Humanities & Social Sciences Communications

ARTICLE

https://doi.org/10.1057/s41599-023-01694-y OPEN

Releasing survey microdata with exact cluster locations and additional privacy safeguards

Till Koebe[1,4]✉, Alejandra Arias-Salazar[2,4] & Timo Schmid[3]

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)-anonymized 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.

[1] Saarland Informatics Campus, Saarland University, Saarbrücken, Germany. [2] School of Statistics, University of Costa Rica, San José, Costa Rica. [3] Institute of Statistics, Otto Friedrich University Bamberg, Bamberg, Germany. [4] These authors contributed equally: Till Koebe, Alejandra Arias-Salazar. ✉email: till.koebe@uni-saarland.de

Introduction

Statistics play an essential role in the quantitative study of phenomena that affect human societies. Official statistics provide data as a public good, thus facilitating research in the field of humanities as well as to define, monitor, and evaluate public policies. However, data access is in many cases regulated by privacy legislation. This leaves data producers in a difficult situation: finding the balance between disclosing unit-level data (i.e., containing individual- or household-specific information, commonly described as 'microdata') to foster use and to support knowledge generation and complying with relevant regulations. How this privacy-utility nexus is handled by data producers greatly affects the conditions under which quantitative research in the humanities can be performed, starting from available sample sizes over how measurement uncertainty needs to be addressed down to the reliability of derived *p*-values.

Since almost 100 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 (Leasure et al., 2020; Stevens et al., 2015), socio-demographic analysis (Chi et al., 2022; Fatehkia et al., 2020; Pokhriyal and Jacques, 2017; Schmid et al., 2017; Subash et al., 2018), policy targeting (Aiken et al., 2022; Blumenstock, 2018), environmental mapping (Grace et al., 2019) and health research (Arambepola et al., 2020; Brown et al., 2014). 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), de Jonge and de Wolf (2019), Templ (2017)).

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 (Blankespoor et al., 2021; Elkies et al., 2015; Hunter et al., 2021; Warren et al., 2016).

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 multi-dimensional joint distribution into conditional univariate distributions. Alfons et al. (2011b) 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), Rocher et al. (2019), Sun et al. (2019), Torkzadehmahani et al. (2019), Xu et al. (2019), Zhang et al. (2017), and others have used Bayesian networks, Generative Adversarial Networks or copulas to capture the underlying linear and non-linear relationships between the attributes.

Recently, also national statistical offices have started to experiment with synthetic data to overcome this privacy-utility nexus. For example, the US Census Bureau used the American Community Survey (ACS) (U.S. Census Bureau, 2022a)—a large socio-economic household survey interviewing a quarter million US households every month—to produce synthetic small-area estimates for historically under-counted communities such as American Indians and Native Alaskans and to protect the respondents' privacy. In addition, the US Census Bureau uses synthetic data based on the Survey on Income and Program Participation (SIPP) to create small-area estimates on poverty (U.S. Census Bureau, 2022b). Both surveys provide the quantitative baseline for numerous studies in the humanities, especially in the fields of cultural studies, demography, education studies, policy analysis, sociology and urban studies (e.g. Bokányi et al. (2016), Chan et al. (2018), Mitra and Brucker (2017), Spjeldnes and Choi (2008) and Topaz et al. (2022)). Furthermore, the Office for National Statistics of the UK published a case study in 2021 in which they experimented with a synthetic version of the Labor Force Survey—the largest household survey in the UK—so users can test the feasibility of their envisioned analysis before traveling to one of the Secure Research Services sites for microdata access (Bates et al., 2019). In 2023, the United Nations Economic Commission of Europe (UNECE)—with contributions from various national statistical offices—published 'Synthetic Data for Official Statistics—A Starter Guide' outlining implementation options and usage recommendations for the official statistics community (United Nations Economic Commission for Europe, 2022).

Going further, 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, by complementing existing approaches and 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. (2011b) to generate the synthetic population data AAT-SILC (*Artificial Austrian Statistics on Income and Living Conditions*(Alfons et al., 2011a). 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.

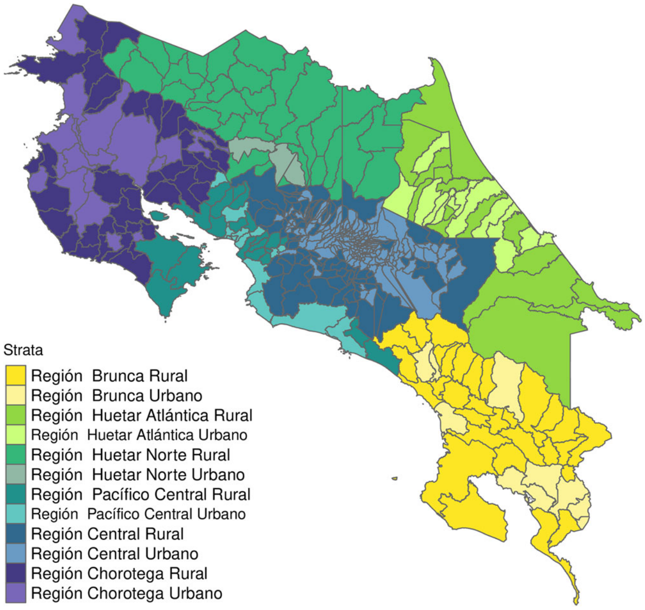

Strata
🟨 Región Brunca Rural
Región Brunca Urbano
Región Huetar Atlántica Rural
Región Huetar Atlántica Urbano
Región Huetar Norte Rural
Región Huetar Norte Urbano
Región Pacífico Central Rural
Región Pacífico Central Urbano
Región Central Rural
Región Central Urbano
Región Chorotega Rural
Región Chorotega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

Fig. 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.

Unsatisfied basic needs in Costa Rica

As our reference dataset in this project, we use data from Costa Rica—notably the Xth Population and VIth 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 10 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. Figure 1 shows the highest level of disaggregation (districts) of Costa Rica together with the 12 strata used in this paper.

For our experiment, we use a 10% random sample of the original 2011 census, which can be obtained from Instituto Nacional de Estadistica y Censos (2022) as a pseudo-population (see Table 1). The smallest geographical information available in this dataset are the 473 districts. In the first stage, we select districts as our PSUs

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

N	*n*	# of all PSUs	# of PSUs in *D*	# of attributes
427830	[7638; 11914]	767	123	106

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.

To exemplify the importance of an alternative microdata dissemination strategy, we select the Unsatisfied Basic Needs index (*Necesidades Básicas Insatisfechas (NBI)*)—a composite indicator similar to the multidimensional poverty index (MPI) (Alkire et al., 2019, Méndez and Bravo, 2011) used as a key statistical indicator in Costa Rica—as our target variable for survey augmentation in section "Utility for survey augmentation".

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., 1 year age difference) in one of the 19 underlying variables can turn a NBI-negative to a NBI-positive survey respondent. In 2011, 24.6% of Costa Rican households had one or more (out of four) unsatisfied basic needs, ranging from 9% in San Vicente, Santo Domingo to 90% in Cirripó, Turrialba (Méndez and Bravo, 2011). Since then, district-level NBI estimates have been derived from household sample surveys only, thus leaving out-of-sample districts without recent data and therefore less suitable to inform targeted policy interventions.

As auxiliary information, we use covariates derived from satellite imagery. Specifically, 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 (for methodological details see Supplementary Information section 1.2.) 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.

Results

We consider a survey D_{true} as a random sample with sample size n from a given population of size N. 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 $\mathbf{x} = X_1, \ldots, X_m$. Obfuscated attributes are denoted as $\mathbf{y} = Y_1, \ldots, Y_m$ 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 $D_{no} := (X_2, \ldots, X_m)$ 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:

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

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 2 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 -- the '*No Zip Code*' survey D_{no}. 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 attributes 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.

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
        $r_{x,v}^{\text{masked}} \leftarrow r_{x,v}$ + dist $* \cos$(angle) ;              /* Displace x-coordinate ($r_{x,v}$) */
        $r_{y,v}^{\text{masked}} \leftarrow r_{y,v}$ + dist $* \sin$(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 $D_{\text{geo}} := Y_{\text{zip}}, X_2, \ldots, X_m$.

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.

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 (Mansfield et al., 1977), 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 on the data transformation using frequency encoding are described in Algorithms 1 and 2 in section 1.1. of the Supplementary Information) using a Gaussian copula model is summarized in Algorithm 2.

Algorithm 2: Geomasked survey: DHS cluster displacement algorithm.

```
Input $\tilde{D}_{\text{true}} = (\tilde{X}_1, \dots, \tilde{X}_m)$
Output $\tilde{D}_{\text{syn}} = (\tilde{Y}_1, \dots, \tilde{Y}_m)$, 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 */
    $C_\Sigma^G(u_1, \dots, u_m) \leftarrow \phi_\Sigma(\phi_1^{-1}(u_1), \dots, \phi_m^{-1}(u_m))$ ;      /* 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 */
                $\tilde{\mathbf{y}}^{\{i\}} \leftarrow F^{-1}(\phi_2(w_2), \dots, \phi_m(w_m))$ ;       /* 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
```

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 $\mathbf{\Psi}$ and the multivariate Gaussian copula $C_\Sigma^G(u_1, \ldots, u_m)$ 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 $c_\Sigma^G(\mathbf{u})$ 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 section 1 of the Supplementary Information and in Nelsen (2007).

Figure 3 provides a first impression on the overall goodness-of-fit of the three different survey datasets (cf. with the evaluation step in Fig. 2). Specifically, Fig. 3a–c 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

$$Z_{KL}(f_{m,k}(X_{m,k}) || f_{m,k}(Y_{m,k})) = \frac{1}{1 + \delta_{KL}(f_{m,k}(X_{m,k}) || f_{m,k}(Y_{m,k}))}, \tag{2}$$

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 (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 modeling 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.

Risk of re-identifying private geocodes. To investigate the first shortcoming mentioned in section "Introduction", 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 $D_{\text{re}} := (\hat{X}_{\text{zip}}, X_2, \ldots, X_m)$ 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 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

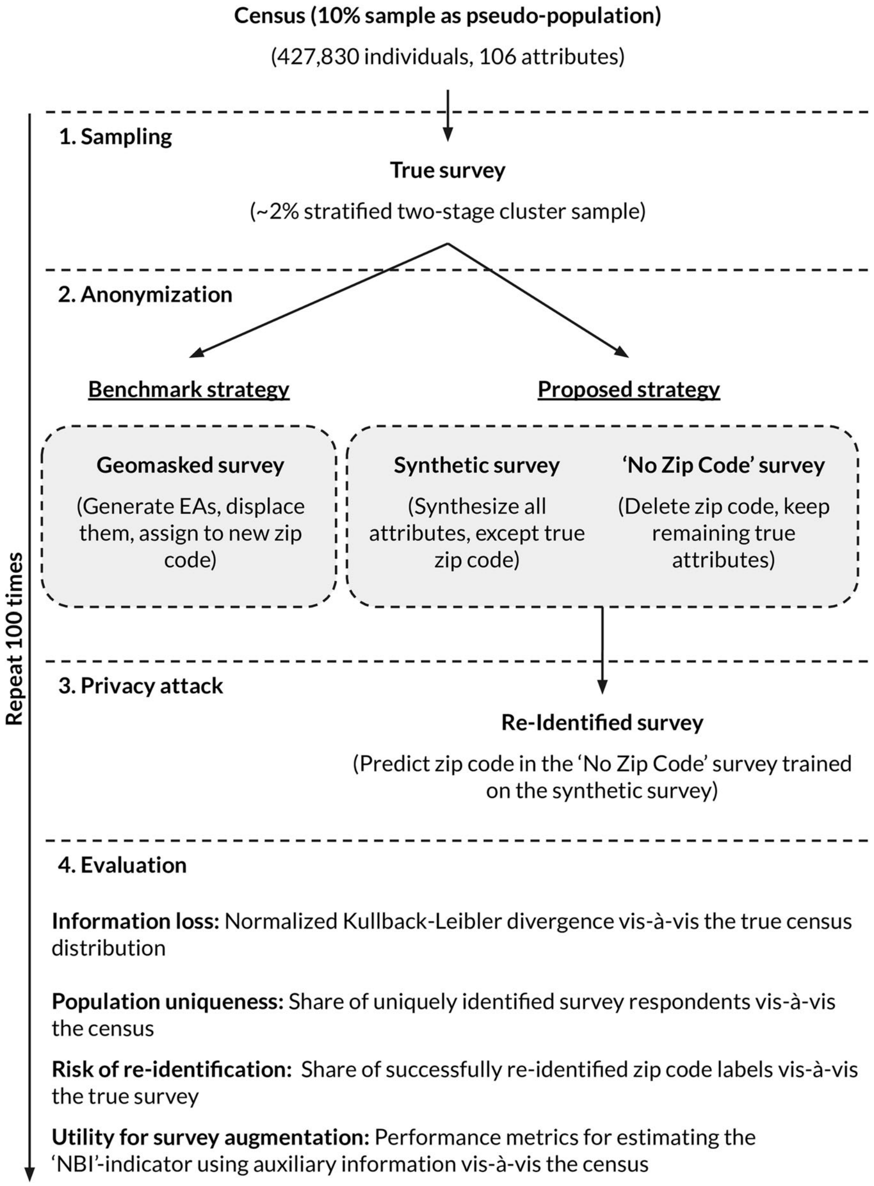

Census (10% sample as pseudo-population)
(427,830 individuals, 106 attributes)
1. Sampling
True survey
(~2% stratified two-stage cluster sample)
2. Anonymization
Benchmark strategy
Proposed strategy
Geomasked survey
(Generate EAs, displace them, assign to new zip code)
Synthetic survey
(Synthesize all 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

Fig. 2 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 Fig. 4), the resulting dataset is not further analyzed in the remaining study for the sake of readability.

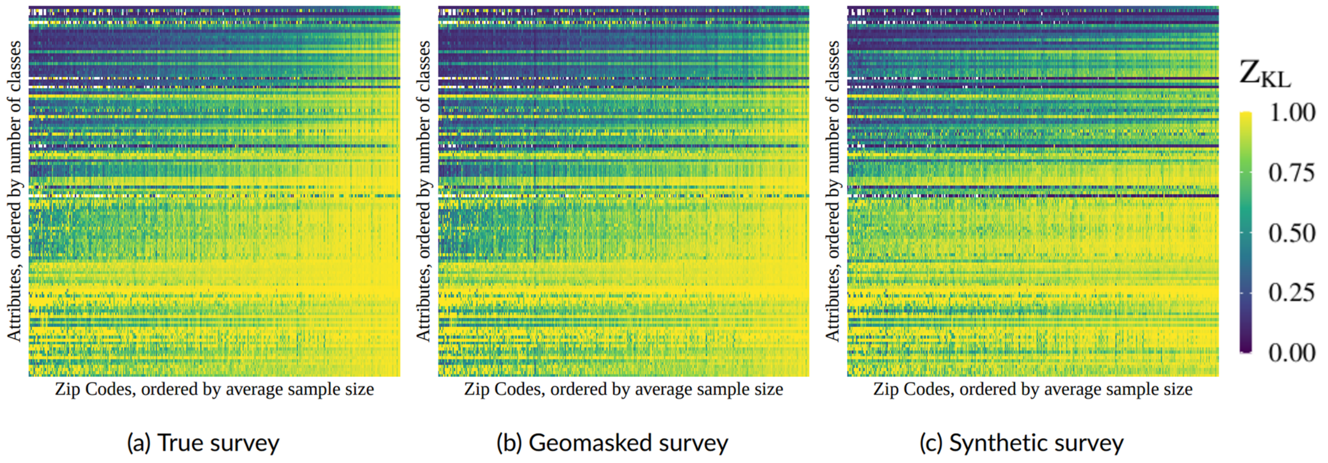
hosting only two small areas in some simulation runs, giving the random guess also a good chance to predict correctly. Recalling that roughly 70% of the displaced clusters stay within the same zip code in the geomasked survey, even predicting the sensitive attribute for strata hosting as little as two 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.

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. We denote the subsets with $D'(t)$, t with $1 \le t \le m$ being the number of attributes used for calculating the population uniqueness.

$$\Xi_t = \frac{1}{n}\sum_{i}^{n} \mathbb{1}_{i(t)} \text{ with } \quad \mathbb{1}_{i(t)} = \begin{cases} 1, & \text{if } i(t) \in D'(t) \text{ unique in population} \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

Figure 5 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%



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

Fig. 3 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 for each panel (**a**)–(**c**) 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 (**a**) scores best with $Z_{KL} = 0.76$ in total, followed by the synthetic survey (**c**) with $Z_{KL} = 0.74$ and the geomasked survey (**b**) at $Z_{KL} = 0.73$.

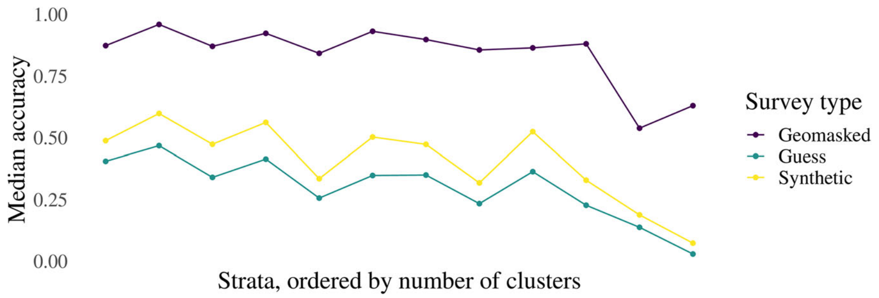


Fig. 4 Re-identification of the zip code as private attribute in the true survey for each stratum across 100 simulation runs. Accuracy is measured by the share of successfully re-identified zip code labels in the true survey. A random forest model is trained on perturbed data, i.e., the geomasked and 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.

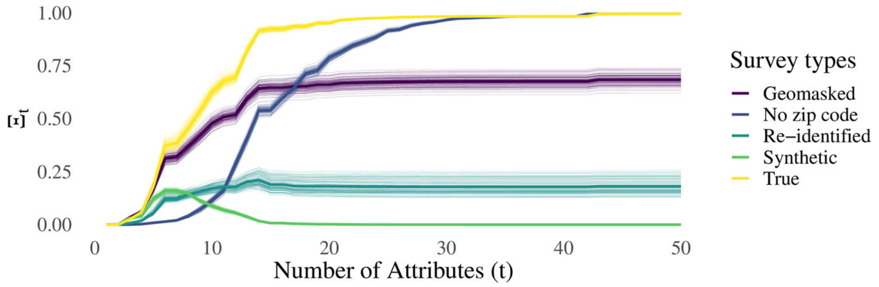
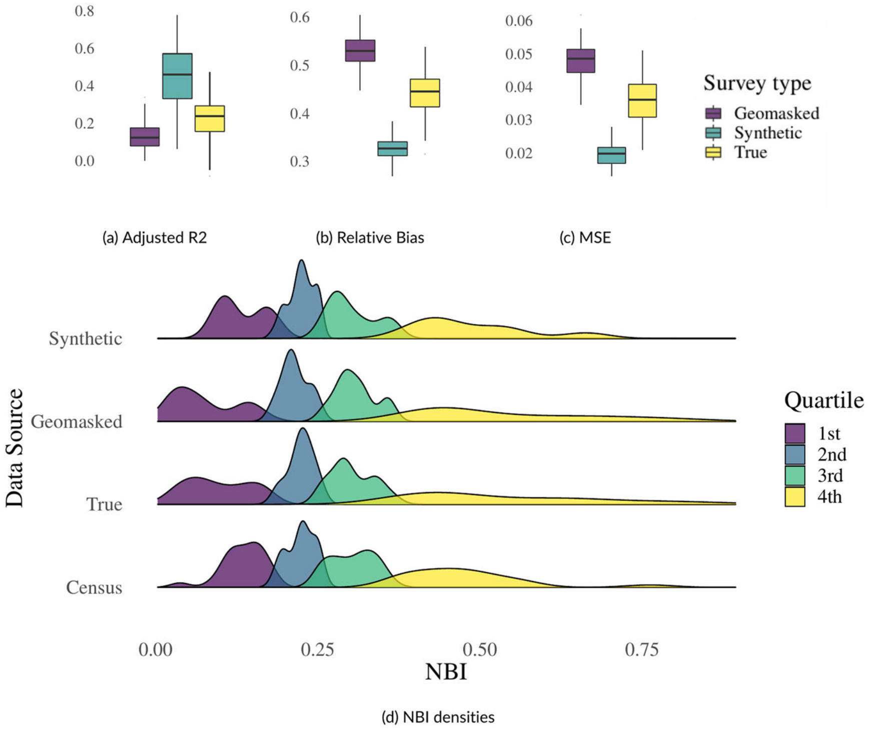


Fig. 5 Population uniqueness across survey types. Share of population-unique survey respondents for 100 simulation runs with a given number of attributes. Geographic identifiers are considered as part of the set of attributes. The thick lines represent the average population uniqueness across the 100 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.

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



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

(d) NBI densities

Fig. 6 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.

representing an actual population unique in a small (area) sample size setting. Therefore, Fig. 5 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 $\Xi_t \sim$ Binomial$\,(\mathbb{1}_{i(t)}, n)$ 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.

Utility for survey augmentation. To give an indication about the utility of the different anonymization approaches for survey data augmentation, we use a setup common in recent academic literature (cf. Leasure et al. (2020); Pokhriyal and Jacques (2017); 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 described in section "Unsatisfied basic needs in Costa Rica", we select the NBI as our target variable. We evaluate our predictions against the census in terms of adjusted R^2, bias and the Mean Squared Error (MSE). Figure 6a–c show the performance along these three evaluation criteria across 100 simulation runs.

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 Fig. 6d 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.

Generally, two strategies for computed indicators exist to create synthetic counterparts: (a) directly synthesize the computed indicators (previously shown in Fig. 6) 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. Table 2 shows that this not fully holds for the NBI.

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. Table 2).

Figure 7 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.

Discussion

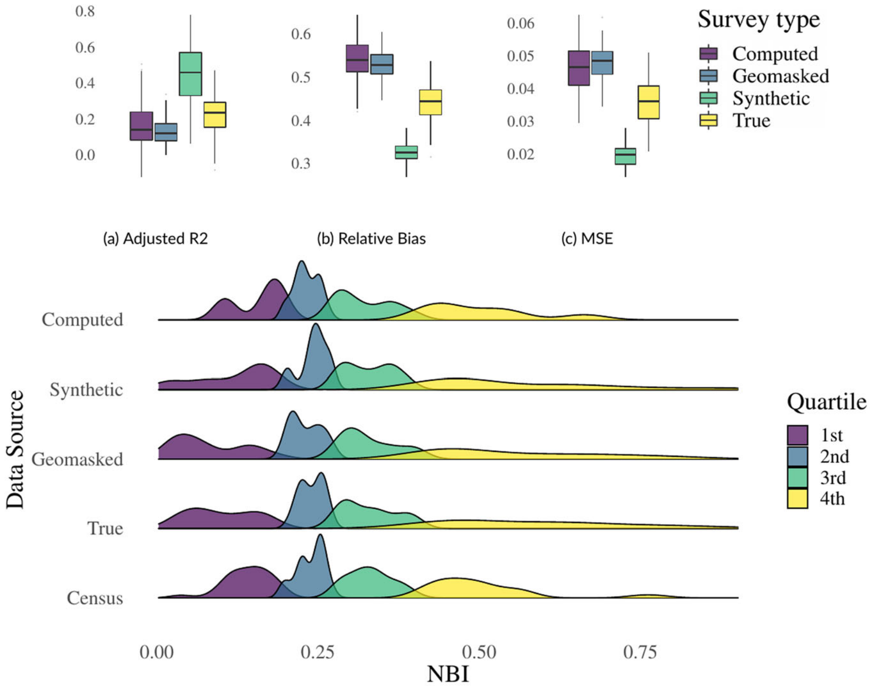
In this paper, we proposed and evaluated an alternative data dissemination strategy for micro-level survey data that improves the trade-off between privacy risk and data utility. Specifically, we showed that by publishing two datasets, namely the original survey data with limited geographic identifiers 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. Besides enabling data producers such as statistical offices and other survey programs (e.g., the Crime Victim Survey of the United Nations Office on Drugs and Crime (United Nations Office on Drugs and Crime, and United Nations Economic Commission for Europe, 2010), The Living Standard Measurement Study of World Bank (The World Bank Group, 2023), The Harmonized European Time Use Surveys (Eurostat, 2023) or the Latin American Public Opinion Project (Vanderbilt University, 2023) to expand the use cases of their data products, this methodology could especially 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. Looking at data use, the proposed methodology supports applications using fine-granular geolocated survey data in two ways: First, it helps to improve data access for users as better privacy protection of the survey data through synthetic data fosters regulatory compliance. Second, it avoids uncertainties where geospatial accuracy is crucial, e.g., connecting mobile network antenna to survey cluster characteristics or matching crop shares to satellite imagery on agricultural fields. Therefore, we regard our proposed microdata dissemination strategy as a way forward to ensure data users we still be able to access rich microdata without jeopardizing the respondents' privacy even under increasingly strict privacy legislation.

Table 2 Relationship between synthetic and computed NBI indicators across 100 simulation runs.

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

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.

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. 1 summarizes the results for our copula model being fitted on



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

(d) NBI densities

Fig. 7 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.

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. Supplementary Information, section 1.2.)—the shrinkage factor γ. Supplementary Fig. 2 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, 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. 3 shows that neither the encoding scheme nor the specification of the marginal distributions have large effects on the quality of the synthetically generated data.

Lastly, we show that our results are already stable after 50 simulation rounds (see Supplementary Fig. 4).

Nevertheless, our approach is not without limitations. As synthetic data generation is in its essence a modeling 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, the 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 modeled 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 modeled non-parametrically, hierarchical structures can be accounted for more rigorously by either modeling the hierarchies separately as suggested by Templ (2017) or by modeling 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.

Data availability

- The 10% sample of the Costa Rican census dataset is available from the microdata catalog of the national statistical office of Costa Rica—INEC—under a licensing agreement at: http://sistemas.inec.cr/pad5/index.php/catalog/113.
- The satellite-derived covariates are openly available from the WorldPop repository at: https://doi.org/10.5258/SOTON/WP00644.
- The shapefile of Costa Rica is available upon request from the authors.
- Computations were done with the statistical software *R* and the *sdv* package. The code to reproduce the study is available on GitHub: https://github.com/tilluz/survey_releases.

Received: 9 June 2022; Accepted: 17 April 2023;
Published online: 09 May 2023

References

Aiken E, Bellue S, Karlan D, Udry C, Blumenstock JE (2022) Machine learning and phone data can improve targeting of humanitarian aid. Nature 603:864–870. https://www.nature.com/articles/s41586-022-04484-9

Alfons A, Filzmoser P, Hulliger B, Kolb J-P, Kraft S, Münnich R, Templ M (2011a) Synthetic data generation of SILC data. Research Project Report WP6, D6.2. Tech. Rep., The AMELI Project. https://www.uni-trier.de/fileadmin/fb4/projekte/SurveyStatisticsNet/Ameli_Delivrables/AMELI-WP6-D6.2-240611.pdf

Alfons A, Kraft S, Templ M, Filzmoser P (2011b) Simulation of close-to-reality population data for household surveys with application to EU-SILC. Stat Methods Appt 20:383–407. https://doi.org/10.1007/s10260-011-0163-2

Alkire S, Kanagaratnam U, Suppa N (2019) The Global Multidimensional Poverty Index (MPI) 2019. OPHI MPI Methodological Note 47. Tech. Rep., Oxford Poverty and Human Development Initiative, University of Oxford. https://www.ophi.org.uk/wp-content/uploads/OPHI_MPI_MN_47_2019_vs2.pdf

Andrés ME, Bordenabe NE, Chatzikokolakis K, Palamidessi C (2013) Geo-indistinguishability: Differential privacy for location-based systems. In Proc. 2013 ACM SIGSAC Conf. Comput. Commun. Secur. 901–914. https://doi.org/10.1145/2508859.2516735

Arambepola R, Keddie SH, Collins EL, Twohig KA, Amratia P, Bertozzi-Villa A, Chestnutt EG, Harris J, Millar J, Rozier J et al. (2020) Spatiotemporal mapping of malaria prevalence in madagascar using routine surveillance and health survey data. Sci Rep 10:18129. https://doi.org/10.1038/s41598-020-75189-0

Armstrong MP, Rushton G, Zimmerman DL (1999) Geographically masking health data to preserve confidentiality. Stat Med 18:497–525. https://doi.org/10.1002/%28SICI%291097-0258%2819990315%2918%3A5%3C497%3A%3AAID-SIM45%3E3.0.CO%3B2-%23

Bates AG, Špakulová I, Dove I, Mealor A (2019) ONS methodology working paper series number 16—Synthetic data pilot. https://www.ons.gov.uk/methodology/methodologicalpublications/generalmethodology/onsworkingpaperseries/onsmethodologyworkingpaperseriesnumber16syntheticdatapilot#authors

Blankespoor B, Croft T, Dontamsetti T, Mayala B, Murray S (2021) Spatial anonymization: Guidance note prepared for the Inter-Secretariat working group on household surveys. Tech. Rep., UN Inter-secretariat Working Group on Household Surveys Task Force on Spatial Anonymization in Public-Use Household Survey Datasets. https://unstats.un.org/iswghs/task-forces/documents/Spatial_Anonymization_Report_submit01272021_ISWGHS.pdf

Blumenstock JE (2018) Estimating economic characteristics with phone data. AEA Pap Proc 108:72–76. https://www.aeaweb.org/articles?id=10.1257/pandp.20181033

Bokányi E, Kondor D, Dobos L, Sebők T, Stéger J, Csabai I, Vattay G (2016) Race, religion and the city: twitter word frequency patterns reveal dominant demographic dimensions in the united states. Palgrave Commun 2:1–9. https://doi.org/10.1057/palcomms.2016.10

Brown ME, Grace K, Shively G, Johnson KB, Carroll M (2014) Using satellite remote sensing and household survey data to assess human health and nutrition response to environmental change. Popul Environ 36:48–72. https://doi.org/10.1007/s11111-013-0201-0

Burgert CR, Colston J, Roy T, Zachary B (2013) Geographic displacement procedure and georeferenced data release health surveys. DHS spatial analysis reports. Tech. Rep. 7, ICF International, USAID, Calverton, Maryland, USA. https://dhsprogram.com/pubs/pdf/SAR7/SAR7.pdf

Chan SS, Gindling TH, Miller NA (2018) The effect of the affordable care act's dependent coverage provisionon health insurance gaps for young adults with specialhealthcare needs. J Adolesc Health 63:445–450. https://www.sciencedirect.com/science/article/pii/S1054139X18301952

Chi G, Fang H, Chatterjee S, Blumenstock JE (2022) Microestimates of wealth for all low- and middle-income countries. Proc Natl Acad Sci USA 119:e2113658119. https://doi.org/10.1073/pnas.2113658119

de Jonge E, de Wolf P-P (2019) sdcSpatial: Statistical Disclosure Control for Spatial Data. https://CRAN.R-project.org/package=sdcSpatial. R package version 0.1.1

Drechsler J, Dundler A, Bender S, Rässler S, Zwick T (2008) A new approach for disclosure control in the iab establishment panel-multiple imputation for a better data access. Adv Stat Anal 92:439–458. https://doi.org/10.1007/s10182-008-0090-1

Dwork C (2008) Differential privacy: a survey of results. In: Theory and applications of models of computation. TAMC 2008. Lecture notes in computer science, vol. 4978, 1–19. Springer, Berlin, Heidelberg

Elkies N, Fink G, Bärnighausen T (2015) "Scrambling" geo-referenced data to protect privacy induces bias in distance estimation. Popul Environ 37:83–98. https://doi.org/10.1007/s11111-014-0225-0

Eurostat The Harmonised European Time Use Surveys (HETUS) (2023) https://ec.europa.eu/eurostat/web/time-use-surveys

Fatehkia M, Coles B, Ofli F, Weber I (2020) The relative value of facebook advertising data for poverty mapping. Proc Int AAAI Conf Web Soc Media 14:934–938. https://ojs.aaai.org/index.php/ICWSM/article/view/7361

Grace K, Nagle NN, Burgert-Brucker CR, Rutzick S, Van Riper DC, Dontamsetti T, Croft T (2019) Integrating environmental context into DHS analysis while protecting participant confidentiality: a new remote sensing method. Popul Dev Rev 45:197–218. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6446718/

Granello DH, Wheaton JE (2004) Online data collection: Strategies for research. J Couns Dev 82:387–393. https://doi.org/10.1002/j.1556-6678.2004.tb00325.x

Heldal J, Iancu D-C (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. https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2019/mtg1/SDC2019_S1_Norway_Heldal_Iancu_AD.pdf

Hunter LM, Talbot C, Twine W, McGlinchy J, Kabudula CW, Ohene-Kwofie D (2021) Working toward effective anonymization for surveillance data: innovation at South Africa's Agincourt Health and Socio-Demographic Surveillance Site. Popul Environ 42:445–476. https://doi.org/10.1007/s11111-020-00372-4

ICF The DHS Program Spatial Data Repository (2022). https://spatialdata.dhsprogram.com/home/

Instituto Nacional de Estadistica y Censos X Censo Nacional de Población y VI de Vivienda (2022) Catálogo central de datos. http://sistemas.inec.cr/pad5/index.php/catalog/113

Janke T, Ghanmi M, Steinke F (2021) Implicit generative copulas. In: Adv. Neural Inf. Process. Syst. In: Ranzato M, Beygelzimer A, Dauphin Y, Liang P, Vaughan JW (eds), vol. 34, 26028–26039. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2021/file/dac4a67bdc4a800113b0f1ad67ed696f-Paper.pdf

Jeong B, Lee W, Kim D-S, Shin H (2016) Copula-based approach to synthetic population generation. PLoS ONE 11:e0159496. https://doi.org/10.1371/journal.pone.0159496

Jordon J, Yoon J, Van Der Schaar M (2019) PATE-GaN: generating synthetic data with differential privacy guarantees. In: International Conference on Learning Representations. https://openreview.net/forum?id=S1zk9iRqF7

Kamthe S, Assefa S, Deisenroth M (2021) Copula flows for synthetic data generation. https://arxiv.org/abs/2101.00598

Koebe T (2020) Better coverage, better outcomes? Mapping mobile network data to official statistics using satellite imagery and radio propagation modelling. PLoS ONE 15:e0241981. https://doi.org/10.1371/journal.pone.0241981

Koebe T, Arias-Salazar A, Rojas-Perilla N, Schmid T (2022) Intercensal updating using structure-preserving methods and satellite imagery. J R Stat Soc Ser A Stat Soc 185:S170–S196. https://doi.org/10.1111/rssa.12802

Kroll M, Schnell R (2016) Anonymisation of geographical distance matrices via Lipschitz embedding. Int J Health Geogr 15:1–14. https://doi.org/10.1186/s12942-015-0031-7

Leasure DR, Jochem WC, Weber EM, Seaman V, Tatem AJ (2020) National population mapping from sparse survey data: a hierarchical Bayesian modeling framework to account for uncertainty. Proc Natl Acad Sci USA 117:24173–24179. https://doi.org/10.1073/pnas.1913050117

Li H, Xiong L, Jiang X (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. https://doi.org/10.5441/002/edbt.2014.43

Mansfield P, Maudsley AA (1977) Medical imaging by NMR. Br J Radiol 50:188–194

Méndez F, Bravo, O (2011) Costa Rica Mapas de Pobreza 2011. Tech. Rep., INEC Costa Rica, San José, Costa Rica. https://www.inec.cr/sites/default/files/documentos/pobreza_y_presupuesto_de_hogares/pobreza/metodologias/documentos_metodologicos/mepobrezacenso2011-01.pdf.pdf

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

Mitra S, Brucker DL (2017) Income poverty and multiple deprivations in a high-income country: the case of the United States. Soc Sci Q 98:37–56

Nelsen RB (2007) An introduction to copulas. Springer Science & Business Media

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

Pokhriyal N, Jacques DC (2017) Combining disparate data sources for improved poverty prediction and mapping. Proc Natl Acad Sci USA 114:E9783–E9792. https://doi.org/10.1073/pnas.1700319114

Reiter JP (2005) Using CART to generate partially synthetic public use microdata. J Off Stat 21:441–462. https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/using-cart-to-generate-partially-synthetic-public-use-microdata.pdf

Rocher L, Hendrickx JM, de Montjoye YA (2019) Estimating the success of re-identifications in incomplete datasets using generative models. Nat Commun 10:3069. https://doi.org/10.1038/s41467-019-10933-3

Schmid T, Bruckschen F, Salvati N, Zbiranski T (2017) Constructing sociodemographic indicators for national statistical institutes by using mobile phone data: estimating literacy rates in Senegal. J R Stat Soc Ser A Stat Soc 180:1163–1190. https://doi.org/10.1111/rssa.12305

Sklar A (1959) Fonctions de répartition à n dimensions et leurs marges. Publ Inst Statist Univ Paris 8:229–231

Spjeldnes S, Choi J-K (2008) Direct and indirect effects of interparental relationship quality on child behavior problems in low-income, black, single-mother families. Marriage Fam Rev 44:411–438. https://doi.org/10.1080/01494920802453910

Stevens FR, Gaughan AE, Linard C, Tatem AJ (2015) Disaggregating census data for population mapping using Random forests with remotely-sensed and ancillary data. PLoS ONE 10:e0107042. https://doi.org/10.1371/journal.pone.0107042

Subash SP, Kumar RR, Aditya KS (2018) Satellite data and machine learning tools for predicting poverty in rural India. Agric Econ Res Rev 31:231–240. https://ageconsearch.umn.edu/record/284254

Sun Y, Cuesta-Infante A, Veeramachaneni K (2019) Learning vine copula models for synthetic data generation. Proc AAAI Conf Artif Intell 33:5049–5057. https://doi.org/10.1609/aaai.v33i01.33015049

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

Templ M, Meindl B, Kowarik A, Dupriez O (2017) Simulation of synthetic complex data: The R package simPop. J Stat Softw 79:1–38. https://www.jstatsoft.org/index.php/jss/article/view/v079i10

The World Bank Group Living Standards Measurement Study (LSMS) (2023) https://www.worldbank.org/en/programs/lsms

Topaz CM, Higdon J, Epps-Darling A, Siau E, Kerkhoff H, Mendiratta S, Young E (2022) Race-and gender-based under-representation of creative contributors: art, fashion, film, and music. Humanit Soc Sci Commun 9:221. https://doi.org/10.1057/s41599-022-01239-9

Torkzadehmahani R, Kairouz P, Paten B (2019) Dp-cgan: Differentially private synthetic data and label generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. https://openaccess.thecvf.com/content_CVPRW_2019/html/CV-COPS/Torkzadehmahani_DP-CGAN_Differentially_Private_Synthetic_Data_and_Label_Generation_CVPRW_2019_paper.html

United Nations Economic Commission for Europe (2022) Synthetic Data for Official Statistics: A Starter Guide. Geneva, Switzerland: United Nations. https://unece.org/statistics/publications/synthetic-data-official-statistics-starter-guide

United Nations Office on Drugs and Crime, and United Nations Economic Commission for Europe Manual on victimization surveys (2010) https://www.unodc.org/unodc/en/data-and-analysis/Manual-on-victim-surveys.html

U.S. Census Bureau American Community Survey (2022a) https://www.census.gov/programs-surveys/acs

Survey of Income and Program Participation (2022b) https://www.census.gov/programs-surveys/sipp.html

Vanderbilt University (2023) The Latin American Public Opinion Project (LAPOP). https://www.vanderbilt.edu/lapop/

Wang H, Reiter JP (2012) Multiple imputation for sharing precise geographies in public use data. Ann Appl Stat 6:229–252. https://doi.org/10.1214/11-AOAS506

Warren JL, Perez-Heydrich C, Burgert CR, Emch ME (2016) Influence of demographic and health survey point displacements on distance-based analyses. Spat Demogr 4:155–173. https://doi.org/10.1007/s40980-015-0014-0

West BT, Kirchner A, Hochfellner D, Bender S, Nichols EM, Mulry MH, Childs JH, Holmberg A, Bycroft C, Benson G, Hubbard F (2017) Establishing infrastructure for the use of big data to understand total survey error. chap. 21, 457–485. John Wiley & Sons, Ltd

WorldPop Global High Resolution Population Denominators Project (2018) www.worldpop.org

Xu L, Skoularidou M, Cuesta-Infante A, Veeramachaneni K (2019) Modeling tabular data using conditional GAN. Adv Neural Inf Process Syst 32:1–11. https://proceedings.neurips.cc/paper/2019/hash/254ed7d2de3b23ab10936522dd547b78-Abstract.html

Zhang J, Cormode G, Procopiuc CM, Srivastava D, Xiao X (2017) Privbayes: private data release via bayesian networks. ACM Trans Database Syst 42:1–41. https://doi.org/10.1145/3134428

Zhang Z, Wang T, Li N, Honorio J, Backes M, He S, Chen J, Zhang Y (2021) PrivSyn: differentially private data synthesis. In: Proceedings of the 30th USENIX Security Symposium, pp. 929–946. https://www.usenix.org/system/files/sec21fall-zhang-zhikun.pdf

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Ethical approval

This article does not contain any studies with human participants performed by any of the authors.

Informed consent

This article does not contain any studies with human participants performed by any of the authors.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1057/s41599-023-01694-y.

Correspondence and requests for materials should be addressed to Till Koebe.