Assessing the impact of measurement error in crime data

Applications to regression modelling and geographic crime analysis

David Buil-Gil[1], Jose Pina-Sánchez[2], Ian Brunton-Smith[3], and Alexandru Cernat[1]

[1]University of Manchester, UK
[2] University of Leeds, UK
[3]University of Surrey, UK

Overview

1.- Crime data bias and crime analysis

2.- Research questions

4.- Results of the simulation study

5.- Conclusions

Crime statistics and crime analysis

- Police-recorded crimes are used by:

  Police forces → Design and evaluate policing strategies

  Policy makers → Design and evaluate crime prevention policies

  Academics → Develop and test theories of crime and deviance

- However… police statistics are affected by:

  Willingness to report crimes to police (varies by sex, age, ethnic group…)

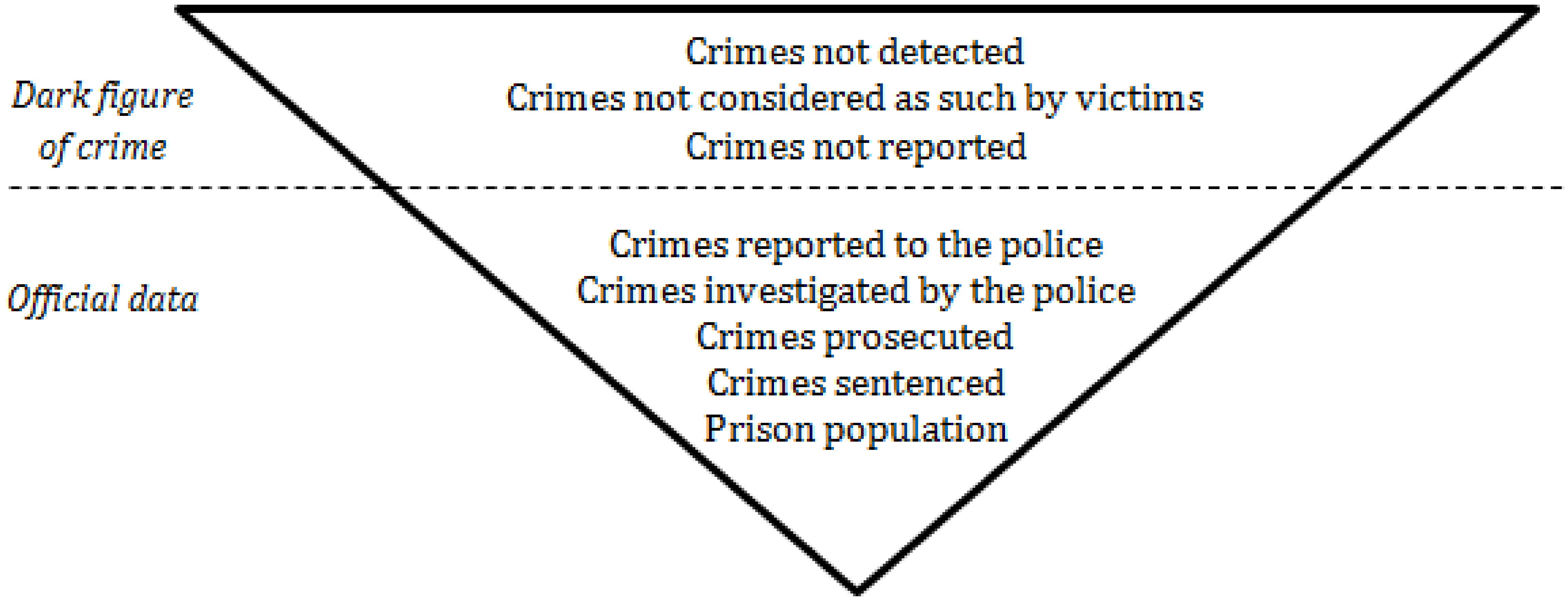
  Police control over areas (likelihood to witness crimes)

  Counting rules

  Dark figure of crime

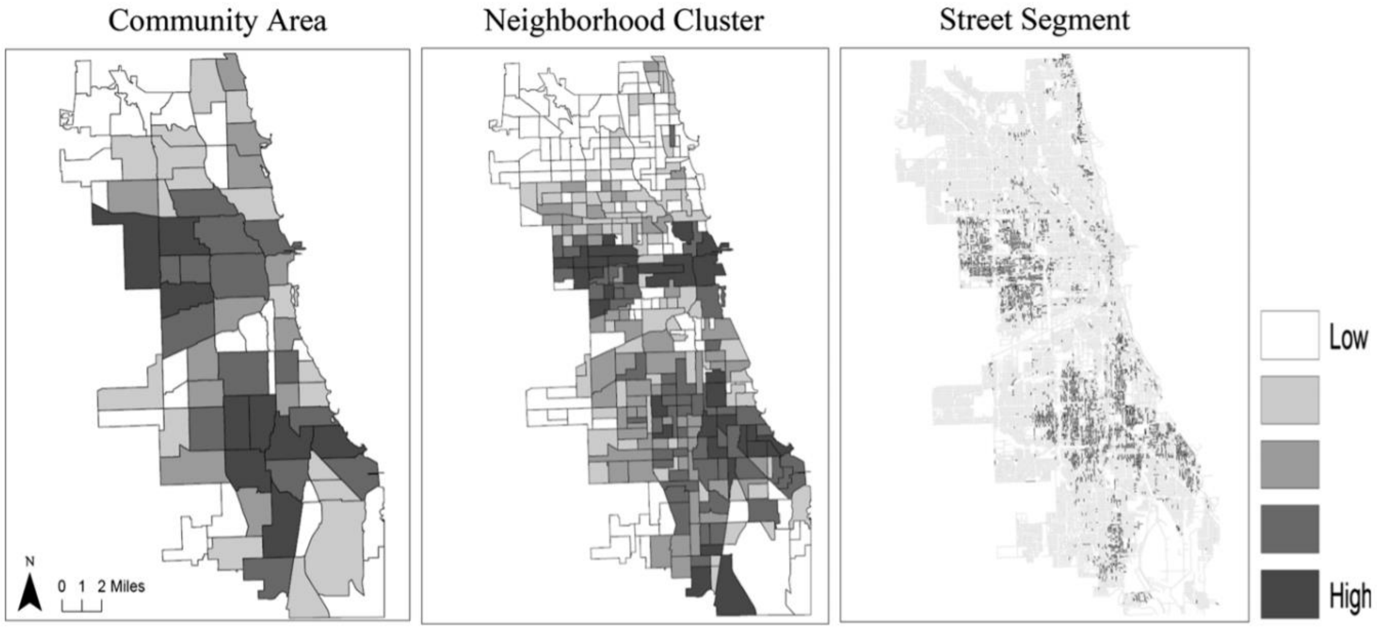
- This may not be necessarily a problem if the proportion of crimes missing in police statistics is equal across areas – this is not the case!

FIGURE X. The criminal justice data funnel.



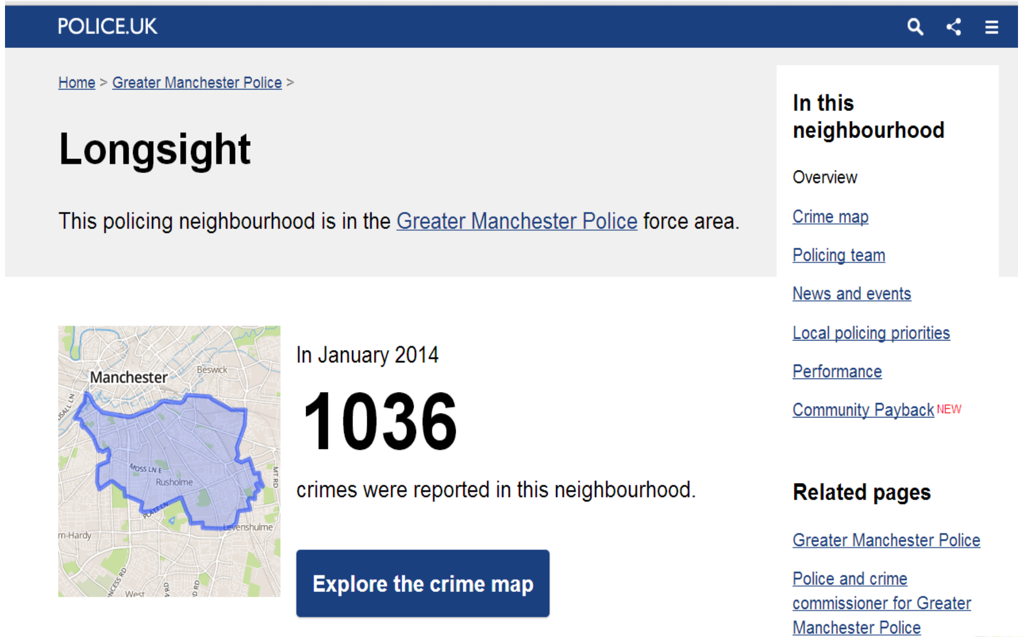
Moreover…

- Since the 1980s, move towards mapping police statistics at micro places…



Schnell et al. (2017)

- … and micro places are defined by socially homogeneous communities, while larger scales are more heterogeneous.
- The dark figure of crime may vary widely across micro places.



POLICE.UK

Home > Greater Manchester Police >

Longsight

This policing neighbourhood is in the Greater Manchester Police force area.

In January 2014

1036

crimes were reported in this neighbourhood.

Explore the crime map

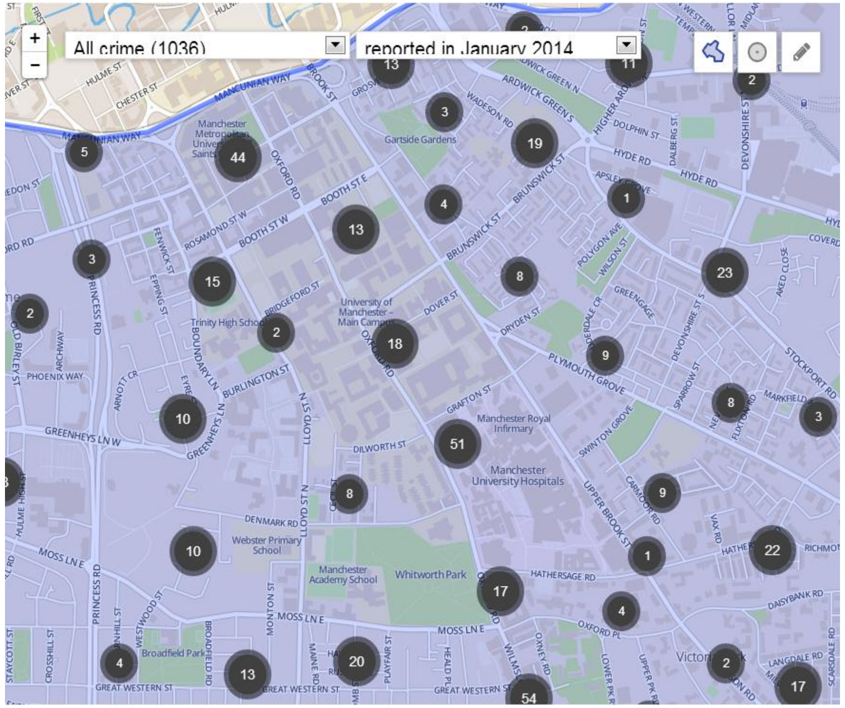
Contact your local policing team

In this neighbourhood

- Overview
- Crime map
- Policing team
- News and events
- Local policing priorities
- Performance
- Community Payback NEW

Related pages

- Greater Manchester Police
- Police and crime commissioner for Greater Manchester Police



All crime (1036)

reported in January 2014

Manchester Metropolitan University All Saints
Gartside Gardens
University of Manchester - Main Campus
Trinity High School
Manchester Royal Infirmary
Manchester University Hospitals
Webster Primary School
Manchester Academy School
Whitworth Park
Broadfield Park
44
19
13
15
18
51
23
17
20
13
10
10
54

Research questions

RQ1 - Are crime maps produced at smaller, more socially homogeneous spatial scales, at a larger risk of bias compared to maps produced at larger, more socially heterogeneous scales?

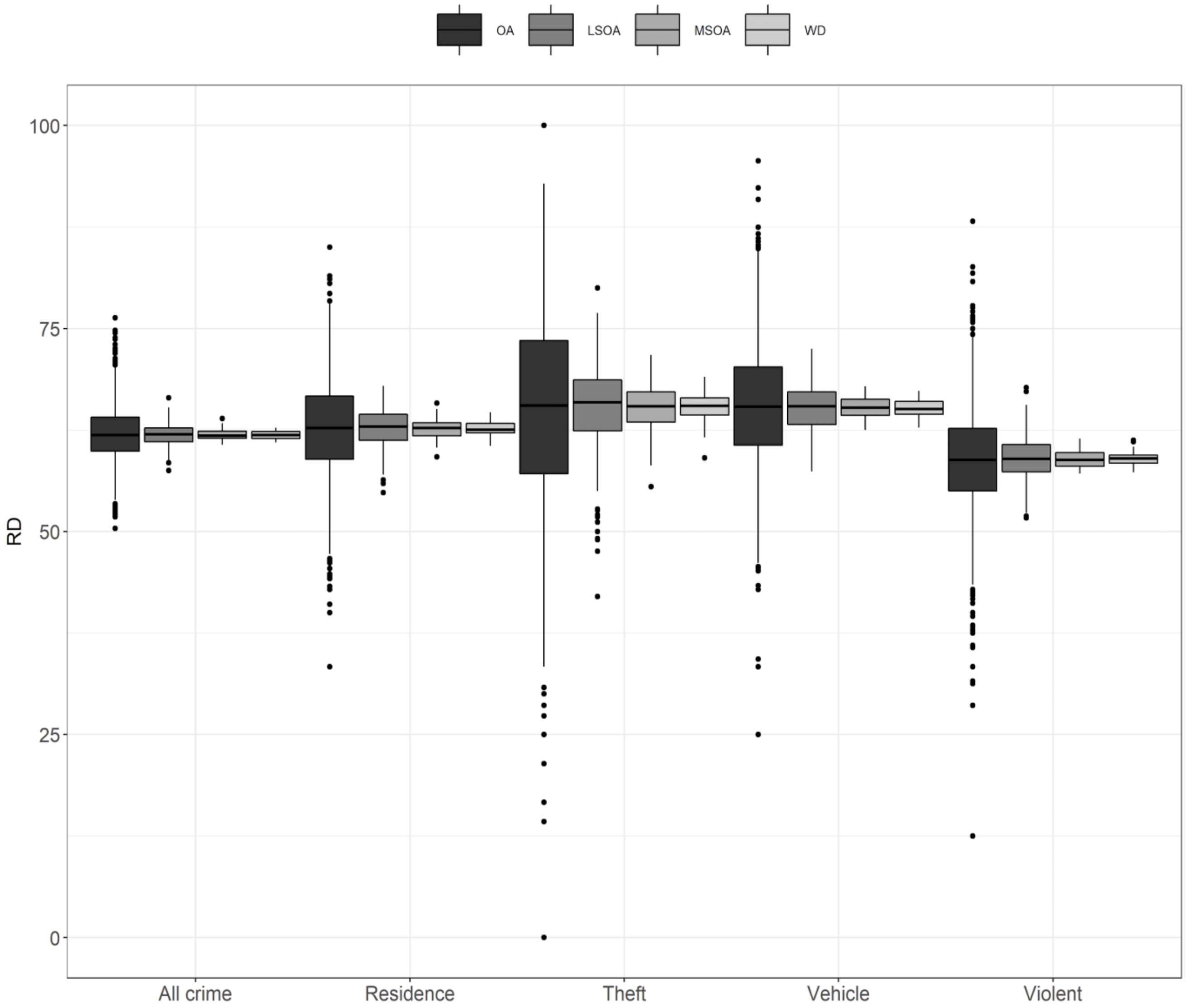
RQ2 -To what extent does measurement error in police recorded crime rates impact the estimates of regression models exploring the causes and consequences of crime?

RQ1.- Measurement error and geographic analysis

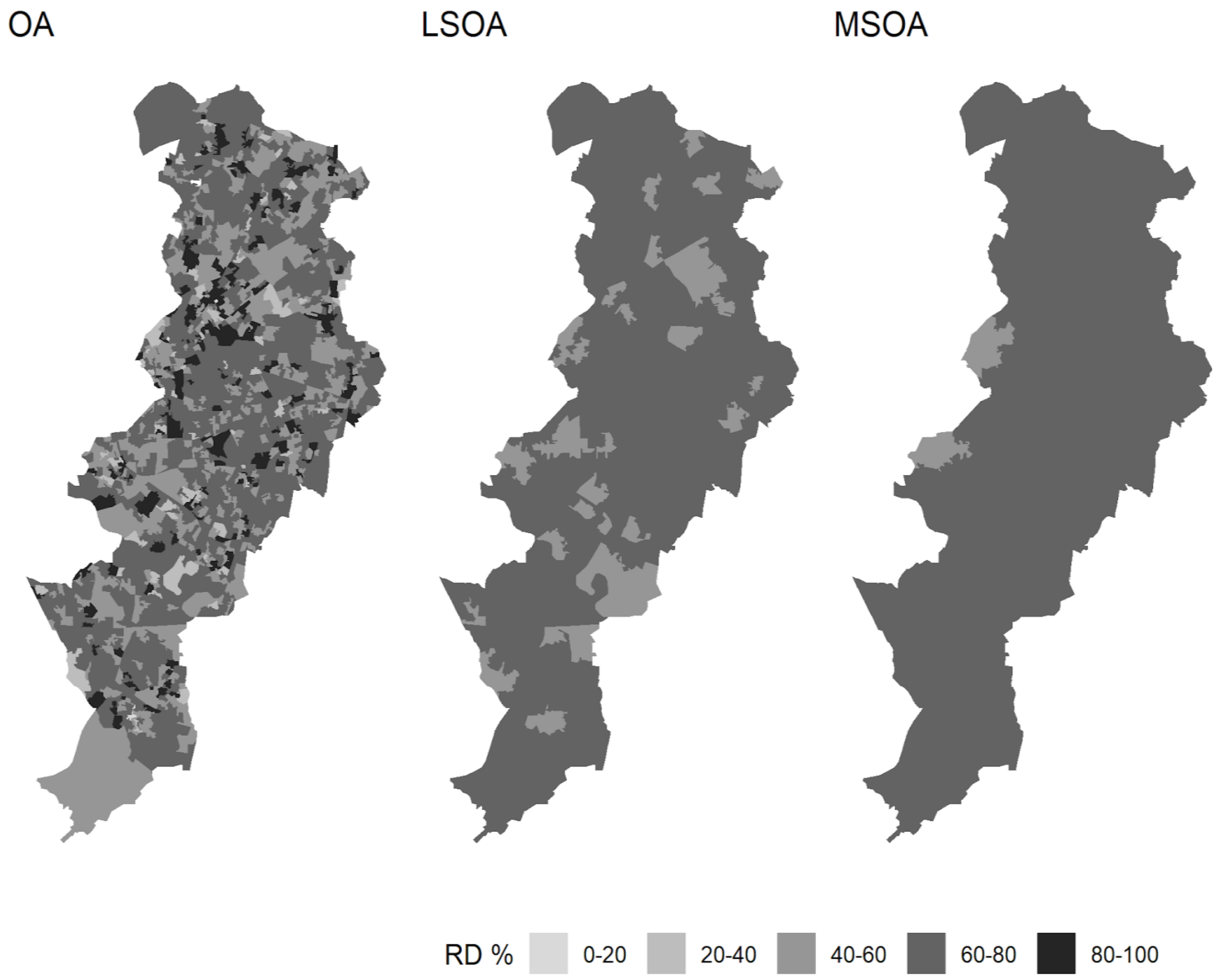
Measures of absolute RD% and absolute RB% between crimes known to police and all crimes

		OA	LSOA	MSOA	Ward
RD%	Mean	62.0	61.9	61.9	61.9
	SD	3.5	1.4	0.7	0.6
	Min	50.4	57.5	60.7	61.0
	Max	76.3	66.5	63.9	62.8
RB%	Mean	165	163	163	163
	SD	25.7	9.6	4.8	3.8
	Min	101	135	154	156
	Max	322	198	177	169

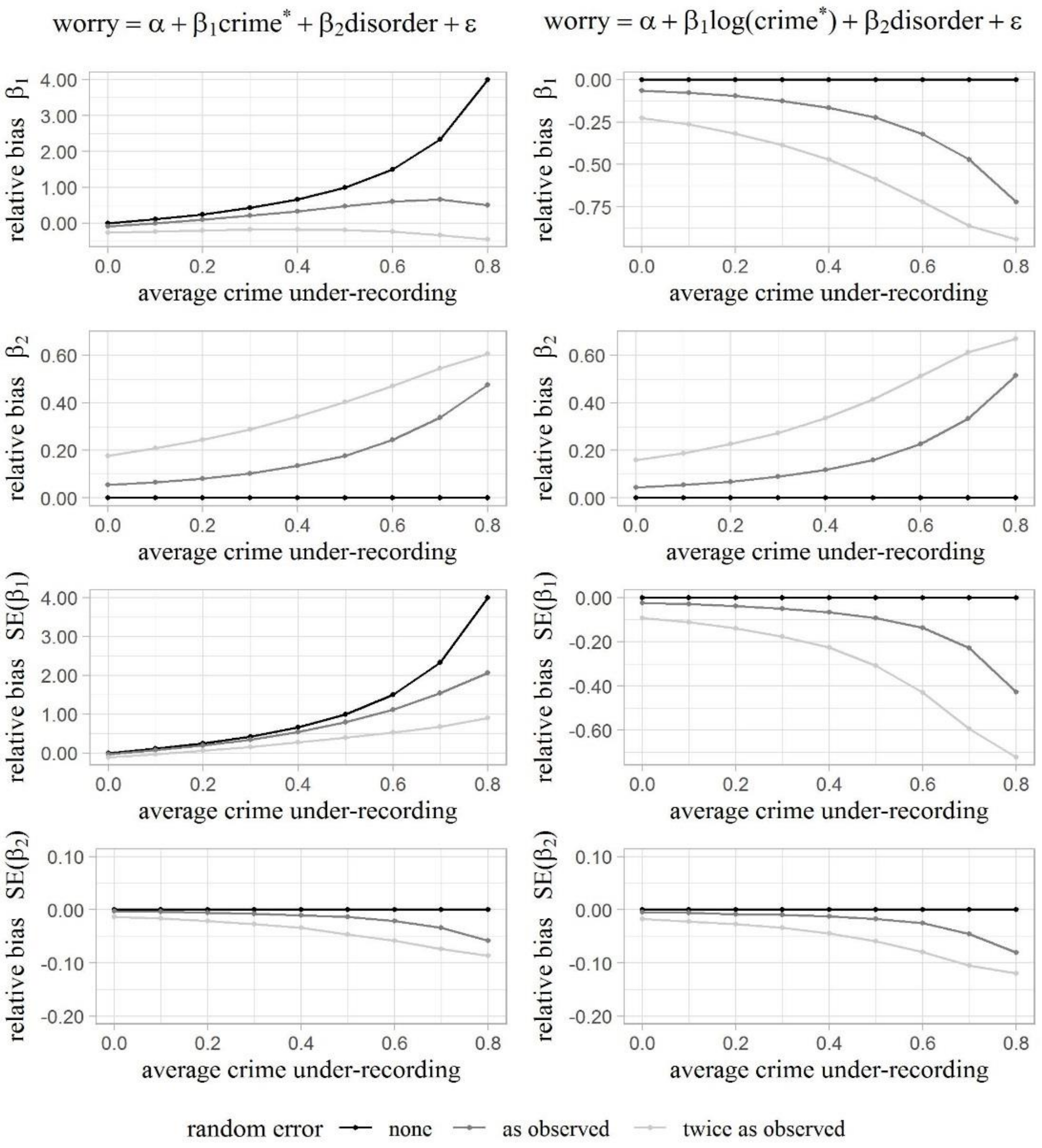
Boxplots of RD% between all crimes and crimes known to police at the different spatial scales



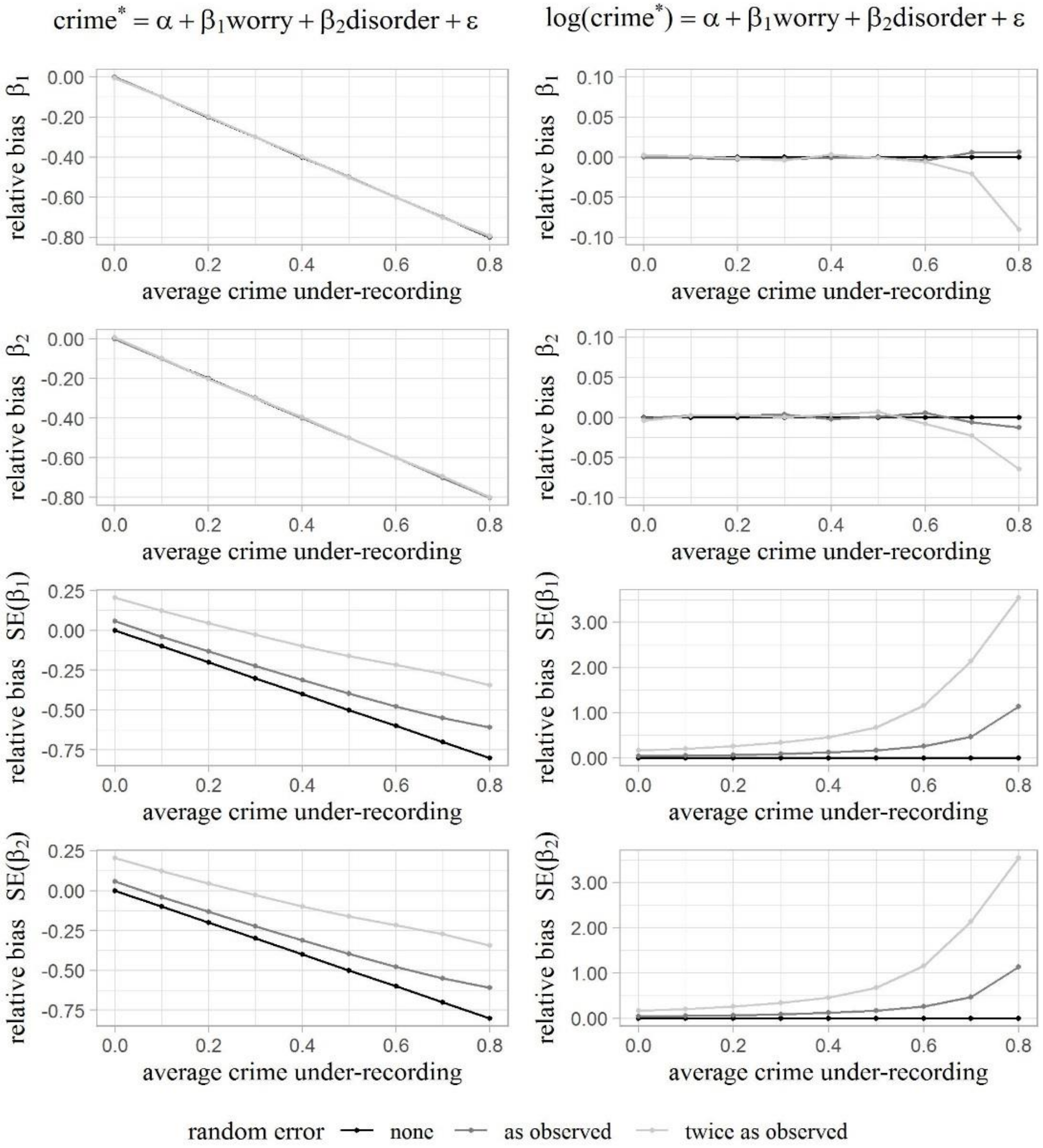
Maps of RD% between all property crimes and property crimes known to police at the different spatial scales



RQ2.- Measurement error and regression modelling



RQ2.- Measurement error and regression modelling



Conclusions and limitations

- Aggregating crimes known to police at very detailed levels of analysis increases the risk of inaccurate maps
- Maps of police-recorded crimes produced for neighbourhoods and wards (larger scales) show a more accurate image of the geography of crime
- Linear regression models that use police-recorded crime rates are biased
- The bias can be mitigated by log-transforming crime rates

For more information:

- Paper measurement error and regression modelling:
  - Preprint published in SocArxiv:
    https://osf.io/preprints/socarxiv/ydf4b/
  - Codes published in GitHub:
    https://github.com/davidbuilgil/crime_simulation2

- Paper measurement error and geographic crime analysis:
  - Preprint published in SocArxiv:
    https://osf.io/preprints/socarxiv/myfhp/
  - Codes published in OSF:
    https://osf.io/kv3sc

MANCHESTER
1824
The University of Manchester

Thank you for your attention!

david.builgil@manchester.ac.uk

Twitter: @DavidBuil

Method: Generating a synthetic population

Simulation steps (4 steps):

1. Simulating a synthetic population of Manchester from Census 2011
   - Download census data aggregated in Output Areas
   - Obtain empirical parameters of age, sex, income, education and ethnicity
   - Generate synthetic population from empirical parameters in each area

2. Simulating crime victimisation from Crime Survey for England and Wales 2011/12
   - Estimate Negative Binomial regression models at individual level of (i) violent crime, (ii) residence crime, (iii) theft and property crime, and (iv) vehicle crime in CSEW
     - Same independent variables as in Step 1
   - Obtain regression parameter estimates and simulate crime victimisation in synthetic population following Negative Binomial regression models

Method: Generating a synthetic population

Simulation steps (4 steps):

3. Simulating whether each crime is known to the police
   - Estimate logistic regression models of crimes being known to police (0/1) in CSEW dataset of crimes
   - Same independent variables as in Step 1 (Census)
   - Obtain regression parameter estimates and simulate if each crime (synthetic population) is known to police

4. Simulating whether each crime happens in local area or not
   - Same steps as Step 3
   - Then, remove all simulated crimes that did not take place in local area

Final sample of 359,248 crimes across 1,530 OAs in Manchester

Then, we aggregate these in LSOAs, MSOAs and Wards

Assessing the results

In order to know the difference between crimes known to police and all crimes, we calculate the Relative Difference (RD) and Relative Bias (RB).

- RD is calculated for every area d in the specified level of geography (i.e., Geo={OA,LSOA,MSOA,wards}), as follows:

$$RD_d^{Geo} = \left|\frac{E_d \; - \; K_d}{E_d}\right| \times \; 100$$

  where E_d denotes the count of all crimes in area d and K_d is the count of crimes known to police in the same area.

- RB is computed as follows

$$RB_d^{Geo} = \left(\frac{E_d}{K_d} - 1\right) \times \; 100$$

Empirical evaluation

To evaluate our simulated dataset of crimes, we compared:

- Average number of victimisations based on demographic characteristics of victims in our synthetic dataset and the CSEW – very good results
- Proportion of crimes known to police based on demographic characteristics of victims in our synthetic dataset and the CSEW – very good results
- Measures of ranking correlation between simulated crimes and incidents recorded by Greater Manchester Police – good results, but can be improved

		LSOA	MSOA	Ward
All crimes	Spearman's rank correlation	0.36***	0.40**	0.38*
	Global Moran's I	0.36***	0.39***	0.20*
Vehicle crimes	Spearman's rank correlation	0.13*	0.12	0.14
	Global Moran's I	0.30***	0.30***	0.18*
Residence crimes	Spearman's rank correlation	0.29***	0.30*	0.23
	Global Moran's I	0.37***	0.48***	0.31**
Property crimes	Spearman's rank correlation	0.18**	0.30*	0.23
	Global Moran's I	0.33***	0.33***	0.26**
Violent crimes	Spearman's rank correlation	0.34***	0.45***	0.31^{+}
	Global Moran's I	0.28***	0.30***	0.07
Number of areas		282	56	32

*** p-value < 0.001; ** p-value < 0.01; * p-value < 0.05; + p-value < 0.1