Structure preserving estimators to update socio-economic indicators in small areas

Inaugural dissertation to obtain the academic degree Dr. rer. pol. at the School of Business and Economics of Freie Universität Berlin

presented by

Alejandra Arias-Salazar

born in San José, Costa Rica

Berlin, 2022

Dean: Prof. Dr. Dr. Giacomo Corneo

First reviewer: Prof. Dr. Timo Schmid (Otto-Friedrich-Universität Bamberg)

Second reviewer: Prof. Dr. Natalia Rojas-Perilla (United Arab Emirates University)

Date of defense: November 14, 2022

Acknowledgements

I would like to express my deep gratitude to my supervisors, Prof. Dr. Timo Schmid (Otto-Friedrich-Universität Bamberg, Germany) and Prof. Dr. Natalia Rojas-Perilla (United Arab Emirates University), for all their support, guidance, and collaborations.

Special thanks go to my family and friends from *here and there* for being my joy and strength throughout this journey.

Publication List

The publications listed below are the result of the research carried out in this thesis titled, ‘Structure Preserving Estimators to Update Socio-Economic Indicators in Small Areas’.

1. Arias-Salazar, A. (2022). **Updating Intercensal Health Indicators for Small Areas using the R Package spree**, *Working paper*, to be submitted.
2. Arias-Salazar, A. (2022). **Small Area Estimates of Poverty Incidence in Costa Rica under a Structure Preserving Estimation (SPREE) Approach**, submitted to Journal of Official Statistics, minor revision.
3. Koebe, T., Arias-Salazar, A., Rojas-Perilla, N., and Schmid, T. (2022). **Intercensal Updating using Structure Preserving Methods and Satellite Imagery**, Journal of the Royal Statistical Society: Series A (Statistics in Society), 1-27.
4. Koebe, T., Arias-Salazar, A., and Schmid, T. (2022). **Releasing Survey Microdata with Exact Cluster Locations and Additional Privacy Safeguards**, submitted to Humanities and Social Sciences Communications, major revision.

Contents

Introduction **9**

I New Developments in Structure Preserving Estimators **11**

1 Updating Intercensal Health Indicators for Small Areas using the R Package spree **13**

1.1 Introduction 13

1.2 Methodology 15

1.2.1 SPREE-type estimators 15

1.2.2 MSE estimation 19

1.3 The R package spree 20

1.4 Case study: Updating health indicators of the MPI for small areas in Costa Rica 21

1.4.1 Data sources 23

1.4.2 Usage of the R package spree 24

1.5 Conclusions and further developments 33

Supplementary material A **35**

A.1 Iterative proportional fitting algorithm 35

2 Small Area Estimates of Poverty Incidence in Costa Rica under a Structure Preserving Estimation (SPREE) Approach **37**

2.1 Introduction 37

2.2 Data description and definition of poverty 40

2.2.1 Population and housing census 2011 40

2.2.2 National household surveys 2011-2017 40

2.2.3 Demographic projections 41

2.2.4 Definition of poverty 41

2.3 Methodology 42

2.3.1 Structure preserving estimation (SPREE) methods 42

2.3.2 The empirical best predictor method 45

2.3.3 Strategy to estimate and update poverty estimates in Costa Rican cantons 46

2.3.4 Uncertainty of the updated estimates 47

2.4 Results of the application 49

2.4.1 Poverty estimates for the census 49

2.4.2 Updated poverty estimates 51

2.5 Design-based simulation 54
2.6 Conclusions and further steps 57

3 Intercensal Updating using Structure Preserving Methods and Satellite Imagery 59
3.1 Introduction 60
3.2 Methodology 62
3.2.1 Structure preserving estimation (SPREE) 62
3.2.2 SPREE with auxiliary information 64
3.2.3 Assessment of uncertainty 65
3.3 Case study: Multidimensional poverty in Senegal 66
3.3.1 Geography of Senegal 67
3.3.2 Multidimensional poverty in Senegal 67
3.3.3 Data sources 69
3.4 Validation 76
3.5 Application 80
3.6 Conclusion 83

II Microdata Dissemination for Small Area Estimation 85

4 Releasing Survey Microdata with Exact Cluster Locations and Additional Privacy Safeguards 87
4.1 Introduction 87
4.2 Results 89
4.2.1 Geomasking to obfuscate true survey locations 92
4.2.2 Copula-based synthetic data generation 93
4.2.3 Risk of re-identifying private geocodes 95
4.2.4 Population uniqueness of survey respondents 96
4.2.5 Utility for survey augmentation 97
4.3 Discussion 99
4.4 Methods 100
4.4.1 Fitting Gaussian copulas to survey attributes 100
4.4.2 Area-level survey augmentation methods 103

Supplementary material B 105
B.1 Data description 105
B.2 Sensitivity of copula vis-à-vis geographic fitting level 106
B.3 Effects of synthetic sample size on prediction outcomes 107
B.4 Choosing marginal distributions & encoding schemes 108
B.5 Detailed analysis of the NBI as composite indicator 109
B.6 Stability of anonymization approaches across simulation runs 111

Bibliography 113

Summaries **123**

Abstracts in English 123

Kurzzusammenfassungen auf Deutsch 124

Introduction

Official statistics are intended to support decision makers by providing reliable information on different population groups, identifying what their needs are and where they are located. This allows, for example, to better guide public policies and focus resources on the population most in need. Statistical information must have some characteristics to be useful for this purpose. This data must be reliable, up-to-date and also disaggregated at different domain levels, e.g., geographically or by sociodemographic groups (Eurostat, 2017).

Statistical data producers (e.g., national statistical offices) face great challenges in delivering statistics with these three characteristics, mainly due to lack of resources. Population censuses collect data on demographic, economic and social aspects of all persons in a country which makes information at all domains of interest available. They quickly become outdated since they are carried out only every 10 years, especially in developing countries. Furthermore, administrative data sources in many countries have not enough quality to produce statistics that are reliable and comparable with other relevant sources. On the contrary, national surveys are conducted more frequently than censuses and offer the possibility of studying more complex topics. Due to their sample sizes, direct estimates are only published based on domains where the estimates reach a specific level of precision. These domains are called planned domains or large areas in this thesis, and the domains in which direct estimates cannot be produced due to lack of sample size or low precision will be called small areas or domains.

Small area estimation (SAE) methods have been proposed as a solution to produce reliable estimates in small domains. These methods allow improving the precision of direct estimates, as well as providing reliable information in domains where the sample size is zero or where direct estimates cannot be obtained by combining data from censuses and surveys (Rao and Molina, 2015). Thereby, the variables obtained from both data sources are assumed to be highly correlated but the census actually may be outdated.

In these cases, structure preservation estimation (SPREE) methods offer a solution when the target indicator is a categorical variable, with at least two categories (for example, the labor market status of an individual can be categorised as: ‘employed’, ‘unemployed’, and ‘out of labor force’). The population counts are arranged in contingency tables: by rows (domains of interest) and columns (the categories of the variable of interest) (Purcell and Kish, 1980). These types of estimators are studied in Part I of this work. In Chapter 1, SPREE methods are applied to produce postcensal population counts for the indicators that make up the ‘health’ dimension of the multidimensional poverty index (MPI) defined by Costa Rica. This case study is also used to illustrate the functionalities of the R **spree** package. It is a user-friendly tool designed to produce updated point and uncertainty estimates based on three different approaches: SPREE

(Purcell and Kish, 1980), generalised SPREE (GSPREE) (Zhang and Chambers, 2004), and multivariate SPREE (MSPREE) (Luna-Hernández, 2016).

SPREE-type estimators help to update population counts by *preserving* the census structure and relying on new and updated totals that are usually provided by recent survey data. However, two scenarios can jeopardise the use of standard SPREE methods: a) the indicator of interest is not available in the census data e.g., income or expenditure information to estimate monetary-based poverty indicators, and b) the total margins are not reliable, for instance, when changes in the population distribution between areas are not captured correctly by the surveys or when some domains are not selected in the sample. Chapters 2 and 3 offer a solution for these cases, respectively.

Chapter 2 presents a two-step procedure that allows obtaining reliable and updated estimates for small areas when the variable of interest is not available in the census. The first step is to obtain the population counts for the census year using a well-known small-area estimation approach: the empirical best prediction (EBP) (Molina and Rao, 2010) method. Then, the result of this procedure is used as input to proceed with the update for postcensal years by implementing the MSPREE (Luna-Hernández, 2016) method. This methodology is applied to the case of local areas in Costa Rica, where incidence of poverty (based on income) is estimated and updated for postcensal years (2012-2017). Chapter 3 deals with the second scenario where the population totals in local areas provided by the survey data are strengthened by including satellite imagery as an auxiliary source. These new margins are used as input in the SPREE procedure. In the case study in this paper, annual updates of the MPI for female-headed households in Senegal are produced.

While the use of satellite imagery and other big data sources can improve the reliability of small-area estimates, access to survey data that can be matched with these novel sources is restricted for confidentiality reasons. Therefore, a data dissemination strategy for micro-level survey data is proposed in the paper presented in Part II. This strategy aims to help statistical data producers to improve the trade-off between privacy risk and utility of the data that they release for research purposes.

Part I

New Developments in Structure Preserving Estimators

Chapter 1

Updating Intercensal Health Indicators for Small Areas using the R Package spree

1.1 Introduction

The definition, monitoring, and evaluation of public policies based on quality statistics allow decision makers to better guide their actions to combat phenomena such as poverty. The reduction of poverty in all its expressions is relevant for national and international agendas (United Nations, 2019). Due to its complexity, poverty can be approached from different points of view, as well as measurements. For example, from an income-based perspective, the Foster-Greer-Thorbecke (FGT) indicators (Foster et al., 1984) are an extensively used set of indicators that includes, among others, the incidence of poverty or head count ratio, and the poverty gap. Both indicators require a defined poverty line specifying when an individual or household is under poverty.

Another approach is to measure poverty based on deprivations, for instance with the unsatisfied basic needs (USB) index (Feres and Mancero, 2001), and the multidimensional poverty index (MPI) (Alkire and Foster, 2007). These methods intend to represent the non-monetary capacities of the individuals or households in several aspects, that are also highly correlated with their purchasing power. The analysis of both approaches is relevant for most developing economies, e.g., in Latin America, where poverty has been identified as a structural problem (ECLAC, 2014). In this paper, the MPI of Costa Rica is taken as an example. Since one of the main characteristics of this index, is its flexibility to be applicable to specific national realities (Alkire et al., 2019), Costa Rica adopted and adapted the MPI as a complimentary tool along with income-based poverty measures to evaluate and monitor this phenomenon.

This Central American country defines five equally-weighted dimensions (‘housing’, ‘education’, ‘employment’, ‘health’, and ‘social protection’), each one consisting of a set of dichotomous indicators defining the deprivations that a individual or household has. Based on the number of deprivation and a specific national threshold that defines when a person or household is considered multidimensionally poor (INEC, 2015). Although Costa Rica produces and releases information on this index annually, the sample size of the National Household Survey (ENAHO - *Encuesta Nacional de Hogares*) only allows the production of reliable estimates for zone (urban and rural) and six planning regions. Costa Rica faces limitations not only to produce estimates of the MPI for smaller geographic areas, e.g., 81 cantons and 472 districts, but also for its components, i.e., dimensions and indicators. In recent years, 'health' has been found to be one of the dimensions that has more relative contribution to the overall index in this country. For instance, the contribution of this dimension to the index was 22.8% in 2020 (INEC, 2020). However, it remains unknown how of the health indicator behaves in small domains.

Small area estimation (SAE) methods are popularly used as a solution for these types of scenarios. SAE models allow improving the reliability of direct estimates based only on survey data and increasing the resolution of the target geographic scale, by combining data sources and making use of highly correlated auxiliary information, generally from population censuses or administrative records (Pfeffermann, 2013; Rao and Molina, 2015; Tzavidis et al., 2018).

Regarding the availability of auxiliary sources to apply SAE methods, many developing countries do not have administrative records or systems to control their quality. For this reason, population and housing censuses are traditionally considered as auxiliary source although they are carried out only every 10 years. The literature of SAE methods that specifically tackles the time constraint between the survey and the census data is very limited. In this paper the most popular methods in this field are used: the structure preserving-estimation (SPREE) methods. These estimators have been implemented mainly to produce updated population information on labour market status for local areas (Luna-Hernández et al., 2015a), occupation (Berg and Fuller, Berg and Fuller; Hidiroglou and Patak, 2009), and monetary poverty (Isidro et al., 2016; Arias-Salazar, 2022). The SPREE-type estimators permit updating population counts by *preserving* the internal structure of a census. They require the census information to be organised in contingency or multi-way tables: by rows, for instance, domains of interest, and by columns, where each column contains the different categories of the variable of interest, e.g., 'employed', 'unemployed', 'out of labor force'.

The purpose of this paper is twofold. First, to provide updated population counts in the 81 cantons, from 2012 to 2017 in the 'health' dimension of the Costa Rican MPI and its respective indicators. The cantons with the highest proportion of vulnerable households will be identified, i.e., households that have three or more deprivations. The different indicators at zone and planning region level will also be analysed.

The second objective of this work is to present and illustrate the use of the R package **spree** to update intercensal indicators using SPREE-type estimators. To the best of my knowledge, none of these estimators are available in existing R packages.

This paper has the following structure: Section 1.2 describes the SPREE methods as techniques to updated population counts in intercensal years. Section 1.3 introduces the package **spree** and its functionalities. The case study that motivates this paper is presented in Section 1.4. Examples of the implementation of the package in this specific case are also shown in this section, as well as some results of the application. In the last section, conclusions and some recommendations for further research are pointed out.

1.2 Methodology

This section describes the methodology to update intercensal counts for small areas based on SPREE-type estimators behind the R package **spree**. Three main estimators are available in SPREE literature, namely SPREE, generalised SPREE (GSPREE), and multivariate SPREE (MSPREE). All of them are provided in the R package **spree**.

A clarification of the concepts that will be used in the upcoming sections is relevant at this point. In the context of SPREE, the population and sample data are cross-classified tables, also referred as *compositions* where the inner cells are counts. The term *row margins* refer to area sizes, domains sizes or sums by rows, and *column margins* refer to totals by categories or columns. Furthermore, the names *population* and *census data* are used interchangeably as well as *sample* and *survey data*.

1.2.1 SPREE-type estimators

The structure preserving estimator (SPREE)

The first version of the SPREE estimators was proposed by Purcell and Kish (1980) as a solution to obtain counts and proportions for intercensal years, when the variable of interest has at least two categories and the population (census) data can be arranged in a cross-classification table. In this sense, for a population of size N, all units i (e.g., individuals or households) are organised in tables according to the area and category of the variable of interest to which they belong. For the census year (t_0), the table is represented as Z_{aj,t_0}, with $a = 1, \ldots, A$ areas or domains (rows) and $j = 1, \ldots, J$ categories of the variable of interest (columns).

The main assumption behind this proposal is that the census data is reliable in its interactions between rows and columns (i.e., between domains and categories of the variable of interest), but the total of observations that belongs to each domain and category are outdated in the postcensal years.

A two-way contingency table for the census year t_0 can be represented as a saturated log-linear model

$$\log Z_{aj,t_0} = \alpha^Z_{0,t_0} + \alpha^Z_{a,t_0} + \alpha^Z_{j,t_0} + \alpha^Z_{aj,t_0}.$$

The terms $\alpha^Z_{0,t_0}, \alpha^Z_{a,t_0}, \alpha^Z_{j,t_0}$ are called *allocation* structure and the interaction term α^Z_{aj,t_0} *association* structure. These terms can be defined using a centred-constraint parametrization:

$$\alpha^Z_{0,t_0} = \frac{1}{AJ} \sum_{a=1}^{A} \sum_{j=1}^{J} \log Z_{aj,t_0},$$

$$\alpha^Z_{a,t_0} = \frac{1}{J} \sum_{j=1}^{J} \log Z_{aj,t_0} - \alpha^Z_{0,t_0},$$

$$\alpha^Z_{j,t_0} = \frac{1}{A} \sum_{a=1}^{A} \log Z_{aj,t_0} - \alpha^Z_{0,t_0},$$

$$\alpha^Z_{aj,t_0} = \log Z_{aj,t_0} - \alpha^Z_{a,t_0} - \alpha^Z_{j,t_0} - \alpha^Z_{0,t_0},$$

and the constraints:

$$\sum_{a=1}^{A} \alpha^Z_{a,t_0} = \sum_{j=1}^{J} \alpha^Z_{j,t_0} = \sum_{a=1}^{A} \alpha^Z_{aj,t_0} = \sum_{j=1}^{J} \alpha^Z_{aj,t_0} = 0, \tag{1.1}$$

must be fulfilled (Luna-Hernández, 2016). The updated target composition Y_{aj,t_1} can also be defined as a saturated log-linear model (logY_{aj,t_1}) with the same constraints. However, usually only survey data is available in t_1. As aforementioned, official results are only released for planned domains in that period due to sample size limitations.

To obtain an updated population table $\hat{Y}_{aj,t_1}$, SPREE uses the information form survey data that can be assumed reliable: its margins, i.e., the *allocation* structure. For the inner cells or interactions between rows and columns, the *association* structure of the census table is used. Therefore, the assumption of SPREE is that this structure in t_0 remains unchanged in the target year t_1, and is defined as follows:

$$\alpha^Y_{aj,t_1} = \alpha^Z_{aj,t_0}. \tag{1.2}$$

A visual representation of the components and structure of SPREE is shown in Figure 1.1. The target composition $\hat{Y}_{aj,t_1}$ is obtained with census data Z_{aj,t_0} and reliable actual margins Y_{a,t_1} and Y_{j,t_1}.

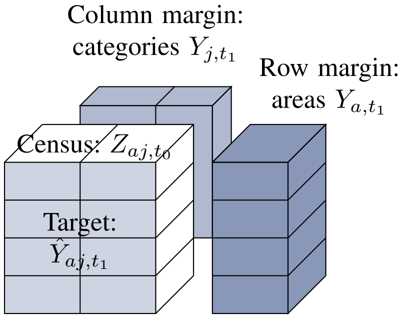


Figure 1.1: Representation of data requirements to apply SPREE based on Purcell and Kish (1980)

The updated target estimates can be obtained via the iterative proportional fitting (IPF) algorithm (Deming and Stephan, 1940), which is a well-known method to update census information (Suesse et al., 2017). This algorithm fits counts of a contingency table based on a set of given margins (Deville and Särndal, 1992). In the context of SPREE, the fitting is done on a census table by keeping a set of updated and reliable margins fixed. See the supplementary material for further details.

The updated SPREE composition $\hat{Y}^S_{aj,t_1}$ obtained using the IPF algorithm is represented by Luna-Hernández (2016) as:

$$\hat{Y}^S_{aj,t_1} = \text{IPF}\big[\exp(\hat{\alpha}^Y_{aj,t_1}), Y_{a,t_1}, Y_{j,t_1}\big], \tag{1.3}$$

with Y_{a,t_1} and Y_{j,t_1} representing the reliable survey margins (rows and columns) and $\hat{a}^Y_{aj,t_1} = a^Z_{aj,t_0}$, the interaction terms that are preserved as in the census data. The superscript *S* is used to denote that SPREE was implemented.

The IPF approach requires some considerations for its implementation: the input data must be arranged by domains and categories, i.e., contingency tables, and both, census and survey data, must have the same dimensions. Furthermore, the variable of interest must be available in the census data with same definition as in survey data, or a highly correlated variable (Green et al., 1998).

The generalised SPREE (GSPREE)

Zhang and Chambers (2004) introduced the GSPREE with the aim of relaxing the assumption that the interactions between domains and categories of the variable of interest in t_1 remain exactly as in the census period, i.e., $\alpha^Y_{aj,t_1} = \alpha^Z_{aj,t_0}$. Using direct estimates from survey data $\hat{Y}^{\text{Dir}}_{aj,t_1}$ with same structure as the census composition (e.g., a contingency table), a *proportionality coefficient* β is included and the structural assumption is now:

$$\alpha^Y_{aj,t_1} = \beta\alpha^Z_{aj,t_0}.$$

Note that SPREE is a particular case of GSPREE when the constant $\beta = 1$. To compute this coefficient, Zhang and Chambers (2004) suggest, among other alternatives, to obtain maximum likelihood estimators (MLE) of β assuming a Poisson or a multinomial distribution for the survey counts for each area, and consequently obtain α^Y_{aj,t_1}. In terms of data requirements, GSPREE needs the availability of a survey composition table, meanwhile the former version of the estimator requires only a set of row and column margins. The IPF algorithm can be used to obtain the target composition $\hat{Y}^G_{aj,t_1}$ as in Equation 1.3.

The multivariate SPREE (MSPREE)

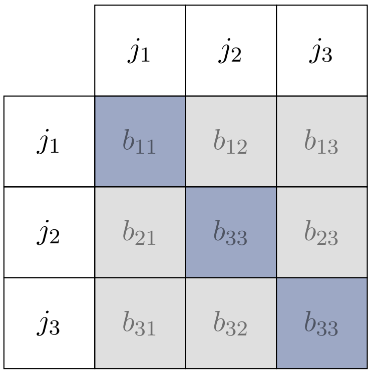
Although GSPREE relaxes the structural assumption of SPREE by adding a coefficient of proportionality gathered from direct estimates, this term is only one constant that is assumed to be equal for all categories of the variable of interest. Luna-Hernández (2016) proposes a novel version that aims to capture the relationships between the categories. The MSPREE uses $\boldsymbol{\beta}$ as proportionality coefficient, which is now a $J \times J$ matrix (with $(J-1)\times(J-1)$ free parameters), and varies inside each area, from one category to another. Here, the structural assumption is:

$$\alpha^Y_{a,t_1} = \boldsymbol{\beta}\alpha^Z_{a,t_0}.$$

Similarly as the previous two cases, the target MSPREE composition $\hat{Y}^M_{aj,t_1}$ can also be obtained via IPF as in Equation 1.3. Notice that as well as in the GSPREE estimator, the $\boldsymbol{\beta}$ in MSPREE requires direct estimates of the counts $\hat{Y}^{\text{Dir}}_{aj,t_1}$, so the relationship between it and the target composition is also captured. If there are only two categories ($J = 2$) the $\boldsymbol{\beta}$ will be only one constant, which would be the case of GSPREE.

The $\boldsymbol{\beta}$ satisfies also the sum-zero constraints similarly as in Equation 1.1, i.e., the sum of coefficients by rows and columns is zero. A scaled-version of $\boldsymbol{\beta}$ in terms of proportional interactions is used in the package **spree** to facilitate its interpretation. The scaled matrix permits to compare the relationship between the categories in the updated (target) table and the census table. This is illustrated in Figure 1.2. The diagonal terms in the scaled matrix represent for each category how the interactions change (or differ) between the target and census table. When these values are close to one, it means that the interactions in these two compositions are basically the same, which is actually the structural assumption in SPREE (Equation 1.2).

The other elements (in gray in the Figure 1.2) are arranged based on the interactions of the other categories with the purpose of satisfying the required constraints. The sum of these off-diagonal terms is zero. The methodology to produce the scaled-version of $\boldsymbol{\beta}$ can be found in Luna-Hernández (2016).



	j_1	j_2	j_3
j_1	b_{11}	b_{12}	b_{13}
j_2	b_{21}	b_{33}	b_{23}
j_3	b_{31}	b_{32}	b_{33}

Figure 1.2: Representation of the scaled-$\boldsymbol{\beta}$ matrix

Survey-based direct estimates are required to compute $\boldsymbol{\beta}$, and several approaches can be implemented: via maximum likelihood (ML) or iterative weighted least squares (IWLS) (Jiang, 2007), in both cases by using a log or a logit link. ML estimates of $\boldsymbol{\beta}$ can be obtained by assuming

$$\hat{Y}^{Dir}_{aj,t_1}|\alpha^{Z}_{a,t_0} \overset{\text{ind}}{\sim} \text{Poisson}(\mu_{aj,t_1})$$

or

$$\hat{Y}^{Dir}_{a,t_1}|\alpha^{Z}_{a,t_0} \overset{\text{ind}}{\sim} \text{Multinomial}\Big(\sum_{j=1}^{J} \hat{Y}^{Dir}_{aj,t_1}, \pi_{a,t_1}\Big),$$

where μ_{aj,t_1} are Poisson expected frequencies and π_{a,t_1} the cell probabilities used under multinomial sampling. The implementation of the ML approach can, however, lead to model misspecification when an informative sampling design is ignored (Zhang and Chambers, 2004). The IWLS algorithm provides a solution under this scenario by including an estimate of the variance-covariance matrix of the target composition $\hat{Y}^{M}_{aj,t_1}$ (see Jiang (2007) for details on this algorithm). Since this matrix is usually not available, Luna-Hernández (2016) proposes a solution by multiplying a design effect with the variance that corresponds to a simple random sampling design without replacement. The estimate of this matrix is defined as

$$\hat{V}_{aj,t_1} = \frac{\text{deff}_{j,t_1}\ \hat{\pi}_{aj,t_1}(1-\hat{\pi}_{aj,t_1})}{n_{a,t_1}}, \tag{1.4}$$

with $\hat{\pi}_{aj,t_1} = \frac{\hat{Y}^M_{aj,t_1}}{Y_{a,t_1}}$ and n_{a,t_1} the area sample sizes. The term deff_{j,t_1} is a vector of size J, containing a design effect for each category.

For both approaches, ML and IWLS, the use of a log or a logit link leads to equal results. For this reason, in the package **spree** only the multinomial model is implemented.

1.2.2 MSE estimation

To estimate a measure of uncertainty of the SPREE-type estimators an unconditional MSE (U-MSE) based on a parametric bootstrap can be implemented. The U-MSE is given by $E[(\hat{Y}_{aj,t_1} - Y_{aj,t_1})^2]$, where the expectation is over Y_{aj,t_1} and the sampling procedure, given a population model for Y_{aj,t_1} (Tzavidis et al., 2018; Luna-Hernández, 2016). Other alternatives such as a finite population MSE (FP-MSE) or an analytical approximation of the variance can be consulted in Luna-Hernández (2016).

Let K represent the SPREE-type estimator: SPREE, GSPREE, and MSPREE, and Ψ represents the proportionality coefficient for each case (1 , β, and $\boldsymbol{\beta}$ respectively). The steps to obtain the U-MSE are summarised as follows:

1. From the point estimate $\hat{Y}^K_{aj,t_1}$ calculate the within-area proportions $\hat{\pi}_{aj,t_1} = \frac{\hat{Y}^K_{aj,t_1}}{Y_{a,t_1}}$, with Y_{a,t_1} reliable domain totals (row margins).
2. Generate B-bootstrap populations Y^{*b}_{aj,t_1} under the assumption that the target estimate across areas has the following distribution:

$$Y_{a,t_1}|\alpha^Z_{a,t_0} \overset{\text{ind}}{\sim} \text{Multinomial}\Big(Y_{a,t_1}, \hat{\pi}_{a,t_1}\Big).$$

If $K =$ GSPREE or MSPREE:

(a) Select a sample y^{*b}_{aj,t_1} from each B population by defining a sampling fraction or setting the same sample size as in the sample data used to compute the point estimate.

(b) For each sample obtain the proportionality coefficient $\hat{\Psi}$ and set the structural assumption:

$$\hat{\alpha}^{Y,b}_{aj,t_1} = \hat{\Psi}\alpha^{Y,*b}_{aj,t_1}.$$

3. Compute B K-SPREE estimators $\hat{Y}^{K,b}_{aj,t_1} = \text{IPF}\big[\exp(\hat{\alpha}^{Y,b}_{aj,t_1}), Y_{a,t_1}, Y_{j,t_1}\big]$.
4. Estimate the U-MSE:

$$\widehat{\text{U-MSE}}(\hat{Y}^K_{aj,t_1}) = \frac{1}{B}\sum_{b=1}^{B}(\hat{Y}^{K,b}_{aj,t_1} - Y^{*b}_{aj,t_1})^2.$$

1.3 The R package spree

In this Section, an overview of the R package **spree** is presented. The package is available from the GitHub folder (https://github.com/AlejandraAriasSalazar/spree_pkg). The main function of the package is spree, which allows obtaining point and uncertainty estimates. The arguments of spree are summarised in Table 1.1.

As input, this function requires two data frames (population_data and

sample_data) with the exact same structure. They must have the same number of rows and columns, and the population_domains, and sample_domains should be the same, as well as the name of the columns in each data set. There are two cases when these requirements are not met:

Case 1. Due to changes in the administrative organisation of the country or region. The user must adjust the data in order to obtain the same structure in both data sets, considering the possibilities based on the data characteristics and the relevance in the application. For example, if in the target year a domain was divided into two, the user could re-construct the original domain in the survey data as in the census data by merging them.

Case 2. Some domains where not included in the sample. A practical way to solve this issue is adding the missing rows with a small value (e.g., 0.0001) (Isidro, 2010), so the updating procedure can be performed. The **spree** package includes the function prep_sample_data that helps to prepare the sample data in case of out-of-sample domains. The output is a data frame that could be directly used as sample_data in the spree function.

Furthermore, as explained in Section 1.2.1, the 'true' total in t_1 by rows and columns must be the same. In this package, the updating process is performed using the totals of the sample_data when no set of margins is provided. If the user wants to specify other margins, both row_margins and column_margins must be provided and the sum of them must be also equal. Column margins can be adjusted based on the sum of row margins and the proportion of observations in each category of the sample data. The function column_tot is an auxiliary function that can help the user to prepare the column_margins based on given row_margins.

The output of the spree function is a list with four elements:

- updated_point: updated population data in t_1. The total by rows and columns are the same as the sample data if not set of margins were provided.
- MSE: mean squared error.
- CV: coefficients of variation.
- Beta: (scaled) proportionality coefficient.

Finally, the function compare_spree allows the user to make a visual comparison (based on box plots) when various methods have been applied, in terms of point estimates (Section 1.2.1), MSEs (Section 1.2.2) and CVs, which are computed as:

$$CV = \sqrt{\widehat{\text{U-MSE}}(\hat{Y}^K_{aj,t_1})}/\hat{Y}^K_{aj,t_1}.$$

Table 1.1: Arguments of `spree`

Argument	Description
`population_domains`	Variable in population data that identifies the domains.
`sample_domains`	Variable in sample data that identifies the domains.
`population_data`	Data frame with A rows and J columns.
`sample_data`	Data frame with A rows and J columns.
`row_margins`	Numeric vector with A elements containing the 'true' domain sizes in t_1. `NULL` is set as default and the sum by rows in `sample_data` is used.
`col_margins`	Numeric vector with J elements containing the 'true' totals for each category. The sum by columns in `sample_data` is used as default.
`type`	SPREE version to implement: 'SPREE', 'GSPREE', 'MSPREE'. If this argument remains empty, the first option is selected. The option 'MSPREE' requires $J \geq 3$. If $J = 2$, the 'GSPREE' will be applied.
`B`	Number of bootstraps used in the MSE estimation.
`method`	Method to obtain estimates of the β. This argument is only required when `type` = 'MSPREE'. The alternatives are 'ML' and 'IWLS'.
`design_effect`	If `type` ='MSPREE' and `method`= 'IWLS' are selected, a design effect for the column totals can be provided. A vector containing 1's is used as default.

1.4 Case study: Updating health indicators of the MPI for small areas in Costa Rica

Administratively, Costa Rica is divided in four levels. Although several changes have occurred in recent years, in the last census year (2011) the division was:

- Level 1: two zones (urban /rural).
- Level 2: six planning regions.
- Level 3: 82 cantons.
- Level 4: 472 districts.

The National Institute of Statistics and Censuses (INEC - *Instituto Nacional de Estadística y Censos*) of Costa Rica produces monetary and non-monetary based indicators of poverty with the aim to study the phenomenon from different perspectives. However, due to sample size limitations of the data available on that matter, INEC releases official statistics only for the first and second administrative level.

As aforementioned, this paper has two objectives: to update the indicators of the 'health' dimension of the Costa Rican MPI, and to show how this process is carried out using the **spree** package. The upcoming subsections describe the target indicator and the data sources to be used in the updating process, followed by illustrations of the use of the R package **spree** to produce the target indicators. The last part of this section provides a more in-detail analysis of the results.

The multidimensional poverty index of Costa Rica

INEC adapted the MPI proposed by Alkire and Foster (2007) as shown in Figure 1.3. The Costa Rican MPI consists of five equally weighted dimensions, and each of them contains four indicators. Each indicator is a dichotomous variable indicating 1 if the household has the deprivation or 0 otherwise. In this paper, the focus is to update the indicators of the 'health' dimension, since unlike other dimensions, all its indicators are in both, census and survey data. The methodology to compute this index can be found in INEC (2015).

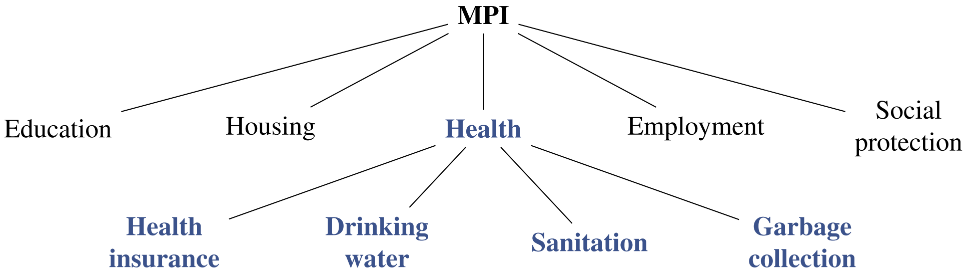


Figure 1.3: Dimensions and indicators of the MPI of Costa Rica

The 'health' dimension has the following indicators:

1. Health insurance: a household is considered deprived if at least one member (18 years old or older) has no health insurance (private or public).

2. Drinking water: access to drinking water and source.

3. Sanitation: access and type of toilet facility.

4. Garbage collection: a household is considered deprived if it does not have a formal garbage collection system.

It is important to emphasise the differences between the results in this paper and the official publications. Results of the MPI released by INEC, taking into account its dimensions and indicators, are only available for two zones (urban and rural) and six planning regions. By zone, results are provided considering both, the total of households and the total of multidimensionally poor households as the population of interest. For planning regions, only results on the total of multidimensionally poor households are published. To produce the results of this paper, the specification of which household is multidimensionally poor is not available. For this reason, all results here provided (by zone, planning region, and cantons) consider the total of Costa Rican households as the population of interest. Thus, results can only be compared with official publications at zone level.

1.4.1 Data sources

The data sources used in this application were produced and provided by INEC of Costa Rica under a confidentiality agreement.

Population and housing census 2011

The population and housing census in Costa Rica is carried out every ten years. The latest data available is from the X$^{\text{th}}$ National Population and VI$^{\text{th}}$ Housing Census (2011) (*Xth Censo Nacional de Población y VIth de Vivienda (2011)*). The primary goal of this statistical operation is to collect information of people, households, and dwellings necessary for the planning, execution, and evaluation of public policies (INEC, 2012). Several aspects of the population are characterised based on this census, for example: access and use of internet, dwelling conditions, employment status, access to social protection, among others. Regarding the MPI, this census does not contain all the required variables for its computation. The dimension ‘health’ is the only one that is available, i.e., all four indicators are included in the census data.

National household surveys 2012-2017

The annual National Household Survey (ENAHO - *Encuesta Nacional de Hogares*) is the primary source for poverty measures based on poverty lines such as the incidence of poverty and the poverty gap, as well as the Gini coefficient (Gini, 1912). Since 2014, the ENAHO includes the MPI as one of the main products, providing information about housing characteristics, education, social protection, health, and employment of the household members. The sampling design used in the ENAHO is a two-stage stratified random sampling where census segments are the first stage units selected with probability proportional to size, and dwellings are defined as the final stage units. To guarantee precision in its estimates, INEC releases results only at national level, for two zones and six planning regions. This institution uses as quality parameter for the main poverty measure (the percentage of household under poverty) a coefficient of variation of 15% or less (INEC, 2017).

In the period considered in this work (2011-2017) one canton was divided in two so the number of domains in the census data and some survey data differs. As one of the requirements to apply SPREE-type estimators is that the dimension in both data sets must be equal, the divided canton was re-grouped as in the census year.

Demographic projections

All SPREE-type estimators require a set of reliable total margins. That means trustworthy and updated population sizes for each domain as well as 'true' totals for each category of the variable of interest.

A possible approach is to use the total provided by the survey data as 'true' set of margins, under the assumption that the inner cells are not reliable, but the totals are. In this application, the ENAHO has not enough sample size to provide reliable margins, indeed, in two years of the period of study there were some cantons out of the sample. Consequently, demographic projections for 2012-2017 are used as 'true' domain sizes. For this application, the procedure to construct household projections was done following the methodology of Sáenz (2002).

1.4.2 Usage of the R package spree

One of the goals of this paper is to produce updated population estimates of health indicators of the MPI for zone, regions, and cantons in Costa Rica, from 2012 to 2017. Two different target compositions are defined:

1. the number of households in each canton with *Zero*, *One*, *Two*, *Three or four* deprivations, and
2. the number of households in each zone and region with deprivation (*Yes*, *No*), in:
    - Health insurance
    - Drinking water
    - Sanitation
    - Garbage collection.

The first result makes it possible to identify the most vulnerable households that suffer from various deprivations. The second result provides information on which aspect of 'health' dimension needs more improvement.

Notice that for the first case MSPREE can be implemented since it requires more than two categories of the variable of interest. Because the second case consists of four different independent dichotomous variables, only the SPREE and the GSPREE estimator can be used.

To illustrate the use of the package **spree**, the procedure to update the number of households in each canton with *Zero*, *One*, *Two*, *Three or four* deprivations, is shown below. Other results that will be presented at the end of this section at the zone and region level were also produced with the help of the **spree** package following the same procedure. The only difference is the specification of the domain of interest.

Preparing data

The minimum inputs required to use the function `spree` are the population and sample data, as well as their respective domain (ID) variable within the data. Both input arguments are data frames with the same number of domains (rows) and variables (categories or columns).

First, it is important to check that the names of the domains and categories are the same. The dimension of the census and survey data must be also equal:

```
R> str(census)
'data.frame':   81 obs. of  5 variables:
 $ Canton    : chr  "101" "102" "103" "104" ...
 $ Zero      : int  75252 15043 52047 6040 2932 13321 ...
 $ One       : int  7914 1658 5877 3013 1094 2437 1111 ...
 $ Two       : int  761 119 561 716 395 484 262 245 161 ...
 $ Three_Four: int  139 19 123 122 211 106 47 28 41 108 ...

R> str(survey_17)
'data.frame':   80 obs. of  5 variables:
 $ Canton    : int  101 102 103 104 105 106 107 108 ...
 $ Zero      : int  443 80 280 57 7 84 58 206 45 128 ...
 $ One       : int  165 30 107 43 6 48 25 72 29 73 ...
 $ Two       : int  9 2 9 6 5 3 2 0 2 3 ...
 $ Three_Four: int  3 0 3 0 2 2 0 0 0 0 ...
```

Notice that the survey data `survey_17` has only 80 domains, meaning that in 2017 one of the cantons were not selected in the sample. The sample table must have the same dimensions, domains and categories as the population table. A solution is to add the missing domain with a small value in each category. This can be done with the function `prep_sample_data`:

```
R> survey_17f <- prep_sample_data(population_domains =
    "Canton", sample_domains = "Canton", population_data =
    census, sample_data = survey_17)

Out of sample domains:
Percentage:  1.23 %
Number:  1
Domains:  117
```

The output shows the number and percentage of domains that are out of sample, as well as the name or identification of them (in this example, the name of the domain is '117'). Now, the output object `survey_17f` is a data frame with the exact same domains and categories as in the census file:

```
R> str(survey_17f)
'data.frame':   81 obs. of  5 variables:
 $ Canton    : chr  "101" "102" "103" "104" ...
 $ Zero      : num  443 80 280 57 7 84 58 206 45 128 ...
 $ One       : num  165 30 107 43 6 48 25 72 29 73 ...
 $ Two       : num  9 2 9 6 5 3 2 0 2 3 ...
 $ Three_Four: num  3 0 3 0 2 2 0 0 0 0 ...
```

As stated in Table 1.1 if no specification for the `row_margins`, and `col_margins` arguments is given, the 'true' margins will be set as in the survey data, i.e., based on row and column sums. For this case-study, household projections (`P2017c`) for each canton are used as ‘true’ row margins:

```
R> str(P2017c)
 num [1:81] 112177 21855 73772 12073 5598 ...
```

To provide ‘true’ column margins (i.e., totals for each category), the function `column_tot` is implemented. Here, the sum of the column margin (total) will be set as the given `total_true` and the totals for each category will be assigned based on the proportions of `data_true`:

```
R> ctotals_17 <- column_tot(data_true= survey_17f,
                domains = "Canton",
                total_true= sum(P2017c))
```

The result is the total number of households for each category:

```
R> ctotals_17
       [,1]
[1,] 949178
[2,] 482892
[3,] 103716
[4,]  32193
```

The sum of these numbers almost equals the total of households provided by the ‘true’ row margin. Small derivations are due to the proportional assignment:

```
R> sum(ctotals_17)
[1] 1567979

R> sum(P2017c)
[1] 1567978
```

Updating procedure

The function `spree` allows to obtain updated population counts by defining the arguments mentioned in Table 1.1. Here, an example of the type ‘MSPREE’ with ‘IWLS’ as the method to estimate β is shown, since this is the most complex alternative in terms of arguments. In this application, no design effects are available and for that reason a simple random sampling design is assumed:

```
R> MSPREE_IWLS_17 <- spree(population_domains = "Canton",
   sample_domains = "Canton",  population_data = census,
   sample_data = survey_17f, row_margins = P2017,
   col_margins = ctotals_2017, type = "MSPREE",
   method = "IWLS", design_effect = c(1,1,1,1))
```

The output is a list with the updated point estimates, MSEs, CVs and the proportionality coefficient.

Comparing methods

A basic overview of the results can be reached, for example, with the `summary` function:

```
R> summary(MSPREE_IWLS_17$updated_point)
    Canton                Zero            One
 Length:81          Min.   :  871   Min.   :  802
 Class :character   1st Qu.: 4276   1st Qu.: 2483
 Mode  :character   Median : 7064   Median : 4112
                    Mean   :11718   Mean   : 5962
                    3rd Qu.:14899   3rd Qu.: 6828
                    Max.   :78072   Max.   :30454
      Two            Three_Four
 Min.   :  71.05   Min.   :  33.33
 1st Qu.: 395.51   1st Qu.: 163.92
 Median : 694.49   Median : 270.81
 Mean   :1280.44   Mean   : 397.44
 3rd Qu.:1864.96   3rd Qu.: 456.88
 Max.   :6879.26   Max.   :2218.16
```

The summary shows that on average, more household have no or only one deprivation than two or more. If other SPREE-type estimators are also performed, the user can make some comparison between them. For instance, the scaled-proportionality coefficients for each SPREE-type can also be extracted from the return object `Beta`. For SPREE and GSPREE, `Beta` is only one value:

```
R> SPREE_17$Beta
[1] 1

R> GSPREE_17$Beta
          [,1]
[1,] 0.5674238
```

For MSPREE, a matrix is returned which can be summarised by the column sums:

```
R> MSPREE_ML_17$Beta
            [,1]         [,2]         [,3]        [,4]
[1,]  0.63789813  0.118702658 -0.136335443  0.01763279
[2,] -0.14718895  0.047786306  0.002263418  0.14492554
[3,]  0.16499605 -0.002437728  0.741091597 -0.16255832
[4,] -0.01780709 -0.116264930  0.134072025  0.81537154

R> colSums(MSPREE_ML$Beta)
[1] 0.63789813 0.04778631 0.74109160 0.81537154

R> MSPREE_IWLS_17$Beta
          [,1]       [,2]        [,3]       [,4]
[1,]  1.1554847  0.2396119 -0.03605342 -0.2035585
[2,] -0.7789303 -1.0248650 -0.27354121  1.0524715
[3,]  0.6727972  0.1761158  0.27959438 -0.8489130
```

```
[4,]  0.1061331 -0.4157277  0.30959463  0.7512027

R> colSums(MSPREE_IWLS$Beta)
[1]  1.1554847 -1.0248650  0.2795944  0.7512027
```

Notice, for example, that in the second category when the MSPREE with IWLS is performed, the scaled-coefficient is close to one. This gives an indication, that this particularly category does not receive much benefit from applying this estimator vs applying SPREE (which has $\beta = 1$). The function `compare_spree` helps to compare visually the different estimators in terms of the point and uncertainty estimations. The instruction requires a minimum of two lists which are the outputs from using the function `spree` :

```
R> plots_spree<- compare_spree(SPREE_17, GSPREE_17,
    MSPREE_ML_17, MSPREE_IWLS_17)
```

As output, three comparative plots are displayed (Figure 1.4).

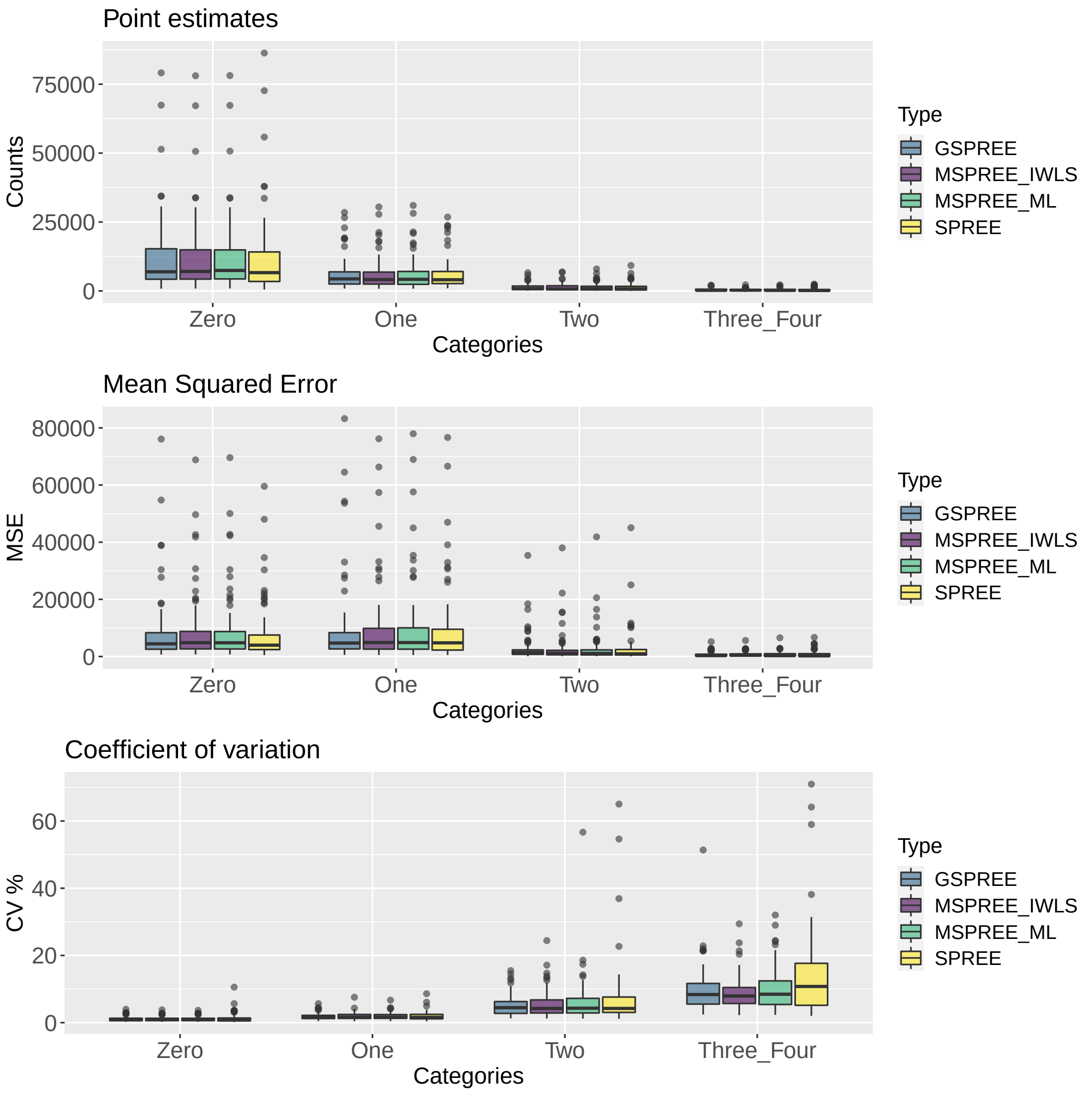


Figure 1.4: Comparison between SPREE-type estimators in terms of point estimates, MSEs and CVs

Unlike other SAE methods based on regression models, there are no model diagnostics for SPREE-type estimators. The decision of which type is more suitable for a specific case study can be made based on the MSEs and CVs.

In this specific case, the uncertainty estimations are relatively similar among estimators. For the last category 'Three or four' deprivations in the dimension 'health', the MSPREE with IWLS offers smaller values of the CV. This estimator is selected for the upcoming results.

Application

The updated tables are provided as `data.frame` and further analysis can be done. Using the results from the MSPREE with IWLS, the five cantons that had the biggest absolute change between 2012 and 2017 in the proportion of households with 'Three or four' deprivations in the dimension 'health' are displayed in Figure 1.5.

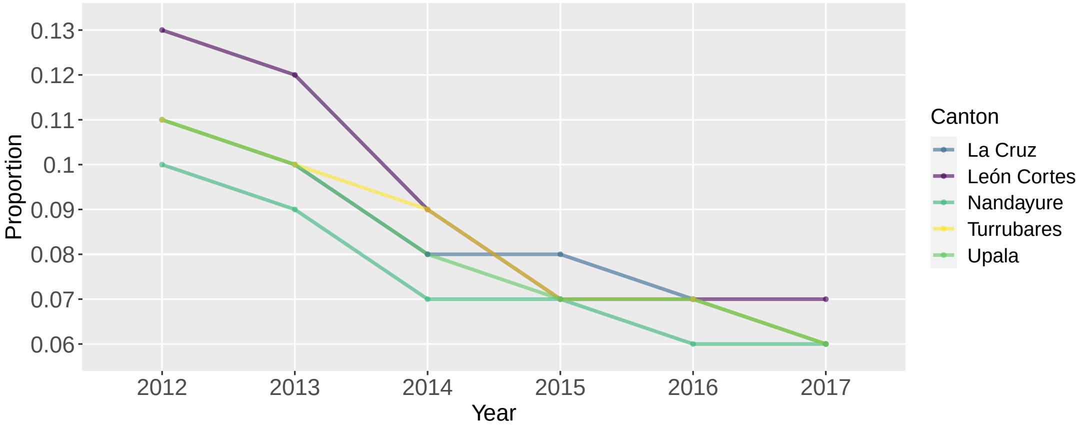


Figure 1.5: Cantons with the biggest reduction in the proportion of households with 'Three or four' deprivations in the 'health' dimension

Figure 1.6 shows the proportion of households with 'Three or four' deprivations in the dimension 'health' in 2012 and 2017 and the exact location of the five cantons with biggest reductions between these years. Although reductions in general can be noticed, the cantons with the highest values of this indicator in 2012 had also the highest values in 2017. Even though this study only focuses on one dimension, results of previous publications on the overall MPI and monetary poverty support these trends. For instance, the cantons on the border with Nicaragua located in the north of the country (e.g., 'Upala' and 'La Cruz') are in general poorer (Méndez and Bravo, 2011; Arias-Salazar, 2022).

An analysis by indicator makes it possible to identify the most common deprivations among Costa Rican households. Figure 1.7 compares the proportion of households with deprivation in each of the four indicators of the 'health' dimension, between 2016 and 2017 for the six planning regions. The indicator 'sanitation' is in general the less problematic and additionally the biggest reductions can be seen except for the Central region. The indicators 'health insurance' and 'garbage collection' show higher proportions of households with these deprivations. The Central region which is predominantly urban shows, as expected, by far the lowest values in the indicator 'garbage collection'.

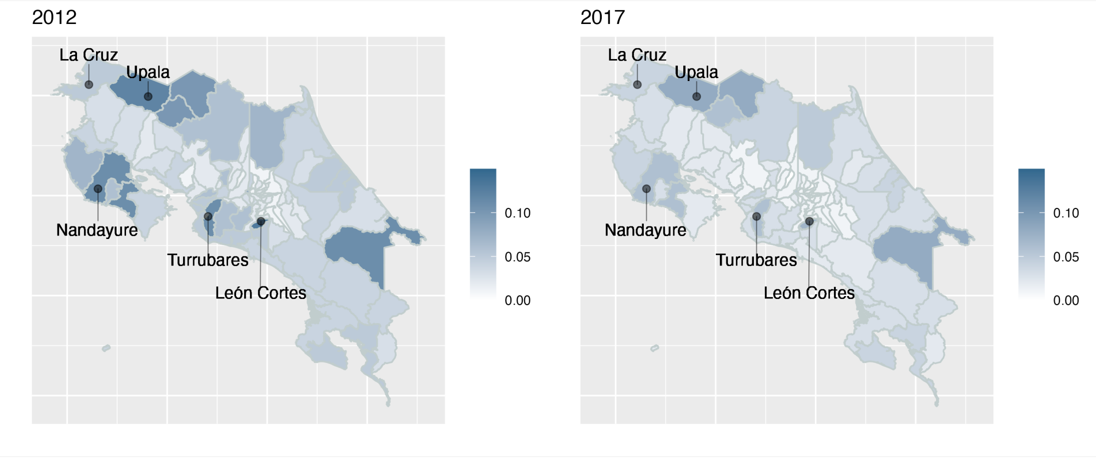


Figure 1.6: Proportion of households with 'Three or four' deprivations in the 'health' dimension by canton, 2012 and 2017

In 2017, the biggest values in terms of incidence (34.7%) and intensity (28.6 %) of the overall MPI were found in the Huetar Norte region (INEC, 2017). This region is made up of six cantons. The analysis of this paper allows to zoom into the changes between 2016 and 2017 in the proportion of households that are considered deprived in each of the indicators of the 'health' dimension which are shown in the Figure 1.8. 'San Carlos' is the only canton that has seen reductions in all the indicators during this period, while 'Upala', 'Los Chiles', and 'Sarapiquí' have seen increases in the proportion of households with deprivations in three of the indicators.

A direct comparison between the updated estimates via SPREE-estimators and the official results is only possible by zone (urban and rural), since this is the only result published by INEC considering the total number of Costa Rican households as the population of interest. Table 1.2 shows that the estimated indicators (i.e., the proportion of private households in each indicator) are very close to those published. In fact, they are all within the 95% confidence interval. This result can be consulted in INEC (2017).

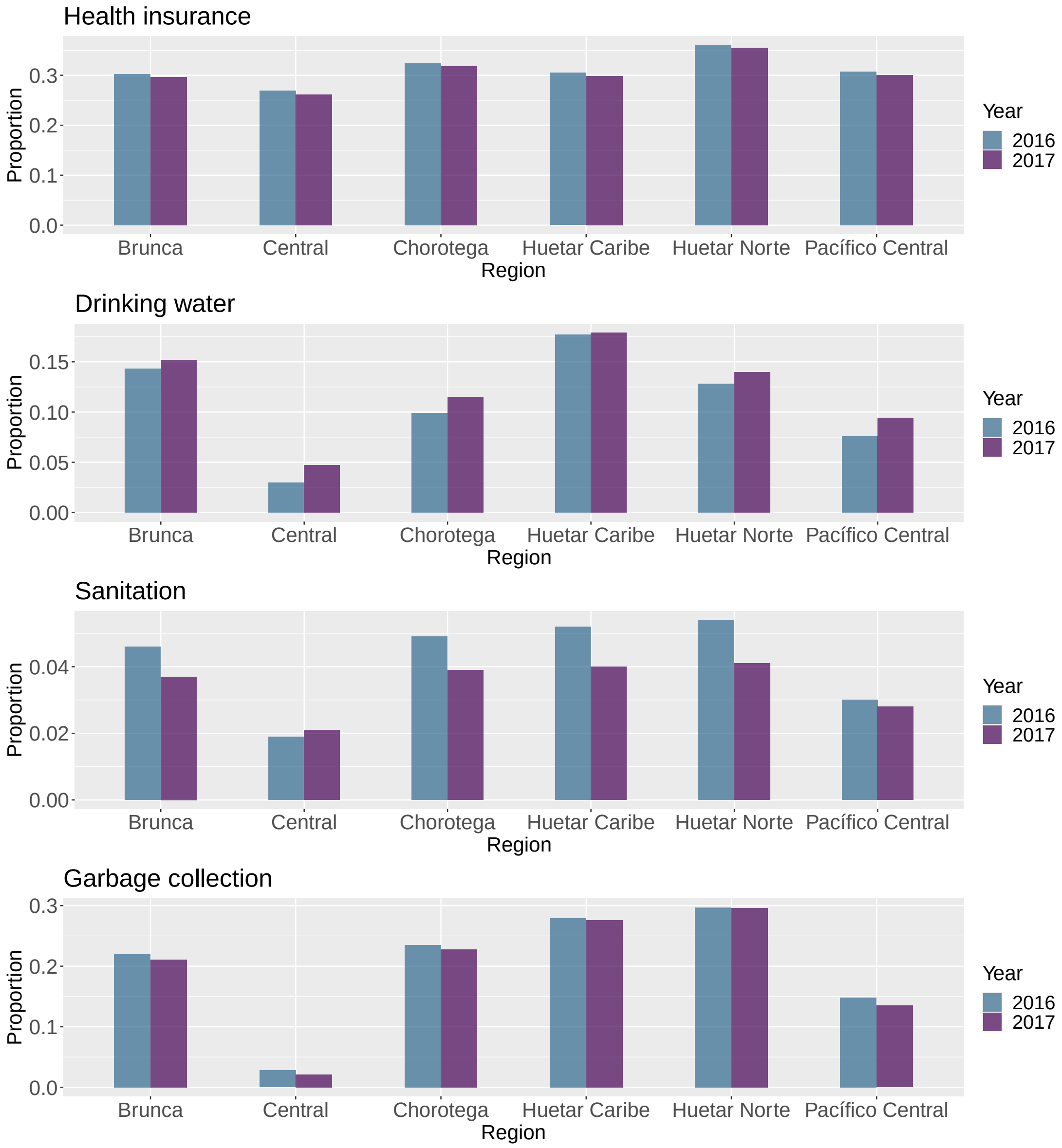


Figure 1.7: Proportion of households with deprivation in the indicators of the ‘health’ dimension by region (2016-2017)

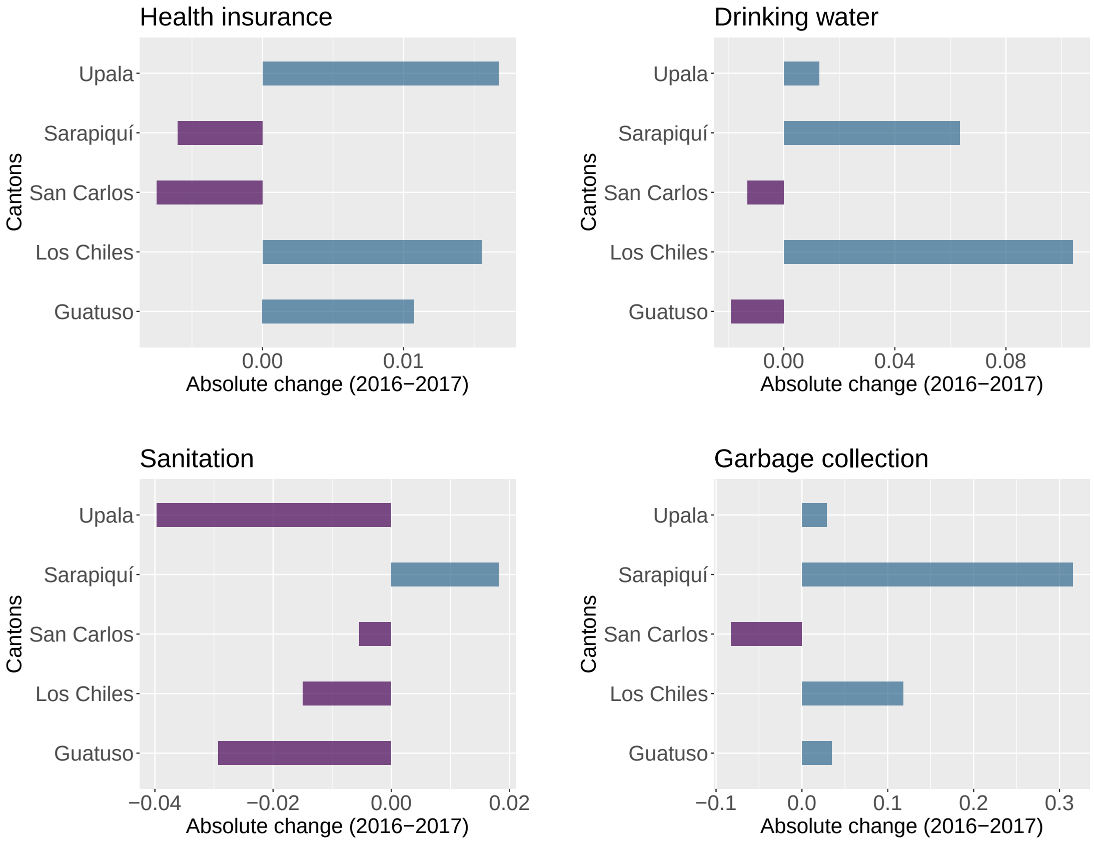


Figure 1.8: Absolute change in the proportion of households with deprivation in each indicator, for the cantons of the Huetar Norte region (2016-2017)

Table 1.2: Comparison between updated estimates via GSPREE and official results. The lower and upper bounds around the estimated proportion are based on a 95% confidence interval

Indicator/	**GSPREE**	**Official results**		
Zone	**Estimation**	**Estimation**	**Lower**	**Upper**
Health insurance				
Urban	0.267	0.265	0.255	0.276
Rural	0.304	0.390	0.292	0.328
Drinking water				
Urban	0.034	0.033	0.028	0.039
Rural	0.178	0.181	0.156	0.206
Sanitation				
Urban	0.018	0.018	0.014	0.228
Rural	0.042	0.043	0.033	0.952
Garbage collection				
Urban	0.024	0.022	0.016	0.029
Rural	0.261	0.267	0.239	0.295

1.5 Conclusions and further developments

In this paper, the health indicators of the multidimensional poverty index of Costa Rica are updated for several postcensal years using SPREE-type estimators. With the help of these techniques, it was possible to identify the cantons with the highest proportion of households that have three or more deprivations in the 'health' dimension, as well as the cantons that achieved the biggest reductions in these numbers from 2012 to 2017. Moreover, updating census counts by planning region made possible to compare the absolute changes in the last two years of the study period, in each of the health indicators.

As mentioned before, a fair and direct comparison with official publications is not possible, because INEC measures the proportion of households with specific deprivations among the multidimensionally poor households at planning region level, and this information is not available in the census data. Nonetheless, the trend is the same, e.g., the regions with the highest incidence in all deprivations are Huetar Norte and Huetar Caribe.

Several ways to improve these results are identified. A higher geographical resolution, for instance for the 472 districts, could provide better inputs to study these indicators and offer better tools for decision makers. The use of auxiliary information can be a good alternative to have better population totals for each subdomain (i.e., row margins), for example through satellite imagery (Koebe et al., 2022). Greater disaggregation by other socio-demographic groups (e.g., sex, education and age of the head of the household) could also be relevant.

In addition, this case study was used to show the functionalities of the R package **spree**. This package is a user-friendly tool to update postcensal counts and also produce uncertainty measures. Currently, this package offers the estimators that can be used when the variable of interest is available in the population data. An extension of this package could include the new SPREE-type estimators that apply a SAE procedure as a previous step (Isidro et al., 2016; Arias-Salazar, 2022). The package could also be improved by providing more flexibility to the user in regarding the structure of the input data. For instance, when changes occur in the definition of domains in postcensal years. Furthermore, the package could support the looping though several intercensal years, providing more analysis and comparison options.

Acknowledgment
The elaboration of this paper was only possible thanks to the data provided by the National Statistical Institute of Statistics and Censuses from Costa Rica. The data access was provided under specific confidentiality conditions. The numerical results are not official estimates and are only produced for illustrating the methods. Furthermore, the author thanks to Angela Luna-Hernández for providing important part of the code with which the **spree** package was built.

Supplementary material A

A.1 Iterative proportional fitting algorithm

Let Y_{aj,t_1} be the unknown (target) table in the current period t_1. The estimated table $\hat{Y}_{aj,t_1}$ is obtained by using a census table Z_{aj,t_0} from an earlier period t_0 and a set of margins from a recent survey Y_{a,t_1} and Y_{j,t_1}.

Using the census Z_{aj,t_0} as a base or starting point, the first cycle of the algorithm is described below:

1. Rescale to the first row margins of Z_{aj,t_0}:

$$\hat{Y}^{(1)}_{aj,t_1} = Z_{aj,t_0} \frac{Y_{j,t_1}}{Z_{j,t_0}}.$$

2. The cells rescaled in the previous step are rescaled again, now with the column margins of $\hat{Y}^{(1)}_{aj,t_1}$:

$$\hat{Y}^{(2)}_{aj,t_1} = \hat{Y}^{(1)}_{aj,t_1} \frac{Y_{a,t_1}}{\hat{Y}^{(1)}_{a,t_1}}.$$

3. Cells are rescaled again, with row margins of $\hat{Y}^{(2)}_{aj,t_1}$:

$$\hat{Y}^{(3)}_{aj} = \hat{Y}^{(2)}_{aj,t_1} \frac{\hat{Y}^{(2)}_{j,t_1}}{Y_{j,t_1}}.$$

In this way, the IPF follows an iterative process, repeating the last two final steps until reaching convergence.

This estimator manages to fulfil two important characteristics, minimising the χ^2 distance:

$$\chi^2 = \sum_{a=1}^{A} \sum_{j=1}^{J} \frac{(Y_{aj,t_1} - \hat{Y}_{aj,t_1})^2}{Y_{aj,t_1}}.$$

Furthermore Ireland and Kullback (1968) show that for positive initial values, this procedure manages to find an optimal solution, according to the Kullback-Leibler (KL) divergence measure, in other words, the IPF manages to minimise the relative entropy:

$$KL = \sum_{a=1}^{A}\sum_{j=1}^{J} Y_{aj,t_1} \log \frac{Y_{aj,t_1}}{\hat{Y}_{aj,t_1}},$$

More information about this algorithm can be found in Bishop et al. (2007) and Zaloznik (2011).

Chapter 2

Small Area Estimates of Poverty Incidence in Costa Rica under a Structure Preserving Estimation (SPREE) Approach

2.1 Introduction

The estimation and monitoring of socio-economic indicators is relevant for decision-making and the development of public policies aimed at improving the conditions of the citizens. Among other characteristics, high-quality statistics must be relevant, accurate, and reliable to use them in the design, development and assessment of programs of social interest (Eurostat, 2017). The success of these plans depends on how they are formulated and oriented, but in many cases, the information available is not enough to achieve this objective. Traditionally, national surveys are carried out every year in many countries to produce an up-to-date status of important topics such as poverty, inequality, and unemployment. This information, which is obtained periodically, usually satisfies the quality requirements, for instance of national statistical offices only at bigger domains. In other words, due to lack of resources, the sample sizes are not large enough to study the problems of interest in detail. For example, in the case of poverty: Where is the most vulnerable population located? Which areas have been improved through the years and which areas have stagnated? Which other conditions (e.g., sex, age, disabilities) are associated with this phenomenon, and in which local areas?

Small area estimation (SAE) methods have the goal of producing reliable estimates in smaller domains, i.e., with adequate precision. Most of these methodologies, usually classified as unit- or area-level models, provide efficiency gains if the correlation between existing auxiliary information and the survey data is sufficient (Pfeffermann, 2013; Rao and Molina, 2015; Tzavidis et al., 2018). In middle-income countries, administrative records are usually not sound enough and therefore censuses are the most important auxiliary source of information for the entire population, with the limitation that it is usually collected every ten years.

The time gap between annual surveys and population censuses is usually ignored in SAE methods. The use of covariates from an earlier period may lead to less reliable indicators than what would be expected from more solid auxiliary information. Academic literature on updating estimates in small areas is limited. Post censal population estimates have been obtained, for example, from traditional procedures in demography, like the component method (United Nations, 1956) and vital rates (Rao, 2003). Emwanu et al. (2006) use panel survey data to obtain small area estimates of welfare in post-census years by regressing recent income (or expenditure) data on household characteristics that are available in both survey and census data. The best-known tool in this field is the structure preserving estimation (SPREE) method, which is also the focus of this paper and will be described in detail in Section 2.3. This technique was originally introduced by Purcell and Kish (1980) to obtain post-census estimates (counts or proportions), arranged by small domains and categories of interest. SPREE have been especially applied for updating demographic information and socio-economic indicators including employment (Berg and Fuller, Berg and Fuller; Hidiroglou and Patak, 2009; Luna-Hernández et al., 2015a) and poverty (Isidro et al., 2016).

Thereafter, several versions of SPREE have been proposed with the aim of improving the method by adding flexibility and reducing bias, namely the Generalized-SPREE (GSPREE) (Zhang and Chambers, 2004) and most recently, the Multivariate-SPREE (MSPREE) (Luna-Hernández, 2016). These SPREE-techniques have specific assumptions and requirements for their implementation. For example, the variable of interest must be categorical, and it must be not only in the survey (most recent) but also in the census data, which for indicators like the poverty rate are usually not available for variables based on income or expenditure. An alternative version called Extended-SPREE (ESPREE) (Isidro et al., 2016) solves this problem by applying a small area estimation technique: the Elbers, Lanjouw and Lanjouw (ELL) method (Elbers et al., 2003), as a previous step to obtain the required information for the census year. Once the estimated census information is obtained, Isidro et al. (2016) perform the original SPREE (Purcell and Kish, 1980) to compute postcensal estimates. Moreover, Luna-Hernández (2016) showed that MSPREE is more efficient compared to SPREE (in terms of lower mean squared errors). Therefore, the current paper extends the framework of Isidro et al. (2016) by allowing for the MSPREE in the updating process. In particular, the paper aims to provide a modern methodology to a) estimate and b) update counts or proportions of relevant indicators in small areas when the information of interest is not available in the census data.

The motivation of the proposal is to offer updated reliable income-based poverty estimates of Costa Rican cantons in three mutually exclusive categories: 'extreme poor', 'poor' (not extreme), and 'not poor' in the postcensal years 2012 to 2017. Due to its political stability and good performance in general macroeconomic aspects, Costa Rica has been for several years an example among other economies in the region (OECD, 2016). Despite this, a point that draws attention and has been the object of study in recent years, is the stagnation of relative poverty that the country has had for more than two decades, unlike other Latin American countries that have achieved greater reductions in their poverty rates (CEPAL & MIDEPLAN, 2016). As well as in the international agenda, previously through the Millennium Development Goals (MDGs) (United Nations, 2015) and currently the Sustainable Development Goals (SDGs) (United Nations, 2019), one of the main concerns specifically in this country is the extreme poverty. Traditionally, and for international comparison, the National Institute of Statistic and Censuses (INEC - *Instituto National de Estadística y Censos*) of Costa Rica measures poverty based on the poverty line method. With this approach, a person or household is considered poor (or extreme poor) if its monthly per capita income is equal or below a specific poverty line. The idea of defining a threshold or line is to set the minimum amount in the per capita income that a person or household requires to satisfy food and non-food needs, included in a basket of goods and services (INEC, 2015). In Costa Rica, extreme poverty had a reduction, from 2004 to 2013, of only 0.8 percentage points in accordance with a diagnosis of structural gaps. Research showed that this status of poverty has three determinants: the adverse home and social environment, the insufficient scope of social programs and the exclusive labor market (CEPAL & MIDEPLAN, 2016). Because of a lack of data, this kind of studies can only be conducted in census years or for larger areas, limiting the possibility of applying targeted policy interventions for specific groups or domains.

As well as other middle-income economies, Costa Rica faces several limitations to obtain small area estimates of poverty: administrative records at the unit level are not available, the census does not contain income or expenditure information to compute poverty estimates via the poverty line method, and the census is carried out only every ten years which can reasonably lead to outdated poverty estimates. The main study previously conducted in Costa Rica to obtain estimates of poverty in local areas was carried out for the same year as the census, using the ELL method. Although this work, developed by Méndez and Bravo (2011), certainly allowed to obtain more detailed information about poverty in local areas (classified as poor and not poor), two aspects can be improved with the proposal presented in this paper: a) provide poverty estimates for non-censal years, and b) produce estimates on extreme poverty as a specific group of interest. The methodology proposed in this paper can also be applied to many other countries that share similar conditions and extended to other relevant demographic and socio-economic (categorical) indicators.

This paper has the following structure: Section 2.2 describes the data sets and explains the definition of poverty used in the application. Section 2.3 introduces the SPREE methods as small area estimation and updating techniques. The strategy proposed to obtain and update poverty estimates in Costa Rican cantons is also explained in this section, as well as the methodology to produce uncertainty measures. Application results are shown in Section 2.4. The results of a simulation study to validate the proposed method are presented in Section 2.5. The last section is dedicated to the conclusions and recommendations for further research.

2.2 Data description and definition of poverty

As will be explained in detail in Section 2.3, the basic approach of the SPREE techniques requires one complete cross-classification table (also known as composition), usually from a census collected in a previous time, and reliable, up-to-date population totals (margins) for the variable of interest and for the area population sizes. Extensions to this methodology including the one implemented here require an updated estimate of the cross-classification of interest that can be obtained from survey data. The aim of this section is to describe the data sources available and explain the definition of poverty considered in the application. Population and housing census, as well as the National Household Survey data sets, were provided by the INEC of Costa Rica, under specific confidential agreements.

2.2.1 Population and housing census 2011

In Costa Rica, the Population and Housing Censuses are carried out every ten years by the INEC. The most recent census (*X^{th} Censo Nacional de Población y VI^{th} de Vivienda* was conducted in 2011 (data collection from 30th, May to 3rd, June) and it collected information of people, households, and dwellings necessary for the planning, execution, and evaluation of public policies (INEC, 2012). With the information collected, it is possible to identify the relevant characteristics of the population such as access to education, employment, social security, technology, and health centres. Although the census 2011 includes questions to compute the unsatisfied basic needs (UBNs) index (Feres and Mancero, 2001), it did not produce information about income or expenditures of the persons and households, which are necessary to calculate the incidence of poverty via the poverty line method. The sampling frame which is needed to conduct national surveys and other statistical operations is constructed based on this population and housing census. With this census, 10,461 primary sampling units (PSUs) and 1,359,168 dwellings were identified.

2.2.2 National household surveys 2011-2017

The National Household Survey (ENAHO - *Encuesta Nacional de Hogares*) is the primary source for poverty and inequality measures in Costa Rica. In this aspect, this survey collects information about housing characteristics, education, social security and employment of the household members. This study is carried out annually (data collection during the month of July). Surveys from 2011 to 2014 used the sampling framework from the previous census 2000, the following surveys used the sampling framework updated with the census 2011. The sampling design used in the ENAHO is a two-stage stratified random sampling where census segments are the first stage units selected with probability proportional to size, and dwellings are defined as the final stage units. Administratively, Costa Rica has four disaggregation levels: two zones, six planning regions, 82 cantons and 473 districts (municipalities). The sampling design specifies twelve strata - each planning region divided by urban and rural areas. In this case, the strata coincide with the study domains. Smaller domains are not considered to guarantee a coefficient of variation less than 15% for the main poverty measure (percentage of household under poverty) (INEC, 2017). For 2011, the ENAHO selected 1120 PSUs and

13,440 dwellings (10.7% and 9.9% of the sampling framework, respectively).

There are two main differences across the survey rounds as well as regarding the census data. On the one hand, the last surveys collected more variables than the previous ones and the census, and the definition of some of the variables also changed. On the other hand, there were land reforms in the period of study: in 2011 there were only 81 cantons but in 2017 another canton was created. To deal with these obstacles, only variables that exist with the same definition in both the census and survey data are included for the analysis. Also, for the last survey, cantons were grouped exactly the same as in the census 2011. This straightforward-aggregation is possible because the new canton was created by dividing one of the existing cantons. In this paper, the objective is to obtain quality poverty information of households for the third administrative level as defined in 2011, i.e., the 81 cantons, which are defined as the target small areas.

2.2.3 Demographic projections

In order to apply a SPREE technique, it is necessary to provide reliable and up-to-date area totals as it will be explained in detail in Section 2.3. Since survey data usually does not produce trustworthy population sizes for small areas (due to sample sizes and out-of-sample areas), demographic projections are used instead in this paper. In Costa Rica, population projections are calculated with the cohort component method (Preston et al., 2001) which considers changes in three components: mortality, fertility and migration. The mortality projection was carried out with an autoregressive integrated moving average (ARIMA) random walk model with drift, and for the fertility and migration components, functional data analysis models were implemented. Further details can be found in INEC & CCP (2013).

Because of population projections in Costa Rica consider persons at an aggregate (e.g., canton) level, but in this application the aim is to update the total of households according to their status of poverty, the headship rate (United Nations, 1973) by sex and age groups is applied in order to get the household projections. A previous implementation of this methodology in this country can be found in Sáenz (2002).

2.2.4 Definition of poverty

In Costa Rica, poverty is measured under different uni- and multidimensional approaches. One of the most important, and that is the focus in this paper, is the (monetary) poverty rate which is based on the poverty line method (using non-equivalised household per capita income). The INEC defines two types of lines or thresholds:

- The indigence or extreme poverty line: set by the per capita costs of a basic food basket. The composition of this basket is defined from the national survey on Income and Expenditure of the Households (ENIGH - *Encuesta Nacional de Ingresos y Gastos de los Hogares*) which is carried out every five years. The value of the basket is updated every month based on the consumer price index. If the monthly per capita income of a household is below this line, it is considered under ‘extreme’ poverty. For the census time (July 2011), the indigence line was 39,428 colones (Costa Rican currency) (INEC, 2011) which was 27.9% of the median per capita income at that time.

- The poverty line: considers additionally non-food basic needs. A household is classified in this category if the monthly per capita income is equal to or below this value but higher than the indigence line. The poverty line in July 2011 was 84,006 colones which is 60.7% of the median per capita income (Méndez and Bravo, 2011).

In this paper, the number of households grouped in three categories of poverty (‘extreme poor’, ‘poor’ (not extreme) and ‘not poor’) is estimated and updated for six postcensal years (2012-2017) in 81 cantons.

2.3 Methodology

This section describes the methodology for estimating and updating counts and proportions for small areas. First, SPREE techniques are introduced because they are the basis of this proposal. Second, the recommended strategy for obtaining and updating estimates is explained. Finally, the steps to produce uncertainty measures are described.

2.3.1 Structure preserving estimation (SPREE) methods

As stated above, SPREE was originally proposed by Purcell and Kish (1980) as a tool to update counts or proportions of a categorical variable of interest according to study domains in intercensal years. The target information of interest in a recent time t_1 is shown as a multi-way contingency table Y_{aj,t_1} grouped by $a = 1, \ldots, A$ areas or domains (rows) and $j = 1, \ldots, J$ categories of the variable of interest (columns) (e.g., poverty status). In other words, for a population of size N, all the units i (e.g., individuals or households) are organised according to the area and category to which they belong. The structure of a two-way contingency table can be represented as a saturated log-linear model:

$$\log Y_{aj,t_1} = \alpha^Y_{0,t_1} + \alpha^Y_{a,t_1} + \alpha^Y_{j,t_1} + \alpha^Y_{aj,t_1}. \tag{2.1}$$

The terms $\alpha^Y_{0,t_1}, \alpha^Y_{a,t_1}, \alpha^Y_{j,t_1}$ and α^Y_{aj,t_1} can be defined using a centred-constraint parametrisation, see e.g., Luna-Hernández (2016). Notice that data from a census (time t_0) can also be arranged as a contingency table (Z_{aj,t_0}) and represented as a saturated log-linear model (logZ_{aj,t_0}) with the same constraints.

For intercensal years, the production of official statistics relies in many cases on survey data. However, due to sample size limitations, trustworthy results are only available for big areas. The SPREE method provides a solution when updated estimates of frequency characteristics are required in smaller domains. The terms $\alpha^Y_{0,t_1}, \alpha^Y_{a,t_1}$ and α^Y_{j,t_1} represent the *allocation* structure which are benchmarked to totals or current margins (A row and J column totals), usually provided by direct survey estimates and/or demographic projections. In this paper, for simplicity, the allocation structure is also referred to as the survey margins, although it is made up of both demographic projections and survey data. It is assumed that these totals are reliable and updated for postcensal years. Furthermore, the method supposes that the interactions between rows and columns of the census Z_{aj,t_0} (inner cells in the contingency table) remain unchanged for the target years. Therefore, the structural assumption is;

$$\alpha^{Y}_{aj,t_1} = \alpha^{Z}_{aj,t_0}. \tag{2.2}$$

This interaction term provided by the census is usually known as the *association* structure.

Following the proposal of Purcell and Kish (1980), the updated estimates are obtained via the iterative proportional fitting (IPF) algorithm (Deming and Stephan, 1940) (also found in literature as *raking* or *contingency table standardisation*). Detailed description of the IPF is available e.g., in Bishop et al. (2007) and Zaloznik (2011). This algorithm fits a census table by keeping reliable survey margins fixed. The process to obtain the updated SPREE of Y_{aj,t_1} is represented by Luna-Hernández (2016) as:

$$\hat{Y}^{S}_{aj,t_1} = \text{IPF}\big[\exp(\hat{\alpha}^{Y}_{aj,t_1}), Y_{a,t_1}, Y_{j,t_1}\big], \tag{2.3}$$

with the superscript *S* to denote that SPREE was applied, Y_{a,t_1} and Y_{j,t_1} represent the reliable survey margins (rows and columns) and $\hat{a}^{Y}_{aj,t_1} = a^{Z}_{aj,t_0}$.

As it will be later described, new versions of SPREE estimators have been proposed and each of them define different assumptions on their association structure $\hat{\alpha}^{Y}_{aj,t_1}$. The process in 2.3 is applied with the defined association structure of each SPREE-type estimator.

In order to apply a fitting strategy via IPF, Koebe et al. (2022) summarise some basic requirements that should be considered:

1. The data to fit must be arranged in categories (e.g., contingency tables).

2. The margins of the census and survey structures must have same length (same number of rows and columns).

3. Totals by rows and columns must be equal.

4. The census data (association structure) must contain the indicator of interest (e.g., poverty status) with the same definition or a highly correlated indicator (Green et al., 1998).

In practice, requirements two and four are not met in the Costa Rican scenario. Due to administrative reforms that occurred in postcensal years (e.g.,merge or split domains), the number of local areas in the census and in the surveys differs. Another situation where requirement two may not be fulfilled is when some areas were not selected in the sample, leading to an incomplete allocation structure. For these cases, several solutions can be considered: complimentary information such as administrative registries or population projections could be used as reliable survey margins, rows that are not in the survey composition can be eliminated from the census, or adding missing rows with small values (e.g., 0.0001) to re-construct the survey compositions in the same way as defined in the census. Since population projections are available, the first alternative is implemented in the current example.

Regarding the fourth requirement, income or poverty information is not obtained directly in the Costa Rican census. A solution for this kind of situation was proposed by Isidro et al. (2016). The so-called ESPREE considers the case when the indicator of interest is not present in the census data. Another small area estimation method, the ELL method, is applied as a first step with the aim of obtaining the required information for the census year. Thereafter, the original SPREE technique is conducted as a second step to proceed with the updating process. Considering the characteristics of the case study exposed in this paper, the methodology of ESPREE is followed to obtain updated poverty estimates.

Another line of research within the SPREE framework is the bias reduction of the estimator. The GSPREE introduced by Zhang and Chambers (2004) makes the structural assumption of SPREE more flexible by adding a *proportionality coefficient* (constant obtained through direct estimates $\hat{Y}^{Dir}_{aj,t_1}$) adjusting Equation 2.2 to:

$$\alpha^{Y}_{aj,t_1} = \beta\alpha^{Z}_{aj,t_0}. \tag{2.4}$$

Note that the SPREE assumption defined previously in Equation 2.2 is the case of Equation 2.4 when $\beta = 1$. Moreover, Luna-Hernández (2016) proposes a version that aims to relax two restrictions of the former methods (SPREE and GSPREE): the relationship between the association structures (interaction terms) of census and target compositions are controlled only by one parameter, and in the case of the GSPREE the proportionality parameter β, is assumed to be the same for all the categories. GSPREE and MSPREE mainly differ from the original SPREE because a multi-way contingency table of direct estimates is also necessary in order to update the association structure, meanwhile the former version only requires the availability of suitable (total) margins. The novel method is called *multivariate* SPREE (MSPREE) because in this case, the target compositions are the interactions within each area. Then, the coefficient $\boldsymbol{\beta}$, similar to the proportionality coefficient specified by Zhang and Chambers (2004), now is represented by a $J \times J$ matrix (with $(J-1) \times (J-1)$ free parameters), and varies inside each area, from one category to another. The main benefit of this proposal is to be able to capture better relationships between categories, instead of assuming that these interactions remain identical over time. With this, the bias that can occur through changes in the association structure (which is not accounted in SPREE and GSPREE), is reduced. Similarly as in Equations 2.2 and 2.4, the MSPREE structural assumption is expressed as:

$$\begin{bmatrix} \alpha^{Y}_{a1} \\ \vdots \\ \alpha^{Y}_{aJ} \end{bmatrix} = \begin{bmatrix} \beta_{11} & \cdots & \beta_{1J} \\ \vdots & \ddots & \vdots \\ \beta_{J1} & \cdots & \beta_{JJ} \end{bmatrix} \begin{bmatrix} \alpha^{Z}_{a1} \\ \vdots \\ \alpha^{Z}_{aJ} \end{bmatrix}$$

which is equivalent to:

$$\boldsymbol{\alpha}^{Y}_{a,t_1} = \boldsymbol{\beta}\boldsymbol{\alpha}^{Z}_{a,t_0}, \tag{2.5}$$

where $a = 1, ..., A$ areas, and for each area, the interaction terms: $\boldsymbol{\alpha}^{Y}_{a,t_1} = \left(\alpha^{Y}_{a1,t_1}, \dots, \alpha^{Y}_{aJ,t_1}\right)$ and $\boldsymbol{\alpha}^{Z}_{a,t_0} = \left(\alpha^{Z}_{a1,t_0}, \dots, \alpha^{Z}_{aJ,t_0}\right)$. The target MSPREE composition $\hat{Y}^{M}_{aj,t_1}$ can also be obtained via IPF as in Equation 2.3. In the same way as in the original SPREE and GSPREE, the reliable total margins are preserved, but the same list of requirements aforementioned must be fulfilled.

Estimates of $\boldsymbol{\beta}$ can been obtained via maximum likelihood (ML) or iterative weighted least squares (IWLS), in both cases by using a log or a logit link. Because a target contingency table can be represented as a log-linear model, SPREE fits within the framework of generalised linear models (Marker, 1999; Noble et al., 2002). Moreover, as Agresti (2002) indicates, a log link with a Poisson response is commonly applied to model cell counts in contingency tables, but he also shows that the Poisson expected frequencies μ_{aj,t_1} are equal to $n\pi_{aj,t_1}$, where π_{aj,t_1} are the cell probabilities used under multinomial sampling and n the sample size. For this reason, when working with log-linear models (as SPREE), the estimation of coefficients $\boldsymbol{\beta}$ can be obtained with a Poisson $\hat{Y}^{Dir}_{aj,t_1}|\alpha^{Z}_{a,t_0} \overset{\text{ind}}{\sim} \text{Poisson}(\mu_{aj,t_1})$ or multinomial $\hat{Y}^{Dir}_{a,t_1}|\alpha^{Z}_{a,t_0} \overset{\text{ind}}{\sim} \text{Multinomial}(\sum_{j=1}^{J}\hat{Y}^{Dir}_{aj,t_1}, \pi_{a,t_1})$ distribution, both leading to similar results. Notice that, similarly to the GSPREE, the computation of $\boldsymbol{\beta}$ requires direct estimates (i.e., not only survey margins but a survey composition $\hat{Y}^{Dir}_{aj,t_1}$).

The second alternative to obtain estimates of $\boldsymbol{\beta}$ is using IWLS. This algorithm is implemented in Zhang and Chambers (2004) and Luna-Hernández (2016) suggests it when the sample is drawn using a complex sampling design because using ML for the parameter estimation can lead to misspecification if sampling design information is ignored. In the case of the survey data used in this paper, the sample was gathered in a two-stage selection process considering unequal selection probabilities, and for this reason, a fully distributional approach should not be assumed. Consequently, in this paper, the parameters of interest are estimated via IWLS. This method requires an estimate of the variance-covariance matrix of the target composition $\hat{Y}^{M}_{aj,t_1}$ which is usually not available. Luna-Hernández (2016) solves this issue by multiplying a design effect with the variance that corresponds to a simple random sample without replacement design. This proposal assumes that samples are independently selected in each area and there are no existing correlations among estimates from different areas. The estimate of the variance-covariance matrix is represented as:

$$\hat{V}_{aj,t_1} = \frac{\text{deff}_{j,t_1}\ \hat{\pi}^{M}_{aj,t_1}(1-\hat{\pi}^{M}_{aj,t_1})}{n_{a,t_1}}, \tag{2.6}$$

with $\hat{\pi}^{M}_{aj,t_1} = \frac{\hat{Y}^{M}_{aj,t_1}}{Y_{a,t_1}}$ and n_{a,t_1} the area sample sizes. Further details about the IWLS algorithm can be consulted in Jiang (2007) and Luna-Hernández (2016).

2.3.2 The empirical best predictor method

The empirical best predictor (EBP) is applied in order to obtain the information of poverty status in the census structure and satisfy requirement four. The EBP methodology proposed by Rao and Molina (2015) implements a unit-level nested error regression model to get estimates of a specific variable of interest in the census, using (usually) survey data that contains this variable. This method has been extensively implemented in SAE problems and also specifically for poverty estimation (see e.g., Pratesi (2016) or Das and Haslett (2019)). The process assumes a random effects model for a finite population of size N:

$$y_{ai} = \mathbf{x}^{T}_{ai}\boldsymbol{\beta} + u_a + e_{ai},$$

where y_{ai} represents the target variable and $\mathbf{x}_{ai}$ the set of covariates for the ith individual or household in ath area, u_a indicates the area-specific random effects and e_{ai}, the unit-level error. The two last terms are assumed to be normal, independent and identically distributed. By using survey data, the following estimates are obtained: $\hat{\boldsymbol{\beta}}, \hat{\sigma}_u^2, \hat{\sigma}_e^2$ and the weighting factors $\hat{\gamma}_a = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \frac{\hat{\sigma}_e^2}{n_a}}$, where n_a denotes the sample size in area a, and σ_u^2 and σ_e^2 indicates the between and within group variance respectively.

Then, $l = 1, \ldots, L$ Monte Carlo simulations to generate a pseudo population are conducted:

$$y_{ai}{}^{(l)} = \mathbf{x}_{ai}^T \hat{\boldsymbol{\beta}} + \hat{u}_a + v_a^{(l)} + e_{ai}^{(l)}, \tag{2.7}$$

where $v_a^{(l)} \overset{iid}{\sim} N\big(0, \hat{\sigma}_u^2(1-\hat{\gamma}_a)\big)$ and $e_{ai}^{(l)} \overset{iid}{\sim} N\big(0, \hat{\sigma}_e^2\big)$ and the predicted random effect $\hat{u}_a$ is defined as $\hat{u}_a = E(u_a|\mathbf{y}_{as})$. The subscript s in $\mathbf{y}_{as}$ denotes the in-sample elements. The final indicator of interest is obtained taking the mean over the L iterations.

Molina and Rao (2010) explain that the EBP estimator can be biased when model error terms depart from normality. When working with income data, a common practice to achieve Gaussian assumptions is by using a logarithmic transformation (Elbers et al., 2003; Molina and Rao, 2010) which is a special case of the Box-Cox transformation (Box and Cox, 1964). In this case study, as will be explained in detail in Section 2.4.1, the incidence of poverty is approximated by modeling income. For this reason, departures from normality are reduced with a Box-Cox transformation. Further details on the performance of the EBP under data-driven transformations can be found in Rojas-Perilla et al. (2020). The EBP was conducted using R, specifically with the Package **emdi** (Kreutzmann et al., 2019).

2.3.3 Strategy to estimate and update poverty estimates in Costa Rican cantons

The goal of this paper is to obtain and update poverty estimates in three categories: 'extreme poor', 'poor' (not extreme), and 'not poor'. However, as noted previously, no poverty information is collected directly from the census, nor income or expenditure data. Thus, the applied methodology considers characteristics of some of the SPREE methods, specifically the ESPREE and the MSPREE. Since the census data does not contain poverty information, this paper adjusts the ESPREE framework. Instead of applying an ELL model to estimate poverty in the census data followed by the original SPREE to update the counts as in Isidro et al. (2016), the EBP (Rao and Molina, 2015) is implemented followed by the MSPREE in this work (for simplicity, also referred as EBP-MSPREE). To the best of my knowledge, MSPREE is the most recent and complete technique mentioned in SPREE literature. This version provides more flexibility and bias reduction compared with the previous versions, therefore it is implemented as the main tool in the updating part of the process.

The estimation and updating strategy can be summarised in the following steps:

1. **Estimating the proxy association structure via EBP.** Considering that census and survey data, both for the same year at individual level are available, the EBP explained in Section 2.3.2 is applied in order to obtain the information of poverty status in the census

structure.

2. **Obtaining the allocation structure.** Household projections as mentioned in Section 2.2.3 provide the total of households in each canton (row margins) and survey data, described in Section 2.2.2, the total of households by poverty status. Both sources are available for postcensal years (2012-2017).

3. **Updating estimates via MSPREE.** Intercensal EBP-MSPREE compositions $\hat{Y}^{EM}_{aj,t_1}$ are estimated considering the outputs from the two previous steps: association structure α^{Z}_{aj,t_0} from Step 1, and row Y_{a,t_1} and column Y_{j,t_1} margins from Step 2, that represent the allocation structure. Direct estimates from survey data and design effects are also required as explained at the end of Section 2.3.1. With these inputs, the procedure can be described as follows:

   (a) The matrix of coefficients $\boldsymbol{\beta}$ is estimated with an IWLS algorithm that requires a variance-covariance matrix. This matrix is approximated using design effects as showed in Equation 2.6 with EBP composition estimated in t_0 and the direct estimate obtained from the survey in t_1.

   (b) $\hat{\alpha}^{Y}_{aj,t_1}$ is estimated with Equation 2.5.

   (c) Taking into account all these elements, the target estimate $\hat{Y}^{EM}_{aj,t_1}$ is finally obtained with Equation 2.3.

2.3.4 Uncertainty of the updated estimates

The benefits of two SPREE versions are used in this paper. To motivate the proposed uncertainty measure $\widehat{\text{MSE}}\big(\hat{Y}^{EM}_{aj,t_1}\big)$, the procedures via bootstrap of the predecessors (MSE of ESPREE and MSPREE) are briefly described in this section. Details about other approaches, e.g., via linearisation methods can be found in Isidro (2010), Isidro et al. (2016) and Luna-Hernández (2016).

Two sources of variation are considered when obtaining estimates via the ESPREE method: survey data (allocation structure) and pseudo-populations (association structure). Being $\hat{Y}^{E}_{aj,t_1}$ the ESPREE estimates, the uncertainty estimate is the sum of two variances: $\text{Var}(\hat{Y}^{E}_{aj,t_1}) = \text{Var}^{\text{survey}}\big(\hat{Y}^{E}_{aj,t_1}\big) + \text{Var}^{\text{census}}\big(\hat{Z}^{E}_{aj,t_1}\big)$. The first variance term Var$^{\text{survey}}$ is obtained by generating $b = 1, \dots, B$ independent bootstrap samples from the original survey data and computing $\hat{Y}^{E,b}_{aj,t_1}$ by keeping fixed the census data (i.e., association component) in every replication.

The second term Var$^{\text{census}}$ is obtained in a similar way. In this case, B ESPREE estimates $\hat{Y}^{E,b}_{aj,t_1}$ are compute based on $b = 1, \dots, B$ bootstrap populations or pseudo-census to account for the uncertainty provided by the association structure, and the allocations structure (survey margins) will be held fixed over the replications. Because in ESPREE, the census values are estimated via ELL (or EBP), the $b = 1, \dots, B$ pseudo-census generated from the ELL process can also be used here. Both, $\widehat{\text{Var}}^{\text{survey}}$ and $\widehat{\text{Var}}^{\text{census}}$ are unconditional variances. As aforementioned, the final uncertainty estimation is only the sum of these two terms, meaning that the authors assume that there is no covariance between both estimators.

Regarding MSPREE uncertainty measures of $\hat{Y}^{M}_{aj,t_1}$, Luna-Hernández (2016) proposed three alternatives: an analytical approximation for the variance of the estimator, a Finite Population MSE (FP-MSE) and an Unconditional MSE (U-MSE). For simplicity and consistency, the last one is described here. From the point estimate $\hat{Y}^{M}_{aj,t_1}$, calculate the within-area proportions $\hat{\pi}_{a,t_1} = \frac{\hat{Y}^{M}_{aj,t_1}}{Y_{a,t_1}}$ for each area $a = 1, \ldots, A$. Generate B-bootstrap populations Y^{*b}_{aj,t_1} under the assumption that the target estimate across areas has the following distribution: $\mathbf{Y}_{a,t_1}|\boldsymbol{\alpha}^{Z}_{a,t_0} \overset{\text{ind}}{\sim} \text{Multinomial}\Big(Y_{a,t_1}, \hat{\pi}_{a,t_1}\Big)$. From each bootstrap population draw a sample and follow the steps of Section 2.3.1 to obtain $\hat{Y}^{M,b}_{aj,t_1}$. Finally, compute $\widehat{\text{U-MSE}}\big(\hat{Y}^{M}_{aj,t_1}\big) = \frac{1}{B}\sum_{b=1}^{B}\big(\hat{Y}^{M,b}_{aj,t_1} - Y^{*b}_{aj,t_1}\big)^2$.

The two procedures to obtain the MSE of $\hat{Y}^{E}_{aj,t_1}$ and $\hat{Y}^{M}_{aj,t_1}$, are briefly described to better contextualise the MSE proposed in this paper, and to point out the differences and similarities between them. Thus, results from these two previously described MSEs are not produced in this paper.

Because the estimation and updating process in this application makes use of the EBP followed by the MSPREE, the sources of uncertainty in all of the steps should be considered. The idea of including the variation of each element is taken from Isidro et al. (2016). The objective is to contemplate that the MSE of the EBP-MSPREE entails three sources of uncertainty: allocation structure, association structure, and the yearly association updating. The difference in this proposal is that instead of calculating each variability term and adding them ($\text{Var}^{\text{survey}} + \text{Var}^{\text{census}} + \text{Var}^{\boldsymbol{\beta}}$), a single bootstrap procedure will be performed, varying each of the required elements in each replication. This decision is made since it cannot be denied that covariances between the estimators exist. This aspect requires special attention and it should be study in further investigations. The steps to obtain the MSE for the EBP-MSPREE are as follows:

1. From Equation 2.7, L Monte Carlo pseudo populations in t_0 are generated. Based on defined thresholds (poverty lines), L cross-classified population tables are created with dimension: A areas and J categories of poverty. Notice that the average across them was used to defined α^{Z}_{aj,t_0} and finally compute the point estimate $\hat{Y}^{EM}_{aj,t_1}$. Now, these L pseudo populations are used to create $\alpha^{Z,b}_{aj,t_0}$ (with $L = B$), to account for the uncertainty that this structure provides, similarly as in Isidro (2010) and Isidro et al. (2016).

2. To take into consideration the uncertainty from the allocation structure, generate B pairs of margins $Y^{b}_{a,t_1}, Y^{b}_{j,t_1}$ from the point estimate $\hat{Y}^{EM}_{aj,t_1}$ assuming a multinomial distribution, i.e., $Y^{b}_{a,t_1} \overset{\text{ind}}{\sim} \text{Multinomial}(\tilde{Y}, \tilde{\pi}_a)$, where $\tilde{\pi}_a = \frac{\hat{Y}_{a,t_1}}{\tilde{Y}}$ and $\tilde{Y} = \sum_{a=1}^{A}\hat{Y}_{a,t_1}$.

3. The uncertainty due to the estimation of $\boldsymbol{\beta}$ required in the MSPREE is obtained following the procedure of the U-MSE previously described: From the point estimate $\hat{Y}_{aj,t_1}$ calculate the within-area proportions $\hat{\pi}_{a,t_1} = \frac{\hat{Y}^{M}_{aj,t_1}}{Y_{a,t_1}}$ for each area $a = 1, \ldots, A$. Generate B bootstrap populations Y^{*b}_{aj,t_1} under the assumption that the target estimate across areas has the following distribution: $\mathbf{Y}_{a}|\boldsymbol{\alpha}^{Z}_{a,t_0} \overset{\text{ind}}{\sim} \text{Multinomial}\Big(Y_{a,t_1}, \hat{\pi}_{a,t_1}\Big)$. From each bootstrap population, draw a sample to get $y^{EM,b}_{aj,t_1}$.

4. With the output of Step 1 and 3, specify the MSPREE structural assumption $\hat{\boldsymbol{\alpha}}^{Y,b}_{a,t_1}$ as in Equation 2.5.

5. With $L = B$ and the output of Step 2 and 4, compute B EBP-MSPREE estimates

$$\hat{Y}^{EM,b}_{aj,t_1} = \text{IPF}\big[\exp\big(\hat{\boldsymbol{\alpha}}^{Y,b}_{a,t_1}\big), Y^{b}_{a,t_1}, Y^{b}_{j,t_1}\big].$$

6. Finally, estimate the MSE:

$$\widehat{\text{MSE}}\big(\hat{Y}^{EM}_{aj,t_1}\big) = \frac{1}{B}\sum_{b=1}^{B}\big(\hat{Y}^{EM,b}_{aj,t_1} - Y^{*b}_{aj,t_1}\big)^2. \tag{2.8}$$

2.4 Results of the application

In this section, the results of the estimation and the updating process are explained. First, the EBP model for obtaining the association structure is described with its corresponding model evaluation and descriptive statistics. Second, some relevant results in the practical sense are shown, e.g., the evolution of (updated) poverty indicators from 2012 to 2017 for selected cantons, and the cantons with the highest poverty rates in the last year of study. Finally, uncertainty measures and validation results are presented.

2.4.1 Poverty estimates for the census

An EBP model to obtain poverty incidence by domain was conducted by setting income per capita of the household as the dependent variable and the socio-economic covariates described in Table 2.1 as predictors. In Costa Rica, urbanity plays an important role in socioeconomic topics. For this reason, the variable zone (urban/rural) is added to the model. The other groups of variables contain information on the head of the household, the household members, and the housing. Most of them have been used in previous studies as predictors of poverty or are part of indexes such as the UBNS and the multidimensional poverty index (MPI) (Alkire and Foster, 2007). Under the model specified here, the sample design is assumed non-informative.

To select the model, several transformations are considered to reduce normality departures of the error terms, but the final version applies a Box-Cox transformation (Box and Cox, 1964) with an optimal lambda (0.1585) using the restricted maximum likelihood (REML) approach. Regarding the normality assumptions of this linear mixed model, formal tests and graphical diagnostic of the residuals are used. Figure 2.1 shows that normality assumptions are rejected for the unit level (Skewness -0.049 and Kurtosis 6.418), but not for the random effects (canton-level). The latter is also confirmed with the Shapiro Wilk test (W= 0.986 and p = 0.529), and Skewness (-0.359) and Kurtosis (3.065) measures. For this example, normality for both, the unit level and the random effects, is assumed. The use of the Box-Cox transformation helps, at least, to get more symmetrical tails (Rojas-Perilla et al., 2020) than other (or no) transformations. Also, the marginal $R^2 = 0.500$ and the conditional $R^2 = 0.508$ were observed.

Table 2.2 shows the summary statistics of the population and sample domains. For 2011, all cantons are in-sample although it is not the case for some postcensal years. Domain sizes

Table 2.1: Covariates included in the EBP model to obtain poverty estimates in census

Category	Variable
Geographical	1. Zone
Head of the household	2. Age
	3. Highest degree of education completed
	4. Sex
	5. Labor condition
Household conditions	6. Proportion of employees in the household
	7. Equivalized size of the household
	8. Overcrowding
	9. At least one member without health insurance
	10. Quantity of economically dependent members
	11. At least one member not attending to formal education
	12. At least one member not with a educational lag
Housing conditions	13. Poor condition of the floor or ceiling
	14. Any member not used internet last 3 months
	15. No garbage disposal system
	16. No exclusive toilet for the household

Notes: Labor condition has three categories: employed, unemployed, out of labor force.
Variable 9 refers to population older than 17 years old.
Variables 11 and 12 refer to population between 5 and 19 years old.
Variable 14 refers to population older than 4 years old.

from the survey data vary from 12 to 877 households, and in the case of the census data, it varies from 1705 to 84,066 households,

In order to study poverty in the three aforementioned interest groups, the two poverty lines described in Section 2.2.4 were implemented (as ‘customised indicators’ in the R package **emdi**, for further details about this functionality see Kreutzmann et al. (2019)). Descriptives of coefficients of variation (CV) for the direct- and model-based estimates obtained via EBP are presented in Table 2.3. As expected, the CVs reflect the lack of precision in the categories ‘extreme poor’ and ‘poor’ (not extreme) for the direct estimates. The improvement when the EBP model is conducted is clear, with a maximum CV of 29.9% and 18.2% for ‘extreme poor’ and ‘poor’ (not extreme) categories respectively. Notice also, that the third quartile of the CV in the category ‘extreme poor’ is below 20%.

Figure 2.2 shows the proportion of households under extreme poverty based on the direct estimates (ENAHO 2011) and the model-based estimates obtained via the EBP model. The

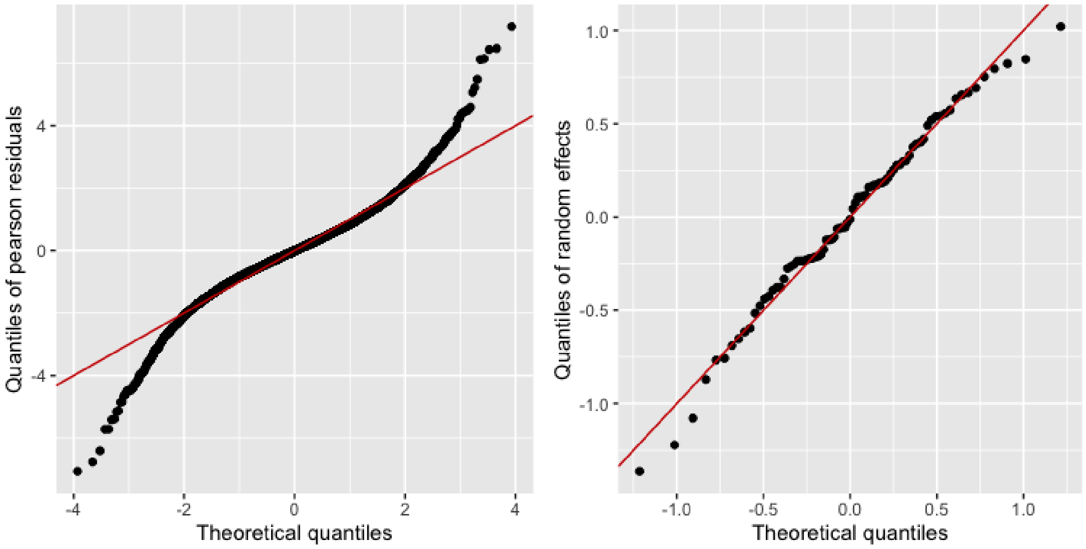


Figure 2.1: Q-Q plots of the unit-level errors and the random effects

Table 2.2: Summary statistics for sample and population sizes

	Min	1st Q	Median	Mean	3rd Q	Max
Sample domains (In-sample: 100%)	12	62	103	144.37	191	877
Population domains	1705	5961	11032	15271.37	17148	84066

maps on the right side give a closer look at the metropolitan area which consists of 31 cantons, covering approximately 60% of the population.

In four of these cantons, no households were identified as ‘extreme poor’ with direct estimates (out-of-sample domains represented in black in Figure 2.2 (a) and (b)). The results are consistent with a previous study (Méndez and Bravo, 2011), where higher levels of poverty are found on the border with Nicaragua (e.g., La Cruz, located in the northwest of the country) and on the border with Panamá (e.g., Buenos Aires, and Talamanca, located in the southeast of the country).

2.4.2 Updated poverty estimates

For the intercensal years, estimates on incidence of extreme poverty are obtained with EBP-MSPREE. To analyse major changes in this indicator between 2011 and 2017 Z-scores are used:

$$Z = \frac{\text{Estimate}_{2011} - \text{Estimate}_{2017}}{\sqrt{(\text{Standard error}_{2011})^2 + (\text{Standard error}_{2017})^2}}.$$

This measure represents the standardised distance between the estimates in both years (Isidro, 2010).

Figure 2.3 presents the three cantons with biggest change in this category of poverty.

Table 2.3: Coefficients of variation of the direct and model-based estimates for poverty status

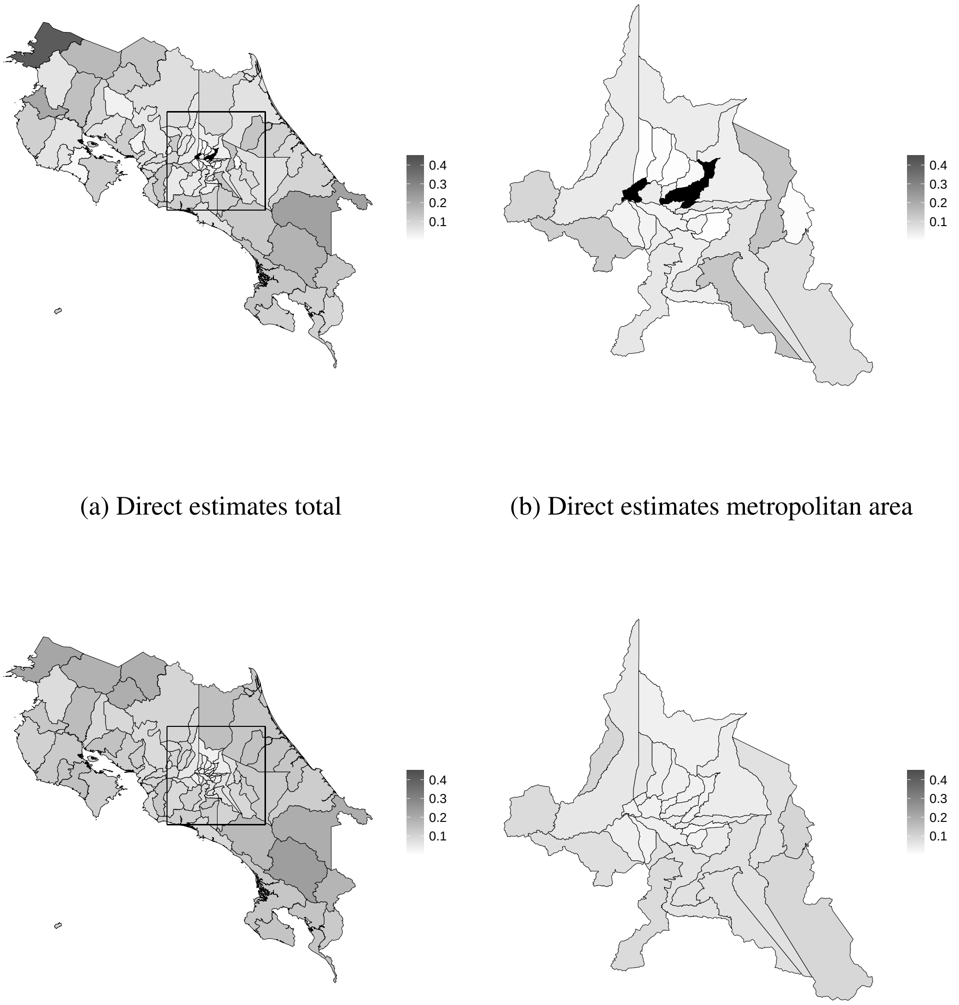
	Direct			EBP		
CV	Extreme	Poor	Not Poor	Extreme	Poor	Not Poor
Min	0.068	0.071	0.033	0.085	0.045	0.012
1st Q.	0.232	0.164	0.086	0.127	0.065	0.023
Median	0.332	0.323	0.110	0.155	0.085	0.033
Mean	0.412	0.271	0.123	0.160	0.089	0.033
3rd Q.	0.524	0.234	0.145	0.181	0.106	0.040
Max	1.000	1.000	0.281	0.299	0.182	0.062

Among the 81 cantons, these are the cantons with absolute Z-scores higher than two. Note that all of them show a reduction in the incidence of poverty between 2011 and 2017 meaning that the biggest changes in this period were actually reductions. It is also important to mention that none of the three cantons are among the poorest, which means that the biggest improvements are not observed in the areas most in need. On the contrary, the cantons Curridabat and Montes de Oca had the highest growth in extreme poverty (although small Z-score values: 0.223 and 0.205, respectively). However, both are among the cantons with the lowest incidence of extreme poverty in both years.

Identifying the poorest cantons is also relevant in order to fight against this phenomenon in a more efficient way. Figure 2.4 shows the small areas with the highest incidence of extreme poverty (in proportions) in the last year (2017). Here, it is important to point out that for all the years of study (2011-2017) the same five cantons remain on this list, indicating that economic conditions of these areas have not been better in comparison with other areas in recent years.

The sources of uncertainty that were explained in Section 2.3.4, are displayed for the last year of study as a coefficients of variation in Figure 2.5. As expected, the category ‘not poor’ is the one with the minimum CV and most of the values are under 20% which is considered ‘acceptable’ according to the parameters for official publications of the national statistical office of Costa Rica (INEC, 2015).

As explained in Section 2.2.2, the target areas in this paper, the 81 cantons, are nested in six planning regions. Because the INEC of Costa Rica publishes official results on poverty only for these planning regions (gathered from the ENAHO), this is the only geographical level where it is possible to make the comparison with the updated estimates via EBP-MSPREE. For this reason, as a way to evaluate the updated estimates of poverty, model-based estimates of cantons are aggregated into the six planning regions and compared with the official publication. It is relevant to mention that three cantons overlap with two regions at the same time. This problem was solved by allocating the estimated counts in proportion to the respective population in each region. For a more practical comparison, proportions instead of counts are shown in Table 2.4. EBP-MSPREE results are satisfactory in terms of their similarity to the direct estimates. Most of the regions show close results to the published one, and the region with the highest discrepancies is the Pacífico Central. This region, however, is the domain with a smaller sample



(a) Direct estimates total

(b) Direct estimates metropolitan area

(c) Model-based estimates total

(d) Model-based estimates metropolitan area

Figure 2.2: Proportion of extreme poverty: direct and model-based estimates, 2011

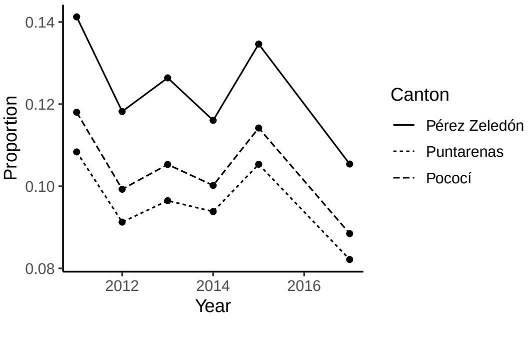


Figure 2.3: Cantons with biggest change in the incidence of extreme poverty from 2011 to 2017

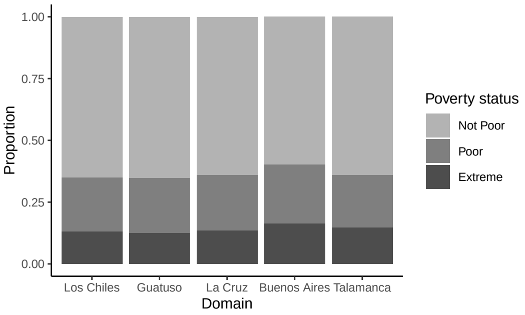


Figure 2.4: Cantons with highest incidence of extreme poverty, 2017

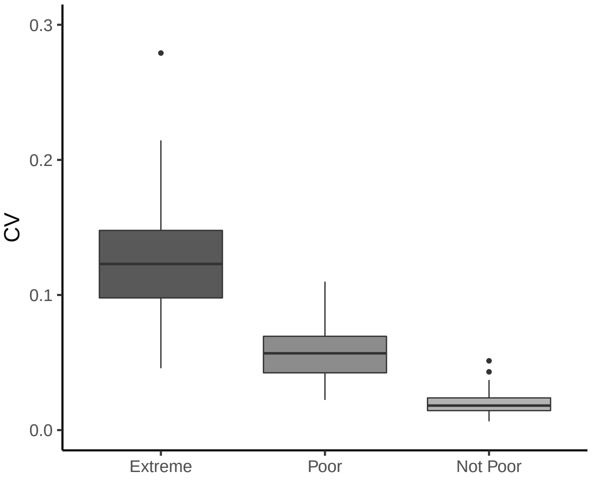


Figure 2.5: Coefficients of variation for 2017

size. Therefore, it is expected to have less accurate results. The opposite is the case of the region Central which has the biggest sample size and results are very close to the ones in the official publication (results available in INEC (2017)).

Table 2.4: Direct and EBP-MSPREE estimates for poverty status (proportions) by planning region 2017

	Direct			EBP-MSPREE		
Regions	Extreme	Poor	Not Poor	Extreme	Poor	Not Poor
Central	3.9	11.9	84.2	3.8	11.7	84.5
Chorotega	5.9	16.5	77.6	8.3	18.0	73.7
Pacífico Central	8.9	21.0	70.1	6.8	16.7	76.5
Brunca	10.4	19.1	70.5	10.7	20.7	68.6
Huetar Caribe	8.9	17.8	73.3	9.3	18.9	71.8
Huetar Norte	9.2	18.3	72.5	7.8	17.6	74.6
Total	5.7	14.3	80.0	5.7	14.3	80.0

2.5 Design-based simulation

In this section, results from a design-based simulation study are presented. The objective is to evaluate the EBP-MSPREE procedure explained in Section 2.3.3 to estimate and update counts on poverty incidence in three categories: ‘Extreme poor’, ‘poor’ (not extreme), and ‘not poor’, as well as assessing the performance of the bootstrap MSE estimator described in Section 2.3.4. To conduct the evaluation, two census compositions are required. One census from time t_0 as the primary input to get EBP-MSPREE estimates but also a second census from time t_1 to compare the updated results. Since a recent census is not available, the survey data described in Section 2.2.2 is used in this experiment. Survey data set from 2011 is used as a census in t_0 (Z_{aj,t_0}) and survey data from 2012 as a census in t_1 (Z_{aj,t_1}). Samples are drawn from them with simple random sampling without replacement, with a sampling fraction of f = 0.2. Since survey data is used as census, there are many domains with only few observations. For this reason, the number of cantons in the simulation is reduced to $A = 23$ (instead of $A = 81$ as in the application), and the biggest domains were selected. This allows having all cells with a positive sample, in this case, with at least 15 observations in each category of poverty for all domains. For the first part of the procedure where point estimates are obtained, i.e., with an EBP model, a Box-Cox transformation is chosen. Also, the income per capita is defined as the dependent variable and a reduced number of covariates are included in the model, namely: the proportion of employees in the household, highest degree of education completed by the head of the household, zone, quantity of economically dependent members in the household, equivalised size of the household, at least one member without health insurance, and at least one member not with an educational lag.

In this simulation study, the performance of an EBP-SPREE (similarly as in Isidro et al. (2016)) is compared with the proposal of this paper, i.e., EBP-MSPREE. A total of R = 500 Monte Carlo iterations are defined, with L = 100 Monte Carlo iterations for implementing the EBP, and B = 100 bootstrap iterations for MSE estimation. The performance of the estimated EBP-MSPREE ($\hat{Y}^{EM}_{aj,t_1}$) is evaluated with the relative bias (RB) and the square root MSE (RMSE), defined as:

$$\text{RB}(\hat{Y}^{EM}_{aj,t_1}) = \frac{1}{R}\sum_{r=1}^{R}\Big(\frac{\hat{Y}^{EM,r}_{aj,t_1} - Z_{aj,t_1}}{Z_{aj,t_1}}\Big)$$

and,

$$\text{RMSE}(\hat{Y}^{EM}_{aj,t_1}) = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\big(\hat{Y}^{EM,r}_{aj,t_1} - Z_{aj,t_1}\big)^2}, \tag{2.9}$$

which is treated as the empirical RMSE. A plot comparing the estimated RMSE (from Equation 2.8) and the empirical RMSE (from Equation 2.9) is used to validate the proposed MSE estimator $\widehat{\text{MSE}}(\hat{Y}^{EM}_{aj,t_1})$. Relative bias and relative RMSE of the estimated RMSE for each area a and category j are also computed as follows:

$$\text{rel.Bias.Est.RMSE} = \Big(\frac{\text{Est.RMSE} - \text{Emp.RMSE}}{\text{Emp.RMSE}}\Big)$$

$$\text{rel.RMSE.Est.RMSE} = \sqrt{\Big(\frac{\text{Est.RMSE} - \text{Emp.RMSE}}{\text{Emp.RMSE}}\Big)^2}.$$

Results of the design-based simulation

Table 2.5 summarises results of the evaluation of the EBP-MSPREE estimator in comparison with a previous version, namely EBP-SPREE. The values of RB and RMSE are averaged over 23 areas for each category of poverty. In general, EBP-MSPREE shows lower values, although the category ‘poor’ (not extreme) has larger RMSEs in comparison with the SPREE version.

Table 2.5: RB and RMSE for the incidence of poverty by status under different SPREE approaches

		Extreme		Poor not extreme		Not poor	
		Median	Mean	Median	Mean	Median	Mean
RB	EBP-SPREE	0.197	0.128	0.019	0.011	-0.022	0.034
	EBP-MSPREE	0.020	0.099	0.032	0.014	-0.013	0.010
RMSE	EBP-SPREE	0.312	0.347	0.127	0.151	0.074	0.106
	EBP-MSPREE	0.278	0.354	0.148	0.167	0.071	0.099

Figure 2.6 displays the estimated and empirical RMSE over the domains and categories of poverty for the EBP-MSPREE and the EBP-SPREE methods. Based on this figure, it is possible to conclude that the estimated RMSE tracks the empirical RMSE better for the EBP-MSPREE procedure, and this can be observed for all of the categories of poverty. A closer look at the performance of the proposed MSE is provided in Table 2.6. The RB-RMSE for the EBP-MSPREE indicates a moderate underestimation in the mean and the median for the ‘extreme’ and ‘poor’ categories and an overestimation for the median of the category ‘poor’. In terms of RB-RMSE and RRMSE-RMSE, the results on the performance of the MSE are favorable for the EBP-MSPREE.

Table 2.6: Performance of the MSE estimator: Mean and median of RB and RMSE by poverty status

		Extreme		Poor not extreme		Not poor	
		Median	Mean	Median	Mean	Median	Mean
RB-RMSE	EBP-SPREE	-0.329	-0.229	0.211	0.363	-0.297	-0.171
	EBP-MSPREE	-0.075	-0.081	-0.077	-0.071	0.139	0.008
RRMSE-RMSE	EBP-SPREE	0.462	0.460	0.379	0.678	0.381	0.448
	EBP-MSPREE	0.381	0.391	0.359	0.367	0.368	0.367

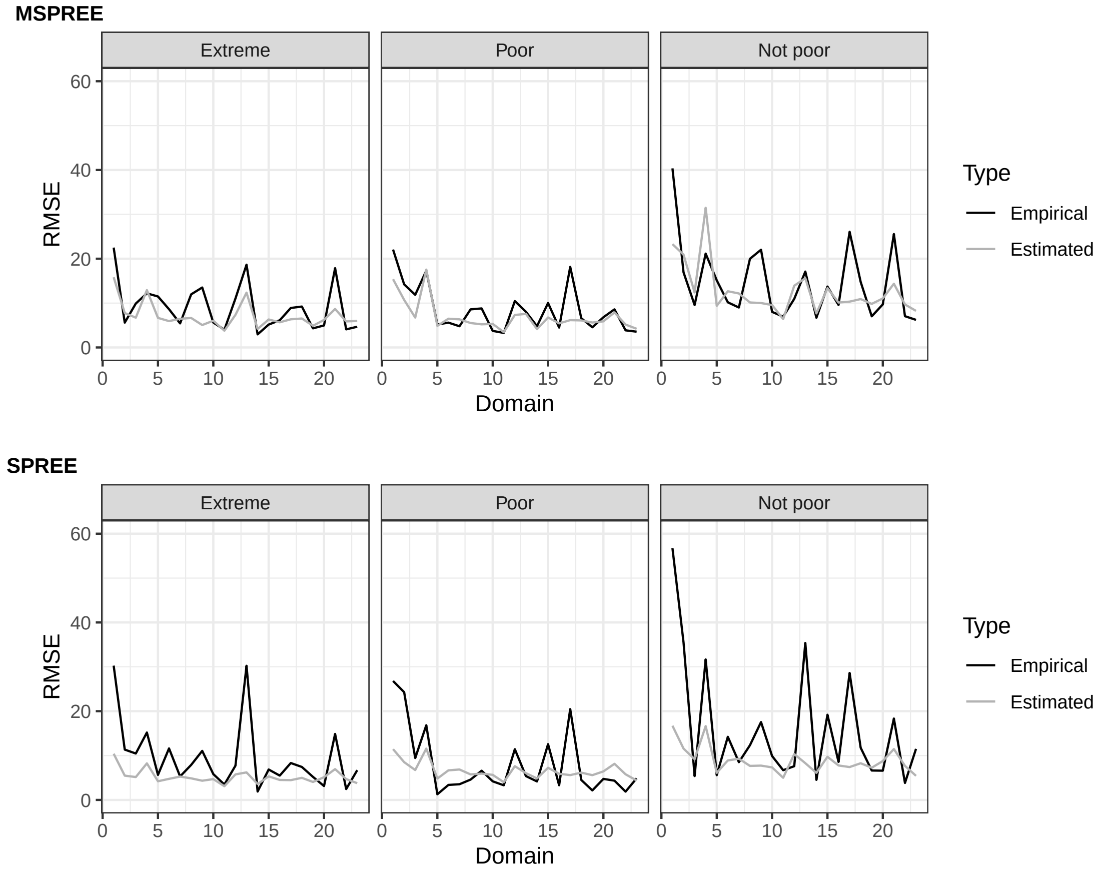


Figure 2.6: Estimated and empirical domain-and category-specific RMSEs of the counts on poverty incidence

2.6 Conclusions and further steps

Public policies require not only accurate and reliable information for decision-making but this information should also be timely. It is common that the production of official statistics faces challenges due to limitations of resources. In this paper, a methodology to obtain reliable and updated estimates in small areas is presented and exemplified with a real-world application. For many developing countries, censuses are conducted every ten years and sample sizes of annual national surveys are not big enough to provide reliable results for small areas. An additional limitation that is considered in this work is that information of interest is not present in census data, as it is required for SPREE methods. The strategy proposed considers two well-known small area estimation techniques. An EBP is conducted to get poverty estimates in the census data, and as a second step, the MSPREE of Luna-Hernández (2016) is applied to update the estimates in postcensal years. Based on the results of the application, it is possible to conclude that the strategy proposed delivers quality results in terms of CVs. The application shows that this methodology gives the opportunity to analyse specific groups of interest, areas, and years. For example, that the poorest cantons in Costa Rica have remained with little overall improvements for the period studied.

Although the methodology proposed allows to obtain the target estimates, there are several aspects that can be improved, especially in the uncertainty estimation. Original SPREE and two other versions (GSPREE and MSPREE) assume that the census has the variable of interest and therefore no uncertainty from the association structure is required. However, when another small area estimation method is needed as a first step to get the census structure (in this case with an EBP), variability from it should be considered. In this methodology, a parametric bootstrap is implemented to get uncertainty from the allocation, the association structure, and the estimation of the β coefficients required in MSPREE. One potential topic for further research is to combine the MSE that is produced directly from the EBP with the one from MSPREE under an analytical approach. Furthermore, the impact that extreme values in the first part of the procedure (e.g., EBP) can have in the final updated estimates deserves also to be investigated.

SPREE methods have other disadvantages that require further study to get a more flexible technique. For example, a potential improvement in the method could be to allow updating more complex indicators or non-categorical indicators such as the Gini index or (mean and/or median) per capita income. The inclusion of associated variables as suggested by Purcell and Kish (1980) can also be beneficial in the estimation procedure, for instance, the inclusion of urbanity (urban and rural) can be relevant when working with poverty status.

The over-shrinking problem present in the context of small area estimation when the expected sample variance is smaller than the true parameter, also deserves to be explored when implementing SPREE-type methods.

Understanding the benefits of SPREE-type methods in comparison with existing models in the small area estimation context requires also further research. Three alternative approaches to deal with the problem of obtaining updated counts or proportions in small domains have been identified and deserve a closer comparison with the SPREE-type methods: 1. The use of the EBP in each year of study with a final benchmark operation performed with MSPREE, 2. exploring potential advantages of using panel survey data or time-series models, for example, with the extension of the Fay-Herriot model proposed by Rao and Yu (1994), or 3. implementation of measurement error models also in the context of area-level models (Ybarra and Lohr, 2008). Finally, it is recommended to study the inclusion of non-traditional information sources (e.g., big data) as proposed in Koebe et al. (2022) since the structures of population censuses can quickly become obsolete. A clear example of this is the socio-economic effect that the COVID-19 pandemic generated in many countries, altering the living conditions of many people in a short period of time.

Acknowledgment

The elaboration of this paper was only possible thanks to the data provided by the National Statistical Institute of Statistics and Censuses from Costa Rica. The data access was provided under specific confidentiality conditions. The numerical results are not official estimates and are only produced for illustrating the methods.

Chapter 3

Intercensal Updating using Structure Preserving Methods and Satellite Imagery

Part II

Microdata Dissemination for Small Area Estimation

Chapter 4

Releasing Survey Microdata with Exact Cluster Locations and Additional Privacy Safeguards

4.1 Introduction

Since almost hundred years, sample surveys are dominating knowledge generation in empirical research. The advantages of survey sampling are obvious: with an appropriate sampling design representative results for a population can be collected by surveying only a fraction of it. With computer assistance, the time from collecting data to publishing results can be sped up significantly (Granello and Wheaton, 2004). Two trends, however, increasingly challenge the way data is collected via surveys. On the one hand, the growing demand for fast and granular information drives up sample size and thus costs. As a response, recent years have seen a large amount of academic research on augmenting surveys with secondary data from non-traditional data sources such as social networks, mobile phones or remote sensing in order to overcome shortcomings in coverage, frequency and granularity with applications in fields as diverse as population dynamics (Stevens et al., 2015; Leasure et al., 2020), socio-demographic analysis (Pokhriyal and Jacques, 2017; Schmid et al., 2017; Subash et al., 2018; Fatehkia et al., 2020; Chi et al., 2022), policy targeting (Blumenstock, 2018; Aiken et al., 2022), environmental mapping (Grace et al., 2019) and health research (Brown et al., 2014; Arambepola et al., 2020). This augmentation is usually done via geographic matching, i.e. combining area-level averages (Koebe, 2020). Since the number of matched areas corresponds to the sample size for subsequent supervised learning tasks, finding the smallest common geographical denominator is essential to avoid running into small sample problems. However, this is not always trivial as sample surveys usually provide data only for a fraction of small geographic areas. On the other hand, digital transformations across various sectors such as health care have led to an explosion of digital personal data. It is the abundance of secondary data that amplifies re-identification risks in published surveys as some of the information could be used to link pseudoanonymized survey responses back to the actual respondents (Armstrong et al., 1999; Kroll and Schnell, 2016; West et al., 2017). Together with new privacy regulations such as the European General

Data Protection Regulation (GDPR) this calls for additional precautionary measures to safeguard the individual's privacy. For aggregated data releases, the introduction of differential privacy has provided a solid mathematical framework to manage re-identification risks independent of a potential attacker's capabilities or prior knowledge (Dwork, 2008). With regard to microdata dissemination strategies, a common de-identification practice today is a combination of deletion and perturbation procedures, which include removing (unique) identifiers such as first and last name and replacing the individual's true location with aggregated (i.e. area-level) and randomized information (see e.g. Andrés et al. (2013); Templ (2017); de Jonge and de Wolf (2019)).

For example, in the Demographic and Health Survey (DHS), a major global household survey program, urban survey clusters are re-located within a 2km-radius and rural clusters within a 5km-, sometimes even 10km-radius (Burgert et al., 2013). This location privacy procedure has two main advantages: it does not affect the quality of the remaining (non-spatial) survey information and it reduces the need for other privacy safeguards, e.g. deleting or perturbing sensitive information. However, it does not provide a similar rigorous measure for privacy protection as already small sets of attributes can quickly increase the chances of re-identification, even in incomplete, pseudonymous datasets (Rocher et al., 2019). In addition, it obviously affects the utility of the published data when it comes to matching with auxiliary data as this type of analysis relies on the congruence of its geographic links (Elkies et al., 2015; Warren et al., 2016; Blankespoor et al., 2021; Hunter et al., 2021).

In that regard, advances in synthetic data generation have introduced new ways to narrow the void between information loss and privacy protection. These methods allow for the generation of synthetic records that resemble the real data by reproducing relationships learned from the latter. While all approaches have in common that they try to capture the joint distribution in the original data, the ways to do so vastly differ. For example, Drechsler et al. (2008) and Heldal and Iancu (2019) use imputation processes to decompose the multidimensional joint distribution into conditional univariate distributions. Alfons et al. (2011) and Templ et al. (2017) use parametric models in combination with conditional re-sampling to synthesize hierarchical relationships. As an alternative to these fully parametric approaches, Reiter (2005) and Wang and Reiter (2012) make use of classification and regression trees (CART), while more recently, Li et al. (2014); Zhang et al. (2017); Rocher et al. (2019); Sun et al. (2019); Torkzadehmahani et al. (2019); Xu et al. (2019) and others have used Bayesian networks, Generative Adversarial Networks or copulas to capture the underlying linear and non-linear relationships between the attributes.

The challenge for data producers is to define adequate microdata dissemination strategies that allow users to satisfy their needs, i.e. release survey microdata that can be used for statistical analysis and that are compatible with other sources of information allowing to answer new and more detailed research questions and – at the same time – it must be ensured that the identities of the respondents are protected. In that regard, the Spatial Data Repository of the DHS program (ICF, 2022) is a good example for facilitating new types of research by combining survey microdata with geospatial covariates and gridded interpolation surfaces. However, also those products are based on perturbed cluster locations, thus incurring a certain information loss.

Hence, as our main contribution of this paper, we propose an alternative microdata dissemination strategy: instead of publishing original microdata with perturbed cluster locations, we investigate the option of publishing two datasets – 1) original microdata stripped of geographic identifiers for which survey results are not considered representative and 2) synthetic microdata with the original cluster locations. The choice is motivated by adopting a user-centric perspective: official household survey publications predominantly report on results up to the strata-level as results below are usually considered not representative. Analysis that benefits from below strata-level data often investigates proximity-related questions such as distances to certain locations and surrounding habitat. For the former, cluster locations are of minor importance, for the latter, however, the spatial perturbation procedure introduces significant levels of uncertainty to the analysis (Warren et al., 2016). The alternative microdata dissemination strategy obviously conserves data utility for analysis on the representative level via the first dataset, while the second dataset allows for the accurate capture of proximity-related information. However, two potential shortcomings need to be considered: first, can we use the synthetic dataset to predict the 'private' attribute in the original dataset, i.e. the small area identifier, thus bypassing the privacy protection measures? Second, is the uncertainty we introduce by synthesizing the non-spatial attributes for spatial analysis smaller than the uncertainty from perturbing the cluster locations?

We show in an experiment using Costa Rican census data from 2011 and satellite-derived auxiliary information from WorldPop (WorldPop, 2018) that we can reduce the re-identification risk vis-à-vis common spatial perturbation procedures, while maintaining data utility for non-spatial analysis and improving data utility for spatial analysis.

From the plethora of options, we choose copulas as our synthetic data generation approach. Copulas facilitate fine-tuning as they allow us to model the marginal distributions separately from the joint distribution. Dating back to 1959 (Sklar, 1959) with diverse applications since, their theoretical properties are well understood. In comparison with alternatives like GANs, copula-based synthetic data generation has lower computational cost (Sun et al., 2019) and it is easier to interpret (Kamthe et al., 2021). Furthermore, the procedure is in general less cumbersome, in comparison with the steps followed by Alfons et al. (2011) to generate the synthetic population data AAT-SILC (*Artificial Austrian Statistics on Income and Living Conditions* (Alfons et al., 2011). Finally, copulas are also attractive for data producers such as National Statistical Offices as only new nationally representative margins are required to update the synthetic microdata file (cf. Koebe et al. (2022)). In addition, well-documented open-source tools such as the Synthetic Data Vault (MIT Data To AI Lab, 2022) are available to users with important features such as data transformation and constraints specification.

4.2 Results

We consider a survey D_{true} as a random sample with sample size n from a given population of size N. For our experiment, we use a 10% random sample of the original 2011 census of Costa Rica, which can be obtained from INEC (2022) as a pseudo-population. In that year, Costa Rica had four administrative disaggregation levels: two zones, six planning regions, 81 cantons and 473 districts. For sampling purposes, enumeration areas (EAs) and strata are defined. Our units of observation are individuals i living together in a household ζ. Each individual is described by a set of attributes denoted as $\mathbf{x} = X_1, \dots, X_m$. Obfuscated attributes are denoted as $\mathbf{y} = Y_1, \dots, Y_m$ in the following. The zip code attribute $X_{\text{zip}} \in \mathbf{x}$ – corresponding to the level of $k = 473$ *districts* in Costa Rica – represents the smallest geographic identifier in this experiment as true locations for the identifier of the census enumeration areas are not available. Consequently, the obfuscated zip code is denoted by $Y_{\text{zip}} \in \mathbf{y}$. Following our proposed data dissemination strategy, we further define the true survey without small-area geographic identifier as *'No Zip Code'* survey $D_{no} := (X_2, \dots, X_m)$ given that $X_1 \leftarrow X_{\text{zip}}$. For notational simplicity, we use X_{zip} and X_1 interchangeably. While different sampling designs are possible, we assume a commonly used complex design for larger household surveys such as the DHS: a stratified two-stage cluster design. In the first stage, the primary sampling units (PSUs) denoted as j – usually enumeration areas from the latest census – are selected for each stratum s with a probability proportional to (population) size Ω_j. In the second stage, households within each selected PSU are sampled with a fixed probability $\Omega_{\zeta|j}$. Consequently, the sampling weights defined as the inverses of the household-level inclusion probabilities are given for each stratum separately by:

$$w = \frac{1}{\Omega_{\zeta j}}, \qquad \Omega_{\zeta j} = \Omega_{\zeta|j} * \Omega_j \quad \text{with} \quad \Omega_j = \frac{n_s}{N_s}, \tag{4.1}$$

with n_s and N_s the sample and population size in stratum s, respectively.

With enumeration area-specific population sizes in the pseudo-population too small to act as survey clusters, we choose the districts (i.e. the zip codes) for each stratum as our PSUs, also called *clusters* in the following. As zip codes can cover both rural and urban areas, there are 767 PSUs in total available in our experiment using Costa Rican census data from 2011. In the following, we describe the original survey attributes as our *true* survey. The true survey builds our starting point for further anonymization approaches, notably the geomasking approach and the copula-based synthetic data generation approach. Figure 4.1 describes the complete experimental setup used in this study.

In the first step, two-stage cluster sampling is used to create household survey microdata (called thereafter the '*True*' survey D_{true}). Randomly sampled point locations within the respective zip codes are assigned to the clusters before displacement. Displaced clusters are allocated to their new zip codes. True survey microdata with (partially) obfuscated zip codes is called '*Geomasked*' survey D_{geo} thereafter and thus constitutes the benchmark anonymization strategy in this experiment. In contrast, the strategy proposed in this paper considers two datasets for dissemination: 1) the '*Synthetic*' survey D_{syn} with original zip codes and remaining attributes being synthetically generated using a copula-based approach, and 2) the true original survey microdata stripped of geographic identifiers below the strata-level (called the '*No Zip Code*' survey D_{no} thereafter). In the third step, an inference attack is designed to disclose the private attribute - i.e. the true zip code - in the geomasked and the 'no zip code' survey, respectively. Similar attacks to disclose private attributes in the synthetic survey could be considered, however, these can be assumed to be comparatively less effective given the amount of true at-

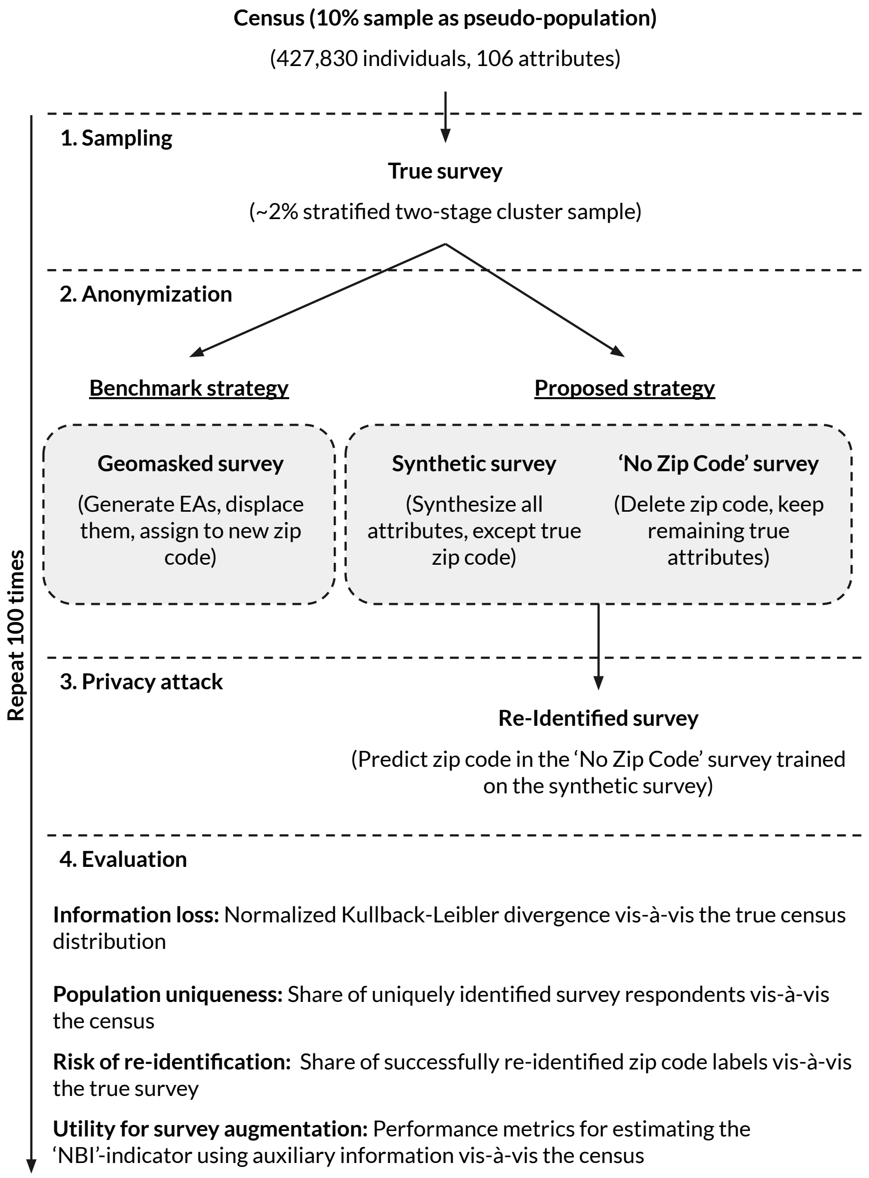

Census (10% sample as pseudo-population)
(427,830 individuals, 106 attributes)
1. Sampling
True survey
(~2% stratified two-stage cluster sample)
2. Anonymization
Benchmark strategy
Proposed strategy
Geomasked survey
(Generate EAs, displace them, assign to new zip code)
Synthetic survey
(Synthesize all attributes, except true zip code)
'No Zip Code' survey
(Delete zip code, keep remaining true attributes)
3. Privacy attack
Re-Identified survey
(Predict zip code in the 'No Zip Code' survey trained on the synthetic survey)
4. Evaluation
Information loss: Normalized Kullback-Leibler divergence vis-à-vis the true census distribution
Population uniqueness: Share of uniquely identified survey respondents vis-à-vis the census
Risk of re-identification: Share of successfully re-identified zip code labels vis-à-vis the true survey
Utility for survey augmentation: Performance metrics for estimating the 'NBI'-indicator using auxiliary information vis-à-vis the census
Repeat 100 times

Figure 4.1: **Workflow diagram of the experiment with census data from Costa Rica.** Geographic identifiers are considered as part of the set of attributes. The attribute 'zip code' represents the smallest geographic identifier in this experiment as true locations of the census enumeration areas are not available. Even though a privacy attack is also performed on the geomasked survey (see Figure 4.3), the resulting dataset is not further analyzed in the remaining study for the sake of readability. A detailed data description can be found in the Supplementary Information.

tributes available to stage such an attack. In order to provide a comprehensive assessment of the risk-utility-trade-off of the two approaches, the evaluation stage is composed of an information loss measure, two measures to assess the privacy risk and three metrics for assessing the utility of the different strategies in a data augmentation setting. Step 1 to 4 are repeated 100 times to get a first understanding of the scale of uncertainty associated with the two approaches.

4.2.1 Geomasking to obfuscate true survey locations

To implement the benchmark strategy in the anonymization step, we follow the geomasking methodology outlined in Burgert et al. (2013) by perturbing the centroids denoted as r of the selected clusters within a given larger administrative area l using a rejection sampling procedure described in Algorithm 1. Even though clusters in our experiment correspond to the zip codes in each stratum, we use available census information on enumeration areas v for the displacement procedure. Since point locations for the corresponding enumeration areas r_v are not available, we randomly sample them from the smallest available area – the zip codes. That way, we can approximate the displacement effect expected when one would sample from the full population using enumeration areas as PSUs.

Algorithm 1: Geomasked survey: DHS cluster displacement algorithm

```
for $v \in D_{true}$ do
    while $r_v^{masked} \notin l_{r_v}$ do
        angle $\leftarrow$ Uniform$_{[0,360]}$ $* \frac{\pi}{180}$ ; /* Random displacement angle */
        if $v$ is Urban then
            dist $\leftarrow$ Uniform$_{[0,2000]}$ ;     /* Random displacement distance
              (in meters) for urban clusters */
        end
        if $v$ is Rural then
            if $v$ is selected as 1% of rural clusters then
                dist $\leftarrow$ Uniform$_{[0,10000]}$ ;           /* Random displacement
                  distance for 1% of rural clusters */
            else
                dist $\leftarrow$ Uniform$_{[0,5000]}$
            end
        end
        $r_{x,v}^{\text{masked}} \leftarrow r_{x,v} + \text{dist} * \cos(\text{angle})$ ;        /* Displace x-coordinate
          ($r_{x,v}$) */
        $r_{y,v}^{\text{masked}} \leftarrow r_{y,v} + \text{dist} * \sin(\text{angle})$ ;        /* Displace y-coordinate
          ($r_{y,v}$) */
    end
end
```

We denote the masked point locations of the sampled EAs with the superscript *masked*. Households with masked EAs now located outside their original zip code, but inside their original larger administrative area l_{r_v} are assigned the respective new zip code. As the overall inclusion probability for a household is not affected by geomasking, direct estimates and corresponding variances for area-level aggregates l (corresponding in case of our experiment to the 81 *cantons* in Costa Rica) and above remain the same. However, this does not hold for area-level aggregates smaller than l. We describe the original survey attributes together with the masked clusters as our *geomasked* survey $D_{\text{geo}} := Y_{\text{zip}}, X_2, \dots, X_m$.

Through the displacement procedure, roughly 30% of the sampled EAs are assigned to a new zip code, representing approx. 30% of the sampled individuals in each simulation round.

4.2.2 Copula-based synthetic data generation

As an alternative to geomasking in the anonymization step, we use synthetically generated survey attributes for protecting the respondents' privacy while keeping the true clusters. To do so, we fit a Gaussian copula model on the transformed attributes denoted with $\tilde{X}_1, \ldots, \tilde{X}_m$ of the original survey and sample from the learned joint distribution for each cluster individually with the original sample size n_j. A copula allows to describe the dependence structure - also called *association structure* - independently from the marginal distributions (also called *allocation structure*). Several copula families are available. We focus on the Gaussian copula that allows us to represent the association structure of random variables irrespective of their true distribution through a multivariate standard normal distribution Patki et al. (2016). Since we also assume the marginals to be normally distributed, which may certainly constitute a mis-specification for some of the variables, we regard the results rather as a lower bound in terms of goodness-of-fit. Further, a copula is uniquely defined only for continuous variables Jeong et al. (2016), meaning that in principle, copulas cannot model non-continuous variables. Since socio-economic surveys are largely made up of categorical variables, data transformation, e.g. via one-hot or frequency encoding, is needed. In addition, we impose constraints on the marginals to account for censoring (e.g. to avoid negative synthetic age records) or between-variable dependencies (e.g. female and male household members need to add up to the total household size) via rejection sampling.

Thus, the process to generate synthetic data $\tilde{D}_{\text{syn}}$ from a survey dataset $\tilde{D}_{\text{true}}$ with transformed categorical attributes $\tilde{X}_1, \ldots, \tilde{X}_m$ (details of the data transformation using frequency encoding are described in Algorithms 3 and 4 in Section 4.4) using a Gaussian copula model is summarized in Algorithm 2.

ϕ_Σ is the cumulative distribution function (cdf) of a multivariate normal distribution with $\mathcal{N}(\mu, \Sigma)$ and ϕ_m the cdf of a standard normal distribution. By fitting our model to the true survey, it learns the parameters of both the allocation and association structure, i.e. of the marginal distributions Ψ and the multivariate Gaussian copula $C^G_\Sigma(u_1, \ldots, u_m)$ built on the probability integral transforms $u_1, \ldots, u_m$. Based on these learned relationships, new synthetic records $\tilde{\mathbf{y}}^{\{i\}}$ are sampled from the multivariate probability function $c^G_\Sigma(\mathbf{u})$ using the inverse probability integral transform for each component $F_m^{-1}(u_m)$ (cf. Janke et al. (2021)). Since we sample in our experiment for each cluster individually to ensure a synthetic cluster-level sample size of exactly n_j, we use the parameters of a conditional multivariate normal distribution. In case no conditions are applied, the scenario is simplified to drawing from a multivariate standard normal distribution. We call the synthetic attributes $Y_2, \ldots, Y_m$ together with the true cluster information X_{zip} our *synthetic* survey $D_{\text{syn}} := X_{\text{zip}}, Y_2, \ldots, Y_m$. Further details about the copula-based synthetic data generation procedure can be found in the Section 4.4 and in Nelsen (2007).

Figure 4.2 provides a first impression on the overall goodness-of-fit of the three different survey datasets (cf. with the evaluation step in Figure 4.1). Specifically, Figures 4.2a - 4.2c show the normalized Kullback-Leibler (KL) divergence Z_{KL} for the survey attributes of D_{true}, D_{geo} and D_{syn} from the true census attributes defined in this case for D_{syn} as

Algorithm 2: Synthetic survey: Copula-based synthetic data generation algorithm

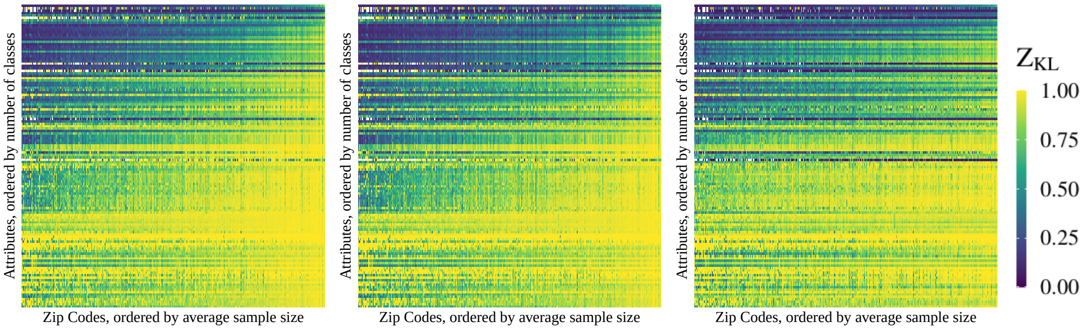
```
Input $\tilde{D}_{\textbf{true}} = (\tilde{X}_1, \dots, \tilde{X}_m)$
Output $\tilde{D}_{\textbf{syn}} = (\tilde{Y}_1, \dots, \tilde{Y}_m)$, with $\tilde{Y}_1 = \tilde{X}_{\text{zip}}$
for $s \in \tilde{D}_{true}$ do
    $\Psi \leftarrow$ Estimated marginal distributions of $\tilde{X}$ with $\Psi_m \sim \mathcal{N}(\mu, \sigma^2)$
    $\Sigma \leftarrow$ Estimated covariance matrix of $\Psi$
    $U \leftarrow F(\Psi)$ ;            /* Probability integral transforms */
    $C^G_\Sigma(u_1, \dots, u_m) \leftarrow \phi_\Sigma\big(\phi_1^{-1}(u_1), \dots, \phi_m^{-1}(u_m)\big)$ ;      /* m-dimensional
     Gaussian copula */
    for $j \in \tilde{D}_{true,s}$ do
        for $i \leftarrow 1$ to $n_j$ do
            while $\tilde{\mathbf{y}}^{\{i\}}$ not meets constraints do
                $\mathbf{w} \sim \mathcal{N}(\bar{\mu}, \bar{\Sigma})$ ;            /* Conditional sampling */
                $\tilde{\mathbf{y}}^{\{i\}} \leftarrow F^{-1}\big(\phi_2(w_2), \dots, \phi_m(w_m)\big)$ ;     /* Convert back to
                 original space */
            end
        end
        $\tilde{D}_{\text{syn},j} \leftarrow (\tilde{Y}_1 = k, \tilde{\mathbf{y}}_j) \quad \forall \quad j \subseteq k$ ;  /* Assign zip code $k$ of the
         respective cluster $j$ */
    end
end
```

$$Z_{KL}(f_{m,k}(X_{m,k})||f_{m,k}(Y_{m,k})) = \frac{1}{1 + \delta_{KL}(f_{m,k}(X_{m,k})||f_{m,k}(Y_{m,k}))}, \tag{4.2}$$

averaged across simulation runs for each attribute m and zip code k, respectively. In general, the KL divergence δ_{KL} measures the difference between two probability distributions, in this case between the census distribution and one of the survey datasets for a given attribute in a given zip code. The better one distribution approximates the other, the smaller δ_{KL}. Therefore, following Equation 4.2, values of the normalized KL divergence Z_{KL} close to 1 indicate a high goodness-of-fit.

Clearly visible is a gradient from the top left to the bottom right indicating that the overall goodness-of-fit of the sample distributions improve the larger the underlying sample sizes and the lower the number of classes per categorical attribute. We expect that high levels of sampling variance usually associated with small samples may also lead to poor outcomes across multiple simulation rounds irrespective the modelling approach. In addition, as expected, attributes with high levels of non-response (visible through the white spots across the horizontal axis) are stronger affected by sampling and anonymization compared to attributes with little or no non-response.

To approach the utility-risk trade-off in (pseudo)-anonymized microdata, we define two risk-related measures: a) the re-identification risk of a sensitive attribute in the original data using the perturbed data, and b) the respondents' re-identification risk, i.e. the population uniqueness of the survey respondents.



(a) True survey (b) Geomasked survey (c) Synthetic survey

Figure 4.2: **Normalized Kullback-Leibler divergence (in bits) from the true census distribution for each attribute and zip code, averaged across 100 simulation rounds.** The attributes on the y-axis are ordered by their respective number of classes, the zip codes on the x-axis are ordered by their average sample size across simulation rounds. Values of Z_{KL} close to one (yellow) represent little divergence from the true census distribution and therefore indicate a high goodness-of-fit. The number of attribute classes range from 2 to 111. Across attributes and zip codes, the true survey scores best with $Z_{KL} = 0.76$ in total, followed by the synthetic survey with $Z_{KL} = 0.74$ and the geomasked survey at $Z_{KL} = 0.73$.

4.2.3 Risk of re-identifying private geocodes

To investigate the first shortcoming mentioned in Section 4.1, we define our first risk-related measure: the re-identification risk of a sensitive attribute in the original data using the perturbed data. In our experiment, we therefore train a random forest model on the small area identifier - the zip code - in the anonymised surveys for each stratum separately. Across the generated sample surveys, the sample sizes by zip code range from 24 to 715 units with mean of 81 and median of 45. We use the trained models on the original data to predict the zip code for each record. We call the 'No Zip Code' survey with the predicted zip codes $\hat{X}_{\text{zip}}$ as *'Re-identified'* survey $D_{\text{re}} := (\hat{X}_{\text{zip}}, X_2, \dots, X_m)$ in the following. Finally, we evaluate our predicted label against the original label. In addition, we compare the outcomes to randomly guessing the correct label in order to account for the number of small areas within each stratum. Figure 4.3 shows the median accuracy of the approaches across 100 simulation runs. While we are able to successfully re-construct the original zip code in most cases for the geomasked survey, it does not work much better for the synthetic data than for the random guess.

In our experiment, only one stratum consequently hosts more than ten small areas across all simulation runs, with one stratum hosting only two small areas in some simulation runs, giving the random guess also a good chance to predict correctly. Recalling that roughly 70% of the displaced clusters stay within the same zip code in the geomasked survey, even predicting the sensitive attribute for strata hosting as little as two small areas, average population uniqueness in the synthetic data would not exceed much the 50/50-chance of the random guess, thus providing better privacy protection in the re-identified original survey than the geomasked alternative.

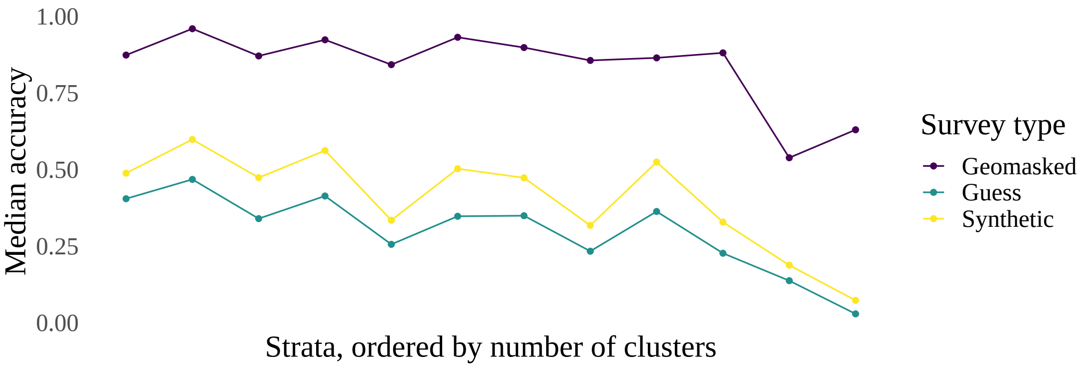


Figure 4.3: **Re-identification of the zip code as private attribute in the true survey for each stratum across 100 simulation runs.**
Accuracy is measured by the share of successfully re-identified zip code labels in the true survey. A random forest model is trained on perturbed data, i.e. the geomasked and the synthetic survey, respectively. We evaluate the results against the true zip code labels in the true survey and compare them against random guesses of the private attribute.

4.2.4 Population uniqueness of survey respondents

Concerning the respondents' re-identification risk, we define population uniqueness Ξ_t as the share of survey respondents being unique in the population for a given (sub-)set of attributes in D_{true}, D_{no}, D_{geo}, D_{syn} and D_{re}, respectively. Similar to algorithm 2, we denote the subsets with $D'(t)$, t with $1 \leq t \leq m$ being the number of attributes used for calculating the population uniqueness.

$$\Xi_t = \frac{1}{n}\sum_{i}^{n} \nVdash_{i(t)} \quad \text{with} \quad \nVdash_{i(t)} = \begin{cases} 1, & \text{if} i(t) \in D'(t) \text{ unique in population} \\ 0, & \text{otherwise.} \end{cases} \tag{4.3}$$

Figure 4.4 shows how Ξ_t changes with the increasing number of attributes t across 100 simulation runs. We kept the order of attributes constant across simulations to improve comparability.

Naturally, the share constantly increases for the true survey with more attributes being available to distinguish between the respondents. For example, there might be 100 women in a country, but likely just one aged 45 with poor eyesight and four children in a specific zip code. For the geomasked survey, the population uniqueness increases to a level of roughly 70%. Recalling that the only difference between the geomasked survey and the true survey is the perturbed zip code, the remaining 30% corresponds to the average number of survey respondents assigned to a new zip code due to the spatial anonymization process. Thus, not considering the zip code (i.e. the 'No Zip Code' setting) lets the population uniqueness of the geomasked survey also converge towards 1 similar to the true survey, even though at a slower rate, which means knowledge on additional attributes is required to compensate for the lack of geographic stratification via the zip code. For the synthetic survey, the curve remains almost flat. The initial bump can largely be explained by the probability of a random combination of attributes representing an actual population unique in a small (area) sample size setting. Therefore, Fig-

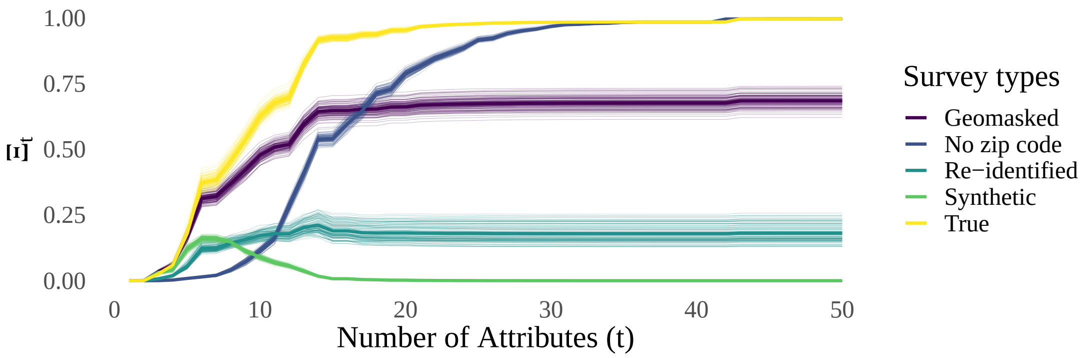


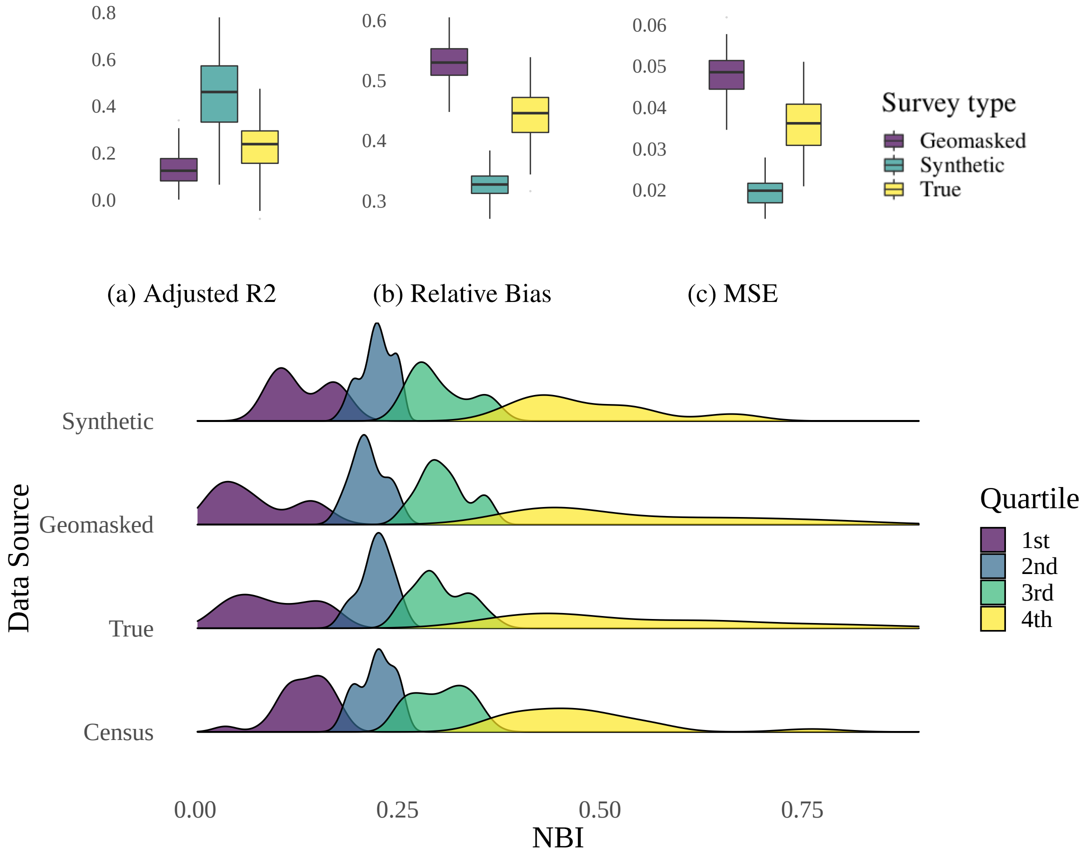
Figure 4.4: **Population uniqueness across survey types.**
Share of population-unique survey respondents for 100 simulation runs with a given number of attributes. Geographic identifiers are considered as part of the set of attributes. The thick lines represent the average population uniqueness across the 100 simulation runs, the thin lines individual simulation runs. In the *true* survey, no attribute is perturbed. In the *geomasked* survey – the benchmark dissemination strategy in this study –, the zip code identifier is perturbed. The *'No Zip Code'* survey corresponds to the true survey, but lacks the geographic identifiers below the strata level. Together with the *synthetic* survey, where all attributes but the zip code identifier are perturbed, it represents the proposed microdata dissemination strategy. In the *re-identified* survey, the synthetic survey is used to predict the "private" attribute – i.e. the zip code – in the 'No Zip Code' dataset as part of a staged inference attack on the proposed microdata dissemination strategy. Both the re-identified and the synthetic survey provide significant privacy gains vis-à-vis the other survey types.

ure 4.4 gives a strong indication that geomasking provides little additional safeguards for the respondents' privacy compared to the true survey in the presence of third-party information on a subset of the contained attributes.

Besides this theoretical argument, synthetic data always provides plausible deniability to the survey respondents. Similarly to our definition, Rocher et al. (2019) use a Gaussian copula model to estimate the empirical likelihood of population uniqueness in incomplete datasets such as D by assuming $\Xi_{\mathbf{t}} \sim \text{Binomial}(\nVdash_{i(t)}, n)$ with $\forall i(t) \in D'(t)$ i.i.d.. While this approach is an excellent alternative to measure the re-identification risk in micro-level survey data when no validation data (in our experiment the 2011 Costa Rican census) is available, it assumes that the individual records are independent and identically distributed, which may be contestable in the presence of hierarchical dependencies and complex sampling designs.

4.2.5 Utility for survey augmentation

To give an indication about the utility of the different anonymization approaches or survey data augmentation, we use a setup common in recent academic literature (cf. Pokhriyal and Jacques (2017); Leasure et al. (2020); Schmid et al. (2017)): we augment the surveys with auxiliary information from geospatial (big) data. Specifically, we construct zip code-level aggregates from gridded satellite-derived features available from the WorldPop repository (WorldPop, 2018) and combine them with zip code-level survey aggregates to provide predictions, especially for areas not sampled in the survey. As our target variable, we select the Unsatisfied Basic Needs index (*Necesidades Básicas Insatisfechas (NBI)*) - a composite indicator similar to the multidimensional poverty index (MPI) (Méndez and Bravo, 2011; Alkire et al., 2019) used as a key statistical indicator in Costa Rica. Details on the index can be found in the Supplementary Information. We evaluate our predictions against the census in terms of adjusted R^2, bias and the Mean Squared Error (MSE). Figures 4.5a - 4.5c show the performance along these three evaluation criteria across 100 simulation runs.



(a) Adjusted R2 (b) Relative Bias (c) MSE

(d) NBI densities

Figure 4.5: **Performance metrics of survey-based NBI estimates on the zip code-level.** (a) Adjusted R^2 is based on the in-sample zip codes. (b) and (c) are based on the full sample and predictions are evaluated against the census across 100 simulation runs. (d) compares zip code-level NBI averages for a single simulation run.

Surprisingly, the synthetic approach not only outperforms the geomasked survey, it also provides predictions more in line with the census results than the true survey. A possible explanation could be that the copula approach reduces the impact of outliers on the zip code-specific NBI sample averages. This explanation is supported by Figure 4.5d that shows the distribution of zip code-level NBI averages grouped into quartiles for one simulation run as both the synthetic survey and the census showcase smaller tails in their distributions, respectively.

We run additional experiments to compare the directly synthesized NBI and its underlying indicators with their counterparts computed from synthetic survey variables (see Supplementary Table B.2 and Supplementary Fig. B.5 in the Supplementary Information).

4.3 Discussion

In this paper, we proposed and evaluated an alternative data dissemination strategy for micro-level survey data that improves the trade-off between privacy risk and data utility. Specifically, we showed that by publishing two datasets, namely the original survey data with limited geographic identifiers and a synthetically-generated survey dataset with the true cluster locations, re-identification risks can be reduced significantly vis-à-vis popular geomasking approaches without incurring additional losses in terms of data utility for survey augmentation. This could help mapping initiatives such as WorldPop or GRID3 to improve their products as more accurate spatial data is available. In addition, by separating the marginals from the dependence structure, it provides data producers such as National Statistical Offices also with a useful tool to update the respective synthetic microdata files for the following years by updating the margins with nationally representative new data as sub-nationally representative surveys may only be conducted every few years. In the Supplementary Information, we further investigate the stability of our results by alternating the experiment design.

First, while we chose the strata for the main analysis as they provide 'large-enough' sample sizes at the same time explicitly accounting for at least high-level regional variation, we study in further experiments whether fitting on smaller or larger geographic levels may better capture local variation at the expense of running into the risk of small sample problems or vice versa. Supplementary Fig. B.2 summarizes the results for our copula model being fitted on the whole survey, the twelve strata and the zip code-level, respectively. It shows that by selecting the strata as our fitting level, we strike a balance between the underlying sample size (usually the larger the better) and capturing regional variation (usually the more disaggregated the better) both in terms of utility and risk. In addition, by using subsets of the full microdata for model fitting, the approach becomes computationally tractable also for larger surveys.

Second, since generative models allow us to sample an arbitrary number of synthetic observations, we look at the impact of the synthetic sample size on the outcomes of the survey augmentation experiment, notably the adjusted R^2 and a measure of confidence in the direct survey estimates of the Fay-Herriot model (cf. Section 4.4.2) - the shrinkage factor γ. Supplementary Fig. B.3 shows that with an increasing sample size, γ increases as well, thus shifting more weight to the direct estimate. Even though intuitive as the sampling variance naturally decreases in n, at some point it may become misleading with potentially negative effects on the model performance as the synthetic data generating process still relies on the same information conveyed in the true survey with sample size n. However, in our experiment the adjusted R^2 does not exhibit a bump, but increases monotonically, thus hinting at little additional explanatory power of our satellite-derived covariates vis-à-vis the area-level direct survey estimate for the in-sample areas.

Third, since our target variable *NBI* is a composite indicator, we compare the different composition levels of the synthetic NBI with the NBI constructed from synthetic data. While the divergence measure shows an overall good fit for the underlying indicators (see Supplementary Table B.2 and Supplementary Fig. B.5), correlations are low, especially for higher-level compositions as the dimensions or the NBI itself.

Lastly, we test alternative encoding schemes for the transformation of categorical data.

Also, we relax our assumption of the normally distributed margins by opening up to a wider group of parametric copulas (such as beta, gamma or uniform distributions) selected for each margin individually based on the two-sample Kolmogorov-Smirnov (KS) statistic to study the effect of the specification choice on the normalized KL divergence. Supplementary Fig. B.4 shows that neither the encoding scheme nor the specification of the marginal distributions have large effects on the quality of the synthetically generated data.

Nevertheless, our approach is not without limitations. As synthetic data generation is in its essence a modelling task by creating an abstract representation of the underlying data, similar rules of thumb apply: a) a model is as good as its underlying data – if the sample is partially skewed due small (class-specific) sample sizes or high levels of non-response, they model might reproduce this skewedness and b) composite indicators have to be treated with care as decomposability of the predictions is not necessarily guaranteed unless explicitly modelled that way. The copula-based approach towards synthetic data generation largely fails to correctly capture lower-level hierarchical relationships such as *individuals - line numbers - households - houses* from the original data. As said before, since we see our analysis using a naïve Gaussian copula model as providing somewhat a lower bound for improving the utility-risk trade-off by adopting the proposed microdata dissemination strategy vis-à-vis common geomasking approaches, there is much room for improvement. To name a few, latent copula designs can be considered to avoid data transformations, marginal distributions can be modelled non-parametrically, hierarchical structures can be accounted for more rigorously by either modelling the hierarchies separately as suggested by Templ (2017) or by modelling the relationships explicitly. In addition, synthetic data may - under some circumstances - leak private information, e.g. through the generated value ranges. As a response, differentially-private implementations of existing generative models have been proposed such as PrivBayes (Zhang et al., 2017), PrivSyn (Zhang et al., 2021) and PATE-GAN (Jordon et al., 2019). That said, it is important to point out that microdata irrespective of the selected dissemination strategy, cannot be considered fully anonymous, but rather pseudonymous, thus requiring the data publisher (e.g. the National Statistical Office) to conduct data protection impact assessments before release – depending on the respective jurisdiction. Lastly, as with most empirical research, it would be interesting to apply the proposed dissemination strategy to other contexts/countries.

4.4 Methods

4.4.1 Fitting Gaussian copulas to survey attributes

As an alternative to geomasking, we use synthetically generated survey attributes for protecting the respondents' privacy while keeping the true point locations of the selected clusters. To do so, we fit a Gaussian copula model on the transformed survey attributes from $\tilde{D}_{\text{true}}$ and sample from the learned joint distribution for each cluster individually with the originally sample size n_j. Therefore, consider our survey $\tilde{D}_{\text{true}}$, where $\tilde{X}_1$ represents a random variable with a continuous marginal cumulative distribution function (cdf) denoted by $F_1(\tilde{x}_1) = P(\tilde{X}_1 \leq \tilde{x}_1)$. For the multivariate case, the joint cdf for $\tilde{D}_{\text{true}}$ can be generalized to $F_{1,\dots,m}(\tilde{x}_1, \dots, x_m) = P(\tilde{X}_1 \leq \tilde{x}_1, \dots, \tilde{X}_m \leq \tilde{x}_m)$.

A copula, firstly introduced in the work of Sklar (1959), is a cumulative density function with uniform marginals between [0,1]. Thus - based on Sklar´s theorem (Sklar, 1959) - when all variables are continuous, the m-dimensional random vector $\tilde{X}_1, \ldots, \tilde{X}_m$ can be defined in a uniform space $[0,1]^m$, creating a random vector $U_1, \ldots, U_m$ via the respective probability integral transforms, e.g. $u_m = F_m(x_m)$. In this case, a unique m-dimensional copula $C(\mathbf{u})$ exists:

$$C(\mathbf{u}) = F\big(F_1^{-1}(u_1), \ldots, F_m^{-1}(u_m)\big). \tag{4.4}$$

As motivated in Section 4.2.2, we account for the fact that household surveys largely consist of categorical variables by applying data transformation. Among the plethora of possible encoding schemes, the most common encoding scheme is one-hot encoding, where for each class of a categorical variable a binary dummy variable is created (Benali et al., 2021). A disadvantage of this option is that it may become computationally challenging and prone to multicollinearity in the presence of variables with a high cardinality, i.e. with a large number of classes, since each possible class creates a new variable (Bourou et al., 2021). Interestingly, there is – to the best of our knowledge – little comprehensive, comparative and conclusive scientific evidence on the properties and performance of different categorical encoding schemes. Therefore, we explore two other well-known alternatives with more favourable computation times: ordinal and frequency encoding. Ordinal encoding uses integers to represent each class in a categorical variable. Assigning an unreal order to nominal variables is the main pitfall of this alternative (Jiang et al., 2020). Frequency encoding – as used in medical imaging (Mansfield and Maudsley, 1977) and similar to the concept of *term frequency* in Natural Language Processing (Aizawa, 2003) – assigns an interval in [0,1] to each class based on and ordered by its proportion of occurrence. Then, it uses the middle point of each interval as float representative of the respective class. Back-transformation is done by assigning a new point to a class via the respective interval it falls into. In this sense, this alternative conveys information of the importance of each class (Sabharwal and Agrawal, 2021) without increasing the number of attributes. Based on the results of the different encoding schemes shown in Supplementary Fig. B.4, we opt for the frequency encoding scheme in the following. Consequently, we denote the subset of continuous attributes with P and the subset of non-continuous attributes that require data transformation with Q. Algorithms 3 and 4 provide details on the chosen scheme.

Algorithm 3: Transform categorical variables

```
Input $D_{\textbf{true}} = (X_1, \ldots, X_m)$
Output $\tilde{D}_{\textbf{true}} = (\tilde{X}_1, \ldots, \tilde{X}_m)$
for $X_m \in D_{true}$ do
    if $X_m$ is Continuous then
        $\tilde{X}_p \leftarrow X_m$
    end
    if $X_m$ is Non-continuous then
        $\tilde{X}_q \leftarrow \mathrm{T}(X_m)$
    end
end
$\tilde{D}_{\text{true}} \leftarrow (\tilde{X}_p, \tilde{X}_q) \quad \forall p \in P, q \in Q$ and with $m = p + q$
```

Algorithm 4: Back-transform frequency encoded variables

```
Input $\tilde{D}_{\textbf{syn}} = \tilde{Y}_1, \ldots, \tilde{Y}_m)$
Output
  $D_{\textbf{syn}} = (X_{\textbf{zip}}, X_2, \ldots, Y_m)$
for $\tilde{Y}_m \in \tilde{D}_{true}$ do
    if $\tilde{Y}_m$ is indexed as variable in $P$ then
        $Y_m \leftarrow \tilde{Y}_m$
    end
    if $\tilde{Y}_m$ is indexed as variable in $Q$ then
        $Y_m \leftarrow T^{-1}(\tilde{Y}_m)$
    end
end
$D_{\text{syn}} \leftarrow (X_{\text{zip}}, Y_2, \ldots, Y_m)$ with $Y_1 = X_{\text{zip}}$
```

Thus, the m-dimensional Gaussian copula $C_\Sigma^G(\mathbf{u})$ is defined as the cdf of a multivariate normal distribution $\mathcal{N}(\mu, \Sigma)$ with $\Sigma \in \mathbb{R}^{m \times m}$ represented on the unit cube $[0,1]^m$:

$$C_\Sigma^G(u_1, \ldots, u_m) = \phi_\Sigma\big(\phi^{-1}(u_1), \ldots, \phi^{-1}(u_m)\big). \tag{4.5}$$

The density of a Gaussian copula is then defined as:

$$c_\Sigma^G(\mathbf{u}) = \frac{1}{\sqrt{\det \Sigma}} \exp\Big(-\frac{1}{2}\phi^{-1}(\mathbf{u})^T \cdot (\Sigma^{-1} - I) \cdot \phi^{-1}(\mathbf{u})\Big). \tag{4.6}$$

with $\mathbf{u} \in [0,1]^m$, $I \in \mathbb{R}^{m \times m}$ being the identity matrix, and ϕ^{-1} being the inverse cumulative distribution function of a standard normal distribution. Σ is a positive semi-definite covariance matrix that we estimate based on Pearson's correlation coefficient ρ (Li et al., 2014).

As noted in Section 4.2.2, we sample for each cluster individually with a sample size of n_j. While rejection sampling could be an option for ensuring only synthetic rows with the respective cluster identifier are selected, it proves computationally inefficient. With copulas being multivariate cdfs, we introduce conditions instead. Hence, we sample from a multivariate normal distribution conditional on cluster j. Thus, our transformed dataset $\tilde{D}_{\text{true}}$ with one conditional variable becomes $\tilde{D}_{\text{true}} = (\tilde{\mathbf{X}}_a | \tilde{\mathbf{X}}_b)$ with $\tilde{\mathbf{X}}_a := \tilde{X}_2, \ldots, \tilde{X}_m$ being the transformed attributes to be synthesized and $\tilde{\mathbf{X}}_b := \tilde{X}_1$ being the transformed cluster identifier. The parameters of the respective multivariate normal distributions are thus partitioned into:

$$\tilde{D} = \begin{bmatrix} \tilde{\mathbf{X}}_a \\ \tilde{\mathbf{X}}_b \end{bmatrix}, \ \mu = \begin{bmatrix} \mu_a \\ \mu_b \end{bmatrix} \text{ and } \Sigma = \begin{bmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{bmatrix} \tag{4.7}$$

with $\mu_a \in \mathbb{R}^{m-1}$ and $\mu_b \in \mathbb{R}^1$ and $\Sigma_{aa} \in \mathbb{R}^{(m-1)\times(m-1)}$, $\Sigma_{ab} \in \mathbb{R}^{(m-1)\times 1}$, $\Sigma_{ba} \in$

$\mathbb{R}^{1\times(m-1)}$, and $\Sigma_{bb} \in \mathbb{R}^{1\times 1}$ being the means and positive semi-definite covariance matrices, respectively. Following Algorithm 2, the parameters of our estimated marginal distributions Ψ and of the copula $C^G_\Sigma(\mathbf{u})$ need to be adapted to mirror the conditionality such that $\Psi_{a|b}(\tilde{\mathbf{X}}_a|\tilde{\mathbf{X}}_b)$ and $C^G_\Sigma(\mathbf{u}_a|\mathbf{u}_b)$.

Consequently, we sample from $\sim \mathcal{N}(\bar{\mu}, \bar{\Sigma})$ with:

$$\bar{\mu} = \mu_a + \Sigma_{ab}\Sigma_{bb}^{-1}(X_b - \mu_b) \in \mathbb{R}^{m-1} \tag{4.8}$$

and

$$\bar{\Sigma} = \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba} \in \mathbb{R}^{(m-1)\times(m-1)}. \tag{4.9}$$

We iterate the copula-based fitting and sampling procedure for every stratum separately as it allows to better capture sub-national variation using representative sub-samples and as it proves computationally more tractable. For sampling designs with varying household- or individual-level inclusion probabilities (e.g. in the DHS, women - in comparison to men - are usually oversampled), Templ (2017) suggests to sample a synthetic population and re-iterate the sampling procedure to produce valid synthetic sampling weights. As in our design sampling weights are identical across households for a given PSU due to the systematic sampling approach in the second stage, the original sampling weights remain valid. The virtue in our model choice is the relative simplicity, little requirements in terms of ex-ante knowledge about the individual distributions $\tilde{X}_m$ and its computational efficiency. For further experiments on the robustness and sensitivity of our modelling choices, we refer to the Supplementary Information.

4.4.2 Area-level survey augmentation methods

Survey data can be augmented with the use of area-level models, e.g. the Fay-Herriot model (Fay and Herriot, 1979) by linking direct estimators gathered from survey data to relevant auxiliary information. Both, direct estimators, and auxiliary data are aggregated on k areas. Traditionally, these auxiliary covariates $\mathbf{x}_k$ are obtained from recent censuses, administrative records or other geospatial (big) data sources. In this paper, we make use of satellite imagery features as area-level covariates. The Fay-Herriot is a two-level model, the first part is composed by the sampling model:

$$\hat{\theta}_k^{\text{Dir}} = \theta_k + e_k, \quad e_k \sim N(0, \sigma^2_{e_k}), \tag{4.10}$$

where the sampling error is represented by e_k and $\hat{\theta}_k^{\text{Dir}}$ is the direct estimator of θ_k (e.g. sample mean). The linking model provides the second part, where relevant area-level covariates are considered:

$$\theta_k = \mathbf{x}_k'\hat{\beta} + u_k. \tag{4.11}$$

Here, the random area effects u_k are assumed to be independent with mean 0 and variance σ^2_u. The empirical best linear unbiased predictor (EBLUP) estimator is given by:

$$\hat{\theta}_k^{\text{FH}} = \gamma_k \hat{\theta}_k^{Dir} + (1-\gamma_k)\mathbf{x}'_k\hat{\beta} = \mathbf{x}'_k\hat{\beta} + \hat{u}_k, \tag{4.12}$$

with $\gamma_k = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_{e_k}^2}$ denoting the shrinkage factor for each area k. The parameter estimates of this model can be obtained via maximum likelihood (ML) or restricted ML (REML). Note that the shrinkage factor allows to weight in favor of the direct estimator when sampling variances are small; on the contrary the synthetic estimator $\mathbf{x}'_k\hat{\beta}$ receives more weight when the sampling variance is larger. Results on an experiment studying the sensitivity of the shrinkage factor and adjusted R^2 for varying synthetic sample sizes are shown in Supplementary Fig. B.3. Further details on the Fay-Herriot model can be found in Rao and Molina (2015).

Supplementary material B

B.1 Data description

As our reference dataset in this project, we use data from Costa Rica – notably the X$^{\text{th}}$ Population and VI$^{\text{th}}$ Housing Census of Costa Rica, 2011 (Censo Nacional de población y Viviendas de Costa Rica 2011) – to produce three different data file types: First, we draw survey samples from a census population using a stratified two-stage cluster sample design without applying any statistical disclosure control mechanisms. We use these survey samples (called *true* surveys in the study) as starting point for creating file types two and three: By re-assigning clusters to new zip codes based on the displacement algorithm described in Algorithm 1 , we perturb the zip code identifier in the true surveys, thereby creating the *geomasked* surveys. Again based on the true surveys, we apply the copula-based synthetic data generation algorithm described in Algorithm 2 to generate synthetic data for each attribute except the zip code, which keeps it original structure. In addition, in order to test the robustness of our specifications, we create additional datasets with alternating data generating process designs. The censuses are carried out every ten years by the national statistic office of Costa Rica (INEC) and collect information of people, households, and dwellings on topics such as access to education, employment, social security, technology necessary for the planning, execution, and evaluation of public policies (Méndez and Bravo, 2011).

Administratively, Costa Rica had in 2011 four disaggregation levels: two zones, six planning regions, 81 cantons and 473 districts (municipalities). The sampling design used for the main National Household Survey (Encuesta Nacional de Hogares, ENAHO) specifies twelve strata - each planning region divided by urban and rural areas. In this case, the strata coincide with the study domains. Supplementary Fig. B.1 shows the highest level of disaggregation (districts) of Costa Rica, with the 12 strata since are the disaggregation levels used in this paper.

For our experiment, we use a 10% random sample of the original 2011 census, which can be obtained from INEC (2022) as a pseudo-population. The smallest geographical information available in this dataset are the 473 districts. In the first stage, we select districts as our PSUs for each stratum separately with a selection probability proportional to population size. In the second stage, we select a minimum of 10 households in each PSU by using simple random sampling without replacement. PSUs with less than 10 households are discarded from this procedure, affecting roughly 4% of all PSUs.

As auxiliary information, we use covariates derived from satellite imagery. Specifically,

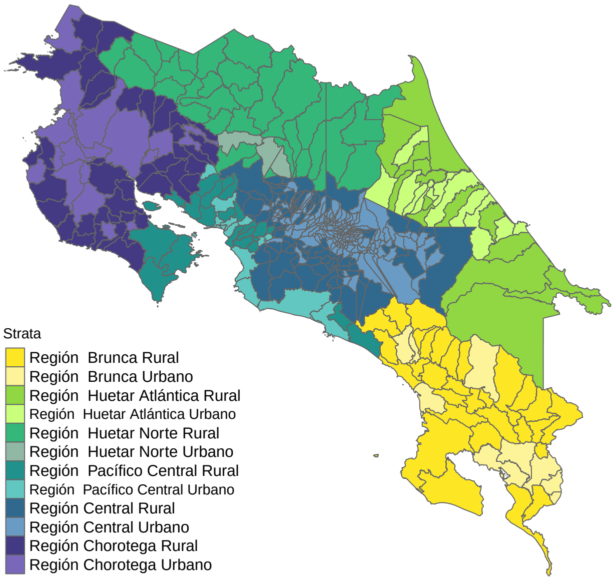

Strata
Región Brunca Rural
Región Brunca Urbano
Región Huetar Atlántica Rural
Región Huetar Atlántica Urbano
Región Huetar Norte Rural
Región Huetar Norte Urbano
Región Pacífico Central Rural
Región Pacífico Central Urbano
Región Central Rural
Región Central Urbano
Región Chorotega Rural
Región Chorotega Urbano

Figure B.1: Administrative disaggregation of Costa Rica. Overlay of 473 districts (zip codes) and 12 strata from the Xth Population and VIth Housing Census of Costa Rica, 2011.

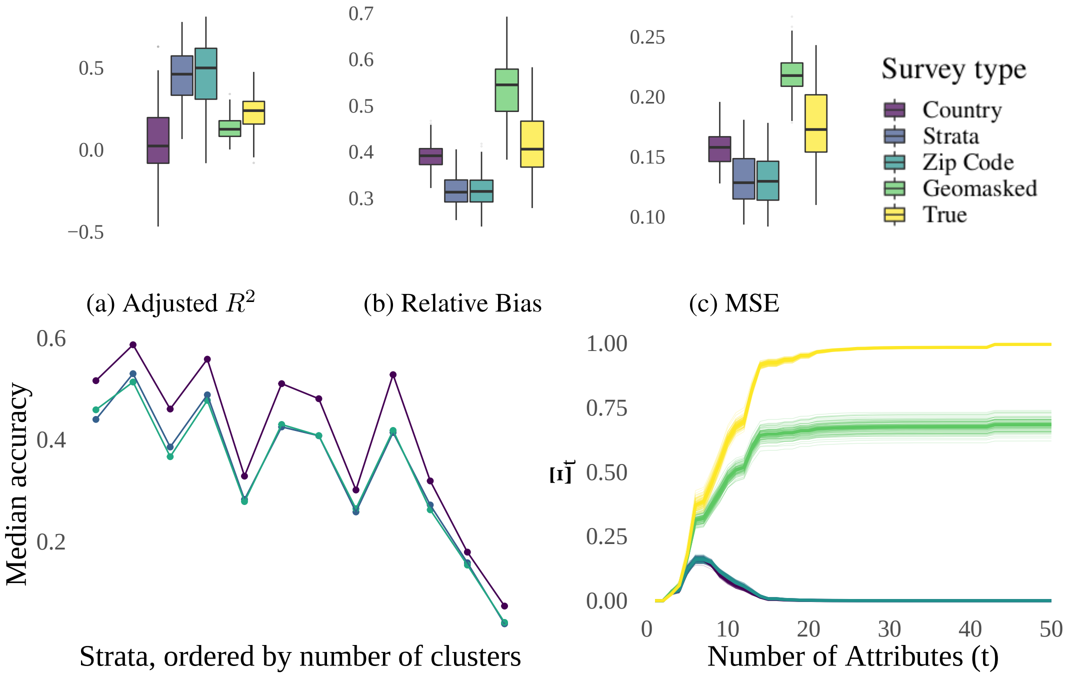
N_C	n_D	# of all PSUs	# of PSUs in D	# of attributes
427830	[7638; 11914]	767	123	106

Table B.1: **Descriptive statistics on the census-derived data across 100 simulation runs**

we use features derived from satellite imagery provided by WorldPop (2018) in our survey augmentation setup. The advantages of using satellite imagery here are five-fold: Data with virtually global coverage at high spatial resolutions for frequent time intervals on human-made impact provided in a structured format enables us to extract covariates for all administrative areas in Costa Rica at the time of the census. Therefore, we can use area-level survey augmentation (cf. Methods Section) to provide estimates, especially for areas not covered by the respective survey. WorldPop data are provided in the tagged image file format (TIFF) with a pixel representing roughly a 100m × 100m grid square in an open data repository under CC4.0 licence (WorldPop (2018)). Pixel values are aggregated to the administrative areas of Costa Rica via their centroids. Specifically, we generate area-level averages for the distances to different types of natural areas (e.g. cultivated, woody-tree, and shrub areas, coastlines etc.) and to infrastructure such as roads and waterways, the intensity of night-time lights, topographic information and information on the presence of human settlements.

B.2 Sensitivity of copula vis-à-vis geographic fitting level

In order to study the effect of the geographic level on the copula modelling performed for synthetic data generation, we run Algorithm 2 on the whole survey (‘Country), the twelve strata (‘Strata’) and the roughly 110 zip code areas (‘Zip Code’), respectively. Results are provided in Supplementary Fig. B.2. It appears that fitting the copula model on the whole survey limits the ability of the approach to capture regional variations. On the other hand, model fitting on the zip code-level does neither increase the re-identification risk of the zip code identifier as a private attribute and nor affect the overall prediction performance of the outcome variable, hinting at overfitting not being a problem on that level. Striking a balance between underlying sample size and a certain level of disaggregation shows better results. Also, it allows to scale computations to settings with larger samples and more attributes.



(a) Adjusted R^2 (b) Relative Bias (c) MSE

(d) Re-identification Risk (e) Population uniqueness across attributes

Figure B.2: **Evaluation metrics for different geographical copula fitting levels.**
(a) - (c) The copula model is fitted on the whole survey ('Country'), for each of the twelve strata ('Strata') and for each of the roughly 110 zip codes ('Zip Code') separately. As a reference, the metrics for the geomasked and the true survey are provided as well. (d) The accuracy to successfully re-identify the zip code as a private attribute in the original data using a random forest model trained on synthetic data across fitting levels remains similar. (e) The share of population-unique survey respondents is virtually not affected by the copula fitting level.

B.3 Effects of synthetic sample size on prediction outcomes

Generative models can be used to create synthetic samples of an arbitrary size regardless the amount of underlying data. While the advantages of that are similar to those of other resampling procedures such as bootstrapping (i.e. to estimate the precision of the sample statistics or to perform cross-validation), it can also mislead modelling approaches that 'borrow strength' from auxiliary data by overestimating the strength of the synthetic direct estimates eventually resulting in losses of explanatory power of the model. In our survey augmentation setup, the shrinkage factor γ indicates whether final estimates rather rely on the direct estimates from the synthetic survey or on the satellite-derived covariates for the in-sample predictions depending on the sampling variance. Supplementary Fig. B.3 shows that larger sample sizes lead to increasing gamma values (via decreasing sampling variances of the direct estimator), however, not incurring losses in the goodness-of-fit of our estimation model. This hints at the fact that the contribution of the auxiliary information to the explanatory power of the model for the in-sample predictions is negligible.

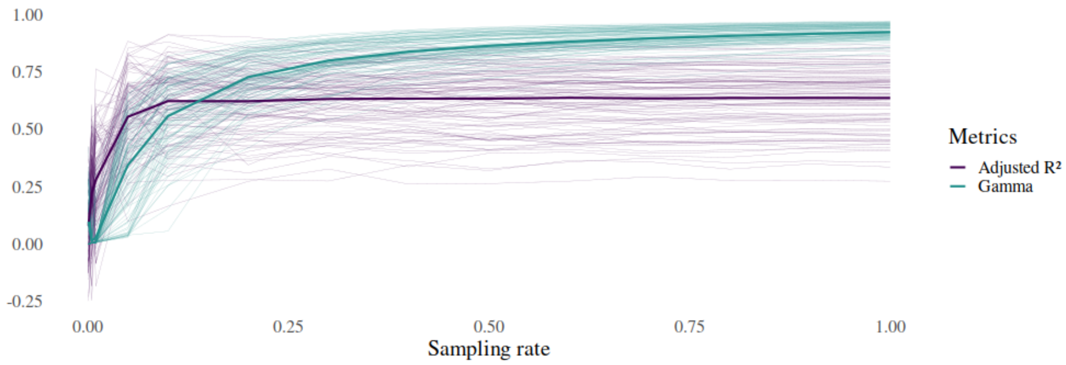


Figure B.3: **Sensitivity of model performance on changes in synthetic sample size.** Samples are drawn from a synthetic population. The synthetic population is generated using the copula-based approach described in the Results Section. Sample sizes are determined by the sampling rate (shown on the x-axis). Results are evaluated against the true census population. The shrinkage factor γ is averaged across zip codes. The thick lines represent the metric averages across the 100 simulation runs, the thin lines individual simulation runs.

B.4 Choosing marginal distributions & encoding schemes

As already mentioned in the Results section, assuming normally-distributed margins may represent a misspecification of the true univariate distribution of X_{dm}. In addition, computationally tractable alternatives to one-hot encoding exist. We compare two different ways to model the marginal distributions together with two different encoding schemes. The results are presented in Supplementary Fig. B.4. Measured by the normalized KL divergence averaged across 100 simulation runs, frequency encoding produces slightly better goodness-of-fit of the synthetic data ($Z_{KL} = 0.74$ for frequency encoding versus $Z_{KL} = 0.72$ for ordinal encoding with gaussian marginals). Surprisingly, the naïve assumption of normally distributed marginals outperforms the KS-based parametric marginals with $Z_{KL} = 0.74$ and $Z_{KL} = 0.70$, respectively.

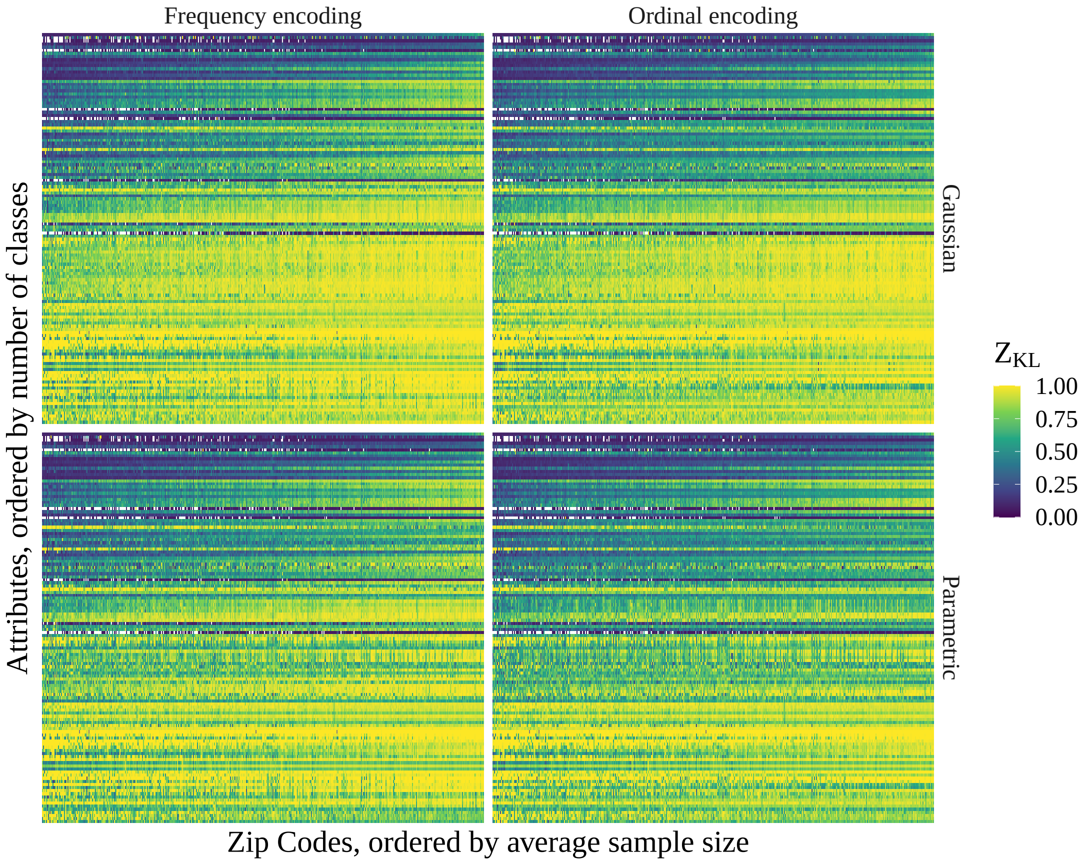


Figure B.4: **Effect of encoding schemes and marginal distribution choice on the overall goodness-of-fit of the synthetic data measured by the normalized KL divergence Z_{KL} (in bits).**
The attributes on the y-axis are ordered by their respective number of classes, the zip codes on the x-axis are ordered by their average sample size across simulation rounds. Values close to one (yellow) represent little divergence from the true census distribution and therefore indicate a high goodness-of-fit.

B.5 Detailed analysis of the NBI as composite indicator

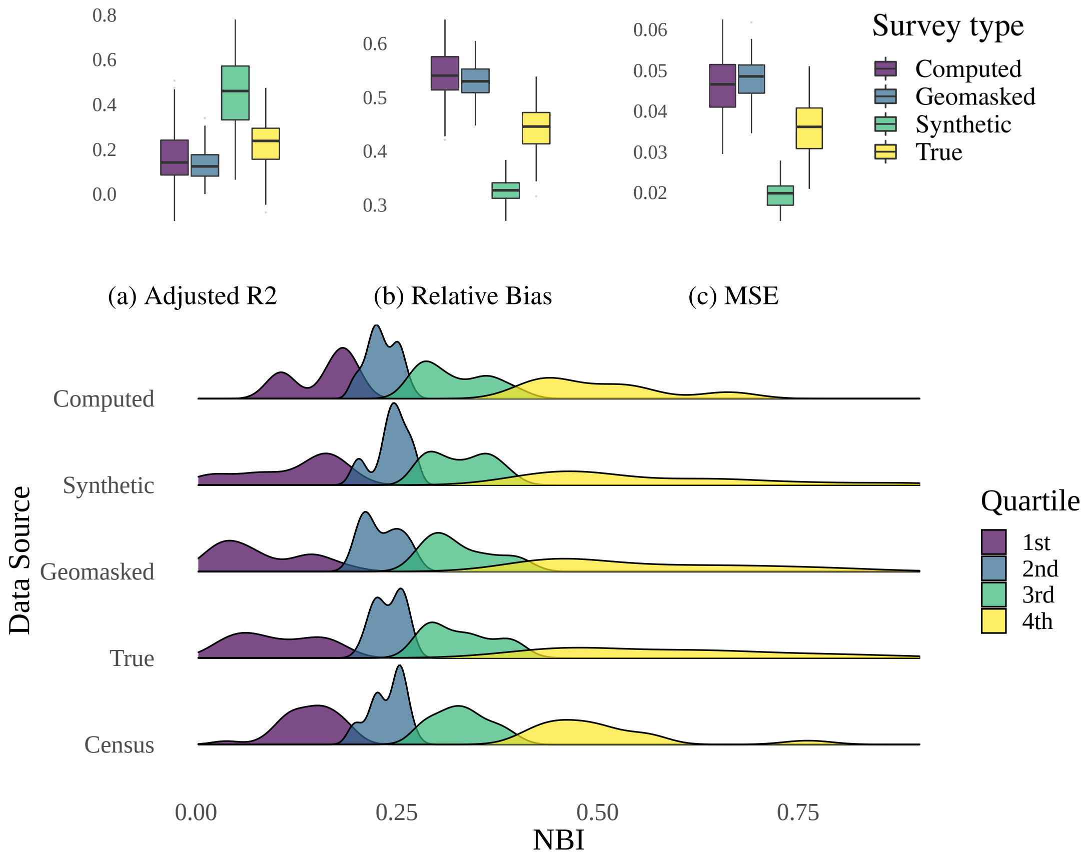
The NBI is a composite indicator computed from approx. 20 underlying survey variables grouped into four dimensions (i.e. access to decent housing (*Acceso albergue digno*), access to a healthy life (*Acceso a vida saludable*), access to knowledge (*Acceso al conocimiento*) and access to other goods and services (*Acceso a otros bienes y servicios*)) using 19 indicators in total. All indicators and dimensions are binary (yes/no). An identified need in one of the indicators leads to a positive needs status in higher dimensions. The sensitivity for false positives is thus assumed to be high for the NBI as a small change (e.g. one year age difference) in one of the 19 underlying variables can turn a NBI-negative to a NBI-positive survey respondent.

Generally, two strategies for computed indicators exist to create synthetic counterparts: a) directly synthesize the computed indicators or b) re-construct the indicator based on synthetic survey variables. While the former is more likely to reflect the original distribution, it may not be consistently decomposable into its underlying indicators; vice-versa holds for the latter. The strength of these effects are largely determined by the complexity and sensitivity of the composite indicator and the overall goodness-of-fit of the synthetic data. Thus, if both approaches produce similar compositions, it can be regarded as a strong indication that the underlying synthetic data also successfully captures relationships across multiple variables in the dataset, not only the composite index. Supplementary Table B.2 shows that this not fully holds for the NBI.

Indicators	# of indicators	Pearson's ρ	Z_{KL}	Incidence
1.x	5	0.42	0.99	100
Dimension 1		0.24	0.98	647
2.x	5	0.22	0.98	85
Dimension 2		0.19	0.98	455
3.x	2	0.02	0.89	507
Dimension 3		0.02	0.84	1845
4.x	7	0.02	0.99	60
Dimension 4		0.03	1.00	622
Composite NBI	19	0.07	0.97	3253

Table B.2: **Relationship between synthetic and computed NBI indicators across 100 simulation runs.**
Indicator-level results (e.g. 1.x) are averaged across indicators. The incidence describes the average number of respondents across 100 simulated surveys with unsatisfied needs in the respective indicator/dimension.

Although the overall number of survey respondents with unsatisfied needs are captured with a high accuracy as measured by the normalized KL divergence Z_{KL} for binary data, the NBI status on the individual level strongly diverges following Pearson's ρ (cf. Supplementary Table B.2). Supplementary Fig. B.5 shows that the lack of linear correlation is mainly due to improperly captured relationships in the underlying variables than in the synthetic NBI as the former is outperformed by the latter for survey augmentation expressed in terms of adjusted R^2, bias and MSE. However, it remains on par with the geomasked survey at lower privacy risks.

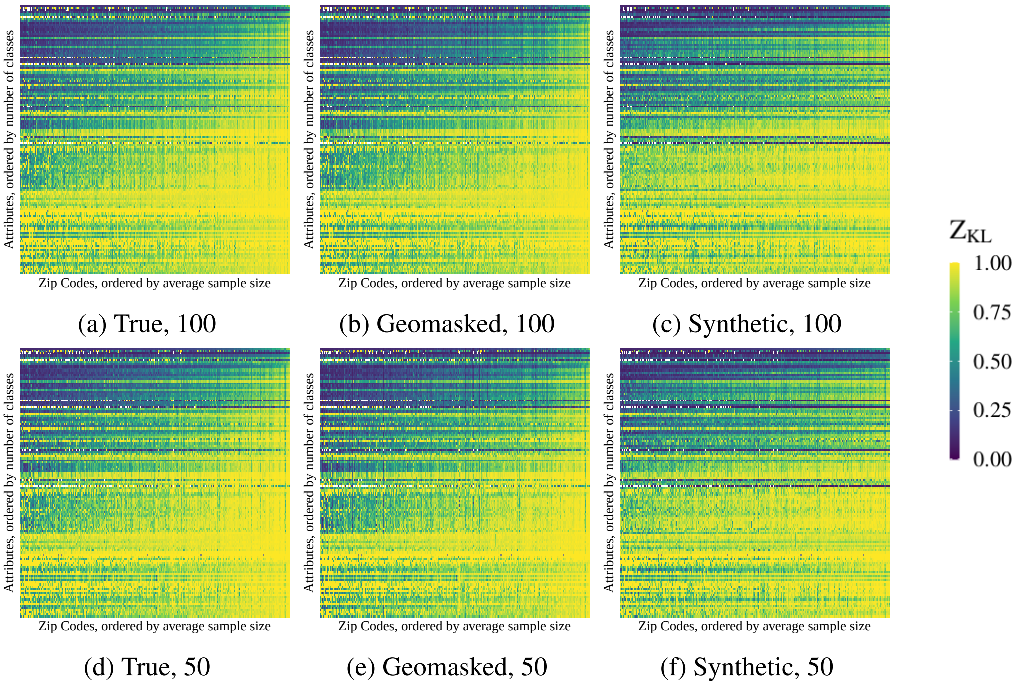


(a) Adjusted R2 (b) Relative Bias (c) MSE

(d) NBI densities

Figure B.5: **Performance of the synthetic vs. computed composite NBI.**
(a) - (c) show of the different survey types in our survey augmentation experiment across 100 simulation runs. (d) shows the densities of the composite NBI by quartiles for one simulation run.

B.6 Stability of anonymization approaches across simulation runs



(a) True, 100 (b) Geomasked, 100 (c) Synthetic, 100

(d) True, 50 (e) Geomasked, 50 (f) Synthetic, 50

Figure B.6: **Stability of normalized KL divergence for 50 and 100 simulation rounds** Normalized Kullback-Leibler divergence (in bits) for the true, geomasked and synthetic survey from the true census distribution for each attribute and zip code, averaged across 100 (B.6a - B.6c) and 50 (B.6d - B.6f) simulation rounds, respectively. The attributes on the y-axis are ordered by their respective number of classes, the zip codes on the x-axis are ordered by their average sample size across simulation rounds. The results give strong indication that the results across 100 simulation rounds can be considered stable.

Bibliography

Agence Nationale de la Statistique et de la Démographie (2013). Rapport Projection de la Population du Senegal 2013-2063. Technical report, Agence Nationale de la Statistique et de la Démographie.

Agence Nationale de la Statistique et de la Démographie (2015). Enquête de Mise à jour du Registre National Unique des Ménages Vulnérables.

Agresti, A. (2002). Categorical data analysis. Hoboken, NJ: John Wiley & Sons, Inc.

Aiken, E., S. Bellue, D. Karlan, C. Udry, and J. E. Blumenstock (2022). Machine learning and phone data can improve targeting of humanitarian aid. Nature 603, 864–870.

Aizawa, A. (2003). An information-theoretic perspective of tf–idf measures. Information Processing & Management 39(1), 45–65.

Alfons, A., P. Filzmoser, B. Hulliger, J.-P. Kolb, S. Kraft, R. Münnich, and M. Templ (2011). Synthetic Data Generation of SILC Data. Research Project Report WP6, D6.2. Technical report, The AMELI Project.

Alfons, A., S. Kraft, M. Templ, and P. Filzmoser (2011). Simulation of close-to-reality population data for household surveys with application to EU-SILC. Statistical Methods & Applications 20(3), 383–407.

Alkire, S. and J. Foster (2007). Counting and multidimensional poverty measures. OPHI working paper 7.

Alkire, S., U. Kanagaratnam, and N. Suppa (2019). The Global Multidimensional Poverty Index (MPI) 2019. OPHI MPI Methodological Note 47. Technical report, Oxford Poverty and Human Development Initiative, University of Oxford.

Andrés, M. E., N. E. Bordenabe, K. Chatzikokolakis, and C. Palamidessi (2013). Geo-indistinguishability: Differential privacy for location-based systems. In Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security, pp. 901–914.

Arambepola, R., S. H. Keddie, E. L. Collins, K. A. Twohig, P. Amratia, A. Bertozzi-Villa, E. G. Chestnutt, J. Harris, J. Millar, J. Rozier, et al. (2020). Spatiotemporal mapping of malaria prevalence in madagascar using routine surveillance and health survey data. Scientific Reports 10(1), 18129.

Arias-Salazar, A. (2022). Small Area Estimates of Poverty Incidence in Costa Rica under a Structure Preserving Estimation (SPREE) Approach. Working Paper.

Armstrong, M. P., G. Rushton, and D. L. Zimmerman (1999). Geographically masking health data to preserve confidentiality. Statistics in Medicine 18(5), 497–525.

Asian Development Bank (2020). Introduction to Small Area Estimation Techniques: A Practical Guide for National Statistics Offices.

Benali, F., D. Bodénès, N. Labroche, and C. de Runz (2021). MTCopula: Synthetic complex data generation using copula. In 23rd International Workshop on Design, Optimization, Languages and Analytical Processing of Big Data (DOLAP), pp. 51–60.

Berg, E. and W. A. Fuller. A spree small area procedure for estimating population counts. Proceedings of the Statistical Society of Canada, Section on Survey Methods. American Statistical Association, pages=3811–3825, year=2009.

Bishop, Y. M., S. E. Fienberg, and P. W. Holland (2007). Discrete multivariate analysis: theory and practice. New York: Springer Science & Business Media.

Blankespoor, B., T. Croft, T. Dontamsetti, B. Mayala, and S. Murray (2021). Spatial anonymization: Guidance note prepared for the Inter-Secretariat working group on household surveys. Technical report, UN Inter-secretariat Working Group on Household Surveys Task Force on Spatial Anonymization in Public-Use Household Survey Datasets.

Blumenstock, J. E. (2018, May). Estimating economic characteristics with phone data. AEA Papers and Proceedings 108, 72–76.

Bonafilia, D., J. Gill, D. Kirsanov, and J. Sundram (2019). Mapping for humanitarian aid and development with weakly-and semi-supervised learning. Technical report, Facebook.

Boo, G., E. Darin, D. R. Thomson, and A. J. Tatem (2020). A grid-based sample design framework for household surveys. Gates Open Research 4.

Bourou, S., A. El Saer, T. H. Velivassaki, A. Voulkidis, and T. Zahariadis (2021, sep). A Review of Tabular Data Synthesis Using GANs on an IDS Dataset. Information 12(9), 375–389.

Box, G. E. and D. R. Cox (1964). An analysis of transformations. Journal of the Royal Statistical Society: Series B (Methodological) 26(2), 211–243.

Brown, M. E., K. Grace, G. Shively, K. B. Johnson, and M. Carroll (2014). Using satellite remote sensing and household survey data to assess human health and nutrition response to environmental change. Population and Environment 36(1), 48–72.

Burgert, C. R., J. Colston, T. Roy, and B. Zachary (2013). Geographic Displacement Procedure and Georeferenced Data Release Health Surveys. DHS Spatial Analysis Reports. Technical Report 7, ICF International, USAID, Calverton, Maryland, USA.

CEPAL & MIDEPLAN (2016). El enfoque de brechas estructurales: análisis del caso de Costa Rica. Santiago: CEPAL.

Chen, X. and W. D. Nordhaus (2011). Using luminosity data as a proxy for economic statistics. Proceedings of the National Academy of Sciences of the United States of America 108(21), 8589–8594.

Chi, G., H. Fang, S. Chatterjee, and J. E. Blumenstock (2022). Microestimates of wealth for all low- and middle-income countries. Proceedings of the National Academy of Sciences of the United States of America 119(3), e2113658119.

Das, S. and S. Haslett (2019). A comparison of methods for poverty estimation in developing countries. International Statistical Review 87(2), 368–392.

de Jonge, E. and P.-P. de Wolf (2019). sdcSpatial: Statistical Disclosure Control for Spatial Data. R package version 0.1.1.

Deming, E. and F. Stephan (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. The Annals of Mathematical Statistics 11(4), 427–444.

Deville, J.-C. and C.-E. Särndal (1992). Calibration estimators in survey sampling. Journal of the American Statistical Association 87(418), 376–382.

Drechsler, J., A. Dundler, S. Bender, S. Rässler, and T. Zwick (2008). A new approach for disclosure control in the iab establishment panel—multiple imputation for a better data access. AStA Advances in Statistical Analysis 92(4), 439–458.

Dwork, C. (2008). Differential privacy: A survey of results. In Theory and Applications of Models of Computation. TAMC 2008. Lecture Notes in Computer Science, Volume 4978, pp. 1–19. Springer, Berlin, Heidelberg.

ECLAC (2014). Social Panorama of Latin America 2015. Santiago, Chile: Economic Commission for Latin America and the Caribbean.

Elbers, C., J. O. Lanjouw, and P. Lanjouw (2003). Micro-level estimation of poverty and inequality. Econometrica 71(1), 355–364.

Elkies, N., G. Fink, and T. Bärnighausen (2015). "Scrambling" geo-referenced data to protect privacy induces bias in distance estimation. Population and Environment 37(1), 83–98.

Emwanu, T., J. G. Hoogeveen, and P. Okiira Okwi (2006). Updating poverty maps with panel data. World Development 34(12), 2076–2088.

European Union (2020). Copernicus: Europe's eyes on Earth.

Eurostat (2017). European statistics code of practice.

Fatehkia, M., B. Coles, F. Ofli, and I. Weber (2020). The relative value of facebook advertising data for poverty mapping. Proceedings of the International AAAI Conference on Web and Social Media 14, 934–938.

Fay, R. E. and R. A. Herriot (1979). Estimates of income for small places: an application of James-Stein procedures to census data. Journal of the American Statistical Association 74(366a), 269–277.

Feres, J. C. and X. Mancero (2001). El método de las necesidades básicas insatisfechas (NBI) y sus aplicaciones en América Latina. CEPAL.

Fick, S. E. and R. J. Hijmans (2017, 10). WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. International Journal of Climatology 37(12), 4302–4315.

Foster, J., J. Greer, and E. Thorbecke (1984). A class of decomposable poverty measures. Econometrica 52(3), 761–766.

Freire, S., K. MacManus, M. Pesaresi, E. Doxsey-Whitfield, and J. Mills (2016). Development of new open and free multi-temporal global population grids at 250 m resolution. AGILE.

Gini, C. (1912). Variabilità e Mutabilità. Contributo allo Studio delle Distribuzioni e delle Relazioni Statistiche. P. Cuppini, Bologna.

Gouvernement du Sénégal (2014). Plan Senegal Emergent. Technical report, Gouvernement du Sénégal.

Gouvernement du Sénégal (2017). Programme National de Bourses de Sécurité Familiale (PNBSF).

Grace, K., N. N. Nagle, C. R. Burgert-Brucker, S. Rutzick, D. C. Van Riper, T. Dontamsetti, and T. Croft (2019, mar). Integrating Environmental Context into DHS Analysis While Protecting Participant Confidentiality: A New Remote Sensing Method. Population and Development Review 45(1), 197–218.

Granello, D. H. and J. E. Wheaton (2004). Online data collection: Strategies for research. Journal of Counseling & Development 82(4), 387–393.

Green, A., S. Haslett, and C. Zingel (1998). Small Area Estimation Given Regular Updates of Census Auxiliary Variables. In Proceedings of the New Techniques and Technologies for Statistics Conference, pp. 206–211.

Groß, M., A. K. Kreutzmann, U. Rendtel, T. Schmid, and N. Tzavidis (2020). Switching between Different Non-Hierachical Administrative Areas via Simulated Geo-Coordinates: A Case Study for Student Residents in Berlin. Journal of Official Statistics 36(2), 297–314.

Harvey, J. T. (2002, 5). Estimating census district populations from satellite imagery: Some approaches and limitations. International Journal of Remote Sensing 23(10), 2071–2095.

Heldal, J. and D.-C. Iancu (2019). Synthetic data generation for anonymization purposes. Application on the Norwegian Survey on living conditions/EHIS. In Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality.

Henderson, J. V., A. Storeygard, and D. N. Weil (2012, 4). Measuring economic growth from outer space. American Economic Review 102(2), 994–1028.

Hidiroglou, M. and Z. Patak (2009). An application of small area estimation techniques to the canadian labour force survey. In Proceedings of the Survey Methods Section, Statistical Society of Canada.

Hidiroglou, M. A. and P. Lavallee (2009). Sampling and Estimation in Business Surveys. In D. Pfeffermann and C. R. Rao (Eds.), Handbook of Statistics: Design, Methods and Applications, Volume 29 A, Chapter Sampling a, pp. 441–470. Elsevier.

Hunter, L. M., C. Talbot, W. Twine, J. McGlinchy, C. W. Kabudula, and D. Ohene-Kwofie (2021). Working toward effective anonymization for surveillance data: innovation at South Africa's Agincourt Health and Socio-Demographic Surveillance Site. Population and Environment 42(4), 445–476.

ICF (2022). The DHS Program Spatial Data Repository.

INEC (2011). Boletín mensual. Costo de la canasta básica alimentaria, Julio 2011.

INEC (2012). Ficha Metodológica: X Censo Nacional de Población y VI de Vivienda 2011. Resultados Generales.

INEC (2015). Índice de Pobreza Multidimensional (IPM). Metodología.

INEC (2017). Encuesta Nacional de Hogares. Julio 2017. Resultados generales.

INEC (2020). Encuesta Nacional de Hogares. Julio 2020. Resultados generales.

INEC (2022). X Censo Nacional de Población y VI de Vivienda. Catálogo central de datos.

INEC & CCP (2013). Estimaciones y Proyecciones de Población por sexo y edad 1950 - 2050. San José: INEC.

Ireland, C. T. and S. Kullback (1968). Contingency tables and with given and marginals. Biometrika 55(1), 179–188.

Isidro, M., S. Haslett, and G. Jones (2016). Extended structure preserving estimation (ESPREE) for updating small area estimates of poverty. The Annals of Applied Statistics 10(1), 451–476.

Isidro, M. C. (2010). Intercensal updating of small area estimates. Ph. D. thesis, Massey University, New Zeland.

Janke, T., M. Ghanmi, and F. Steinke (2021). Implicit generative copulas. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan (Eds.), Advances in Neural Information Processing Systems, Volume 34, pp. 26028–26039. Curran Associates, Inc.

Jean, N., M. Burke, M. Xie, W. M. Davis, D. B. Lobell, and S. Ermon (2016). Combining satellite imagery and machine learning to predict poverty. Science 353(6301), 790–794.

Jeong, B., W. Lee, D.-S. Kim, and H. Shin (2016). Copula-based approach to synthetic population generation. PloS ONE 11(8), e0159496.

Jiang, D., W. Lin, and N. Raghavan (2020). A novel framework for semiconductor manufacturing final test yield classification using machine learning techniques. IEEE Access 8, 197885–197895.

Jiang, J. (2007). Linear and generalized linear mixed models and their applications. Springer Science & Business Media.

Jordon, J., J. Yoon, and M. Van Der Schaar (2019). PATE-GaN: Generating synthetic data with differential privacy guarantees. In International Conference on Learning Representations.

Kamthe, S., S. Assefa, and M. Deisenroth (2021). Copula flows for synthetic data generation. arXiv preprint arXiv:2101.00598.

Kish, L. (1965). Survey sampling. New York: John Wiley & Sons.

Koebe, T. (2020). Better coverage, better outcomes? Mapping mobile network data to official statistics using satellite imagery and radio propagation modelling. PLoS ONE 15(11 November), e0241981.

Koebe, T., A. Arias-Salazar, N. Rojas-Perilla, and T. Schmid (2022). Intercensal updating using structure-preserving methods and satellite imagery. Journal of the Royal Statistical Society: Series A (Statistics in Society), 1–27.

Kreutzmann, A.-K., S. Pannier, N. Rojas-Perilla, T. Schmid, M. Templ, and N. Tzavidis (2019). The R package emdi for estimating and mapping regionally disaggregated indicators. Journal of Statistical Software 91(7), 1–33.

Kroll, M. and R. Schnell (2016). Anonymisation of geographical distance matrices via Lipschitz embedding. International Journal of Health Geographics 15(1), 1–14.

Leasure, D. R., W. C. Jochem, E. M. Weber, V. Seaman, and A. J. Tatem (2020). National population mapping from sparse survey data: A hierarchical Bayesian modeling framework to account for uncertainty. Proceedings of the National Academy of Sciences 117(39), 24173–24179.

Leyk, S., A. E. Gaughan, S. B. Adamo, A. de Sherbinin, D. Balk, S. Freire, A. Rose, F. R. Stevens, B. Blankespoor, C. Frye, J. Comenetz, A. Sorichetta, K. MacManus, L. Pistolesi, M. Levy, and A. J. Tatem (2019, 9). The spatial allocation of population: a review of large-scale gridded population data products and their fitness for use. Earth System Science Data 11(3), 1385–1409.

Li, H., L. Xiong, and X. Jiang (2014). Differentially private synthesization of multidimensional data using copula functions. In Advances in database technology: proceedings. International conference on extending database technology, pp. 475–486.

Luna-Hernández, A. (2016). Multivariate structure preserving estimation for population compositions. Ph. D. thesis, University of Southampton.

Luna-Hernández, A., L.-C. Zhang, A. Whitworth, and K. Piller (2015a). Small area estimates of the population distribution by ethnic group in england a proposal using structure preserving estimators. Statistics in Transition new series 16(4), 585–602.

Luna-Hernández, A., L.-C. Zhang, A. Whitworth, and K. Piller (2015b). Small area estimates of the population distribution by ethnic group in England: A proposal using structure preserving estimators. Statistics in Transition 16(4), 585–602.

Mansfield, P. and A. A. Maudsley (1977). Medical imaging by nmr. British Journal of Radiology 50(591), 188–194.

Marker, D. A. (1999). Organization of small area estimators using a generalized linear regression framework. Journal of Official Statistics 15(1), 1–24.

Méndez, F. and O. Bravo (2011). Costa Rica Mapas de Pobreza 2011. Technical report, INEC Costa Rica, San José, Costa Rica.

Méndez, F. and O. Bravo (2011). Mapas de pobreza con datos censales. In Simposio: Costa Rica a la luz del Censo 2011.

MIT Data To AI Lab (2022). The synthetic data vault (SDV). `https://sdv.dev/`.

Molina, I. and J. Rao (2010). Small area estimation of poverty indicators. Canadian Journal of Statistics 38(3), 369–385.

Nelsen, R. B. (2007). An introduction to copulas. New York: Springer Science & Business Media.

Noble, A., S. Haslett, and G. Arnold (2002). Small area estimation via generalized linear models. Journal of Official Statistics 18(1), 45–60.

OECD (2016). OECD Economic Surveys: Costa Rica 2016.

Patki, N., R. Wedge, and K. Veeramachaneni (2016, Oct). The synthetic data vault. In 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA), pp. 399–410.

Pfeffermann, D. (2013). New important developments in small area estimation. Statistical Science 28, 40–68.

Pinkovskiy, M. and X. Sala-i Martin (2016, 5). Lights, Camera ... Income! Illuminating the National Accounts-Household Surveys Debate. The Quarterly Journal of Economics 131(2), 579–631.

Pokhriyal, N. and D. C. Jacques (2017, 11). Combining disparate data sources for improved poverty prediction and mapping. Proceedings of the National Academy of Sciences of the United States of America 114(46), E9783–E9792.

Pratesi, M. (2016). Analysis of poverty data by small area estimation. John Wiley & Sons.

Preston, S., H. P, and G. M (2001). Demography: measuring and modeling population processes. Oxford: Blackwell Publishers Ltd.

Purcell, N. J. and L. Kish (1980). Postcensal estimates for local areas (or domains). International Statistical Review 48(1), 3–18.

Rao, J. N. and M. Yu (1994). Small-area estimation by combining time-series and cross-sectional data. Canadian Journal of Statistics 22(4), 511–528.

Rao, J. N. K. (1986). Synthetic estimators, SPREE and the best model based predictors. In Proceedings of the Conference on Survey Research Methods in Agriculture, pp. 1–16. US Department of Agriculture Washington, DC.

Rao, J. N. K. (2003). Small area estimation. New York: Wiley.

Rao, J. N. K. and I. Molina (2015). Small area estimation. Hoboken: John Wiley & Sons.

Reiter, J. P. (2005). Using CART to generate partially synthetic public use microdata. Journal of Official Statistics 21(3), 441–462.

Rocher, L., J. M. Hendrickx, and Y. A. de Montjoye (2019). Estimating the success of re-identifications in incomplete datasets using generative models. Nature Communications 10(1), 1–9.

Rojas-Perilla, N., S. Pannier, T. Schmid, and N. Tzavidis (2020). Data-driven transformations in small area estimation. Journal of the Royal Statistical Society: Series A (Statistics in Society) 183(1), 121–148.

Sabharwal, N. and A. Agrawal (2021). Introduction to Word Embeddings, pp. 41–63. Berkeley, CA: Apress.

Schmid, T., F. Bruckschen, N. Salvati, and T. Zbiranski (2017). Constructing sociodemographic indicators for national statistical institutes by using mobile phone data: estimating literacy rates in Senegal. Journal of the Royal Statistical Society: Series A (Statistics in Society) 180(4), 1163–1190.

Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. Publications de l'Institut Statistique de l'Universit´e de Paris 8, 229–231.

Steinnocher, K., A. De Bono, B. Chatenoux, D. Tiede, and L. Wendt (2019). Estimating urban population patterns from stereo-satellite imagery. European Journal of Remote Sensing 52(sup2), 12–25.

Stevens, F. R., A. E. Gaughan, C. Linard, and A. J. Tatem (2015). Disaggregating census data for population mapping using Random forests with remotely-sensed and ancillary data. PLoS ONE 10(2), e0107042.

Subash, S. P., R. R. Kumar, and K. S. Aditya (2018). Satellite data and machine learning tools for predicting poverty in rural India. Agricultural Economics Research Review 31(2), 231–240.

Suesse, T., M.-R. Namazi-Rad, P. Mokhtarian, and J. Barthélemy (2017). Estimating cross-classified population counts of multidimensional tables: An application to regional australia to obtain pseudo-census counts. Journal of Official Statistics 33(4), 1021–1050.

Sun, Y., A. Cuesta-Infante, and K. Veeramachaneni (2019). Learning vine copula models for synthetic data generation. Proceedings of the AAAI Conference on Artificial Intelligence 33(01), 5049–5057.

Sundsøy, P. (2016). Can mobile usage predict illiteracy in a developing country? arXiv preprint arXiv:1607.01337.

Sáenz, I. (2002). Estimación de la cantidad de viviendas y consumo de agua. Master's thesis, University of Costa Rica.

Templ, M. (2017). Statistical disclosure control for microdata. Cham: Springer.

Templ, M., B. Meindl, A. Kowarik, and O. Dupriez (2017). Simulation of synthetic complex data: The R package simPop. Journal of Statistical Software 79, 1–38.

Torkzadehmahani, R., P. Kairouz, and B. Paten (2019). Dp-cgan: Differentially private synthetic data and label generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops.

Tzavidis, N., L. C. Zhang, A. Luna-Hernández, T. Schmid, and N. Rojas-Perilla (2018, 10). From start to finish: a framework for the production of small area official statistics. Journal of the Royal Statistical Society: Series A (Statistics in Society) 181(4), 927–979.

United Nations (1956). Manual III. Methods for population projections by sex and age.

United Nations (1973). Manual VII. Methods of projecting households and families.

United Nations (2015). The Millennium Development Goals Report 2015.

United Nations (2019). The Sustainable Development Goals Report 2019.

United Nations Department of Economic and Social Affairs (2009). Handbook on geospatial infrastructure in support of census activities (ST/ESA/STA ed.). New York, USA: United Nations Publication.

United Nations General Assembly (2015). Res 70/1. Transforming Our World: The 2030 Agenda for Sustainable Development. Technical report, United Nations General Assembly.

Wang, H. and J. P. Reiter (2012). Multiple imputation for sharing precise geographies in public use data. Annals of Applied Statistics 6(1), 229–252.

Warren, J. L., C. Perez-Heydrich, C. R. Burgert, and M. E. Emch (2016). Influence of demographic and health survey point displacements on distance-based analyses. Spatial Demography 4(2), 155–173.

Weidmann, N. B. and S. Schutte (2017, 3). Using night light emissions for the prediction of local wealth. Journal of Peace Research 54(2), 125–140.

West, B. T., A. Kirchner, D. Hochfellner, S. Bender, E. M. Nichols, M. H. Mulry, J. H. Childs, A. Holmberg, C. Bycroft, G. Benson, and F. Hubbard (2017). Establishing Infrastructure for the Use of Big Data to Understand Total Survey Error, Chapter 21, pp. 457–485. John Wiley & Sons, Ltd.

White, I. R., P. Royston, and A. M. Wood (2011, 2). Multiple imputation using chained equations: Issues and guidance for practice. Statistics in Medicine 30(4), 377–399.

WorldPop (2018). Global High Resolution Population Denominators Project.

Xu, L., M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni (2019). Modeling tabular data using conditional GAN. Advances in Neural Information Processing Systems 32, 1–11.

Ybarra, L. M. and S. L. Lohr (2008). Small area estimation when auxiliary information is measured with error. Biometrika 95(4), 919–931.

Zaloznik, M. (2011). Iterative proportional fitting - theoretical synthesis and practical limitations. Ph. D. thesis, University of Liverpool.

Zhang, J., G. Cormode, C. M. Procopiuc, D. Srivastava, and X. Xiao (2017). Privbayes: Private data release via bayesian networks. ACM Trans. Database Syst. 42(4), 1–41.

Zhang, L.-C. and R. L. Chambers (2004). Small area estimates for cross-classifications. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 66(2), 479–496.

Zhang, Z., T. Wang, N. Li, J. Honorio, M. Backes, S. He, J. Chen, and Y. Zhang (2021). PrivSyn: Differentially private data synthesis. In Proceedings of the 30th USENIX Security Symposium, pp. 929–946.

Summaries

Abstracts in English

Abstract: Updating Intercensal Health Indicators for Small Areas using the R Package spree

Due to the growing need to obtain quality estimates for small domains, several small area estimation models have been proposed. Most of these models require the use of census data, which in many cases are collected only every 10 years. Structure preserving estimation methods provide a solution to produce updated population characteristics by domains of interest in non-census years. In this paper, health indicators of the multidimensional poverty index of Costa Rica are estimated for planning regions and cantons from 2012 to 2017 using structure preserving estimation methods. Furthermore, this work shows how the process to update these indicators in non-census years is carried out with the help of the R package **spree**. This package permits the use of different structure preserving estimation methods to produce point and uncertainty estimates via parametric bootstrap. In addition, the user is provided with tools to prepare the data sets as needed for the updating process, and to compare different methods in a visual way.
Keywords: structure preserving estimation, official statistics, small area estimation, multidimensional poverty

Abstract: Small Area Estimates of Poverty Incidence in Costa Rica under a Structure Preserving Estimation (SPREE) Approach

Obtaining reliable estimates in small areas is a challenge because of the coverage and periodicity of data collection. Several techniques of small area estimation have been proposed to produce quality measures in small areas, but few of them are focused on updating these estimates. By combining the attributes of the most recent versions of the structure preserving estimation methods, this paper proposes a new alternative to estimate and update cross-classified counts for small domains, when the variable of interest is not available in the census. The proposed methodology is used to obtain and update estimates of the incidence of poverty in 81 Costa Rican cantons for six postcensal years 2012 - 2017. As uncertainty measures, mean squared errors are estimated via parametric bootstrap, and the adequacy of the proposed method is assessed with a design-based simulation.
Keywords: extreme poverty, intercensal updating, small area estimation, log-linear models

Abstract: Intercensal Updating using Structure Preserving Methods and Satellite Imagery

Censuses are fundamental building blocks of most modern-day societies, yet collected every ten years at best. We propose an extension of the widely popular census updating technique *structure preserving estimation* by incorporating auxiliary information in order to take ongoing subnational population shifts into account. We apply our method by incorporating satellite imagery as additional source to derive annual small-area updates of multidimensional poverty indicators from 2013 to 2020 for a population at risk: female-headed households in Senegal. We evaluate the performance of our proposal using data from two different census periods.
Keywords: multidimensional poverty, official statistics, small area estimation, SPREE

Abstract: Releasing Survey Microdata with Exact Cluster Locations and Additional Privacy Safeguards

Household survey programs around the world publish fine-granular georeferenced microdata to support research on the interdependence of human livelihoods and their surrounding environment. To safeguard the respondents' privacy, micro-level survey data is usually (pseudo)-anonymised through deletion or perturbation procedures such as obfuscating the true location of data collection. This, however, poses a challenge to emerging approaches that augment survey data with auxiliary information on a local level. Here, we propose an alternative microdata dissemination strategy that leverages the utility of the original microdata with additional privacy safeguards through synthetically generated data using generative models. We back our proposal with experiments using data from the 2011 Costa Rican census and satellite-derived auxiliary information. Our strategy reduces the respondents' re-identification risk for any number of disclosed attributes by 60-80% even under re-identification attempts.
Keywords: generative models, statistical disclosure control, geomasking, copula, official statistics, satellite imagery

Kurzzusammenfassungen auf Deutsch

Zusammenfassung: Kleinräumigen Aktualisierung von Gesundheitsindikatoren mit Hilfe des R Pakets spree

Aufgrund der zunehmenden Notwendigkeit qualitativ hochwertige Schätzungen für kleine Regionen zu erhalten, werden vermehrt Small-Area-Methoden vorgeschlagen. Die meisten dieser Modelle erfordern die Verwendung von Zensusdaten, die in vielen Fällen nur alle 10 Jahre erhoben werden. Structure PREserving Estimation Methoden bieten eine Lösung, um aktualisierte Bevölkerungsmerkmale bestimmter Gruppen oder Regionen in den Jahren nach der Volkszählung zu erstellen. In dieser Arbeit werden Gesundheitsindikatoren des multidimensionalen Armutsindexes von Costa Rica für Planungsregionen und Kantone von 2012 bis 2017 mit Structure Preserving Estimation Methoden geschätzt. Darüber hinaus zeigt diese Arbeit, wie der Prozess zur Aktualisierung dieser Indikatoren in den Jahren nach der Volkszählung mit dem R Paket **spree** durchgeführt werden kann. Dieses Paket ermöglicht die Verwendung verschiede-

ner strukturerhaltender Schätzmethoden zur Erstellung von Punkt- und Unsicherheitsschätzer mittels parametrischer Bootstraps. Des Weiteren werden Funktionen zur Verfügung gestellt, die den Anwender unterstützen, die für den Aktualisierungsprozess erforderlichen Datensätze vorzubereiten und verschiedene Methoden visuell zu vergleichen.
Schlüsselwörter: Structure Preserving Estimation Methoden, amtliche Statistik, kleinräumige Schätzung, multidimensionale Armut

Zusammenfassung: Kleinräumige Schätzungen der Armutsquote in Costa Rica mit Structure Preserving Estimation (SPREE) Methoden

Die Erstellung zuverlässiger Schätzungen von kleinräumigen Indikatoren ist aufgrund des Erfassungsumfangs und der Periodizität amtlicher Datenerhebung eine Herausforderung. Verschiedene Methoden, wie Indikatoren in kleine Gebieten zuverlässig geschätzt werden können, werden in der Literatur vorgeschlagen aber nur wenige fokussieren die Aktualisierung von Schätz. Durch die Kombination der Eigenschaften der neuesten Structure PREserving Estimation Methoden wird in dieser Arbeit eine neue Alternative zur Schätzung und Aktualisierung von kreuzklassifizierten Anzahlten für kleine Gebiete vorgeschlagen, wenn die interessierende Variable im Zensus nicht verfügbar ist. Die vorgeschlagene Methode wird verwendet, um Schätzungen der Armutsquote in 81 costa-ricanischen Kantonen für die sechs Jahre nach der Erhebung des Zensus, 2012 bis 2017, zu erhalten. Als Unsicherheitsmaße werden mittlere quadratische Fehler mittels parametrischer Bootstraps geschätzt, und die Güte der vorgeschlagenen Methode wird mit einer designbasierten Simulation untersucht.
Schlüsselwörter: Extreme Armut, Aktualisierung zwischen den Zensus, Schätzung für kleinräumige Indikatoren, log-lineare Modelle

Zusammenfassung: Zensusdaten aktualisieren mittels strukturerhaltender Methoden und Satellitenbildern

Volkszählungen sind grundlegende Bausteine der meisten modernen Gesellschaften, werden aber bestenfalls alle zehn Jahre erhoben. Wir schlagen eine Erweiterung der weit verbreiteten Technik zur Aktualisierung von Volkszählungen Structure Preserving Estimation vor, indem wir Hilfsinformationen einbeziehen, um laufende subnationale Bevölkerungsverschiebungen zu berücksichtigen. Wir wenden unsere Methode an, indem wir Satellitenbilder als zusätzliche Quelle einbeziehen, um jährliche kleinräumige Aktualisierungen multidimensionaler Armutsindikatoren von 2013 bis 2020 für eine gefährdete Bevölkerungsgruppe abzuleiten: von Frauen geführte Haushalte im Senegal. Wir bewerten den Mehrwert unseres Vorschlags anhand von Daten aus zwei verschiedenen Zählperioden.
Schlüsselwörter: Mehrdimensionale Armut, Amtliche Statistik, Kleinräumige Schätzung, SPREE

Zusammenfassung: Amtliche Mikrodaten veröffentlichen mit genauen Datenerhebungsstandorten und zusätzlichem Privatsphärenschutz

Programme zu Haushaltsbefragungen auf der ganzen Welt veröffentlichen detaillierte georeferenzierte Mikrodaten, um Forschung über die Abhängigkeit menschlicher Lebensumstände und ihrer Umgebung zu unterstützen. Um die Privatsphäre der Befragten zu schützen, werden Umfragedaten normalerweise (pseudo-)anonymisiert, indem Lösch- oder Störungsverfahren wie die Verschleierung des wahren Ortes der Datenerhebung durchgeführt werden. Dies stellt jedoch neue Ansätze, die Erhebungsdaten mit Hilfsinformationen auf lokaler Ebene ergänzen, vor eine Herausforderung. Hier schlagen wir eine alternative Veröffentlichungsstrategie für Mikrodaten vor, die den Nutzen der ursprünglichen Mikrodaten weitestgehend erhält und mit zusätzlichen Datenschutzvorkehrungen durch synthetisch generierte Daten unter Verwendung generativer Modelle schützt. Wir untermauern unseren Vorschlag mit Experimenten unter Verwendung von Daten aus der Volkszählung von 2011 in Costa Rica und von Satelliten abgeleiteten Hilfsinformationen. Unser Vorschlag reduziert das Reidentifikationsrisiko der Befragten für eine beliebige Anzahl von offengelegten Merkmalen um 60-80%, selbst nach Reidentifikationsversuchen.

Schlüsselwörter: Generative Modelle, Statistische Offenlegungskontrolle, Geomasking, Copula, Amtliche Statistik, Satellitenbilder

Declaration of Authorship

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name, in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text.

Berlin, November 24, 2022

Alejandra Arias-Salazar
November 24, 2022