MANCHESTER 1824
The University of Manchester

BigSurv18
BIG DATA MEETS SURVEY SCIENCE

Non-parametric bootstrap and small area estimation to mitigate bias in crowdsourced data

Simulation study and application to perceived safety

David Buil-Gil, Reka Solymosi
Centre for Criminology and Criminal Justice (University of Manchester)
and **Angelo Moretti**
Social Statistics Department (University of Manchester)
david.builgil@manchester.ac.uk

Table of contents

1. The rise of crowdsourcing and implications

Introduction

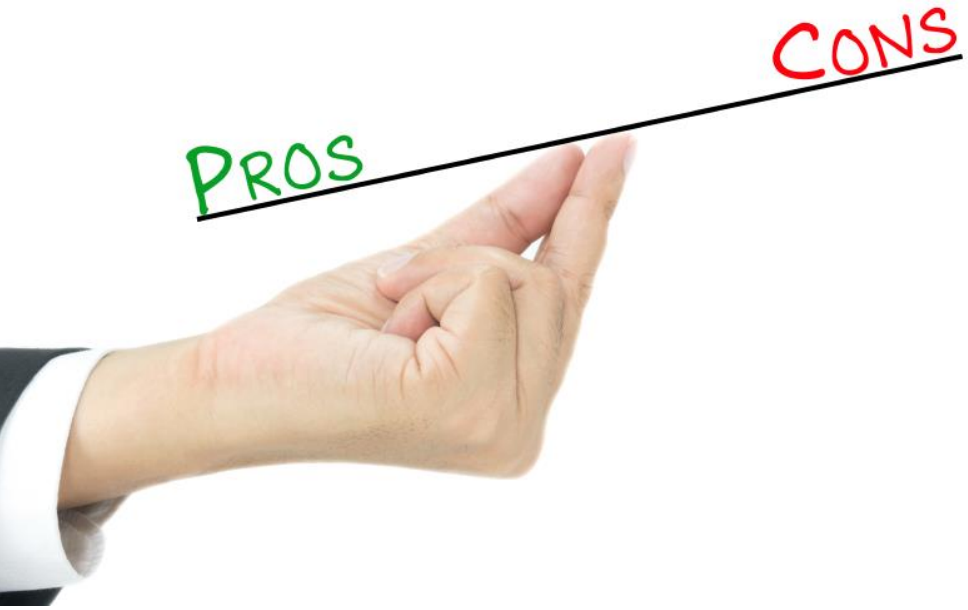
- Crowdsourcing techniques:
  - …methods that allow obtaining open information by enlisting the services of large crowds of people into one collaborative project (Howe, 2006, 2008).

- A growing body of social researchers is using crowdsourced data to analyse:
  - Inequality
  - Poverty
  - **Crime**
  - **Perceived safety**

- Some include accurate geographical information.

Advantages over traditional approaches to data collection

- Reduced cost and big datasets.
- Volunteered Geographical Information (VGI):
  - Participatory mapping
  - Precise spatial information
- Temporal information
- Some examples:
  - Noise pollution (Becker et al., 2013)
  - Emergencies: wildfires in Santa Barbara, US (Goodchild and Glennon, 2010)
- … but biases!

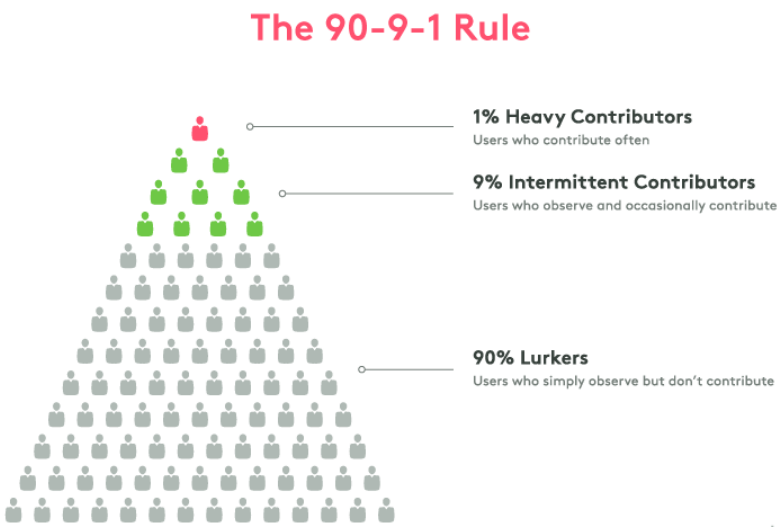


2. Crowdsourcing data to analyse social phenomena: limitations

a. Self-selection bias

- Certain socioeconomic groups are overrepresented:
  - Men tend to participate more than women (e.g. 78% males in Place Pulse 1.0 dataset);
  - Employed people;
  - Aged 20-50;
  - University degree;
  - Community-level deprivation…

b. Unequal participation



c. Under-representation of certain areas and times

- Cluster in urban areas and sparse coverage in rural areas (Picasa and Flickr) (Antoniou et al., 2010)
- Avoidance of areas perceived to be unsafe (Doran and Burgess, 2012)
- Participation higher at noon and almost nonexistent at night (Blom et al., 2010)

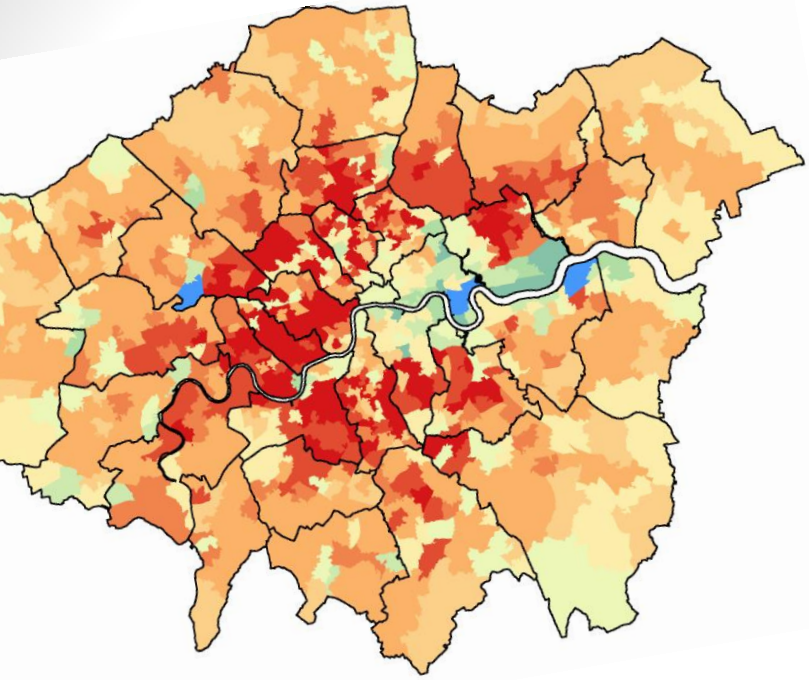
d. Unreliable direct estimates

- Due to these biases, it becomes probable that aggregating responses and producing area-level direct estimates from crowdsourced data might lead to biased and unreliable estimates.

3. Previous approaches for reweighting crowdsourced data

Previous approaches

- Datasets that record auxiliary information from participants:
  - Calibration from benchmarking (Kraemer et al., 2017)
  - Synthetic estimation from logistic regression (Boboth et al., 2007)
  - Propensity Score Adjustment (Lee, 2006)
  - Least Angle Shrinkage and Selection Operator (LASSO) (Chen, 2016)
- When crowdsourced platforms do not record participants' auxiliary information:
  - Arbia et al. (2018) two phase approach:
    - Outliers (also spatial outliers) are detected and removed
    - Reweight responses to let the data resemble an optimal spatial sample design
  - A new approach…



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

Step 1: Non-parametric bootstrap

1. From an observed non-probability sample s selected from a finite population U, draw a sample for each area $d = 1, \ldots, D$ 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. The sample sizes per area selected in the bootstrap are obtained via the simplified optimal sample size (Yamane, 1967, p. 886): $n_d^{Yamane} = \frac{N_d}{1+N_d(h)^2}$, where N_d is the population size in area d and h is the chosen margin of error.
2. Estimate the pseudo-sampling weights in each b^{th} replicate, obtained as the inverse of first-order inclusion probabilities in each replication:

$$w_{di}^{*(b)} = [1 - (1 - \frac{1}{n_d})^{n_d^{Yamane}}]^{-1}$$

   where n_d is the recorded sample size in area d and n_d^{Yamane} refers to the calculated simplified optimal size in area d.
3. The calibrated estimates of $\bar{Y}_d$ in each b^{th} replication are obtained by

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

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

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

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

Step 2: area-level model-based SAE

The original EBLUP makes use of the Horvitz-Thompson estimator and their errors e_d. In this work, however, we make use of the bootstrap estimate and assume

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

where $\psi_d$is the variance of bootstrap estimates in area d.

Then, we assume $\bar{Y}_d$ to be linearly related to a set of area-level covariates $\boldsymbol{x'_d}$:

$$\bar{Y}_d = x'_d\beta + v_d, \qquad v \sim N(0, A), \qquad d = 1, \ldots, D$$

where $v_d$is independent from e_d.

Thus,

$$\hat{\bar{Y}}_d^{Boot} = x'_d\beta + v_d + e_d, v_d \sim N(0, A), e_d \sim N(0, \psi_d), d = 1, \ldots, D$$

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

$$\hat{\bar{Y}}_d^{BLUP} = \hat{\bar{Y}}_d^{Boot} - \frac{\psi_d}{A + \psi_d}\left[\hat{\bar{Y}}_d^{Boot} - x'_d\hat{\beta}(A)\right]$$

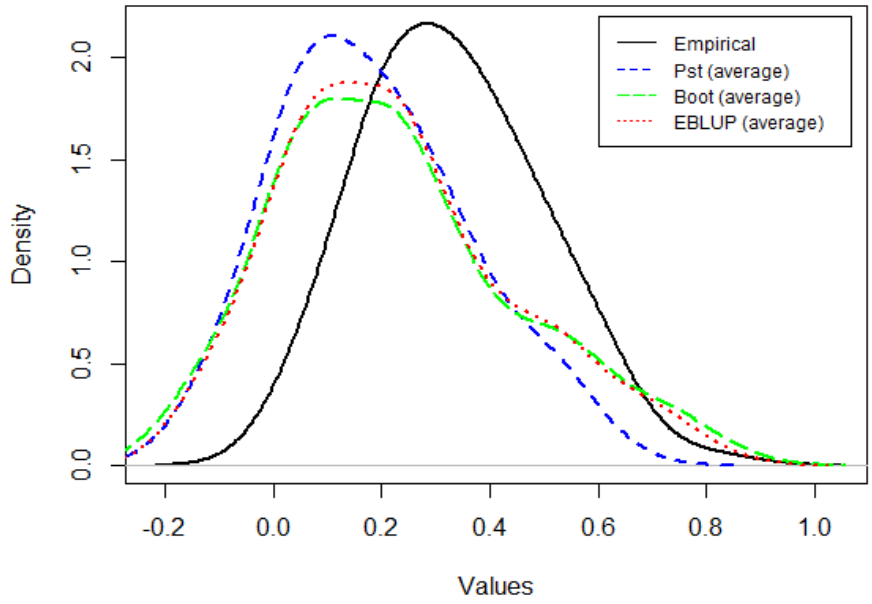
where $\hat{\beta}(A)$is the maximum likelihood estimator of β.

If we replace $\gamma_d(A) = \psi_d/(A + \psi_d)$, then:

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

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

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



5. Simulation study

Population generation

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.

Sample selection and simulation steps

1. Selection of t=1,...,T (T=500) samples from two-stage SSRSWR and 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: 78.3% males and 21.7% females and median age was 38 years in Place Pulse 1.0.
   - Reproduce two of the self-selection mechanisms observed in crowdsourced samples.
   - Sample sizes are drawn with the only constraint of two units selected per area.
2. In each sample, post-stratified unweighted estimates are computed, as well as the bootstrap estimates from b=1,...,B (B=500) replicates and the area-level EBLUP estimates.
3. The results are then assessed by the Bias and the Root Mean Squared Error.

Results 1/3

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

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

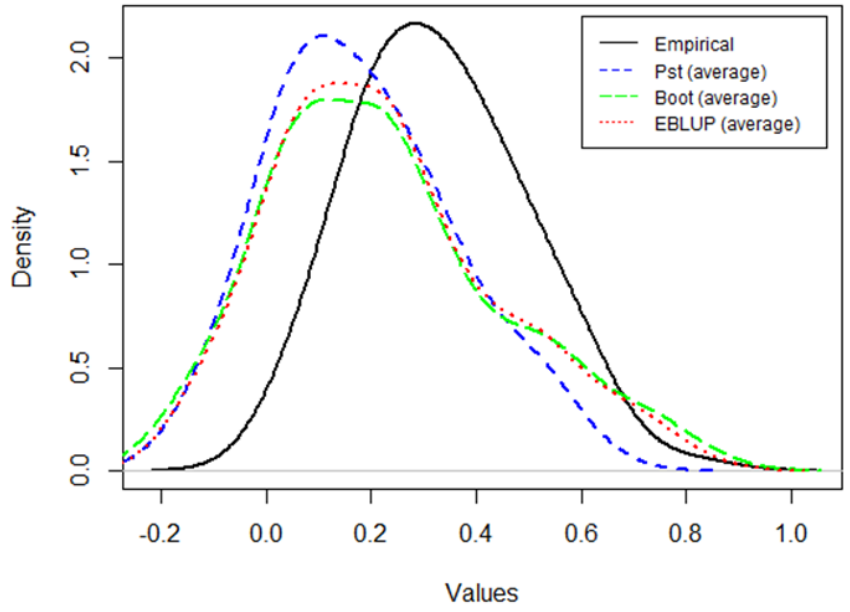


Figure 1. Kernel density plot of empirical values $\bar{Y}_d$, and $\hat{\bar{Y}}_d(pst)$, $\hat{\bar{Y}}_d^{Boot}$ and $\hat{\bar{Y}}_d^{EBLUP}$ estimates across the areas.

Results 2/3

Quality measure	$\hat{\bar{Y}}_d(pst)$	$\hat{\bar{Y}}_d^{Boot}$	$\hat{\bar{Y}}_d^{EBLUP}$
$\overline{Bias}$	-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.

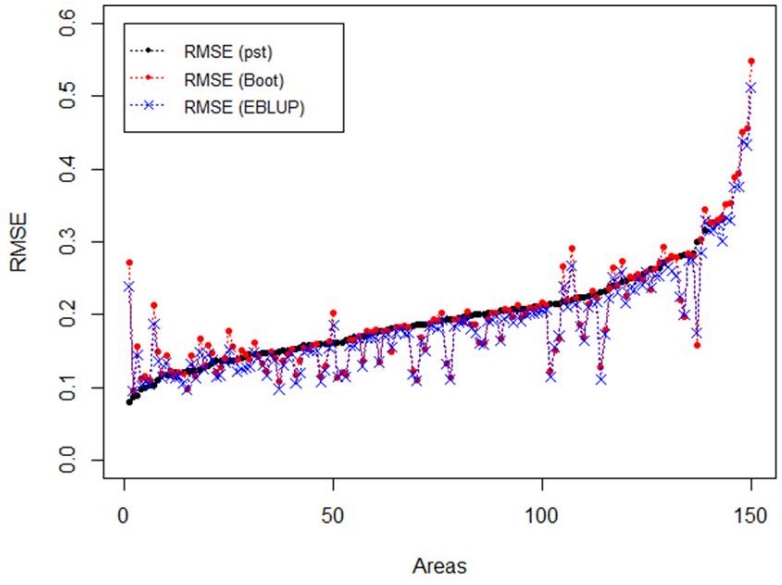


Figure 2. RMSE of post-stratified, bootstrap and EBLUP estimates (ordered by the post-stratified estimates' RMSE).

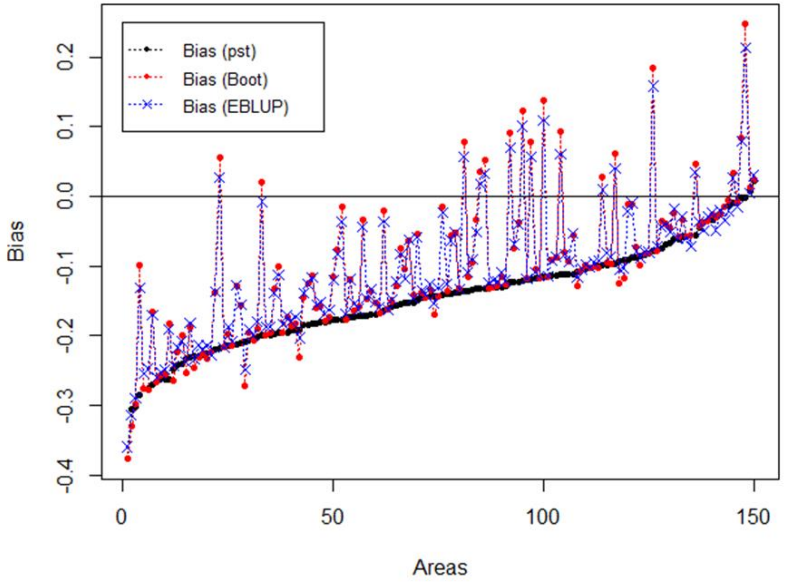
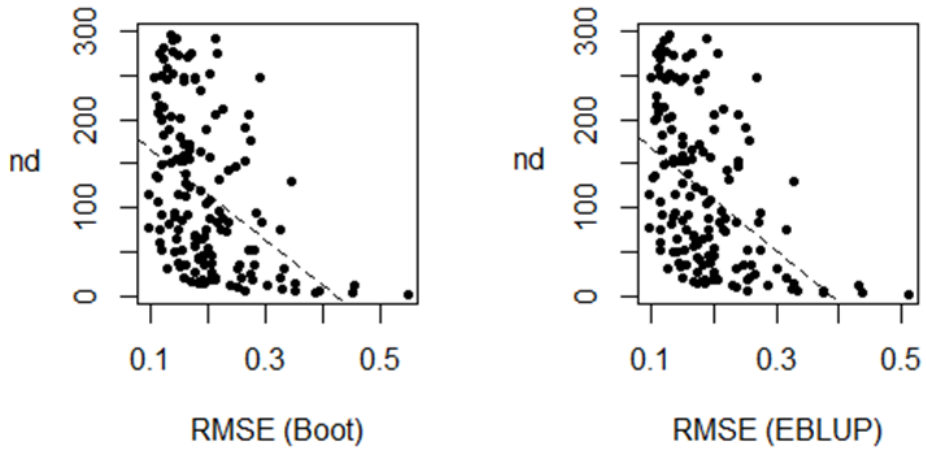


Figure 3. Bias of the post-stratified, bootstrap and EBLUP estimates (ordered by the post-stratified estimates' RMSE).

Results 3/3



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

Bootstrap: $\rho = -0.49$ $(p - value < 0.001)$

EBLUP: $\rho = -0.53$ $(p - value < 0.001)$

6. Case study: safety perceptions in London

Crowdsourcing safety perceptions

- 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).
- Numerous researchers have explored the use of crowdsourced samples to map worry about crime crime and perceived safety.
- Crime and safety perceptions are unequally distributed across cities.
- Severe negative effects for certain communities.
- By mapping crime and security perceptions, researchers are able to analyse their causes at their precise geographies, and to design spatially targeted interventions.

Data and methods

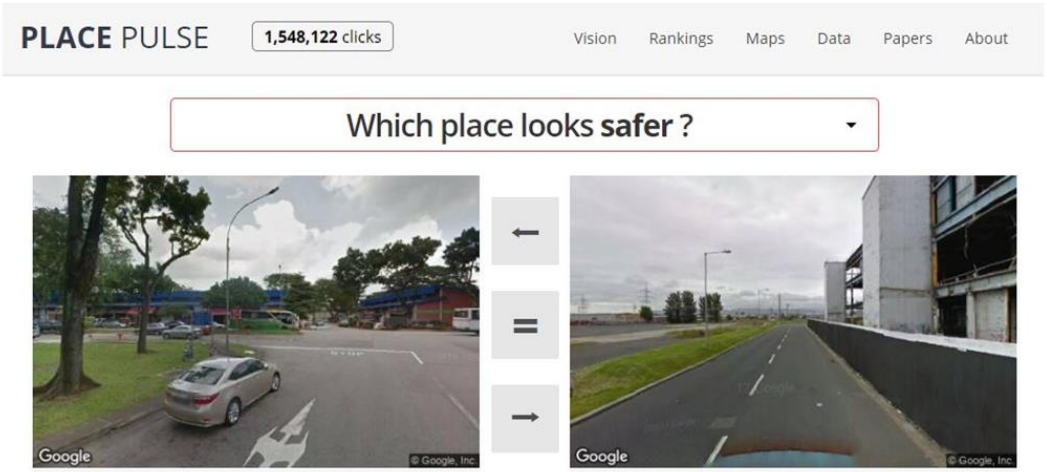


Figure 6. Place Pulse 2.0 website.

- No auxiliary information is provided apart from the users' response and the geographical information of each image.
- Data about perceived safety in Greater London.
- 17,766 responses distributed across 1368 LSOAs.
- $\bar{n}_d = 12.99$, minimum 1 (in 35 areas) and maximum 91
- Reliable estimates of the proportion of 'safer' reports per area (coded as 1).
- No estimated measure of error has been developed yet.

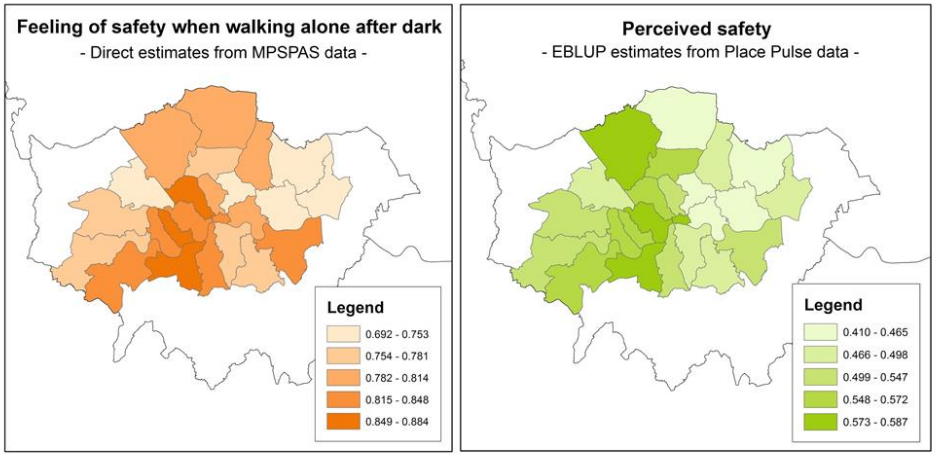
Data and methods

- Area-level covariates:
  1. Proportion of black and minority ethnic citizens (BIME) 2011,
  2. Crimes rate 2012,
  3. Income deprivation score,
  4. Employment deprivation score, and
  5. Education, skills and training deprivation score

- Data about perceived safety in Greater London.
- 17,766 responses distributed across 1368 LSOAs.
- $\bar{n}_d = 12.99$, minimum 1 (in 35 areas) and maximum 91
- Reliable estimates of the proportion of 'safer' reports per area (coded as 1).
- No estimated measure of error has been developed yet.

Model diagnostics and external validation

- No estimated measure of error has been developed yet.
- Model diagnostics:
  - Shapiro-Wilk test to check normality of standardised residuals suggests no rejection of the null hypothesis of normal distribution ($W = 0.957,\ p - value = 0.612$).
- External validation:
  - Reliable estimates of perceived safety obtained from the Metropolitan Police Service Public Attitudes Survey (MPSPAS).
  - $\rho = 0.54,\ p - value < 0.05$



Figures 8 and 9. Direct estimates of 'feeling of safety when walking alone after dark' from MPSPAS data (left) against EBLUP estimates of perceived safety from Place Pulse data (right).

Mapping perceived safety

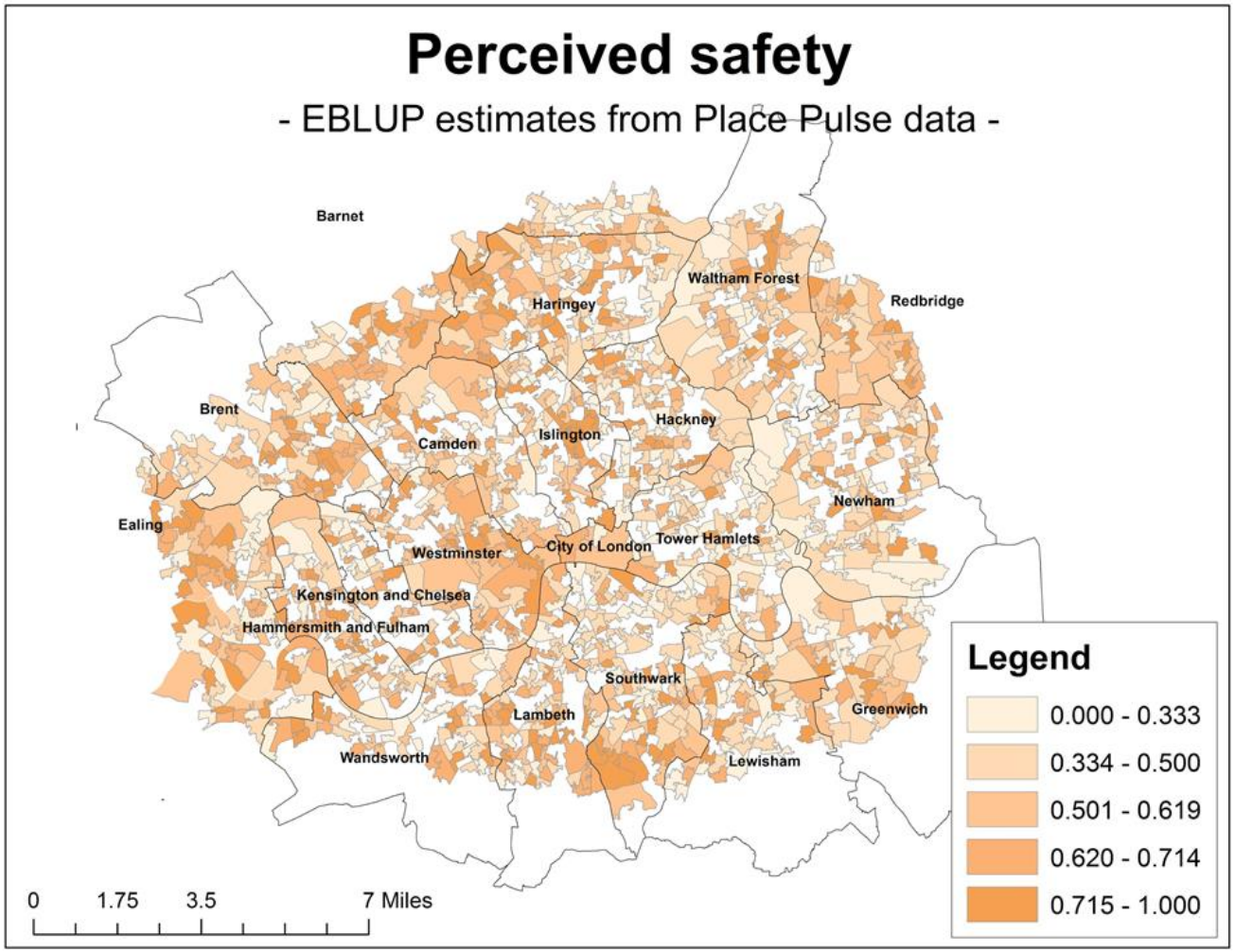


Figure 10. Estimates of perceived safety at LSOA level (division in quantiles).

3. Discussion and conclusions

Conclusions and future work

- The EBLUP approach under the non-parametric bootstrap shows promising results both under the simulation experiment and under a real crowdsourced data.
- Further simulation experiments with more complex sampling designs are needed to examine whether our method also produces more reliable estimates when the sample biases are higher, smaller or different.
- A measure of uncertainty needs to be developed to estimate the RMSE of the EBLUP estimates under the non-parametric bootstrap.

MANCHESTER 1824
The University of Manchester

BigSurv18
BIG DATA MEETS SURVEY SCIENCE

Thank you for your attention!

david.builgil@manchester.ac.uk
reka.solymosi@manchester.ac.uk
a.moretti@manchester.ac.uk