LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

INSTITUT FÜR STATISTIK

Julia Plass, Aziz Omar, Thomas Augustin

# *Draft:* Towards a Cautious Modelling of Missing Data in Small Area Estimation

Technical Report Number XXXX, 000
Department of Statistics
University of Munich

http://www.stat.uni-muenchen.de

# 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)*
*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 behavior in an adequate way. While we introduce a cautious version of the so-called LGREG-synthetic estimator in the context of model-based estimators, we elaborate why the approach above does not directly extend to model-based estimators and proceed with some first studies investigating the obtained proportions under 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 (SAE); 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 errors and non-sampling errors (cf., e.g. Biemer, 2010). Sampling errors occur, 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. The 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 encountering 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 assumption that the missingness is occurring randomly (MAR), i.e. independently of the true underlying value of the variable of interest. Since the MAR assumption is generally not testable and wrongly imposing

this assumption 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 is not damned to insist on strong assumptions to obtain a result at all, but the allowance for partially identified parameters enables to incorporate tenable knowledge only. In this way, one receives credible, but imprecise results, which can be refined if additional knowledge about the missingness were available. Against this background, 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 generally may not conjure information about the missingness process (unless model assumptions are used, cf. e.g. Couso and Sánchez, 2016; Hüllermeier, 2014), uncertainty due to nonresponse has to be interpreted as lack of knowledge, such that these 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 here considered setting, before we give a basic overview about prominent SAE estimators applicable in our situation in Section 3. Relying on the cautious likelihood approach developed in Plass et al. (2015), the classical LGREG-estimator from SAE is studied under nonresponse in Section 4. The results are illustrated by means of the German General Social Survey, introduced in Section 5, where auxiliary information in terms of totals is inferred from a data report by the German Federal Statistical Office. Further prominent estimators of SAE, which can not be directly modified by the cautious likelihood approach in an analogous way, are discussed in Section 6, where some first studies are completed. Section 7 concludes by summarizing the major points and giving some remarks on further research.

## 2. Setting

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 $j$, $j = 1, \ldots, N_i$ and $N_i$ is 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 $X_i^{[g]}$, $g = 1, \ldots, v$. To infer about $\pi_i$, the probability of a certain category of $Y$ in area $i$, a sample $s \in S$ 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 $s_{i,mis}^{[g]}$.

## 3. Theoretical Background

SAE techniques result in producing estimators $\hat{\pi}_i$ for area of interest $i$, $l = 1, \ldots, M$, that are either design-based or model-based.[1] Here we aim at studying prominent estimators of both types of estimators.

Design-based estimators are either direct estimators that only use data from the targeted area, or indirect estimators that use data from other areas as well. The inclusion of data from areas other that the specified one is justified under the assumption of similarity between the areas, an assumption made to *borrow strength* from other areas. The so-called *synthetic estimator* is a design-based indirect estimator that estimates the area specific probability $\pi_i$ as

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

merging all subsamples together and including sample information about the response variable only. Since there is no differentiation between areas, it merely serves as a basis for further estimators.

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), due to the nature of the binary response variable of interest. Applying the LGREG-synthetic estimator is split in two steps: Firstly, the regression coefficients $\beta_0$, $\beta_1$,. .., $\beta_k$ are estimated by means of a logistic regression model

$$\pi_{ij} = \frac{\exp(\beta_0 + \beta_1 x_{1ij} + \ldots \beta_k x_{kij} + I)}{1 + \exp(\beta_0 + \beta_1 x_{1ij} + \ldots \beta_k x_{kij} + I)} \ , \tag{2}$$

linking $\pi_{ij}$, 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, where $I$ represents the term of potential interactions (see later). In this way, we are borrowing strength, receiving global regression coefficients. Frequently in the area of official statistics, weights are used to correct a deliberate over/ under-representation of certain respondents with specific characteristics. In these cases, weighted logistic regression should be chosen. Area-specific information is only used in a second step: In our setting considering categorical covariates exclusively, the original LGREG-estimator (cf. e.g. Lehtonen and Veijanen, 1998, p.52) for a certain area of interest $i$ can be expressed as

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

where $\hat{\pi}^{[g]}$ is the probability of interest in subgroup $g$, $g = 1, \ldots, v$. The subgroup specific representation in (3) will turn out to be beneficial in context of developing a cautious version (cf. Section 4). Due to the strict monotonicity of the response function in (2), in our case of categorical covariates

1. While properties of design-based estimators (e.g. bias and variance) are evaluated under sampling distribution over all samples in $S$ 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).

the unique relation between $\hat{\beta}_0, \hat{\beta}_1, \ldots \hat{\beta}_k$ $(I)$ and $\hat{\pi}^{[g]}$ can be exploited (as e.g. explained in Plass et al., 2017). Consequently, we can drop the first step, where the regression coefficients themselves have to be estimated and directly calculate $\hat{\pi}^{[g]}$ by involving all samples , hence borrowing strength. Proceeding this way, generally refers to the choice of a model explicitly accounting for all interactions. This is desirable, since only then the full information about the subgroup specific information, also provided by the auxiliary information in terms of totals, is used.

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 an area specific random effects $u_i \sim^{iid} N(0, \sigma_u^2)$ into the linear predictor in (2). In this way, the between area variation not explained by auxiliary data is represented, while simultaneously borrowing strength expressed by the global regression coefficients. Based on this model, we can make predictions contributing to the final model-based estimators.

## 4. Cautious Version of Design-based Estimators in Case of Nonresponse

Since the already established ways of dealing with nonresponse in SAE require strong assumptions, we aim at adapting the presented prominent estimators by striving for a proper reflection of the available information on the missingness process. For this purpose, we make use of the cautious approach developed for the more general case of coarse[2] categorical data in Plass et al. (2015) and further extended in Plass et al. (2017). There, 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 $\mathcal{Q}$ is determined by the set of missingness parameters $q_{na|y}^{[g]}$, i.e. the probability associated with refusing the answer ("na"), given a certain subgroup $g$ and the true category[3] $y \in \Omega_Y = \{0, 1\}$ of the response variable $Y$. In the spirit of partial identification, one can start by incorporating no assumptions[4] on $q_{na|y}^{[g]}$, then restricting them successively by certain conceivable conditions. The gMAR assumption[5], requiring $q_{na|1}^{[g]} = q_{na|0}^{[g]}$ and point-identifying the parameter of main interest $\pi^{[g]}$ may be regarded as an extreme case in this context. The opportunity to integrate partial knowledge in terms of comparative statements about the missingness processes' magnitude (also cf. Nordheim, 1984), such as $q_{na|1}^{[g]} \leq q_{na|0}^{[g]}$ turns out to be beneficial in many situations having only a sketchy idea of the missing data situation.

The cautious approach includes this observation model into a classical categorical likelihood problem. Therefor, a connection between the parameter $\pi^{[g]}$ and $p_{\mathfrak{y}}^{[g]}$ is established via the observation model, where $p_{\mathfrak{y}}^{[g]}$ refers to the observed values $\mathfrak{y} \in \Omega_{\mathcal{Y}} := \Omega_Y \cup \{na\}$, thus summarizing the missing values as a category of its own. The invariance of the likelihood allows to rewrite the log-

2. 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).
3. Referring to the framework of analyzing contingency tables, it is natural to drop the reference to individual $j$.
4. In fact, we confine ourselves to very general assumptions detailed in Plass et al. (2017)
5. Conditioning on subgroup $g$ generalizes the typical MAR assumption.

likelihood in terms of $p_{\mathfrak{V}}^{[g]}$, which can be uniquely maximized, in terms of the parameters of interest by relying on the theorem of total probability, receiving

$$\begin{aligned}\ell(\pi^{[g]},\, q_{na|0}^{[g]},\, q_{na|1}^{[g]}) =& n_1^{[g]}\Big(\ln(\pi^{[g]}) + \ln(1-q_{na|1}^{[g]})\Big) + n_0^{[g]}\Big(\ln(1-\pi^{[g]}) + \ln(1-q_{na|0}^{[g]})\Big) \\ &+ n_{na}^{[g]}\Big(\ln(\pi^{[g]}q_{na|1}^{[g]} + (1-\pi^{[g]})q_{na|0}^{[g]}\Big)\,, \end{aligned} \tag{4}$$

where $n_1^{[g]}$, $n_0^{[g]}$ and $n_{na}^{[g]}$ refer to the respective observed cell counts within subgroup $g$. By maximizing the log-likelihood in (4), we determine the generally set-valued[6] estimators. Representing the one-dimensional projection of these estimators by the bounds of intervals, for the case of no assumptions on the missingness process, one obtains[7]

$$\underline{\hat{\pi}}^{[g]} = \frac{n_1^{[g]}}{n_{na}^{[g]} + n_1^{[g]} + n_0^{[g]}}\,,\ \overline{\hat{\pi}}^{[g]} = \frac{n_1^{[g]} + n_{na}^{[g]}}{n_{na}^{[g]} + n_1^{[g]} + n_0^{[g]}}\,,\quad \underline{\hat{q}}_{na|y}^{[g]} = 0\,,\ \overline{\hat{q}}_{na|y}^{[g]} = \frac{n_{na}^{[g]}}{n_{na}^{[g]} + n_y^{[g]}}\,, \tag{5}$$

$y \in \{0,1\}$. By considering $q_{na|1}^{[g]} = Rq_{na|0}^{[g]},\ R \in \mathbb{R}_0^+$, assumptions about the missingness can be incorporated. Specific values about $R$ are associated with a particular missingness scenario, thus point-identifying $\pi^{[g]}$. In this way, partial assumptions, like including $R \in [\underline{R},\ \overline{R}]$ into (4), refine the intervals in (5). While $R \in [0,\ 1]$ corresponds to $q_{na|1}^{[g]} \leq q_{na|0}^{[g]}$, $R \in [1-\tau_1,\ 1+\tau_2],\ \tau_1, \tau_2 \geq 0$, gives us a cautious version of gMAR, where the degree of cautiousness is given by the definition of the neighborhood $\tau_1, \tau_2$ (cf. Plass et al. (2017)). In Section 5, results from partial assumptions of that kind are compared with the ones obtained from imposing no assumptions at all.

We now apply this approach to SAE. By refraining from a subgroup specific analysis and including sampling weights $w_{ij}$ into the result in (5), for the case of including no assumptions we directly receive the lower and upper bound of the synthetic estimator in (1):

$$\underline{\hat{\pi}}_{i,SYN} = \frac{1}{N}\sum_{i=1}^{M}\sum_{j\in s_{i,obs}} w_{ij}y_{ij}\,, \qquad \overline{\hat{\pi}}_{i,SYN} = \frac{1}{N}\sum_{i=1}^{M}\Big(\sum_{j\in s_{i,obs}} y_{ij}w_{ij} + \sum_{j\in s_{i,mis}} w_{ij}\Big)\,. \tag{6}$$

While the upper bound $\overline{\hat{\pi}}_{i,SYN}$ regards all missing values as $y_{ij} = 1,\ \forall j \in s_{i,mis},\ i = 1,\ldots,M$, throughout the lower $\underline{\hat{\pi}}_{i,SYN}$ does not (i.e. $y_{ij} = 0,\ \forall j \in s_{i,mis},\ i = 1,\ldots,M$).

Thus far, we ignored the information from other variables, but now we investigate the LGREG-synthetic estimator, hence considering the setting that sample data and area-specific auxiliary information in terms of known totals of the joint distribution are available. In order to study the bounds $\underline{\hat{\pi}}_{i,LGREG}$ and $\overline{\hat{\pi}}_{i,LGREG}$, as a starting point, it turns out to be beneficial to express $\hat{\pi}^{[g]}$ in terms of $y_{ij}$ and to break the summation over all areas into a term for area $i^*$ [8] of interest and a summation over all other areas $i \neq i^*$. With the regularity condition that sampling weights within area $i$ are

6. The mapping relating $\hat{\pi}^{[g]}$ to $\hat{p}_{\mathfrak{V}}^{[g]}$ is generally not injective.

7. The result for $\pi^{[g]}$, sometimes called best-worst interval, conforms to the one e.g. obtained from cautious data completion (i.e. plugging in all potential precise values consistent with the observation) or sensitivity analysis (cf., e.g. Kenward et al., 2001).

8. 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.

equal such that $w_{ij} = w_i, \forall j = 1, \ldots, n_i$, and defining $n^{[g]}$ and $n_i^{[g]}$ to be respectively the number of units in $s$ and $s_i$ existing in subgroup $g$, $g = 1, \ldots, v$, $i = 1, \ldots, M$, we have

$$\hat{\pi}_{i^*,LGREG} = \sum_{g=1}^{v} \left( \Big( \sum_{i=1}^{M} \sum_{j \in s_i^{[g]}} \frac{y_{ij}}{n^{[g]}} \Big) \Big( X_{i^*}^{[g]} - \sum_{j \in s_{i^*}^{[g]}} w_{i^*j} \Big) + \sum_{j \in s_{i^*}^{[g]}} w_{i^*j} y_{i^*j} \right) / N_{i^*} \tag{7}$$

$$= \sum_{g=1}^{v} \left( \Big( \sum_{\substack{i=1 \\ i \neq i^*}}^{M} \sum_{j \in s_i^{[g]}} \frac{y_{ij}}{n^{[g]}} \Big) \Big( X_{i^*}^{[g]} - n_{i^*}^{[g]} w_{i^*} \Big) + \sum_{j \in s_{i^*}^{[g]}} \frac{y_{i^*j}}{n^{[g]}} \Big( X_{i^*}^{[g]} - w_i^* (n_{i^*}^{[g]} + n^{[g]}) \Big) \right) / N_i^*. \tag{8}$$

All terms including $y_{ij}$ are split for units $j \in s_{i,obs}^{[g]}$ and $j \in s_{i,mis}^{[g]}$ in a next step, where the problem consists of finding the values of $y_{ij}$ for the nonrespondents that minimize (maximize) Equation (8). Since Equation (8) is a sum of subgroup specific quantities, optimization for each subgroup 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 as many (few) as possible – restricted by the included missingness assumption – the $y_{ij}$'s, $j \in s_{i,mis}$ are equal to zero (one). Otherwise, the other extreme allocation of zeros and ones should be chosen to obtain the minimum (maximum). Analogous considerations can be accomplished in the term associated with area $i^*$, now based on the condition $X_{i^*}^{[g]} \geq w_{i^*}(n_{i^*}^{[g]} + n^{[g]})$.

For the situation without imposing missing assumptions, the representation in terms of Equation (8) is sufficient, since here the extreme allocations of nonrespondents to the values of $y_{ij}$, $s_{i,mis}^{[g]}$, are clearly seen. To guarantee compliance with the considered missingness assumption beyond these extreme cases, i.e. when partial assumptions in the sense of $R \in [\underline{R}, \overline{R}]$ are tenable, it is useful to express the LGREG estimator in terms of our estimators from the cautious likelihood approach. Since $\sum_{\substack{i=1 \\ i \neq i^*}}^{M} \sum_{j \in s_i^{[g]}} \frac{y_{ij}}{n^{[g]}}$ and $\sum_{j \in s_{i^*}^{[g]}} \frac{y_{i^*j}}{n^{[g]}}$, appearing in Equation (8) cannot be regarded as estimated probabilities due to the different reference in numerator and denominator, we cannot apply the cautious likelihood approach in its standard form by exploiting the idea of the representation in terms of (8). Instead, we split the sums in Equation (8) for $j \in s_{i^*,obs}^{[g]}$ (and $j \in s_{i,obs}^{[g]}$) as well as $j \in s_{i^*,mis}^{[g]}$ (and $j \in s_{i,mis}^{[g]}$), receiving the bounds

$$\underline{\hat{\pi}}_{i^*,LGREG} \text{ or } \overline{\hat{\pi}}_{i^*,LGREG} = \sum_{g=1}^{v} \Bigg( [\underline{\hat{\pi}}^{[g]} | \overline{\hat{\pi}}^{[g]}] (X_{i^*}^{[g]} - n_{i^*}^{[g]} w_{i^*}) \tag{9}$$

$$+ w_{i^*} \sum_{j \in s_{i^*,obs}^{[g]}} y_{i^*j} + [\underline{\hat{q}}_{na|i^*1}^{[g]} | \overline{\hat{q}}_{na|i^*1}^{[g]}] [\underline{\hat{\pi}}_{i^*}^{[g]} | \overline{\hat{\pi}}_{i^*}^{[g]}] \sum_{j \in s_{i^*,mis}^{[g]}} w_{i^*j} \Bigg) / N_i^* \,,$$

where $\underline{\hat{q}}_{na|i^*1}^{[g]} \overline{\hat{\pi}}^{[g]} \sum_{j \in s_{i^*}^{[g]}} w_{i^*j}$ is the expected number of nonrespondents in $s_{i^*}^{[g]}$ with $y_{ij} = 1$ under the missingness assumption in focus.

To shortly explain this formula, we here turn to the notation of the background of the cautious approach (cf. Plass et al., 2017), where the variable giving the observations in sample space $\Omega_{\mathcal{Y}} = \Omega_Y \cup \{na\}$ is denoted by $\mathcal{Y}$.

$$
\begin{aligned}
& \hat{P}(Y=1|X_g,i,\mathcal{Y}=na)\cdot\textstyle\sum_{j\in s_{i,g,mis}} w_{ij} \\
= \quad & \frac{\hat{P}(Y=1,X_g^+,i,\mathcal{Y}=na)}{\hat{P}(\mathcal{Y}=na,X_g^+,i)}\cdot\textstyle\sum_{j\in s_{i,g}} w_{ij} \\
= \quad & \frac{\hat{P}(\mathcal{Y}=na|Y=1,X_g,i)\cdot\hat{P}(Y=1,X_g^+,i)}{P(\mathcal{Y}=na,X_g^+,i)}\cdot\textstyle\sum_{j\in s_{i,g}} w_{ij} \\
= \quad & \frac{\hat{q}_{na|gi1}\cdot\hat{\pi}_{gi}}{\frac{\sum_{j\in s_{i,g,mis}} w_{ij}}{\sum_{j\in s_{i,g}} w_{ij}}}\cdot\textstyle\sum_{j\in s_{i,g}} w_{ij} \\
= \quad & \hat{q}_{na|gi1}\cdot\hat{\pi}_{gi}\cdot\textstyle\sum_{j\in s_{i,g}} w_{ij}
\end{aligned}
$$

By $[\underline{\hat{\pi}}^{[g]}|\overline{\hat{\pi}}^{[g]}]$, $[\underline{\hat{\pi}}_{i^*}^{[g]}|\overline{\hat{\pi}}_{i^*}^{[g]}]$ and $[\underline{\hat{q}}_{na|i^*1}^{[g]}|\overline{\hat{q}}_{na|i^*1}^{[g]}]$ we denote that the bound has to be chosen minimizing (maximizing) the whole term for subgroup $g$ and for the corresponding units in area $i^*$, respectively, thereby accounting for the fact that $\underline{\hat{\pi}}_{i^*}^{[g]}$ (or $\overline{\hat{\pi}}_{i^*}^{[g]}$) can be combined with $\underline{\hat{q}}_{na|i^*1}^{[g]}$ (or $\overline{\hat{q}}_{na|i^*1}^{[g]}$) only (cf. Plass et al., 2017).

Next, we study whether several constellations have to be excluded: The inclusion of $\hat{\pi}_i^{[g]}$ is in compliance with the conclusions from Equation (8) that extreme points are attained if boundary allocations are regarded for $j \in s_{i^*,mis}$, such that $\underline{\hat{\pi}}_{i^*}^{[g]}$ and $\overline{\hat{\pi}}_{i^*}^{[g]}$ in fact produce the extreme scenarios. Against this, the incorporation of all units into $\hat{\pi}^{[g]}$ is different from the idea of complete separation in Equation (8). Since we are forced to give up the disjoint representation from (7), we have to ensure that units $j \in s_{i^*,mis}^{[g]}$, included into the calculation of $\hat{\pi}^{[g]}$ and $\hat{\pi}_i^{[g]}$, are not supposed to have different values $y_{i^*j}$ in $\hat{\pi}^{[g]}$ and $\hat{\pi}_{i^*}^{[g]}$, which is not guaranteed if all combinations of bounds are admitted, such that the associated constellation of bounds has to be excluded. To avoid the inconsistent allocation of units $j \in s_{i^*,mis}^{[g]}$, we now restrict to incorporating $\underline{\hat{\pi}}^{[g]}$ (or $\overline{\hat{\pi}}^{[g]}$) together with $\underline{\hat{\pi}}_{i^*}^{[g]}$ and $\underline{\hat{q}}_{na|i^*1}^{[g]}$ (or $\overline{\hat{\pi}}_{i^*}^{[g]}$ and $\overline{\hat{q}}_{na|i^*1}^{[g]}$). Since – as stressed earlier – indeed $\underline{\hat{\pi}}_{i^*}^{[g]}$ and $\overline{\hat{\pi}}_{i^*}^{[g]}$ is required to produce a bound for the LGREG, we can conclude that the suggested constellations – and not estimators referring to an allocation between – has to be chosen. Considering (9), we can infer that $(X_{i^*}^{[g]} - n_{i^*}^{[g]} w_{i^*})$ again plays the decisive role in judging whether $\underline{\hat{\pi}}^{[g]}$ or $\overline{\hat{\pi}}^{[g]}$ is needed for the minimum. Thus, the lower bound of the LGREG estimator should be given by plugging $\underline{\hat{\pi}}^{[g]}$, $\underline{\hat{\pi}}_{i^*}^{[g]}$ and $\underline{\hat{q}}_{na|i^*1}^{[g]}$ into (9), provided $(X_{i^*}^{[g]} \geq n_{i^*}^{[g]} w_{i^*})$; for $(X_{i^*}^{[g]} < n_{i^*}^{[g]} w_{i^*})$, constellation $\overline{\hat{\pi}}^{[g]}$, $\overline{\hat{\pi}}_{i^*}^{[g]}$ and $\overline{\hat{q}}_{na|i^*1}^{[g]}$ has to be chosen. The upper bound is determined analogously.

To obtain $\hat{\pi}^{[g]}$, we refer the to likelihood making use of values from all samples, thus one implicitly does not condition on the area within $\hat{q}_{na|1}^{[g]}$ as well. Against this, data from samples of area $i^*$ only is used to determine $\hat{\pi}_{i^*}^{[g]}$ such that also the respective missingness process is assumed to be dependent on this area. One could take the idea of "borrowing strength" seriously and allow for this different handling. Alternatively, if this is perceived as an inconsistency, one could avoid it by generally refraining from this dependence in the estimation of $\hat{\pi}^{[g]}$.

## 5. Application

### 5.1 The German General Social Survey

In order to illustrate the results, we rely on the data of the German General Social Survey (GGSS; German abbreviation: ALLBUS) to estimate the model based on the sample as well as on the (German) Federal Office of Statistics' data report giving the auxiliary data in terms of totals.
From 1980 on, the GGSS is biennially collected, where we here refer to the latest accessible wave of year 2014 (GESIS Leibniz Institute for the Social Sciences, 2016), containing information about $3471$ observations of $861$ variables. Due to the former division of Germany, substantial differences in certain issues are notable between the western and the eastern part of Germany, wherefore comparative studies are worthwhile. To receive meaningful results for social groups in the eastern subsample, respondents from this part are oversampled, such that weights are required in the analysis (weights in GGSS 2014: 0.564 (East Germany), 1.205 (West Germany), cf. Koch et al. (1994)). Additionally, since the GGSS is not a simple random sample, and the exact values of the design-weights $w_{ij}, j = 1, \ldots, n_i, i = 1, \ldots, M$, are not available for reasons of data protection, we used a post-stratification weighting scheme that accounts for states' different population sizes. Therefore, $w_{ij}$ are computed based on population size from each state attributed to state's own sample size in the GGSS, accounting also for the East-West weights.
The considered models aim at explaining the response variable, highly affected by nonresponse, via categorical, completely observed demographic variables. In the application we are interested in the estimated lower and upper bound of the area-specific poverty risk, i.e. $\hat{\underline{\pi}}_i$ and $\hat{\overline{\pi}}_i$, resulting from taking into account all possible missingness assumptions compatible with the observed data. In this way, we construct the binary response variable with values "poor and "rich" by comparing the collected equivalent income measured on the OECD modified scale (V497) with the poverty risk threshold given by $60\%$ of the median net equivalent income, i.e. $986.65$€ for year 2014 DESTATIS, Statistisches Bundesamt (2016). The poverty variable shows $454$ missing values. Comparing the proportion of poor respondents obtained from a complete-case analysis, i.e. $21.0\%$, with the poverty risk published in the DESTATIS data report, i.e. $16.7\%$[9], the result might either indicate a comparative large amount of poor respondents in the GGSS or suggest the assumption that rich respondents tend to refuse to answer. Since in the considered research question one usually is not aware of any comparables, we refrain from over-weighting the rich respondents, only considering the mentioned missingness assumption, since this effect is also well known from former studies (cf., e.g. Tourangeau and Yan, 2007). As covariates we use the highest school leaving certificate (V86), which – for ease of presentation – is dichotomized, thus distinguishing between categories "no Abitur"[10] (coded by 0) and "Abitur" (coded by 1) only, as well as sex (V81, coded with 0: male, 1: female). In the GGSS the most fine-grained information about the region is given by the federal state[11] the respondents live in. In this way, the federal states form our small areas representing a complete partition (i.e. $m = M = 17$) of the overall domain "Germany", where we only observe subsamples from each state. Since our theoretical considerations refer to the situation of precisely observed values of the covariate(s), we exclude five cases from the original data, thus

9. This percentage is calculated based on the EU-SILC survey. Since both surveys, the EU-SILC and the GGSS, rely on the same definition of net equivalized income, both percentages can be compared indeed.
10. The "Abitur" is the general qualification for university entrance in Germany.
11. Although Germany is divided into 16 federal states, the GGSS differentiates between 17 ones, additionally distinguishing between "former East-Berlin" and "former West-Berlin".

| | no assum. | | | assum. 1 | | assum. 2 | | | |
|---|---|---|---|---|---|---|---|---|---|
| sex, Abitur | $\underline{\hat{\pi}}^{[g]}$ | $\overline{\hat{\pi}}^{[g]}$ | $\overline{\hat{q}}^{[g]}_{na\|p}$ | $\overline{\hat{\pi}}^{[g]}$ | $\overline{\hat{q}}^{[g]}_{na\|p}$ | $\underline{\hat{\pi}}^{[g]}$ | $\overline{\hat{\pi}}^{[g]}$ | $\underline{\hat{q}}^{[g]}_{na\|p}$ | $\overline{\hat{q}}^{[g]}_{na\|p}$ |
| 0, 0 | 0.18 | 0.30 | 0.40 | 0.20 | 0.12 | 0.19 | 0.22 | 0.04 | 0.17 |
| 0, 1 | 0.10 | 0.22 | 0.57 | 0.11 | 0.13 | 0.10 | 0.12 | 0.04 | 0.20 |
| 1, 0 | 0.25 | 0.38 | 0.35 | 0.29 | 0.14 | 0.26 | 0.32 | 0.05 | 0.19 |
| 1, 1 | 0.11 | 0.27 | 0.60 | 0.13 | 0.16 | 0.11 | 0.14 | 0.05 | 0.25 |

Table 1: Subgroup specific estimator $\hat{\pi}_{[g]}$ with different assumptions on the missingness behaviour

basing the analysis on the sample with $|s| = 3466$, $|s_{obs}| = 3012$, $|s_{mis}| = 454$. The DESTATIS data report provides the needed domain-specific totals split by the values of the covariate, i.e. $X_i^{[g]}$, $i = 1, \ldots, m$, $g = 1, \ldots, v$.[12]

### 5.2 Results

The area-specific poverty rate is the focus of our illustration. Yet, we explicitly avoid making conclusions on the poverty in a substance matter sense, considering this application example as a first illustration of technical aspects of the elaborated cautious estimators only. If no assumption about the missingness is imposed, we can apply Equation (6) to the (weighted) marginal sample data, thus obtaining the bounds of the cautious synthetic estimate $\underline{\hat{\pi}}_{i,SYN} = 0.17$ and $\overline{\hat{\pi}}_{i,SYN} = 0.30 \; \forall i = 1, \ldots, M$. Since we refrain from making assumptions, the estimated true poverty rate $\hat{\pi}$ is included in the interval characterized by the bounds. If we were additionally concerned with a representative sampling process with regard to "poverty", even the true parameter $\pi$ should be covered. In order to determine the LGREG-synthetic estimator, we consider the model

$$\pi_{ij} = \frac{\exp(\beta_0 + \beta_1 \, x_{1ij} + \beta_2 \, x_{2ij} + \beta_{1:2} x_{1ij} x_{2ij})}{1 + \exp(\beta_0 + \beta_1 \, x_{1ij} + \beta_2 \, x_{2ij} + \beta_{1:2} x_{1ij} x_{2ij})} \; , \tag{10}$$

for the sample data, where $x_{1ij}$ and $x_{2ij}$ are the binary variables referring to variable "Abitur" and "sex", respectively, and $\beta_{1:2}$ is the interaction, modelling the joint effect when both binary variables are equal to one. Table 1 shows the lower and upper bounds for the subgroup specific estimators, also under partial assumptions[13]. Exemplary, we assumed that rich respondents tend to refuse the income question compared to poor ones, i.e. $R \in [0, \; 1]$ (assum. 1; the upper bound representing the MAR case), as well as a cautious version of MAR, here incorporating $R \in [0.3, \; 1.7]$ (assum. 2). Although subgroup specific assumptions were feasible, we here impose the same missingness assumption on all subgroups. These practically weak assumptions already induce a remarkable refinement of the intervals obtained under no assumptions. Including these bounds of $\hat{\pi}^{[g]}$ and the

12. 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.

13. Note that $\underline{\hat{q}}^{[g]}_{na|p} = 0$, $\forall g = 1, \ldots, v$, if the first assumption is considered and that $\underline{\hat{\pi}}^{[g]}$ and $\underline{\hat{q}}^{[g]}_{na|p}$ of the second assumption throughout equal the estimations of the first one.

| | 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 | 0.19 | 0.36 | 0.19 | 0.22 | 0.20 | 0.24 |
| BY | 0.12 | 0.19 | 0.12 | 0.13 | 0.12 | 0.13 |
| HB | 0.16 | 0.35 | 0.16 | 0.19 | 0.17 | 0.19 |
| HH | 0.08 | 0.21 | 0.08 | 0.10 | 0.09 | 0.12 |

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

area-specific totals from the data report into (9), accounting for the preceding considerations made there, the area-specific bounds of the LGREG-synthetic estimator in Table 2 [14] are obtained.

## 6. First Studies Towards a Cautious Model-based Estimator

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, mixed models establish a basis (cf. Section 3). Since we aim at applying the cautious likelihood approach, we consider the likelihood function in the mixed model context first. Generally, the marginal likelihood of the $i$-th area, is received by averaging over the distribution of the random effects $u_i$, that is,

$$L_i(\beta_0, \ldots, \beta_k, \sigma_u^2) = \int p(y_{i1}, \ldots, y_{in_i} | \beta_0, \ldots, \beta_k, u_i) \cdot f(u_i |\, \sigma_u^2) du_i \;, \tag{11}$$

where $p(y_{i1}, \ldots, y_{in_i} | \beta_0, \ldots, \beta_k, u_i)$ is the Bernoulli probability distribution in our case and $f(u_i|\sigma_u^2)$ is the density of the random effect with parameter $\sigma_u^2$ (cf., e.g. Booth and Hobert, 1999). The full marginal likelihood function from all $M$ areas is then determined by $\prod_{i=1}^{M} L_i(\beta_0, \ldots, \beta_k, \sigma_u^2)$. Almost always $L_i(\beta_0, \ldots, \beta_k, \sigma_u^2)$ in (11) involves intractable integrals, wherefore numerical methods are required to maximize the likelihood. Consequently, the cautious likelihood approach that served as the basis for a careful inclusion of auxiliary information on the missingness process in Section 4 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 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 data bases classifying the federal states into four regions ("northeast",..., "southwest"), thus substantially reducing the scenarios that have to be considered within an corresponding missing type. Moreover restricting to covariate "Abitur" (yes/no), we investigate the impact of for two different

14. 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.

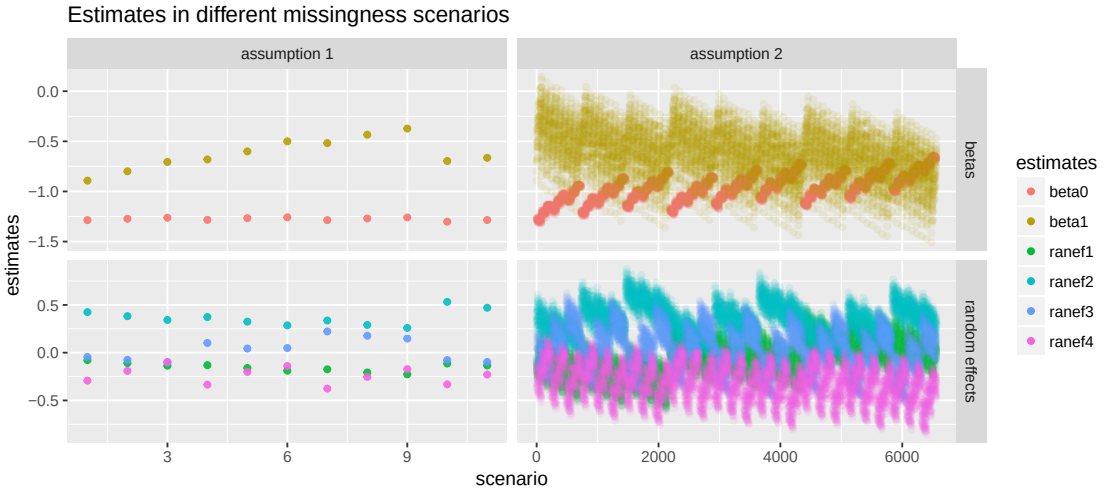


Figure 1: Sensitivity analysis under different types of missingness scenarios

missing types over a grid of values: The first missing type requires independence on the values of the covariates, whereas the second types 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 (cf. Figure 1), 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 $\beta$, can be calculated.

## 7. Conclusion

By exploiting the cautious likelihood approach (cf. Plass et al., 2015), we presented 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 original results. 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. The cautious likelihood approach shows to be promising in

this context, providing an intuitive possibility to adapt classical confidence intervals for partially identified parameters (cf. Plass et al., 2017).

## Acknowledgments

The first author thanks the LMUMentoring program, providing financial support for young, female researchers. The second author is supported by a joint fund from the government of Egypt and the German Academic Exchange Service (DAAD).

## References

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. Reasoning*, 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. Reasoning*, 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. Reasoning*, 55:1535–1547, 2014.

DESTATIS, Statistisches Bundesamt. EU-SILC 2014 - DESTATIS: Living conditions, risk of poverty, 2016. `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, doi:10.4232/1.12437.

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

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

M. Kenward, E. Goetghebeur, and G. Molenberghs. Sensitivity analysis for incomplete categorical data. *Stat. Modelling.*, 1:31–48, 2001.

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. Econom.*, 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. Reasoning*, 56:224–248, 2015.

R. Tourangeau and T. Yan. Sensitive questions in surveys. *Psychol. Bull.*, 133:859–883, 2007.

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.