Non-parametric bootstrap and small area estimation to mitigate bias in crowdsourced data. Simulation study and application to perceived safety

David Buil-Gil¹, Reka Solymosi¹ and Angelo Moretti²

¹Centre for Criminology and Criminal Justice, University of Manchester

²Social Statistics Department, University of Manchester

Abstract

Open and crowdsourced data are becoming prominent in social sciences research. Crowdsourcing projects harness information from large crowds of citizens who voluntarily participate into one collaborative project, and allow new insights into people's attitudes and perceptions. However, these are usually affected by a series of biases that limit their representativeness (i.e. self-selection bias, unequal participation, underrepresentation of certain areas and times). In this chapter we present a two-step method aimed to produce reliable small area estimates from crowdsourced data when no auxiliary information is available at the individual level. A non-parametric bootstrap, aimed to compute pseudo-sampling weights and bootstrap weighted estimates, is followed by an area-level model-based small area estimation approach, which borrows strength from related areas based on a set of covariates, to improve the small area estimates. In order to assess the method, a simulation study and an application to safety perceptions in Greater London are conducted. The simulation study shows that the area-level model-based small area estimator under the non-parametric bootstrap improves (in terms of bias and variability) the small area estimates in the majority of areas. The application produces estimates of safety perceptions at a small geographical level in Greater London from Place Pulse 2.0 data. In the application, estimates are validated externally by comparing these to reliable survey estimates. Further simulation experiments and applications are needed to examine whether this method also improves the small area estimates when the sample biases are larger, smaller or show different distributions. A measure of reliability also needs to be developed to estimate the error of the small area estimates under the non-parametric bootstrap.

Key words

EBLUP, modelling, Place Pulse, fear of crime, open data, reliability

Acknowledgements

The authors would like to thank Natalie Shlomo for comments that greatly improved the manuscript.

Full reference: Buil-Gil, D., Solymosi, R., & Moretti, A. (2020). Non-parametric bootstrap and small area estimation to mitigate bias in crowdsourced data. Simulation study and application to perceived safety. . In C. Hill, P. Biemer, T. Buskirk, L. Japiec, A. Kirchner, S. Kolenikov & L. Lyberg (Eds.), *Big data meets survey science*. Wiley.

1. Introduction

Open and crowdsourced data are shaping a new revolution in social research methods. A growing body of research in social sciences is applying crowdsourcing techniques to collect open data on social problems of great concern for governments and societies, such as crime and perceived safety (Salesses, 2009; Salesses et al., 2013; Solymosi and Bowers, 2018; Solymosi et al., 2017; Williams et al., 2017). Crowdsourcing techniques are defined here as methods for obtaining information by enlisting the services of large crowds of people into one collaborative project (Howe, 2006, 2008). Data generated through people's participation in these (generally) online platforms serving a variety of functions allow for analysing social problems, examining their causal explanations and even exploring their spatial and temporal patterns.

Such data already offer many advantages over traditional approaches to data collection (see Brabham, 2008; Goodchild, 2007; Haklay, 2013; Surowiecki, 2004). Some are highlighted later in this chapter (e.g. reduced cost of data collection, spatial information). It could even be suggested that crowdsourced data provide cheaper and more accurate geographical information than most traditional approaches (e.g. sample surveys). However, to reliably use these data, we must be confident in addressing the biases introduced through their unique mode of production.

Crowdsourced data have been repeatedly criticised due to biases arising from participants' self-selection and consequent non-representative data (Nielsen, 2006; Stewart et al., 2010). Studies looking into unequal participation in crowdsourced data have found systematic over-representation of certain groups: men tend to participate more than women in such activities, as well as employed people, citizens between ages 20-50, and those with a university degree are all more likely contributors (Blom et al., 2010; Solymosi and Bowers, 2018). Moreover, small groups of users are sometimes responsible for most observations (Blom et al., 2010; McConnell and Huba, 2006). As a consequence, although crowdsourced data allow renewed exploratory approaches to social problems, the level of representativeness of such data might be too small and the biases too large to produce direct analyses from these. Thus, new methods are required to analyse representativeness in crowdsourced data and to reduce their bias.

Some model-based techniques have been explored to increase the representativeness of crowdsourced samples, but most of these assume the availability of individual-level auxiliary information (e.g. age, gender, nationality, education level) about participants, which is needed to fit unit-level models (see Elliott and Valliant, 2017). While some crowdsourcing platforms record large samples of highly relevant variables, users do not provide auxiliary individual information apart from the measure of interest and the

geographical information. Some examples are: Place Pulse 2.0, which records data from respondents answering “Which place looks safer?” between two images from Google Street View (Salesses et al., 2013); FixMyStreet, a platform for reporting environmental issues, where over 90% of participations are anonymous and no auxiliary information is provided (Solymosi et al., 2017); and other online pairwise wiki surveys (Salganik and Levy, 2015).

In this research, we propose an innovative approach to reduce biases in crowdsourced data when there is no auxiliary information –with the exception of geo-location– available at the individual level. This chapter presents a non-parametric bootstrap followed by an area-level model-based small area estimation approach, which aims to increase the precision and accuracy of area-level estimates obtained from non-probability samples in crowdsourced data. First, we make use of a non-parametric bootstrap to estimate pseudo-sampling weights and produce area-level bootstrap weighted estimates. The non-parametric bootstrap reduces the implicit bias in crowdsourced data to allow for more reliable estimates. Second, by fitting an area-level model with available area-level covariates and producing Empirical Best Linear Unbiased Predictor (EBLUP) estimates, we borrow strength from related areas and produce estimates with increased precision (Fay and Herriot, 1979; Rao and Molina, 2015). In order to evaluate our approach, we conduct a simulation study and an application. The simulation study is based on a synthetic generated population, while in the application we produce estimates of perceived safety in Greater London from the Place Pulse 2.0 dataset (Salesses, 2009; Salesses et al., 2013).

This chapter is organised as follows. In section 2 we introduce the rise of crowdsourcing and emphasise the implications for its use in social science research. In section 3 we examine the main limitations associated with non-probability samples generated through crowdsourcing. Section 4 briefly introduces some of the main approaches explored to reduce the bias in crowdsourced data, most of which rely on the availability of respondents’ auxiliary information. Section 5 presents the non-parametric bootstrap approach followed by the area-level EBLUP. Section 6 is devoted to the simulation study, including the method to simulate the population and the evaluation of the estimator. In section 7 we apply the new method to estimate perceived safety in Greater London. Finally, section 8 draws conclusions and suggests future work.

2. The rise of crowdsourcing and implications

Crowdsourcing is a term that has gained reasonable traction since it was coined in 2006 by Jeff Howe, referring to harnessing information and skills from large crowds into one collaborative project (Howe, 2006, 2008). Since crowdsourcing originated in the open source movement in software, its definitions are rooted in online contexts, generally referring to it

as an online, distributed problem-solving and production model (Brabham, 2008). An early example of crowdsourcing is the photo-sharing website Flickr (www.flickr.com), where people upload their photographs and tag them with keywords. Others visiting the site can search through pictures using the assigned keywords. What is novel about the mode of production of these projects is that it is not reliant on a specific person to work or collect data until they meet certain requirements expected of them, but instead anyone can participate as much as they want. Then, the crowd's participation adds up to a complete output (Surowiecki, 2004).

A specific subset of crowdsourcing projects encourages people to submit spatial information about their local areas onto a combined platform, resulting in spatially explicit data. Such data is referred to as Volunteered Geographical Information (VGI), where various forms of geodata are provided voluntarily by individuals (Goodchild, 2007). The mechanism behind the creation of such VGI is 'participatory mapping', which refers to the practice of map making by people who contribute to the creation of a map to represent the topic of their expertise. People contribute their insight to collaboratively produce a representation of an area (Haklay, 2013).

Such community-based participatory research has been used to better understand social problems, and it has gained respect for aiming to highlight everyone's experiences in a space equally. These data collection approaches are not one-sided, instead they also serve to collect data to influence direct decision making. The outputs from such data can be used to lobby for changes in their neighbourhoods, contributing to a reversal of the traditional top-down approach to the creation and dissemination of geographic information (Goodchild, 2007). For example, citizens involved with collecting data about noise pollution in their area can use that information as evidence-base when lobbying for interventions by local authorities (Becker et al., 2013). VGI created by citizens can provide an alternative to traditional authoritative information from mapping agencies, and it can even be used for emergency management. During wildfires in Santa Barbara, California, in 2007-2009, volunteer maps online (some of which accumulated over 600,000 hits) provided essential information about the location of the fire, evacuation orders, emergency shelters, and other useful information (Goodchild and Glennon, 2010).

The above examples illustrate some benefits of the mode of production of data generated by these projects, alongside the bonus of their eliciting participation in large numbers. However, they also incur many biases in the sample of participants, which need to be taken into account, especially if such data are going to be used for research purposes. Traditional approaches to data collection for the purposes of drawing statistical inference have paid careful attention to addressing these biases. It is important that if crowdsourced data are used to answer research questions, then similar care should be taken. To support

this, the next section discusses some of the limitations of crowdsourced data from the viewpoint of possible biases in the non-probability samples of participants who generate the content in such projects.

3. Crowdsourcing data to analyse social phenomena: limitations

Researchers are making increasing use of data produced via crowdsourcing, innovating in various fields across the social sciences. Some of these papers also acknowledge the biases inherent from the mode of production of these data (e.g. Malleon and Andresen, 2015; Williams et al., 2017). While often acknowledged, these issues are usually lightly touched upon in a limitations section, and raised as something to be ‘kept in mind’. However, processes to understand and account for these biases are required to make the best possible use of these data. To better understand their effect, we first consider some sources of bias in crowdsourced data.

3.1 Self-selection bias

Participation in crowdsourcing activities is driven by a variety of factors, some discussed above. Therefore, crowdsourced data might be affected by biases arising from people’s self-selection: the sample that contributes to such data is self-selected, giving way for people more motivated to speak about the issue. As noted by Longley (2012), “self-selection is an enemy of robust and scientific generalisation, and crowdsourced consultation exercises are likely to contain inherent bias” (p. 2233).

Beyond motivation as a driver of this bias, an entire body of work has explored the impacts of the digital divide, which refers to certain socioeconomic groups being overrepresented in these data due to technological literacy (e.g. Yu, 2006; Fuchs, 2008). These systematic biases need to be accounted for when analysing crowdsourced data. Gender bias has been found, showing that men tend to participate more in such activities than women: Salesses et al. (2013) examined Place Pulse 1.0 data and found that the 78.3% of participants who reported their gender were males. Further work on VGI participation has also shown unequal participation along many socio-demographic characteristics: employed people, citizens aged between 20 and 50, and those with a university degree are most likely to participate (Haklay, 2010).

Further, area-level characteristics also have an effect; who participates and where people participate are influenced by various external factors. Mashhadi et al. (2013) find that socio-economic factors, such as population density, dynamic population, distance from the

centre and poverty, all play an important role to explain unequal participation in Open Street Map; while analyses of data from FixMyStreet show that the number of reports is positively correlated with neighbourhood-level measures of deprivation (Solymosi et al., 2017).

3.2 Unequal participation

In crowdsourcing projects, it is often observed that few users are responsible for most crowdsourced information, while the majority participate only a few times. This concept is known as participation inequality. In economics and social sciences, this is sometimes referred to as the Pareto principle, which states that approximately 80% of the observed effect comes from 20% of the units observed (Sanders, 1987). The concentration is also observed in other social sciences, such as criminology, where crime calls concentrate in small units: 3.5% of the addresses in Minneapolis produced 50% of all calls to the police in a single year (Weisburd, 2015).

In crowdsourced projects, this discrepancy is even greater, as participation inequality has been noted to follow a 90-9-1 rule. Stewart et al. (2010) identified that about 90% of users are ‘outliers’, who read or observe, but do not contribute to the project. Then, 9% of users contribute occasionally (contributors), and 1% of users account for almost all the contributions (super contributors). For example, in 2006, Wikipedia had only 68,000 active contributors, which was 0.2% of the 32 million visitors it had in the United States, and the most active 1,000 people (0.003% of its users) contributed about two-thirds of the site’s edits (Nielsen, 2006). Furthermore, Dubey et al. (2016) show that 6,118 of the 81,630 users of Place Pulse 2.0 participated only once, while 30 users participated more than 1,000 times and one user provided 7,168 contributions. This is an extreme distribution of the Pareto principle, and it has been termed the “1% rule of the Internet” by McConnell and Huba (2006).

3.3 Under-representation of certain areas and times

Interestingly, there is another bias that is introduced by the under-representation of certain areas and times. In VGI projects, users decide when and where to submit reports, and these decisions are reflected in the under and over-representation of certain areas and times in the sample. For example, Antoniou et al. (2010) looked at the geographical distribution of geotagged photos uploaded to platforms such as Picasa and Flickr, and they found that these cluster in urban areas and tourist attractions, with sparse coverage in rural areas. Furthermore, crowdsourcing applications that wish to gain insight into people’s perception of safety can also suffer from people’s avoidance of areas which they perceive to be most unsafe (Solymosi et al., 2017). With respect to the under-representation of certain times,

Blom et al. (2010) note that participation is five times higher at noon, while the number of participants during the night is almost nonexistent.

3.4 Unreliable area-level direct estimates and difficulty to interpret results

Due to the biases described in this section, and other possible sources of bias such as nonresponse and attrition (see Elliott and Valliant, 2017), it becomes probable that aggregating responses and producing area-level direct estimates from crowdsourced data might lead to biased and unreliable estimates. Such estimates are not only difficult to interpret, but also can contribute to erroneous and spurious theoretical explanations of social phenomena. As crowdsourcing is a growing methodological approach, it becomes important to address these issues, in order to create a refined methodology. In the next section we discuss previous approaches to reweighting crowdsourced data, before we introduce a non-parametric bootstrap algorithm followed by an area-level EBLUP as one possible approach to address these biases when individual auxiliary information is not available.

4. Previous approaches for reweighting crowdsourced data

In cases of crowdsourced datasets that record auxiliary information from participants (e.g. gender, age, income, education level), different approaches have been used to reduce their sample bias and adjust the non-probability samples to the target population distributions (see Elliott and Valliant, 2017). Most of these approaches estimate pseudo-sampling weights to correct for the bias in non-probability samples (e.g. Baker et al., 2013; Elliott, 2009; Elliott et al., 2010). Selection bias in web-surveys can be corrected following a quasi-randomisation approach (Valliant et al., 2013). Moreover, a reference survey with the same covariates of the non-probability survey can be used to make statistical inference possible (Schonlau et al., 2007). Another strategy is sample matching (Baker et al., 2013). Sample matching can be performed at individual or aggregate level. Hierarchical regression modelling may also be used (see Elliot and Valliant, 2017). Wang et al. (2015) propose a multilevel regression and poststratification (MRP) method, which is an extension of the hierarchical regression modelling. Other common techniques used to correct for selection bias are Bayesian Additive Regression Trees (BART), Inverse Probability Bootstrapping (Nahorniak et al., 2015), Propensity Score Adjustment (Lee, 2006), and the Least Absolute Shrinkage and Selection Operator, LASSO (Chen, 2016).

However, some crowdsourcing platforms do not record participants' auxiliary information beyond the target variable and the geographical information of the target place or responding person (e.g. Place Pulse 2.0). For such cases, Arbia et al. (2018) propose a two-phase approach, which does not make use of individual-level auxiliary information, to reduce the bias and allow for statistical inference from crowdsourced data. In the first phase, which aims to reduce non-sampling errors, standard and spatial outliers are detected, removed and replaced with the average of the neighbouring observations. Spatial outliers are defined here as values that exceed r times the standard deviation of the average values in each area. The second post-sampling phase aims to reweight the responses to let the data resemble an optimal spatial sample design. In each area, pseudo-sampling weights are calculated as the ratio between the number of observations available and the number of observations required by an optimal sampling design. The estimation of the outcome measure in each area is finally obtained as a weighted average using the pseudo-sampling weights. Here, we suggest and explore a different approach and present a non-parametric bootstrap algorithm followed by a model-based area-level small area estimation approach.

5. A new approach: small area estimation under a non-parametric bootstrap estimator

In order to reduce the biases in crowdsourced data and produce more reliable area-level estimates when no individual auxiliary information –besides the geographical information– is available, we introduce a non-parametric bootstrap followed by an area-level small area estimation approach. This is based on the non-parametric bootstrap technique studied in general by Efron and Tibshirani (1993), as well as the inverse probability bootstrap approach studied by Nahorniak et al. (2015). This method is designed to produce small area estimates from crowdsourced datasets that record only the outcome variable and the geographical information (of target place or respondent), but no other individual auxiliary information is available.

Let U be the finite target population, which is partitioned into D areas, $U_1, \dots, U_D$, of sizes $N_1, \dots, N_D$. Our aim is to estimate the population mean of a variable of interest Y given by the following formula:

$$\bar{Y}_d = \frac{\sum_{i \in U_d} y_{di}}{N_d}, d = 1, \dots, D, \quad (1)$$

where y_{di} is the observation of the variable of interest Y for unit i from area d , and N_d is the dimension population in area d .

Traditionally, the Horvitz-Thompson estimator (Horvitz and Thompson, 1952) is used to provide unbiased direct estimates of $\bar{Y}_d$. This is defined as follows:

$$\hat{\bar{Y}}_d^{HT} = \frac{\sum_{i \in s_d} w_{di} y_{di}}{\sum_{i \in s_d} w_{di}} \quad (2)$$

where w_{di} is the survey weight of unit i in area d given by the inverse of the first-order inclusion probability of unit i . Unfortunately, the Horvitz-Thompson estimator cannot be directly used due to the non-probabilistic nature of crowdsourced data. In order to directly adjust the crowdsourced sample to the target population, unit-level auxiliary information (e.g. age, gender, ethnicity) is needed to calculate the pseudo-sampling weights (e.g. Elliott and Valliant, 2017). Thus, in cases where there is no individual auxiliary information in crowdsourced data, we suggest following two steps to reduce the unrepresentativeness: a non-parametric bootstrap algorithm (Step 1) followed by an area-level EBLUP (Step 2).

5.1 Step 1: Non-parametric bootstrap

First, a non-parametric bootstrap approach, which draws *stratified simple random samples with replacement* (SSRSWR) based on simplified optimal sample sizes per area (Yamane, 1967), is used to estimate bootstrap pseudo-sampling weights and bootstrap weighted estimates. Weights are computed as the inverse of the first-order inclusion probability (Särndal et al., 1992). The non-parametric bootstrap estimates are the average bootstrap weighted estimate across all bootstrap replicates.

Nahorniak et al. (2015) use pseudo-sampling weights to generate weighted bootstrap samples, and they show that unequal probability samples can be transformed into equal probability data by using the inverse of the original sample inclusion probabilities in a bootstrapping process. In our case, no auxiliary information (apart from the geographies) is available, and thus we estimate pseudo-sampling weights as the inverse of the first-order inclusion probability based on simplified optimal sample sizes per area in each bootstrap replicate.

The bootstrap algorithm steps are listed below:

1. From an observed non-probability sample s selected from a finite population U , draw a sample for each area $d = 1, \dots, D$ using SSRSWR and obtain $y_{di}^{*(b)}$, which denotes the observation of variable Y for unit i in area d for the b^{th} bootstrap replicate. Note that certain units of the original sample may be missing in the SSRSWR, and other elements might be present two or more times; this is due to the nature of sampling designs with repetition. The sample sizes per area selected in each replicate are

obtained via the simplified optimal sample size: $n_d^{Yamane} = \frac{N_d}{1+N_d(h)^2}$, where N_d is the population size in area d and h is the chosen margin of error (Yamane, 1967, p. 886). Here we suggest $h = 0.01$ (99% confidence interval) to maximise the bootstrap performance: this will be the chosen margin of error in the simulation study and application shown below. By selecting SSRSWR with n_d equal to the ideal sample size in each stratum, we adjust the bootstrap method to the optimal sample size in each area, and control that the new method can be applied regardless the size of the target population. The pseudo-sampling weights are calculated as the inverse of first-order inclusion probability (see Särndal et al., 1992 for details about first-order inclusion probabilities in case of sampling designs with repetitions). These are denoted by w_{di}^{boot} :

$$w_{di}^{boot} = [1 - (1 - \frac{1}{n_d})^{n_d^{Yamane}}]^{-1}, \quad (3)$$

where n_d is the original sample size in area d and n_d^{Yamane} refers to the calculated simplified optimal size in area d .

2. The adjusted estimates of $\bar{Y}_d$ in each b^{th} replication are obtained by

$$\hat{Y}_d^{*(b)} = \frac{\sum_{i \in s_d} w_{di}^{boot} y_{di}^{*(b)}}{\sum_{i \in s_d} w_{di}^{boot}}. \quad (4)$$

3. Repeat steps 1 and 2 for $b = 1, \dots, B$ replicates and obtain the following Monte-Carlo approximation of the non-parametric bootstrap estimator:

$$\hat{Y}_d^{Boot} = B^{-1} \sum_{b=1}^B \hat{Y}_d^{*(b)}, \quad (5)$$

which is the non-parametric bootstrap estimator of $\bar{Y}_d$.

5.2 Step 2: Area-level model-based small area estimation

Second, the traditional area-level EBLUP estimator, which is based on the Fay-Herriot model (Fay and Herriot, 1979), is used to borrow strength from available area-level auxiliary information. In small area estimation, area-level models relate the area means or totals (in this case, the bootstrap estimates) to area-level covariates (Rao and Molina, 2015). Thus, available area-level covariates with strong relations with our variable of interest are needed to increase the precision of our estimates (Rao and Molina, 2015): this step relies on the availability of covariates strongly related to our outcome measure.

The original area-level EBLUP makes use of the Horvitz-Thompson estimator given in Eq. 2 and its variance. In this work, however, we make use of the bootstrap estimate (Eq. 5) and assume

$$\hat{Y}_d^{Boot} = \bar{Y}_d + e_d, \quad e_d \sim N(0, \psi_d), \quad d = 1, \dots, D, \quad (6)$$

where ψ_d is the variance of bootstrap estimates (Eq. 4) in area d . Then, we assume $\bar{Y}_d$ to be linearly related to a set of area-level covariates $\mathbf{x}'_d$:

$$\bar{Y}_d = \mathbf{x}'_d \boldsymbol{\beta} + v_d, \quad v_d \sim N(0, A), \quad d = 1, \dots, D, \quad (7)$$

where v_d is independent from e_d . Thus,

$$\hat{Y}_d^{Boot} = \mathbf{x}'_d \boldsymbol{\beta} + v_d + e_d, \quad v_d \sim N(0, A), \quad e_d \sim N(0, \psi_d), \quad d = 1, \dots, D. \quad (8)$$

The area-level Best Linear Unbiased Predictor (BLUP) of $\bar{Y}_d$ is computed as

$$\hat{Y}_d^{BLUP} = \hat{Y}_d^{Boot} - \frac{\psi_d}{A + \psi_d} [\hat{Y}_d^{Boot} - \mathbf{x}'_d \hat{\boldsymbol{\beta}}(A)]. \quad (9)$$

where $\hat{\boldsymbol{\beta}}(A)$ is the maximum likelihood estimator of $\boldsymbol{\beta}$. If we replace $\gamma_d(A) = \psi_d / (A + \psi_d)$, then:

$$\hat{Y}_d^{BLUP} = [1 - \gamma_d(A)] \hat{Y}_d^{Boot} + \gamma_d(A) \mathbf{x}'_d \hat{\boldsymbol{\beta}}(A). \quad (10)$$

Since in real applications A is unknown, we need to replace it by an estimator $\hat{A}$. In this case, $\hat{A}$ is obtained via Restricted Maximum Likelihood method (REML). After we replace A by $\hat{A}$ we obtain the EBLUP (Rao and Molina, 2015):

$$\hat{Y}_d^{EBLUP} = [1 - \gamma_d(\hat{A})] \hat{Y}_d^{Boot} + \gamma_d(\hat{A}) \mathbf{x}'_d \hat{\boldsymbol{\beta}}(\hat{A}). \quad (11)$$

6. Simulation study

This simulation study is designed to explore the performance of the bootstrap (Eq. 5) and EBLUP estimators (Eq. 11) in terms of bias and mean squared error. The study is based on generating one fixed population and drawing random samples with replacement, which is a mixture between a design and model-based simulation approach.

6.1 Population generation

The population is generated from the following unit-level linear mixed-effect model (Battese et al., 1988):

$$y_{di} = x_{di1}\beta_1 + x_{di2}\beta_2 + e_{di} + u_d, \quad (12)$$

where x_{di1} and x_{di2} are the values of the first and second covariates for unit i in area d , β_1 and β_2 are the regression coefficients of covariates 1 and 2, e_{di} refers to the individual error of unit i in area d , and u_d denotes the area effects of area d . The parameters used to fit the linear mixed-effect model have been obtained from a unit-level linear model of perceived safety using data from the European Social Survey 5 (ESS), in which we make use of two covariates (age and gender). Age and gender have been highly analysed in safety perceptions research and are known to be related to our outcome measure (see Hale, 1996). The simulation parameters are then: $\beta_1 = 0.004$, $\beta_2 = 0.50$, $\sigma^2 = 0.50$ and $\sigma_u^2 = 0.02$. $e_{di} \sim N(0, \sigma^2)$ and $u_d \sim N(0, \sigma_u^2)$. x_{di1} values are produced from a normal distribution using parameters from the age distribution in the European Social Survey 5 ($\bar{x}_1 = 48.34$ and $sd(x_1) = 46.69$), while x_{di2} values are produced from a Bernoulli distribution with parameter 0.5 (equal probabilities for males and females). The population size is $N = \sum_{d=1}^D N_d = 30046$, in which N_d is produced from a uniform distribution between 100 and 300 ($\min(N_d) = 100$, $med(N_d) = 195$, $\bar{N}_d = 200.3$, $\max(N_d) = 298$), and $D = 150$. Table 1 shows a summary of the quantities used in the computations for generating the population.

Quantity	Description
d	Values between 1 and 150, in which each value refers to an area d . The population size per area is produced from a uniform distribution between 100 and 300.
x_{di1}	Normal distribution from $\bar{x}_1 = 48.34$ and $sd(x_1) = 46.69$ (obtained from ESS data).
x_{di2}	Bernoulli distribution with parameter 0.5.
β_1	0.004 (obtained from model fitted from ESS data).
β_2	0.50 (obtained from model fitted from ESS data).
σ^2	0.50 (obtained from model fitted from ESS data).
σ_u^2	0.02 (obtained from model fitted from ESS data).
e_{di}	Normal distribution from $\bar{e} = 0$ and $sd(e) = \sqrt{\sigma^2}$.
u_d	Normal distribution from $\bar{u} = 0$ and $sd(u) = \sqrt{\sigma_u^2}$.
y_{di}	$y_{di} = x_{di1}\beta_1 + x_{di2}\beta_2 + e_{di} + u_d$.

Table 1. Summary of the quantities used to generate the population.

6.2 Sample selection and simulation steps

The simulation consists in the following steps:

1. Selection of $t = 1, \dots, T$ ($T = 500$) samples from two-stage SSRSWR and an unequal probability selection design. Sampling probabilities were computed from the calibration of the proportion of units according to their age group and gender to such proportion in a real exemplar crowdsourced dataset: Place Pulse 1.0. Note that unlike Place Pulse 2.0, which does not record participants' auxiliary information, the Place Pulse 1.0 platform asked participants about their age, gender and others. 76% from the 97.1% respondents who reported their gender at Place Pulse 1.0 were males, while 21.1% identified themselves as females (78.3% males and 21.7% females); and the median age was 38 years (Salesses et al., 2013, p. 8). Let p_k be proportion of units in Place Pulse 1.0 falling within class k (defined by age group and gender) and P_k the proportion of simulated population in the same class. Thus, we compute the sampling probabilities as p_k/P_k in order to select non-probability samples as a function of gender and age. These sampling probabilities reproduce two of the self-selection mechanisms observed in crowdsourced samples, where males are more represented than females and where young and middle-age citizens are more represented than children and seniors. Sample sizes are drawn with the only constraint of two units selected per area ($\min(n_d) = 2$, $\text{med}(n_d) = 93.5$, $\bar{n}_d = 117.9$, $\max(n_d) = 296$). Then, we select $T = 500$ samples biased according to the sample distribution noted by Salesses et al. (2013) in Place Pulse 1.0. Hereby, we take control of the bias to make sure that post-stratified estimates suffer from self-selection bias and low reliability as in real crowdsourced data.
2. In each sample, post-stratified unweighted estimates are computed, as well as the bootstrap estimates (Eq. 5) from $b = 1, \dots, B$ ($B = 500$) replicates and the area-level EBLUP estimates (Eq. 11). Area-level models, which are used to produce EBLUP estimates, are fitted from the area-level averaged gender and age obtained from the original population. The post-stratified estimator (direct estimator, $\hat{Y}_d(pst)$), which is expected to produce highly biased and unreliable estimates, is given by:

$$\hat{Y}_d(pst) = \sum_{i \in s_d} \frac{y_{di}}{n_d}. \quad (13)$$

3. The results are then assessed by the empirical Bias and the empirical Root Mean Squared Error, denoted by $\text{Bias}_d(\hat{Y}_d)$ and the $\text{RMSE}_d(\hat{Y}_d)$ (Petrucchi and Salvati, 2006), which are computed as:

$$Bias_d(\hat{Y}_d) = \frac{1}{T} \sum_{t=1}^T (\hat{Y}_{dt} - \bar{Y}_d), \quad (14)$$

$$RMSE_d(\hat{Y}_d) = \sqrt{\frac{1}{T} \sum_{t=1}^T (\hat{Y}_{dt} - \bar{Y}_d)^2}, \quad (15)$$

where $\hat{Y}_d$ denotes each estimate (either post-stratified, bootstrap or EBLUP) in area d , $\hat{Y}_{dt}$ denotes each estimate in area d and sample t , and $\bar{Y}_d$ is the true value observed in the population in area d .

Then, summary statistics across the small areas are calculated and shown in the next section. The simulation experiment has been coded in *R* software and ‘sae’ package has been used to produce the EBLUP estimates (Molina and Marhuenda, 2015).

6.3 Results

Table 2 shows the summary of the empirical values and the summary of the three estimates averaged across samples (i.e. post-stratified, bootstrap and EBLUP estimates). Figure 1 shows the Kernel density distribution of the empirical values and the three estimates across areas. Both Table 2 and Figure 1 show that while the post-stratified estimator is skewed towards lower values due to the bias introduced in our samples, both the non-parametric bootstrap and the EBLUP estimator shrink the estimates towards the empirical mean; and their mean and median are closer to the empirical measures of central tendency. The minimum and the maximum values are also improved by the use of the bootstrap and EBLUP estimators in comparison with the original post-stratified estimates, which show a large bias.

	Min	First quart	Mean	Median	Third quart	Max
$\bar{Y}_d$	-0.012	0.206	0.330	0.319	0.444	0.837
$\hat{Y}_d(pst)$	-0.182	0.052	0.184	0.168	0.299	0.639
$\hat{Y}_d^{Boot}$	-0.191	0.058	0.227	0.209	0.360	0.847
$\hat{Y}_d^{EBLUP}$	-0.168	0.065	0.226	0.211	0.353	0.814

Table 2. Summary of empirical values $\bar{Y}_d$, and $\hat{Y}_d(pst)$, $\hat{Y}_d^{Boot}$ and $\hat{Y}_d^{EBLUP}$ estimates across the areas.

Figure 1. Kernel density plot of empirical values $\bar{Y}_d$, and $\hat{Y}_d(pst)$, $\hat{Y}_d^{Boot}$ and $\hat{Y}_d^{EBLUP}$ estimates across the areas.

In order to assess the performance of the non-parametric bootstrap and EBLUP estimators, the estimates' median empirical $\overline{Bias}$ (Eq. 14) and $\overline{RMSE}$ (Eq. 15) are produced and shown in Table 3. The bootstrap estimator produces better estimates than the post-stratified, both in terms of $\overline{Bias}$ and $\overline{RMSE}$, reducing these from $\overline{Bias}(\hat{Y}_d(pst)) = -0.142$ to $\overline{Bias}(\hat{Y}_d^{Boot}) = -0.115$ and from $\overline{RMSE}(\hat{Y}_d(pst)) = 0.192$ to $\overline{RMSE}(\hat{Y}_d^{Boot}) = 0.178$, respectively. In addition, after fitting the area-level models and producing the EBLUP estimates, both measures of precision and reliability decrease slightly ($\overline{Bias}(\hat{Y}_d^{EBLUP}) = -0.113$, $\overline{RMSE}(\hat{Y}_d^{EBLUP}) = 0.173$), showing a better performance than the two previous estimators.

Quality measure	$\hat{Y}_d(pst)$	$\hat{Y}_d^{Boot}$	$\hat{Y}_d^{EBLUP}$
$\overline{Bias}$	-0.142	-0.115	-0.113
$\overline{RMSE}$	0.192	0.178	0.173

Table 3. Estimates' median Bias and RMSE across the small areas.

However, area-level measures of Bias and RMSE are also needed in order to examine the level of accuracy and precision of our estimates in each area, as our estimator might produce better estimates in some areas but not in others. Figure 2 shows the Bias of the three estimates obtained in each area, and Figure 3 shows the area-level RMSE of the three estimates. Figure 2 shows that the bootstrap estimates' Bias is smaller than the post-stratified estimates' Bias in 119 areas out of 150 (73.9%). At the same time, EBLUP estimates' bias is smaller than the bootstrap estimates' bias in 103 of the 150 areas under study (68.7%). If we compare the final EBLUP estimates' bias against the original post-stratified estimates' bias, we observe that the bias has been reduced in 124 areas in total (82.7%), and such reduction is larger than the 25% in 49 areas.

Figure 2. Bias of the post-stratified, bootstrap and EBLUP estimates (ordered by the post-stratified estimates' Bias).

Figure 3 shows that the bootstrap estimates' empirical RMSE is reduced in 83 areas out of the 150 as compared to the post-stratified estimates' RMSE, while the RMSE of the bootstrap estimates is slightly larger than the post-stratified estimates' RMSE in 67 areas. Although the bootstrap estimates' median measure of RMSE is improved with respect to the post-stratified estimates' RMSE (see Table 3), the bootstrap estimator does not provide

better estimates (in terms of RMSE) than the post-stratified estimator in 44.7% of areas. However, if we compare the final EBLUP with the original post-stratified estimates, we obtain that the empirical RMSE is improved in 127 out the 150 areas (the 84.7%). The RMSE has been increased by more than 25% in three areas, and it has been reduced by more than 25% in 17 areas. The EBLUP estimates' RMSE is better than the bootstrap estimates' RMSE in 141 of the areas under study (94% of the total).

Figure 3. RMSE of post-stratified, bootstrap and EBLUP estimates (ordered by the post-stratified estimates' RMSE).

Finally, Figures 4 and 5 plot the RMSE of the bootstrap and EBLUP estimates, respectively, against the number of units sampled per areas, in order to examine if these estimators perform better when the area sample size increases. Both plots show a significant negative Spearman's rank correlation, denoted as ρ , between the sample size per area and the RMSE, which is $\rho = -0.49$ ($p - value < 0.001$) in the case of the bootstrap estimates and $\rho = -0.53$ ($p - value < 0.001$) in the case of the EBLUP estimates. Thus, there is a direct relation between the areas sample size and the estimators performance.

Figures 4 and 5. Sample size per area plotted against bootstrap and EBLUP estimates' RMSE.

7. Case study: Safety perceptions in London

Crowdsourced data can be used to “study people’s perception of crime, disorder and place at a resolution at which data were previously unavailable” (Solymosi et al., 2017, p. 964). Indeed, numerous researchers have explored the use of crowdsourced samples to map the worry about crime and perceived safety (e.g. Candeia et al., 2017; Harvey et al., 2015; Salesses et al., 2013; Solymosi and Bowers, 2018). Nevertheless, conclusions drawn from crowdsourced samples are likely affected by the biases discussed above (see Solymosi and Bowers, 2018). The method outlined above can be used to reduce such biases. In this section we use data from the Place Pulse 2.0 platform to produce and map the bootstrap (Eq. 5) and EBLUP (Eq. 11) estimates of safety perceptions in Greater London.

7.1 The spatial study of safety perceptions

Social scientists are increasingly interested in examining the geographical distribution of crime and perceptions of security at a detailed geographical level (e.g. Solymosi et al., 2017; Weisburd, 2015). Both crime and safety perceptions are unequally distributed across cities, and their negative effects disproportionately affect certain areas and communities more than others. By mapping these perceptions, researchers can better understand their causes, and ultimately design spatially targeted interventions to mitigate their effects. Section 2 has shown that crowdsourced data provide new insights into the spatial distribution of social perceptions, and section 3 has presented some of the biases that limit their representativeness. For these reasons, innovative approaches are needed to produce more reliable estimates from crowdsourced data. Then, such estimates can be mapped to examine

the spatial distribution of the target parameter. In order to fit area-level models to produce reliable model-based estimates, significant covariates at a small area level should be available. Furthermore, the selection of our covariates must be oriented by previous research results.

Prior research has shown that perceptions of security are driven by a series of individual factors that explain differences between citizens' perceived vulnerability, such as gender, age, employment status, education level or income (Farrall et al., 1997; Hale, 1996; Pantazis, 2000). However, the unequal geographical distribution of perceived security has also been explained by a series of neighbourhood variables that shape citizens' urban perceptions. First, higher crime rates have been associated with lower perceptions of security (Breetzke and Pearson, 2014; Liska et al., 1982; Rotarou, 2017); although other researches show little or no relation between those. Second, different relative measures of deprivation, poverty and socioeconomic development are known to be related to the geographical distribution of perceived safety: these perceptions tend to arise in economically and socially deprived areas (Pantazis, 2000; Rotarou, 2017). According to Pantazis (2000), people living in poverty suffer the greatest from a range of insecurities "that relate to crime and the prospect of experiencing a number of non-criminal incidents including job loss, financial debts, and illness" (p. 433). And third, the ethnic composition of the area, and more particularly the proportion of ethnic minorities living in each neighbourhood, is known to be a predictor of the spatial distribution of perceived insecurity (Liska et al., 1982). These findings influence our choice of area-level covariates for our model.

7.2 Data and methods

7.2.1 Place Pulse 2.0 dataset

Place Pulse 2.0 platform records data from the question 'Which place looks safer / wealthier / more beautiful / more boring / livelier / more depressing?', in which respondents are shown two random images from 56 cities across 28 countries (see Figure 6). Place Pulse 2.0 images are taken from Google Street View and were originally captured between 2007 and 2012. Place Pulse is hosted in an open website (<http://pulse.media.mit.edu/>) and anyone can participate. The images are geocoded, and users answer either by choosing one of the two images or clicking on 'equal'. Respondents provide no auxiliary information about themselves (see Dubey et al., 2016). This platform functions in a very similar way than other online pairwise wiki surveys, such as All Our Ideas (<http://www.allourideas.org/>) (Salganik and Levy, 2015).

Figure 6. Place Pulse 2.0 website.

In this application we only use Place Pulse 2.0 data recorded to measure perceived safety in Greater London. Reports were recorded between 2013 and August 2018. The total sample size is 17,766 responses distributed across 1,368 Lower Super Output Areas (LSOAs) in Greater London. LSOAs are geographic units of analysis designed to improve the reporting of small area statistics in England and Wales: each LSOA contains between 1,000 and 3,000 citizens, and between 400 and 1,200 households. The average number of responses per LSOA is $\bar{n}_d = 12.99$, while the minimum sample size per small area is 1 (in 35 areas) and the maximum is 91.

The aim of this application is to produce reliable estimates of perceived safety at a low geographical level in Greater London. Each response is then a unit grouped within a LSOA area: ‘safer’ reports are coded as 1, while ‘less safe’ and ‘equal’ responses are coded as 0. Then, we produce post-stratified estimates of the proportion of ‘safer’ responses per area ($min = 0.0$, $Q1 = 0.5$, $\bar{x} = 0.6$, $Q2 = 0.6$, $Q3 = 0.7$, $max = 1.0$), as well as non-parametric bootstrap and EBLUP estimates from $B = 500$ replicates. By estimating the proportion of ‘safer’ responses per area, not only will we be able to examine the performance of the EBLUP (Eq. 11) under the non-parametric bootstrap (Eq. 5) when applied to a crowdsourced dataset; but also to produce a map of perceived safety at a low spatial level and to analyse neighbourhood predictors of perceived safety. Note that previous research suggests making use of 0 to 10 Q-scores per image, which are produced as fractions of times each image get selected over another image, corrected by the ‘win’ and ‘loss’ ration of all images with which it was compared (see Candeia et al., 2017; Harvey et al., 2015; Salesses et al., 2013). In this research, we chose a more straightforward approach to estimate the proportion of ‘safer’ responses per area. This reduces the computation time for data analysis,

as only images from Greater London have been used (instead of safety assessments for each image paired at least once to an image from London), and it allows for a better control of the method's performance and for direct interpretations of final model-based estimates.

Due to the novelty of the EBLUP approach under the non-parametric bootstrap, no measure of error (e.g. mean squared error) has been developed yet to analyse the estimates' reliability. Future research will need to develop new methods to estimate the mean squared error of this estimator: this is topic of our future research. However, in order to externally validate our results, we will be able to compare the estimates produced in this research to reliable estimates of perceived safety obtained from the Metropolitan Police Service Public Attitudes Survey (MPSPAS) 2011/12 (only available at the borough level). We expect a high correlation between the EBLUP estimates produced from crowdsourced Place Pulse 2.0 data and the direct estimates obtained from the MPSPAS dataset.

7.2.2 Area-level covariates

In order to fit an area-level model and produce the EBLUP estimates, we select a set of five available covariates in line with the discussion earlier in this section: *(i)* Proportion of black and minority ethnic citizens (BAME) 2011, *(ii)* crimes rate 2012, *(iii)* Income deprivation score, *(iv)* Employment deprivation score and *(v)* Education, skills and training deprivation score (see Table 4). The proportion of BAME is obtained from the UK Census 2011, and the crime data are provided by the Metropolitan Police Service. The scores for Income deprivation, Employment deprivation and Education, skills and training deprivation are items of the English Index of Multiple Deprivation 2015, which provides statistics about measures of relative deprivation in the small areas in England. After fitting the area-level model with our five covariates, all of them show significant negative beta coefficients to estimate the area-level perceived safety.

	Min	First quart.	Mean	Median	Third quart.	Max	Spearman's rank correlation with perceived safety
Proportion BAME	0.02	0.22	0.39	0.37	0.54	0.96	-0.25***
Crimes rate	0.02	0.09	0.18	0.13	0.19	10.23	-0.19***
Income deprivation	0.01	0.09	0.17	0.15	0.23	0.47	-0.22***
Employment deprivation	0.00	0.07	0.11	0.10	0.15	0.36	-0.18***
Education, skills and training deprivation	0.01	5.45	14.09	12.14	20.45	64.03	-0.21***

*** $\alpha < 0.001$, ** $\alpha < 0.01$, * $\alpha < 0.05$

Table 4. Summary measures of area-level covariates and correlation coefficients with the bootstrap estimates of perceived safety.

7.3 Results

In subsection 7.3.1 we will assess our estimates, both internally and externally: first, we will present the model diagnostics to assess our EBLUP model and, second, we will compare our estimates (first produced at a borough level) with estimates of perceived security obtained from the MPSPAS. In subsection 7.3.2 we will present the EBLUP estimates of perceived safety and visualise these on the Greater London map.

7.3.1 Model diagnostics and external validation

Figure 7 shows the normal Q-Q plot of the area-level model standardised residuals, computed as in Petrucci and Salvati (2006). Most residuals follow a normal distribution, though there are some outliers at both tails. The Shapiro-Wilk test to check the normality of the standardised residuals also suggests no rejection of the null hypothesis of normal distribution ($W = 0.957$, $p - value = 0.612$).

Figure 7. Normal Q-Q plot of standardised residuals of the EBLUP model.

In order to externally validate our estimates of perceived safety, we first produce EBLUP estimates at a borough level in Greater London, and compare these to the direct estimates of ‘perceived safety when walking alone after dark’ obtained from MPSPAS 2011/2012 data. The MPSPAS records large representative samples at borough level in Greater London, with an average of 745.09 citizens sampled per area (min=676, max=792). MPSPAS 2011/2012 data were recorded between January 2011 and December 2012, but the proportion of MPSPAS respondents who feel ‘very safe’ or ‘fairly safe’ when walking alone after dark appears to be quite stable over time in most London boroughs (see Figure 8). Time stability shown in Figure 8 demonstrates that external validation results are likely to be similar regardless of the MPSPAS edition that is used. Thus, we feel confident comparing the direct estimates of perceptions of security obtained from the MPSPAS 2011/12 to externally validate our EBLUP estimates of perceived safety produced from Place Pulse data (2013-2018). The measure of ‘How safe do you feel walking alone after dark?’ has been highly analysed in criminological literature and, although it has been criticised as a measure of emotional fear of crime (e.g. Farrall et al., 1997), it provides consistent results to examine the geographies of perceived security (e.g. Luo et al., 2016).

Figure 8. Proportion of MPSPAS respondents who feel ‘very safe’ or ‘fairly safe’ across boroughs and years 2010 to 2013 (boroughs included in Place Pulse 2.0).

We produce direct estimates of the proportion of respondents who feel ‘very safe’ or ‘fairly safe’ (coded as 1), while ‘fairly unsafe’ and ‘very unsafe’ responses are coded as 0. ‘Refusal’ to answer, ‘don’t know’ and ‘do not go out’ responses are coded as ‘no answers’ and deleted from the analysis. Then, we rescale estimates obtained from MPSPAS and Place Pulse 2.0 to 0-1 values $\left(\frac{\hat{Y}_d - \min(\hat{Y})}{\max(\hat{Y}) - \min(\hat{Y})}\right)$ to allow comparisons. Figure 9 shows the borough-level differences between MPSPAS direct estimates of feeling of safety when walking alone after dark and Place Pulse EBLUP estimates of perceived safety. Only 6 out of 24 areas show differences larger than -0.25 or 0.25, among which Place Pulse EBLUP estimates are larger than MPSPAS direct estimates in four cases and MPSPAS estimates are larger than Place Pulse estimates in two areas. In most cases, however, differences between both datasets are small. Moreover, there is a medium-high significant Spearman’s rank correlation between both estimates ($\rho = 0.54$, $p - value < 0.05$) and a Bivariate Moran’s I coefficient equal to 0.51 ($p - value < 0.01$). Thus, considering that the correlation is fitted from only 24 boroughs, we can expect our model-based estimates drawn from the Place Pulse data to represent a very similar construct than the variable of ‘feeling of safety when walking alone after dark’ recorded from the MPSPAS survey.

Figure 9. Differences between direct estimates of feeling of safety after dark obtained from MPSPAS data and EBLUP estimates of perceived safety produced from Place Pulse.

7.3.2 Mapping safety perceptions at neighbourhood level

Figure 10 shows the map of the EBLUP estimates of perceived safety in 1,368 LSOAs across Greater London. Lighter colour scales show a lower estimated perceived safety, while darker colour scales show a higher perceived neighbourhood security. Note that white areas indicate 'no data'. The Place Pulse 2.0 dataset only records data in an elliptic area around the London city centre, which covers some complete boroughs and some LSOAs of incomplete boroughs. Estimates vary from a minimum of 0 and a maximum of 1, and the measures of central tendency are 0.49 (mean), 0.50 (median) and 0.50 (mode). The small area estimates show large differences within each Greater London borough. The lowest estimates of perceived safety are found in Eastern neighbourhoods, especially in certain areas of Newham, Waltham Forest and Tower Hamlets; while the highest estimates of perceived safety have been produced in areas of the central boroughs of City of London and Westminster.

Figure 10. Estimates of perceived safety at LSOA level (division in quantiles).

8. Discussion and conclusions

Social science research is increasingly making use of open and crowdsourced data to analyse and map social phenomena, such as crime and safety perceptions (Salesse, 2009; Salesse et al., 2013; Solymosi and Bowers, 2018; Solymosi et al., 2017; Williams et al., 2017). By using crowdsourced data, researchers can obtain larger samples at a cheaper cost than traditional approaches for data collection. In addition, some crowdsourced datasets record VGI that allow examining spatial patterns at a very detailed geographical scale. Although crowdsourced data can provide new insights into people's attitudes and perceptions, these are affected by a series of biases that limit the representativeness of the data (i.e. self-selection bias, unequal participation, underrepresentation of areas and times), and thus it becomes probable that direct estimators might lead to unreliable area-level estimates.

There have been several attempts to reduce biases and increase representativeness in crowdsourced data, most of which assume the availability of individual auxiliary

information. Such approaches make use of individual covariates to allow for various unit-level modelling techniques aimed to compute pseudo-sampling weights and adjust the non-probability samples to the target population (see Elliott and Valliant, 2017). However, not all crowdsourcing projects record individual auxiliary information. In this chapter we have proposed and evaluated a two-phase approach aimed to produce reliable small area estimates from crowdsourced data with no individual auxiliary information (apart from the area of the target place or respondent). First, a non-parametric bootstrap algorithm selects repeated samples using a SSRSWR design and produces pseudo-sampling weights in each bootstrap replicate, computed as the inverse of the first-order inclusion probabilities, to produce bootstrap weighted estimates. Then, the second phase is aimed to borrow strength from related areas by fitting an area-level model and producing EBLUP estimates (Fay and Herriot, 1979; Rao and Molina, 2015).

In order to evaluate this two-phase approach, a simulation study and an application have been conducted. First, the simulation study is based on generating one fixed population and drawing SSRSWR with an unequal probability selection to reproduce the biases in crowdsourced data. Then, post-stratified unweighted estimates are compared to the non-parametric bootstrap and the EBLUP estimates, which are expected to increase the estimates reliability. From the simulation experiment we observe: first, the distribution and the summary measures of the EBLUP estimates are slightly closer to the empirical summary measures than the post-stratified estimates. Second, the median Bias and RMSE are reduced after producing the non-parametric and the EBLUP estimates. And third, the RMSE and the Bias of the final EBLUP estimates are closer to zero (in most areas) than the post-stratified estimates' RMSE and Bias: the final EBLUP estimates are generally more reliable and less biased than the post-stratified estimates. However, the non-parametric bootstrap estimator does not provide better estimates than the post-stratified estimator, in terms of RMSE, in a large number of areas. Such limitation is clearly reduced after fitting the area-level models and producing the EBLUP estimates, which shows the need for the second step of the method.

Then, the EBLUP under the non-parametric bootstrap approach has been applied to produce small area estimates of perceived safety at LSOA level in Greater London from the Place Pulse 2.0 crowdsourced dataset. Our results have been validated externally by comparing these to reliable direct estimates drawn from the MPSPAS dataset. The final EBLUP estimates allow for reliably mapping the perceived safety at a very detailed micro geographical level.

Although the EBLUP approach under the non-parametric bootstrap has shown positive results, further simulation experiments with more complex sampling designs are needed to investigate whether this method produces reliable estimates when the sample

biases are higher, smaller or show different distributions. Moreover, the method needs to be applied to other crowdsourced datasets (e.g. FixMyStreet, All Our Ideas) to assess its performance under different non-probability samples. Then, once the method performance has been assessed under different simulation experiments and real crowdsourced datasets, new software tools can be developed to facilitate and speed the computation of the estimates. A measure of uncertainty also needs to be developed to estimate the RMSE of the EBLUP estimates under the non-parametric bootstrap. Double bootstrap techniques will also be explored to further reduce the estimates' bias.

Small area estimation techniques are now well established not only to produce research results of academic and scientific relevance, but they are also being used by local and national authorities to provide reliable local statistics at a small area level. Here we show an application of the area-level EBLUP (under a non-parametric bootstrap) to reduce the bias and increase the reliability of crowdsourced datasets. Both academics and policy makers might benefit from the development of new methods to successfully bridge the gap between crowdsourcing techniques and small area estimation, as these techniques might be helpful to produce more reliable, spatially and temporally more precise and cheaper small area statistics. Reliable small area estimates are needed to improve our understanding of social dynamics, and to design and evaluate geographically targeted policies.

References

- Antoniou, V.; Jeremy, M. and Mordechai H. (2010). Web 2.0 geotagged photos: Assessing the spatial dimension of the phenomenon. *Geomatica*, 64(1), 99-110.
- Arbia, G.; Solano Hermosilla, G.; Micale, F.; Nardelli, V. and Genovese G. (2018). *Post-sampling crowdsourced data to allow reliable statistical inference: the case of food prices in Nigeria*. Paper presented at XLIX Riunione Scientifica della Società Italiana di Statistica. Retrieved from <http://meetings3.sis-statistica.org/index.php/sis2018/49th/paper/viewFile/1090/64>
- Baker, R.; Brick, J.M.; Bats, N.A.; Battaglia, M.; Couper, M.P., Dever, J.A.; Gile, K. and Tourangeau, R. (2013). *Report of the AAPOR Task Force on non-probability sampling*. Deerfield: American Association for Public Opinion Research.
- Battese, G. E., Harter, R. M. and Fuller, W. A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83(401), 28-36.
- Becker, M.; Caminiti, S.; Fiorella, D.; Francis, L.; Gravino, P.; Haklay, M.; Hotho, A.; Loreto, V.; Mueller, J.; Ricchiuti, F.; Servedio, V.D.P.; Sirbu, A. and Tria, F. (2013). Awareness and learning in participatory noise sensing. *PLoS ONE*, 8(12), e81638.
- Blom, J.; Viswanathan, D.; Go, J.; Spasojevic, M.; Acharya, K.; Ahonius, R. (2010). Fear and the city - Role of mobile services in harnessing safety and security in urban contexts. In *CHI'10 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1841-1850).
- Brabham, D.C. (2008). Crowdsourcing as a model for problem solving. An introduction and cases. *Convergence: The International Journal of Research into New Media Technologies*, 14(1), 75-90.
- Breetzke, G.D. and Pearson, A.L. (2014). The fear factor: Examining the spatial variability of recorded crime on the fear of crime. *Applied Geography*, 46, 45-52.
- Candeia, D.; Figueiredo, F.; Andrade, N. and Quercia, D. (2017). Multiple images of the city: Unveiling group-specific urban perceptions through a crowdsourcing game. In *HT'17 Proceedings of the 28th ACM Conference on Hypertext and Social Media Pages* (pp. 135-144).
- Chen, K.T. (2016). *Using LASSO to calibrate non-probability samples using probability samples*. PhD thesis, University of Michigan.
- Dubey, A.; Naik, N; Parikh, D.; Raskar, R. and Hidalgo, C.A. (2016). Deep learning the city: Quantifying urban perception at a global scale. In *Computer Vision – European Conference on Computer Vision 2016* (pp. 196-212).
- Efron, B. and Tibshirani, R. (1993). *An Introduction to the bootstrap*. London: Chapman and Hall.

- Elliott, M.R. (2009). Combining data from probability and non-probability samples using pseudo-weights. *Survey Practice*, 2(6), 1-7.
- Elliott, M.; Resler, A.; Flannagan, C. and Rupp, J. (2010). Combining data from probability and non-probability samples using pseudo-weights. *Accident Analysis and Prevention*, 42, 530-539.
- Elliott, M.R. and Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science*, 32(2), 249-264.
- Fay, R., and Herriot, R. (1979). Estimates of income for small places. An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.
- Farrall, S.; Bannister, J.; Ditton, J. and Gilchrist, E. (1997). Questioning the measurement of the 'fear of crime'. *British Journal of Criminology*, 37(4), 658-679.
- Fuchs, C. (2008). The role of income inequality in a multivariate cross-national analysis of the digital divide. *Social Science Computer Review*. 27, 41-58
- Goodchild, M.F. (2007). Citizens as sensors: The world of volunteered geography. *GeoJournal*, 69(4), 211-221.
- Goodchild, M.F. and Glennon, J.A. (2010). Crowdsourcing geographic information for disaster response: a research frontier. *International Journal of Digital Earth*, 3(3), 231-241.
- Haklay, M. (2010). How good is Volunteered Geographic Information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environmental and Planning B: Urban Analytics and City Science*, 37(4), 682-703.
- Haklay, M. (2013). Citizen science and volunteered geographic information: Overview and typology of participation. In D. Sui, S. Elwood and M. Goodchild (Eds.) *Crowdsourcing geographic knowledge. Volunteered Geographic Information (VGI) in theory and practice* (pp. 105-122). Dordrecht: Springer.
- Hale, C. (1996). Fear of crime: A review of the literature. *International Review of Victimology*, 4, 79-150.
- Harvey, C.; Aultman-Hall, L.; Hurley, S.E. and Troy, A. (2015). Effects of skeletal streetscape design on perceived safety. *Landscape and Urban Planning*, 142, 18-28.
- Horvitz, D. G., and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260), 663-685.
- Howe, J. (2006). The rise of crowdsourcing. *Wired Magazine*, 14(06), 1-5.
- Howe, J. (2008). *Crowdsourcing. How the power of the crowd is driving the future of business*. London: Random House.
- Lee, S. (2006). Propensity Score Adjustment as a weighting scheme for volunteer panel web surveys. *Journal of Official Statistics*, 22(2), 329-349.

- Liska, A.E.; Lawrence, J.J. and Sanchirico, A. (1982). Fear of crime as a social fact. *Social Forces*, 60(3), 760–770.
- Longley, P.A. (2012). Geodemographics and the practices of geographic information science. *International Journal of Geographical Information Science*, 26(12), 2227–2237.
- Luo, F.; Ren, L. and Zhao, J.S. (2016). Location-based fear of crime: A case study in Houston, Texas. *Criminal Justice Review*, 4(1), 75-97.
- Mashhadi, A.; Quattrone, G. and Capra, L. (2013). Putting ubiquitous crowd-sourcing into context. In *Proceedings of the 2013 conference on Computer supported cooperative work* (pp. 611–622). San Antonio: ACM.
- McConnell, B. and Huba, J. (2006). *The 1% Rule: Charting citizen participation*. Church of the Customer Blog. Retrieved from: https://web.archive.org/web/20100511081141/http://www.churchofthecustomer.com/blog/2006/05/charting_wiki_p.html
- Malleeson, N. and Andresen, M.A. (2015). The impact of using social media data in crime rate calculations: Shifting hot spots and changing spatial patterns. *Cartography and Geographic Information Science*, 42(2), 112-121.
- Molina, I. and Marhuenda, Y. (2015). sae: An R package for small area estimation. *The R Journal*, 7(1), 81-98.
- Nahorniak, M.; Larsen, D.P.; Volk, C. and Jordan, C.E. (2015). Using inverse probability bootstrap sampling to eliminate sample induced bias in model based analysis of unequal probability samples. *PLoS ONE*, 10(6), e0131765.
- Nielsen, J. (2006). *The 90-9-1 rule for participation inequality in social media and online communities*. Retrieved from <https://www.nngroup.com/articles/participation-inequality/>
- Pantazis, C. (2000). ‘Fear of crime’: vulnerability and poverty. *British Journal of Criminology*, 40, 414-436.
- Petrucci, A. and Salvati, N. (2006). Small area estimation for spatial correlation in watershed erosion assessment. *Journal of Agricultural, Biological, and Environmental Statistics*, 11(2), 169-182.
- Rao, J.N.K. and Molina, I. (2015). *Small area estimation. Second edition*. Hoboken: Wiley.
- Rotarou, E.S. (2017). Does municipal socioeconomic development affect public perceptions of crime? A multilevel logistic regression analysis. *Social Indicators Research*, 138(2), 705-724.
- Salesses, M.P. (2009). *Place Pulse. Measuring the collaborative image of the city*. MSc thesis, Massachusetts Institute of Technology.
- Salesses, P.; Schechtner, K. and Hidalgo, C.A. (2013). The collaborative image of the city: Mapping the inequality of urban perceptions. *PLoS ONE*, 8(7), e68400.

- Salganik, M.J. and Levy, K.E.C. (2015). Wiki surveys: open and quantifiable social data collection. *PLoS ONE*, 10(5), e0123483.
- Sanders, R. (1987). The pareto principle: its use and abuse. *Journal of Services Marketing*, 1(2), 37-40.
- Särndal, C.E.; Swensson, B. and Wretman, J. (1992). *Model assisted survey sampling*. New York: Springer-Verlag.
- Schonlau, M.; van Soest, A. and Kapteyn, A. (2007). Are “Webographic” or attitudinal questions useful for adjusting estimates from web surveys using propensity scoring? *Survey Research Methods*, 1(3), 155–163.
- Solymosi, R. and Bowers, K. (2018). The role of innovative data collection methods in advancing criminological understanding. In G.J.N. Bruinsma and S.D. Johnson (eds.) *The Oxford Handbook of Environmental Criminology* (pp. 210-237). New York: Oxford University Press.
- Solymosi, R.; Bowers, K.J. and Fujiyama, T. (2017). Crowdsourcing subjective perceptions of neighbourhood disorder: Interpreting bias in open data. *British Journal of Criminology*, 58(4), 944–967
- Stewart, O.; Lubensky, D. and Huerta, J.M. (2010). Crowdsourcing participation inequality: a scout model for the enterprise domain. In *Proceedings of the ACM SIGKDD Workshop on Human Computation* (pp. 30-33). Washington: ACM.
- Surowiecki, J. (2004). *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations*. New York: Doubleday.
- Valliant, R.; Dever, J.A. and Kreuter, F. (2013). *Practical tools for designing and weighting survey samples*. New York: Springer.
- Wang, W.; Rothschild, D.; Goel, S. and Gelman, A. (2015). Forecasting elections with non-representative polls. *International Journal of Forecasting*, 31(3), 980–991.
- Weisburd, D. (2015). The law of crime concentration and the criminology of place. *Criminology*, 53(2), 133-157.
- Williams, M.L.; Burnap, P. and Sloan, L. (2017). Crime sensing with big data: The affordances and limitations of using open-source communications to estimate crime patterns. *British Journal of Criminology*, 57, 320-340.
- Yamane, T. (1967). *Statistics. An introductory analysis. 2nd edition*. New York: Harper and Row.
- Yu, L. (2006). Understanding information inequality: Making sense of the literature of the information and digital divides. *Journal of Librarianship and Information Science*, 38, 229-252.