\underline{\hat{q}}^{[g],\mathcal{R}}_{na|i^*1}$, $\underline{\hat{\pi}}^{[g],\mathcal{R}}$ are taken in (8). This choice is possible in this case, since individuals $j \in s^{[g]}_{i^*}$ are assumed to have the

13. For more details see the preliminary version of a technical report available at http://jplass.userweb.mwn.de/forschung.html.

14. While in (6) a splitting into terms for area i^* and areas $i \neq i^*$ was achieved, this cannot be accomplished here. Note that $\sum_{\substack{i=1\\ i\neq i^*}}^{M}\sum_{j\in s^{[g]}_i}\frac{y_{ij}}{n^{[g]}}$ and $\sum_{j\in s^{[g]}_{i^*}}\frac{y_{i^*j}}{n^{[g]}}$, appearing in Equation (6), are different from (9) and cannot be regarded as estimated probabilities due to the different reference in numerator and denominator.

same values within both estimated probabilities in (9). Considering the situation of $X_{i^*}^{[g]} < n_{i*}^{[g]}$, this is not the case. While $\hat{\pi}^{[g],\mathcal{R}}$ is supposed to be maximal, $\hat{\pi}_{i^*}^{[g],\mathcal{R}}$ and $\hat{q}_{na|i^*1}^{[g],\mathcal{R}}$ should be minimal to minimize (8). To proceed, we give a reasonable way out of this situation. Thereby, we distinguish between the case (i), where the correction term in (8) is of greater importance compared to the HT-part and case (ii), considering the opposite situation.

Case (i): The lower bound of the LGREG-synthetic estimator should be obtained by selecting $\overline{\hat{\pi}}^{[g],\mathcal{R}}$. In this way, for all individuals $j \in s_{i^*}^{[g]}$ the lowest possible scenario compatible with the partial knowledge is assumed, such that the inclusion of $\overline{\hat{\pi}}^{[g],\mathcal{R}}$ and $\overline{\hat{q}}_{na|i^*1}^{[g],\mathcal{R}}$ directly follows. This is supported by Equation (6), indicating that bounds of $\hat{\pi}_{i^*}^{[g],\mathcal{R}}$ are included instead of estimators referring to a scenario between.[15]

Case (ii): $\underline{\hat{\pi}}^{[g],\mathcal{R}}$, $\underline{\hat{\pi}}_{i^*}^{[g],\mathcal{R}}$ and $\underline{\hat{q}}_{na|i^*1}^{[g],\mathcal{R}}$ are incorporated for $\underline{\hat{\pi}}_{i^*,LGREG}^{\mathcal{R}}$, while $\underline{\hat{\pi}}^{[g],\mathcal{R}}$ is improvable by assuming the upper missingness scenario for individuals from $i \neq i^*$. A practical compromise is the inclusion of a pooled estimator

$$\hat{\pi}_{\text{pooled}}^{[g]} = \left(\overline{\hat{\pi}}_{i\neq i^*}^{[g]} \cdot n_{i\neq i^*}^{[g]} + \underline{\hat{\pi}}_{i^*}^{[g]} \cdot n_{i^*}^{[g]}\right)/n^{[g]} \,, \tag{10}$$

to receive $\underline{\hat{\pi}}_{i^*,LGREG}^{\mathcal{R}}$, where $\hat{\pi}_{i\neq i^*}^{[g],\mathcal{R}}$ can also be obtained from the cautious log-likelihood calculated based on all data except from area i^*. Analogously, a pooled version can be determined for the calculation of $\overline{\hat{\pi}}_{i^*,LGREG}^{\mathcal{R}}$.

Because of the under-/overweighting of certain subgroups in the sample, automatically some $(X_{i^*}^{[g]} - n_{i^*}^{[g]} w_{i^*})$ will be positive and others negative, such that the distinction of different cases can not be avoided. The development of a criterion evaluating the "importance" of the HT-term and the correction term used in our argument should be part of further research. Thereby, also the results and conditions from Section 4.2 should be taken into consideration. Up to then, we choose the minimum of the results from case (i) and (ii) to obtain a suggestion for $\underline{\hat{\pi}}_{i^*,LGREG}^{\mathcal{R}}$.

5. Results from the Application Example

The area-specific poverty rate is the focus of our illustration explained in Section 2. Yet, we explicitly avoid making conclusions on the poverty in a substance matter sense, considering this application as a first illustration of technical aspects of the elaborated cautious estimators only. Here, additionally to the case without assuming anything about the missingness process, we studied the weak assumption that rich respondents tend to refuse the income question more often compared to poor ones, i.e. $R \in [0,\ 1]$ (assum. 1), as well as a cautious version of MAR, here incorporating $R \in [0.3,\ 1.7]$ (assum. 2). Although subgroup specific assumptions were feasible in the context of the LGREG-synthetic estimator, we here impose the same missingness assumption on all subgroups.

By applying Equations (7) and (8) to the (weighted) marginal sample data,[16] we can calculate the cautious synthetic estimator and the LGREG-synthetic estimator for the different situations of

15. From Equation (6) we could conclude that either all or no virtual values $y_{ij}, j \in s_{i^*,mis}$, have to be equal to 1 to obtain $\underline{\hat{\pi}}_{i^*,LGREG}$ and $\overline{\hat{\pi}}_{i^*,LGREG}$ in the case of no assumptions. If partial assumptions are included, this applies in the sense that this does not have to be satisfied for all, but for the minimum/maximum number of virtual values that is consistent with the partial missing assumption ending up with $\underline{\hat{\pi}}_{i^*}^{[g],\mathcal{R}}$ or $\overline{\hat{\pi}}_{i^*}^{[g],\mathcal{R}}$.
16. In the GGSS, respondents from East-Germany are oversampled, such that weights are required in the analysis (0.564 (East Germany), 1.205 (West Germany), cf. Koch et al. (1994)).

	no assum.	assum. 1	assum. 2
$[\underline{\hat{\pi}}_{SYN}, \overline{\hat{\pi}}_{SYN}]$	$[0.167,\ 0.300]$	$[0.167,\ 0.193]$	$[0.175,\ 0.208]$

Table 1: Bounds for the synthetic estimator under various missingness assumptions

	no assum.		assum. 1		assum. 2	
Federal state	$\underline{\hat{\pi}}_{i,LGREG}$	$\overline{\hat{\pi}}_{i,LGREG}$	$\underline{\hat{\pi}}_{i,LGREG}$	$\overline{\hat{\pi}}_{i,LGREG}$	$\underline{\hat{\pi}}_{i,LGREG}$	$\overline{\hat{\pi}}_{i,LGREG}$
BW	0.129	0.366	0.129	0.210	0.141	0.224
BY	0.088	0.233	0.088	0.133	0.091	0.141
HB	0.077	0.405	0.115	0.193	0.125	0.206
HH	0.009	0.196	0.014	0.075	0.019	0.083

Table 2: Bounds for the LGREG-synthetic estimator under various missingness assumptions

partial knowledge (cf. Table 1 and Table 2, respectively). The practically weak assumptions already induce a remarkable refinement of the intervals obtained under no assumptions.[17]. Due to the separate likelihood optimization that in some cases led us to the pooled version, including different bounds for i^* and $i \neq i^*$, the lower bound from "no assum." and "assum. 1" do not necessarily have to coincide here. This gives rise to an overall likelihood approach that admittedly refrains from "borrowing strength" within the missingness process, but implicitely accounts for interrelations.

6. First Studies Towards a Cautious Model-based Estimator under Nonresponse

Until now, we focused on models dealing with the small sample size by incorporating observations from other areas on the one hand and area-specific auxiliary information on the other hand. To account for between-area variation beyond that explained by auxiliary variables, model-based estimators relying on mixed models establish a basis. Model-based estimators incorporate data from different areas through a model that depends on the level of aggregation of the auxiliary variables. The well known Fay-Herriot (FH) area-level model, introduced by Fay III and Herriot (1979) for linear regression, has been further developed for categorical regression by MacGibbon and Tomberlin (1989). By relying on the logistic mixed model, they include area specific random effects $u_i \overset{iid}{\sim} N(0, \sigma_u^2)$ into the linear predictor of a standard logistic regression model. Based on this model, we can make predictions contributing to the final model-based estimators.

Since we aim at applying the cautious likelihood approach, we consider the likelihood in the mixed model context first. Generally, the marginal likelihood of the i-th area is received by averaging over the probability distribution of the random effects u_i (cf., e.g., Booth and Hobert, 1999). Since thereby almost always analytically intractable integrals are involved, numerical methods are required for the maximization. Consequently, the cautious likelihood approach is stretched to the limits of its direct applicability if model-based estimators are of interest.

Nevertheless, we proceed with some studies to get a first impression about the predictions obtained from a mixed model if refrained from strong assumptions on the missingness process. Since

17. We use the official abbreviations of the federal states, here BW and BY for Baden-Wuerttemberg and Bavaria, and HB and HH for the federal city states (hanse town (H)) Bremen and Hamburg.

the random effects u_i and the regression coefficients are estimated simultaneously with the aid of approximation methods, we can no longer establish a direct connection between the subgroup specific probabilities and the regression coefficients, as we did in Section 4. Hence, we here start with a first sensitivity analysis, estimating $\beta_0, \ldots, \beta_k$ and u_i under different types of missingness mechanisms. Since for a part of our research question, i.e. getting a first impression about the bounds of the estimated random effects, an area-specific missingness behaviour is of high interest, we simplify the databases classifying the federal states into four regions ("northeast",..., "southwest"), thus substantially reducing the scenarios that have to be considered within a corresponding missing type. Moreover restricting to the covariate "Abitur" (yes/no), we investigate the impact of two different missing types over a grid of values: The first missing type requires independence of the covariates, whereas the second type depends on the covariate and the area.

While the estimated random effects tend to show no systematic reaction to different missingness scenarios, the regression estimates[18] attain the bounds in the extreme missingness situations. Consequently, by focusing on the scenarios that either regard all or no missing values as $y_{ij} = 1$, we apparently can at least give an estimator based on the best-worst-case estimation of the regression coefficients, here denoted by $\hat{\pi}^{\beta} \in [\underline{\hat{\pi}}^{\beta}, \overline{\hat{\pi}}^{\beta}]$. For this purpose, we use $\hat{\beta}_0, \ldots, \hat{\beta}_k, \hat{u}_i$ obtained for the extreme cases to determine the individual prediction bounds. Again, in our categorical case it turns out to be sufficient to calculate the bounds of $\hat{\pi}^{[g],\beta}$, now not only split by the values of the covariate, but also the region. Using $\hat{\pi}^{[g],\beta}$ and the area-specific totals $X_i^{[g]}$, the bounds of a model-based estimator, relying on the best-worst estimation of β, can be calculated.

7. Conclusion

By exploiting the cautious likelihood approach (cf. Plass et al., 2015), we considered an opportunity to adapt the LGREG-synthetic estimator for nonresponse, without the need of strict and often practically untenable assumptions about the missingness process. The included observation model is a powerful medium to make use of frequently available, partial assumptions about the missingness, where results from the application example corroborated that very weak assumptions may already suffice to substantially refine the results obtained without the inclusion of any missingness assumptions. Further research should be devoted to a more extensive consideration of the here proposed method characterized by separate likelihood optimizations. Although some first investigations of cautious model-based estimators were accomplished, due to the technically different situation, a more detailed study should be part of future research. In addition, comparing the magnitude of both principally differing sources of uncertainty induced by the problems in focus (i.e. sampling uncertainty as well as lack of knowledge associated to SAE and nonresponse, respectively) is notably worthwhile. For this purpose, uncertainty regions (cf. Vansteelandt et al., 2006), covering both types of uncertainties, should be investigated.

Acknowledgments

The first author thanks the LMUMentoring program, providing financial support for young, female researchers. The second author thanks the government of Egypt and the German Academic

18. cf. figure in the prelim. version of a technical report mentioned in footnote 9.

Exchange Service (DAAD) for their joint financial support. We are very grateful for the helpful remarks of all three anonymous reviewers and especially appreciate the constructive suggestions of one rather critical reviewer, improving the presentation.

References

T. Augustin, G. Walter, and F. Coolen. Statistical inference. In T. Augustin, F. Coolen, G. de Cooman, and M. Troffaes, editors, *Introduction to Imprecise Probabilities*, pages 135–189. Wiley, Chichester, 2014.

P. Biemer. Total survey error: Design, implementation, and evaluation. *Public Opin. Q.*, 74:817–848, 2010.

J. Booth and J. Hobert. Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *J. R. Stat. Soc. Series B Stat. Methodol.*, 61:265–285, 1999.

M. Cattaneo and A. Wiencierz. Likelihood-based imprecise regression. *Int. J. Approx. Reason.*, 53: 1137–1154, 2012. [based on an ISIPTA '11 paper].

I. Couso and D. Dubois. Statistical reasoning with set-valued information: Ontic vs. epistemic views. *Int. J. Approx. Reason.*, 55:1502–1518, 2014.

I. Couso and L. Sánchez. Machine learning models, epistemic set-valued data and generalized loss functions: An encompassing approach. *Inf. Sci. (Ny)*, 358:129–150, 2016.

T. Denœux. Likelihood-based belief function: Justification and some extensions to low-quality data. *Int. J. Approx. Reason.*, 55:1535–1547, 2014.

DESTATIS, Statistisches Bundesamt. Micro-census 2014 – DESTATIS: Results: Federal states, year, sex, general school education, 2016a. https://www.genesis.destatis.de [accessed: 04.02.2017].

DESTATIS, Statistisches Bundesamt. EU-SILC 2014 – DESTATIS: Living conditions, risk of poverty, 2016b. https://www.destatis.de [accessed: 04.02.2017].

R. Fay III and R. Herriot. Estimates of income for small places: an application of James-Stein procedures to census data. *J. Am. Stat. Assoc.*, 74:269–277, 1979.

GESIS Leibniz Institute for the Social Sciences. German General Social Survey – ALLBUS 2014. GESIS Data Archive, Cologne, 2016. ZA5242 Data file Version 1.0.0.

D. Heitjan and D. Rubin. Ignorability and coarse data. *Ann. Stat.*, 19:2244–2253, 1991.

D. Horvitz and D. Thompson. A generalization of sampling without replacement from a finite universe. *J. Am. Stat. Assoc.*, 47(260):663–685, 1952.

E. Hüllermeier. Learning from imprecise and fuzzy observations: Data disambiguation through generalized loss minimization. *Int. J. Approx. Reason.*, 55:1519–1534, 2014.

A. Koch, S. Gabler, and M. Braun. Konzeption und Durchführung der "Allgemeinen Bevölkerungsumfrage der Sozialwissenschaften" (ALLBUS) 1994. *ZUMA-Arbeitsbericht*, 94, 1994.

R. Lehtonen and A. Veijanen. Logistic generalized regression estimators. *Surv. Methodol.*, 24: 51–56, 1998.

R. Little and D. Rubin. *Statistical Analysis with Missing Data*. Wiley, 2nd edition, 2014.

B. MacGibbon and T. Tomberlin. Small area estimation of proportions via empirical Bayes techniques. *Surv. Methodol.*, 15:237–252, 1989.

C. Manski. *Partial Identification of Probability Distributions*. Springer, 2003.

C. Manski. Credible interval estimates for official statistics with survey nonresponse. *J. Econometrics*, 191:293–301, 2015.

R. Münnich, J. Burgard, and M. Vogt. Small Area-Statistik: Methoden und Anwendungen. *AStA Wirtschafts-und Sozialstatistisches Archiv*, 6:149–191, 2013.

E. Nordheim. Inference from nonrandomly missing categorical data: An example from a genetic study on Turner's syndrome. *J. Am. Stat. Assoc.*, 79:772–780, 1984.

J. Plass, T. Augustin, M. Cattaneo, and G. Schollmeyer. Statistical modelling under epistemic data imprecision: Some results on estimating multinomial distributions and logistic regression for coarse categorical data. In T. Augustin, S. Doria, E. Miranda, and E. Quaeghebeur, editors, *Proc ISIPTA '15*, pages 247–256. SIPTA, 2015.

J. Plass, T. Augustin, M. Cattaneo, G. Schollmeyer, and C. Heumann. Reliable categorical regression analysis for non-randomly coarsened data. Preliminary version of a technical report available at http://jplass.userweb.mwn.de/forschung.html, 2017.

J. Rao and I. Molina. *Small Area Estimation*. Wiley, 2nd edition, 2015.

C. Särdnal, B. Swensson, and J. Wretman. *Model Assisted Survey Sampling*. Springer, 1992.

G. Schollmeyer and T. Augustin. Statistical modeling under partial identification: Distinguishing three types of identification regions in regression analysis with interval data. *Int. J. Approx. Reason.*, 56:224–248, 2015. [based on an ISIPTA '13 paper].

L. Utkin and F. Coolen. Interval-valued regression and classification models in the framework of machine learning. In F. Coolen, G. de Cooman, T. Fetz, and M. Oberguggenberger, editors, *Proc ISIPTA '11*, pages 371–380. SIPTA, 2011.

S. Vansteelandt, E. Goetghebeur, M. Kenward, and G. Molenberghs. Ignorance and uncertainty regions as inferential tools in a sensitivity analysis. *Stat. Sin.*, 16:953–979, 2006.