CECAN Webinar:

SIPHER Synthetic Population: An Introduction

Wednesday 28th February 2024, 13:00 – 14:00 GMT

Presenter: Nik Lomax (Professor of Population Geography at the University of Leeds, Co-Director of the Consumer Data Research Centre and Co-Investigator for the SIPHER Consortium)

Welcome to our **CECAN Webinar.**

All participants are muted. Only the Presenter & CECAN Host can speak. The webinar will start at **13:00 GMT.**

Nik will speak for around 45 minutes and will answer questions at the end.

Please submit your questions at any point during the webinar via the Q&A box in the Zoom webinar control panel.

Today's webinar will be recorded and made available on the CECAN website.

E Mail: cecan@surrey.ac.uk　　Web: www.cecan.ac.uk

www.facebook.com/CECANEXUS　　Twitter: @cecanexus

SIPHER Synthetic Population: an Introduction

CECAN Seminar
28th February 2024

Nik Lomax
School of Geography
University of Leeds

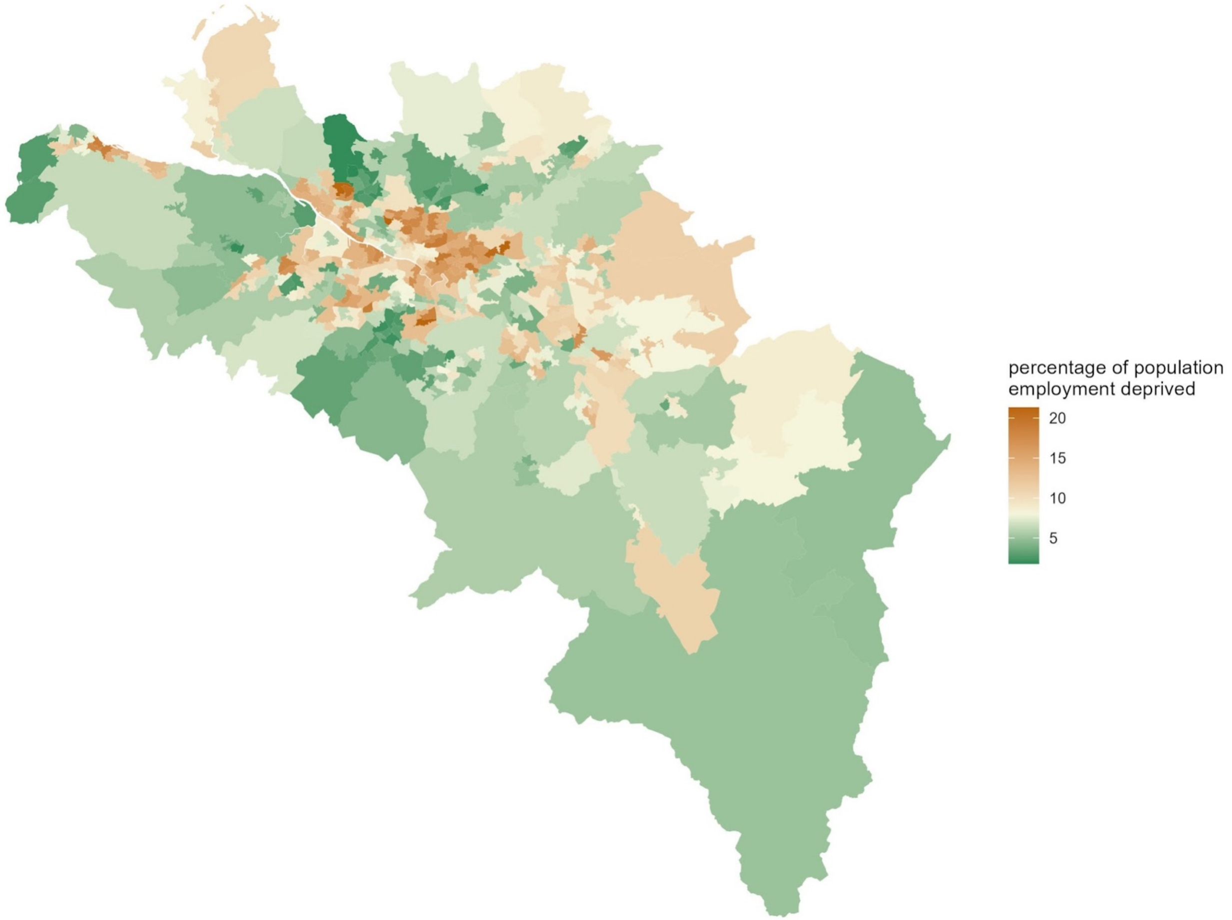
Intro: SIPHER Synthetic Population

SIPHER Synthetic Population for Individuals in Great Britain

"Digital Twin" of the adult population (16+ years) in Scotland, England, and Wales

Created by combining survey data with population statistics for small areas

Representative across a wide number of variables



Plot shows IZ-level for Glasgow City Region (N = 1.5+ Million)
Source: Lomax, Höhn, Heppenstall, et al. (2023)

Rationale

To understand the health outcomes for sub-groups of the population or across different geographies, we need to be able to build bespoke groupings from individual level data.

Individuals possess distinct characteristics, exhibit distinct behaviors and accumulate their own unique history of exposure or experiences.

However, there is a lack of individual level data available outside of secure settings, especially covering large portions of the population.

We create a synthetic dataset of individuals: their detailed attributes can be used to model a wide range of health and other outcomes

The Solution: Microsimulation

Guy Orcutt, an American econometrician was frustrated with the limitation of macroeconomic models for assessing the impacts of policy simulations

He recognised that macro approaches largely ignore any distributional effects

Orcutt argued that theoretical models of socio-economic systems are best applied at the individual level because it is individuals who make decisions within the system

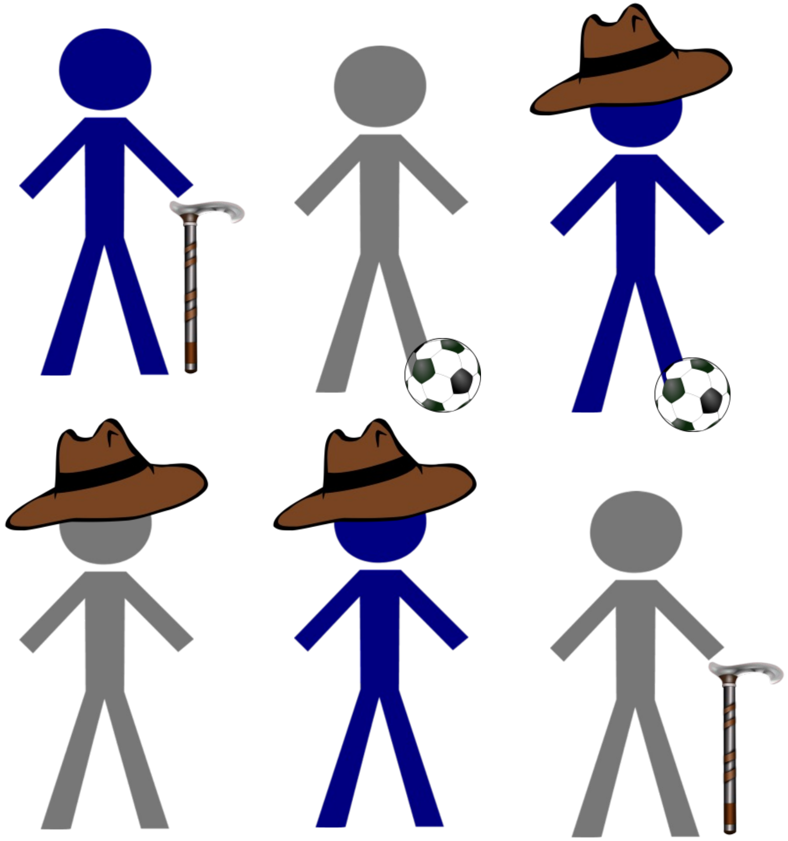
- Orcutt, G.H., 1957. A New Type of Socio-Economic System. Rev. Econ. Stat. 39, 116–123.

Applications of Spatial Microsimulation are varied

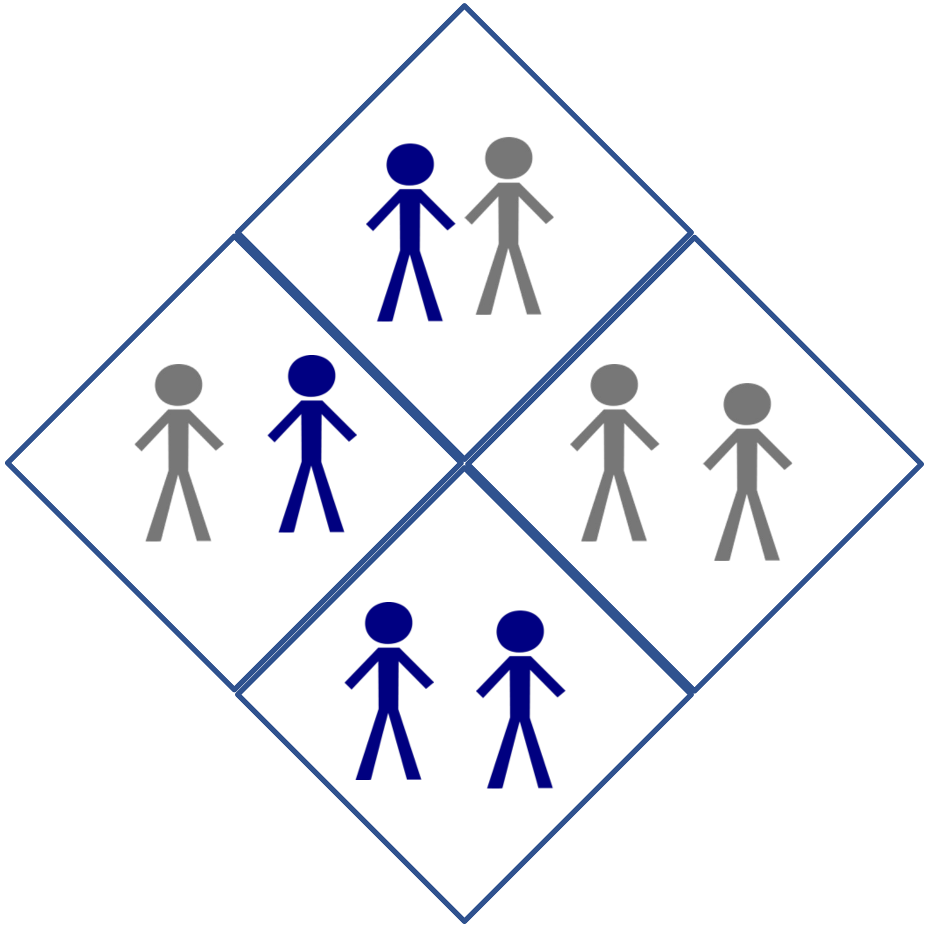
Transport	• Logistics (de Jong et al. 2007) • Commuting (Lovelace et al. 2014)
Health	• Access to GP services (Morrisey et al. 2008) • Estimating elderly morbidity (Clark et al. 2014)
Policy analysis	• Population projection (Harding et al. 2011) • Estimating poverty rates (Tanton et al. 2009)
For further overview see	• Lomax 2022, Ballas (2008)

Spatial Microsimulation

Sample or survey data

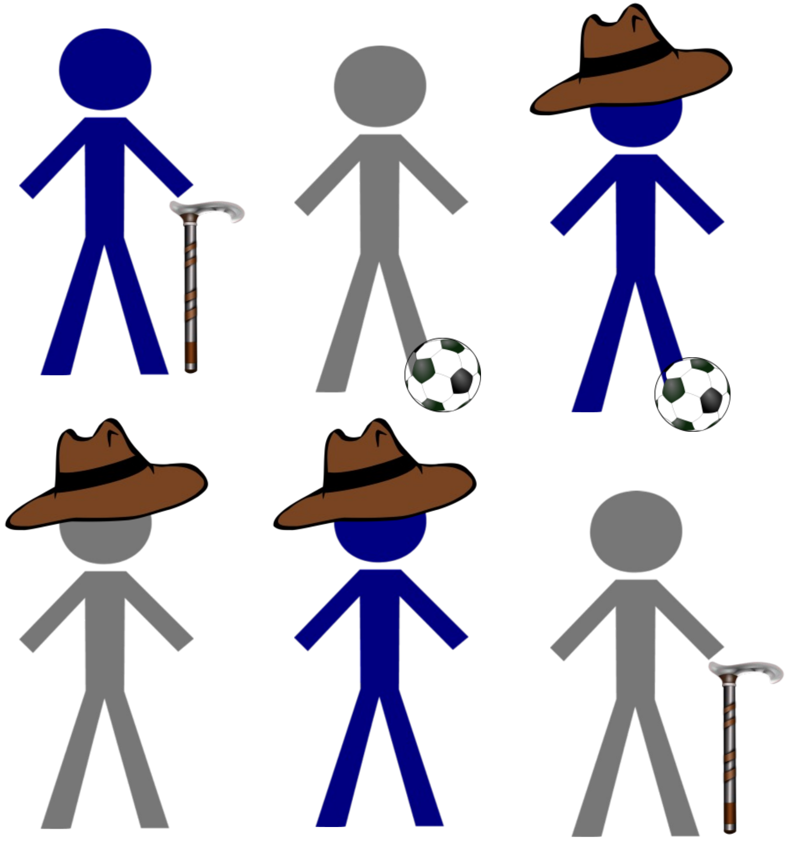


Target or constraining data

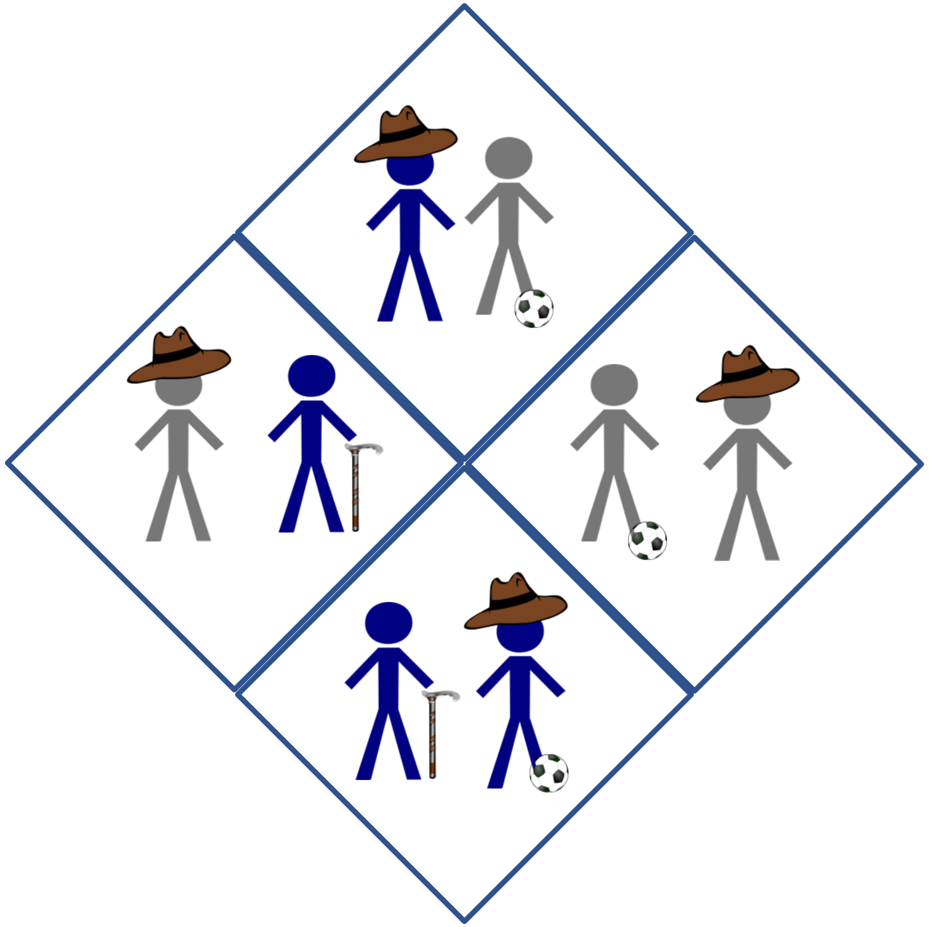


Spatial Microsimulation

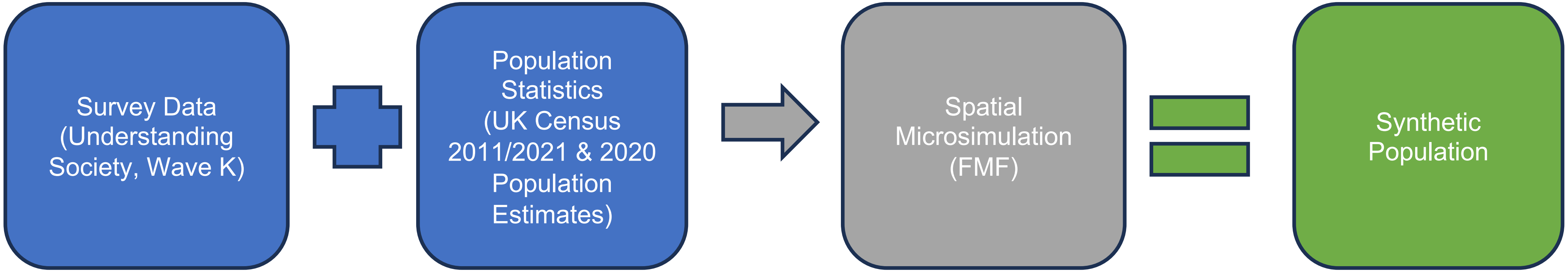
Sample or survey data



Target or constraining data



Intro: SIPHER Synthetic Population



QR Code: Link to Nature Scientific Data paper describing methodology

SIPHER Consortium

Simulated Annealing algorithm

Rutenbar, R.A., 1989. Simulated annealing algorithms: An overview. *IEEE Circuits and Devices magazine*, *5*(1), pp.19-26.

Creation and quality control

Survey Data (Understanding Society, Wave K) + Population Statistics (UK Census 2011/ 2020 Population Estimates)

Understanding Society (UK Household Longitudinal Study)
largest (N = 40,000+)
longest-running (since 2009/2010)
multi-topic (e.g., family, employment, health)
panel study (“repeated visits”)
representative (at the national level)
for the UK (coverage: SCO, E&W, NI)

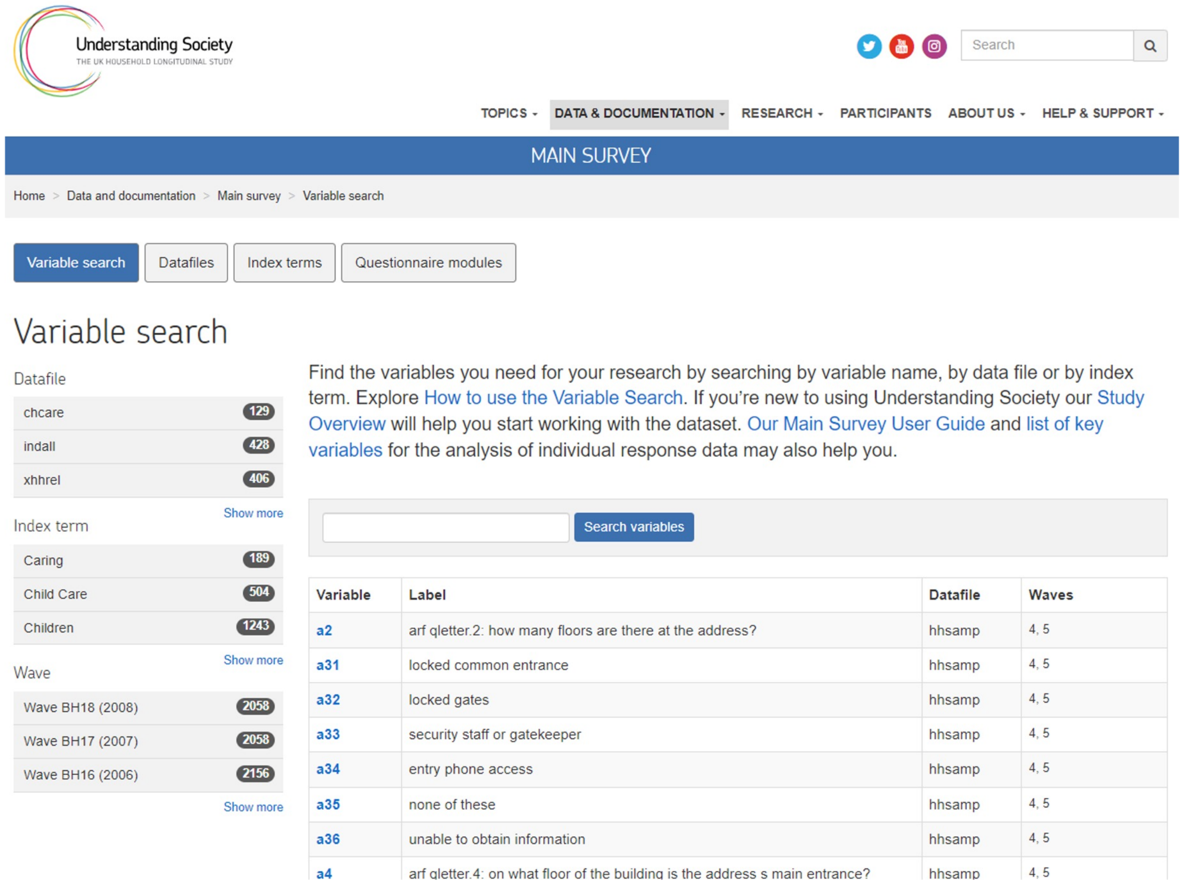
Institute for Social and Economic Research (ISER)
Continuation of BSPS (Waves 1-18, 1991-2009)
Waves available “A” (#1, 2010) to “M” 13 (#13, 2022)
£100 million UKRI Investment for 2023-2032

SIPHER Consortium

Creation and quality control

Survey Data (Understanding Society, Wave K)

Population Statistics (UK Census 2011/ 2020 Population Estimates)



QR Code: Link to Understanding Society variable search online tool

Creation and quality control

Survey Data (Understanding Society, Wave K)

+

Population Statistics (UK Census 2011/ 2020 Population Estimates)

Constraint Dimension	Variables in Understanding Society	Table ID Census 2011
Age/Sex *	age_dv / sex	NOMIS 2020* Population Estimates
Highest qualification	hiqual_dv	QS501EW/SC
Ethnicity	racel_dv	LC6201EW/SC
Marital status	marstat	KS103EW/SC
Economic activity	jbstat	LC6201EW/SC
General health	scsf1	QS302EW/SC
Household tenure	tenure_dv	LC3408EW and QS403SC
Household type (“Composition”)	hhtype_dv	LC1109EW/SC

** Not part of the UK Census 2011*

Aligned categories for household tenure constraint:

(1) owned outright
(2) owned mortgage
(3) rented, social
(4) rented, private
(5) other

Creation and quality control

= Synthetic Population

ZoneID (LSOA / Datazone)	pidp (US id, not unique)
E01004766	1
E01004766	2
E01004766	3
E01004766	4
E01004766	5
E01004766	1
E01004766	7
E01004766	4

Synthetic Population: a two-column file

(1) Columns reflecting area and a non-unique person identifier.

(2) With ca. 55 million rows, one for every synthetic individual

(3) Which can be merged with the Understanding Society survey data sets for individuals and households

Creation and quality control

= Synthetic Population

ZoneID (LSOA / Datazone)	pidp (US id, not unique)	Age	Sex	SF-12 Physical Health Score	HH has problems paying Council Tax
E01004766	1	20	Male	54.12	Yes
E01004766	2	24	Female	47.69	No
E01004766	3	34	Male	37.45	No
E01004766	4	87	Female	51.71	No
E01004766	5	49	Male	52.65	No
E01004766	1	20	Male	54.12	Yes
E01004766	7	54	Male	47.78	No
E01004766	4	87	Female	51.71	No

SIPHER SP **"k_indresp"** **"k_hhresp"**

Creation and quality control

= Synthetic Population

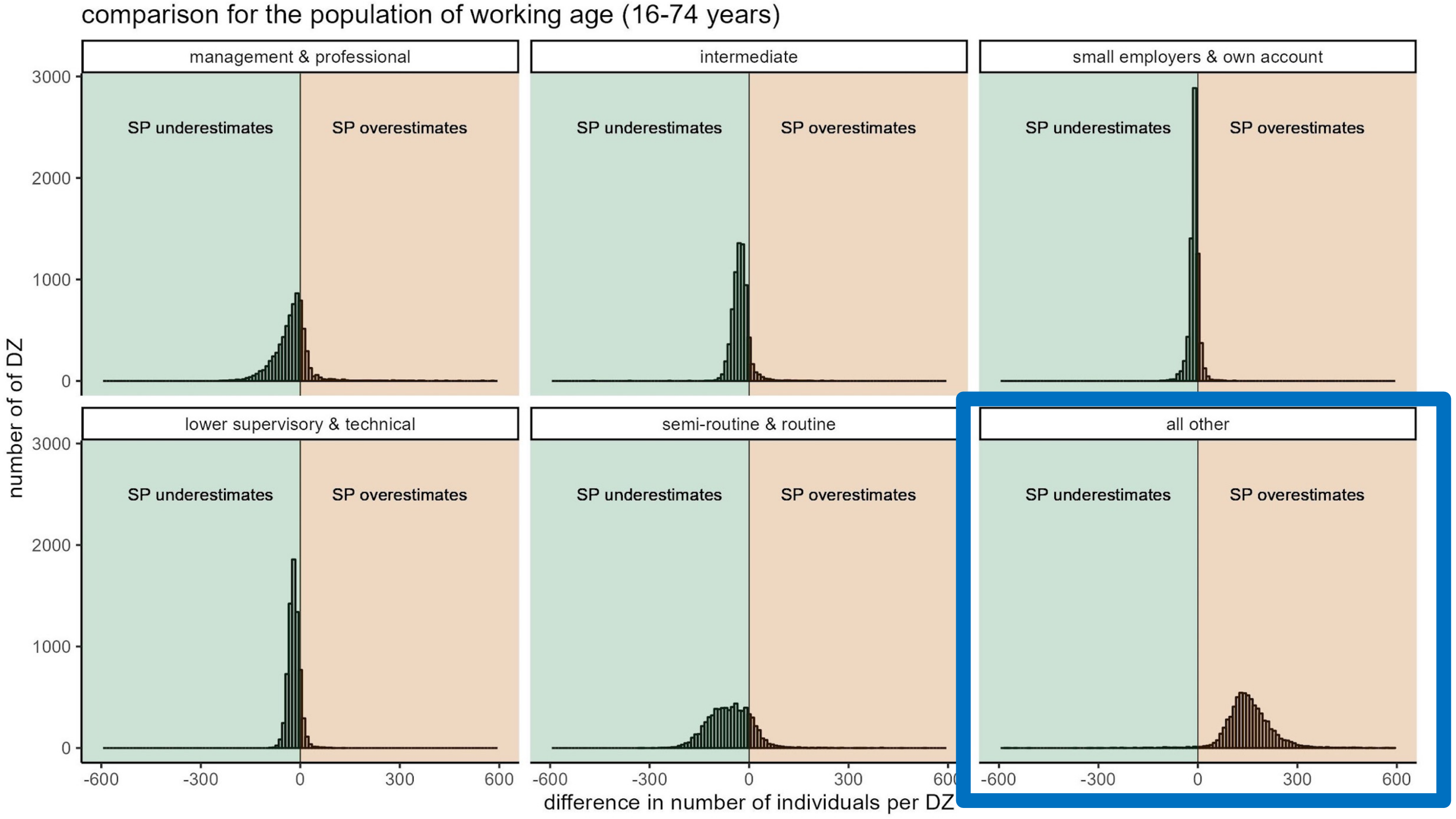
ZoneID (LSOA / Datazone)	pidp (US id, not unique)
E01004766	1
E01004766	2
E01004766	3
E01004766	4
E01004766	5
E01004766	1
E01004766	7
E01004766	4

Synthetic Population: Quality Control

(1) Internal Validation: A check of our data joinery work (e.g., alignment of constraints, major problems of algorithm)

(2) External Validation: Comparison against non-utilised information to assess reliability of created data source (e.g. IMD/SIMD, DWP data)

Creation and quality control: external



Care is required when working with residual categories

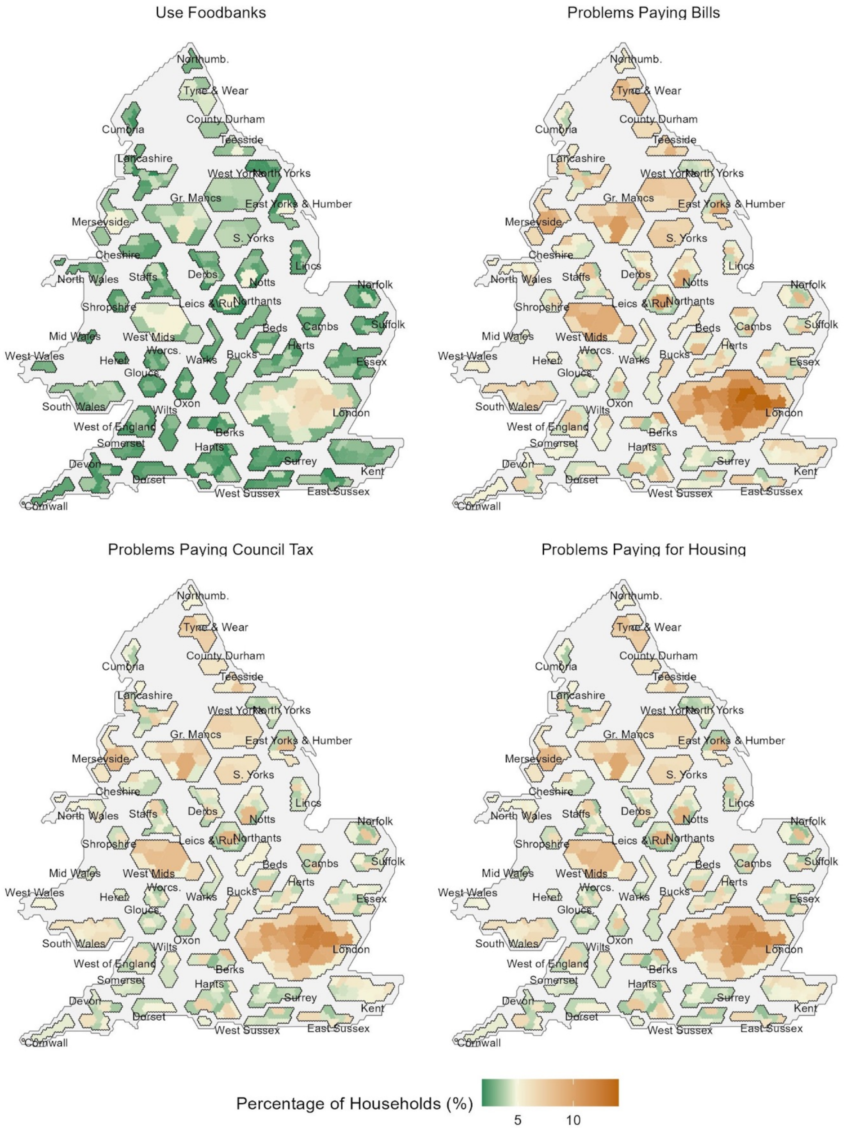
Levels of confidence and limitations

(1) Very high (= <u>all utilised constraints</u>, e.g., age, sex, education, employment)

(2) High, but caution (= <u>strongly associated</u> with utilised constraints e.g., occupational group, financial hardship, health risk factors)

(3) Unknown, likely problematic (= very <u>specific characteristics</u> of individuals or area-level, e.g., swimming in the sea, historic places)

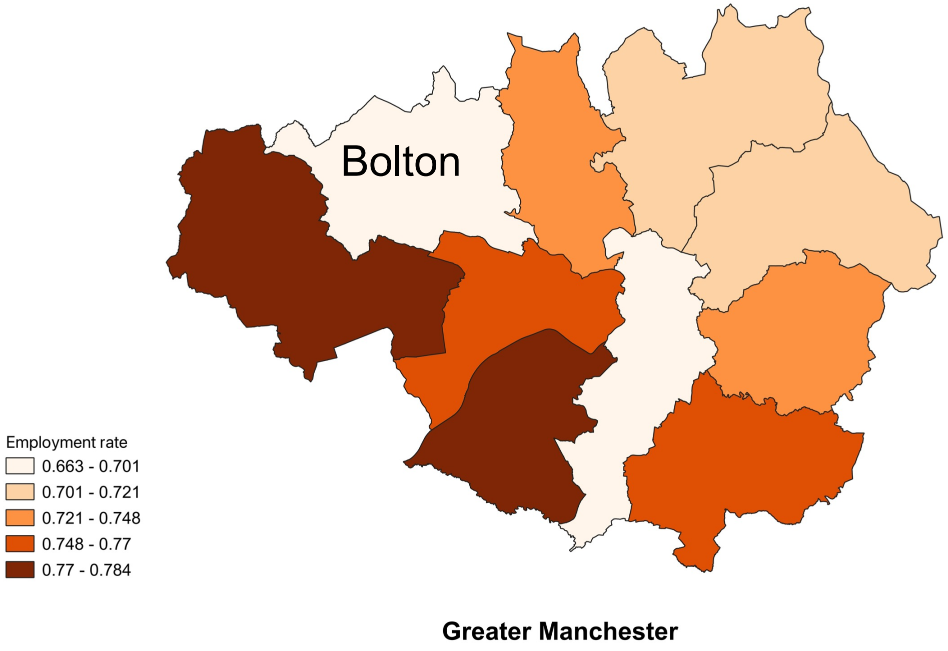
(4) Unknown, but reasonable (= <u>everything else!</u> e.g.: decoration, noisy neighbours)



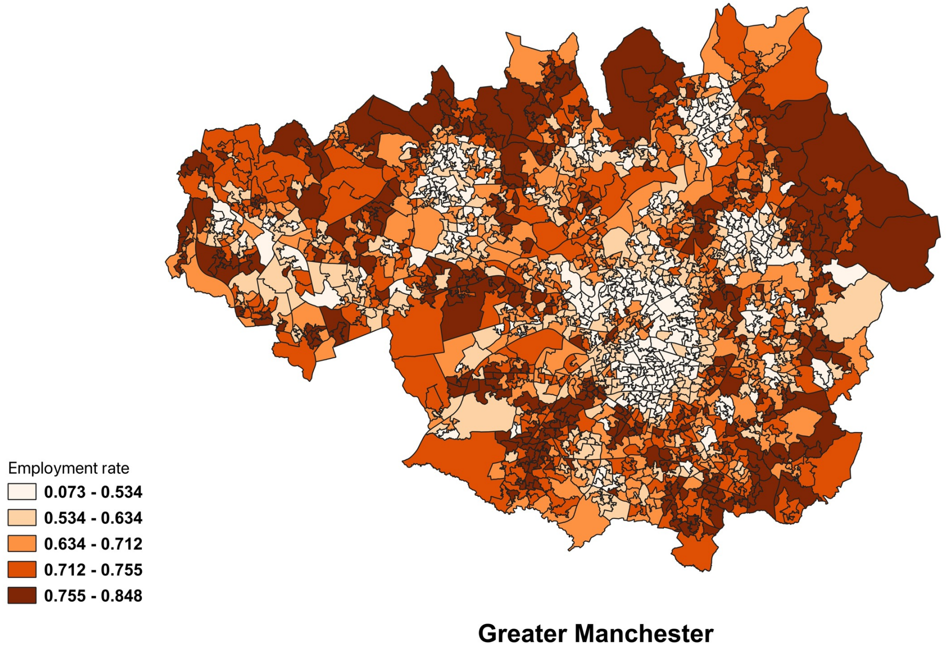
SIPHER Consortium

Part 2:
Some example uses for the SIPHER synthetic population

An example of utility: allows for spatially detailed analysis

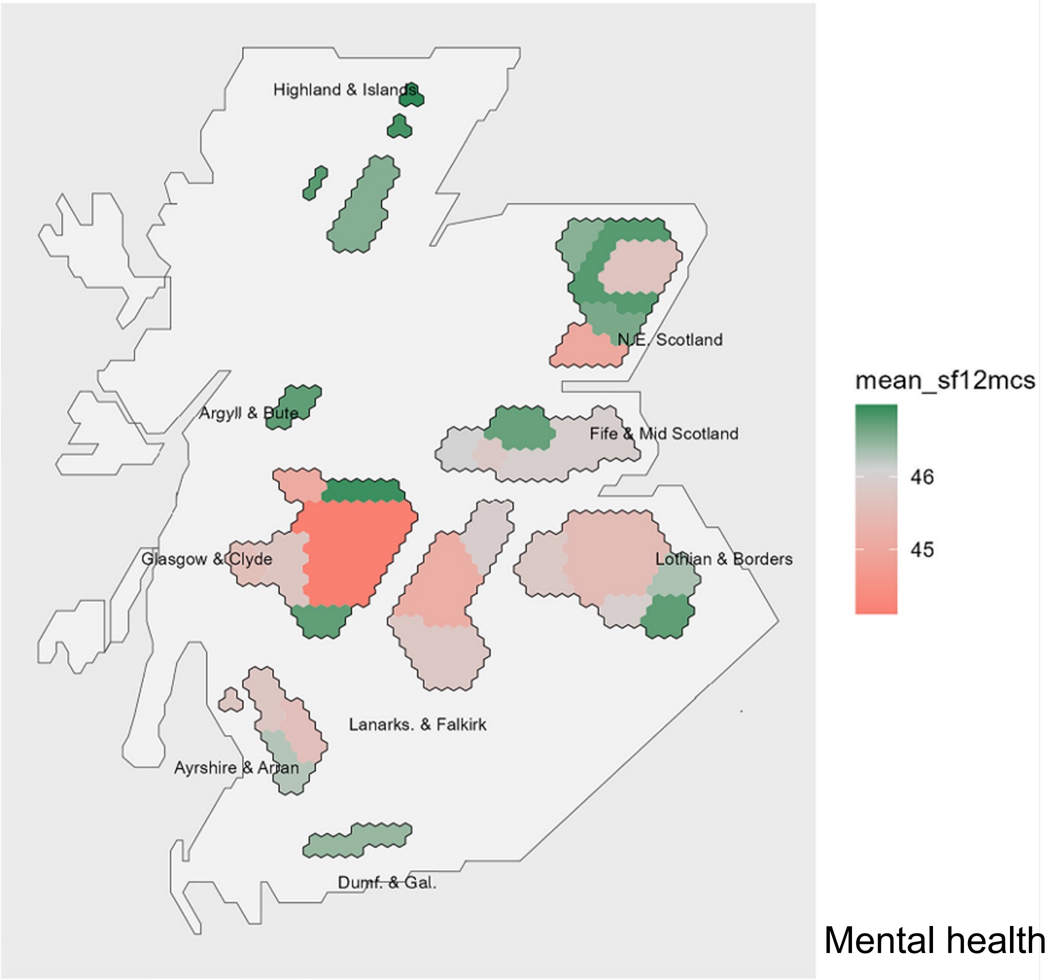


NOMIS (Annual Population Survey)



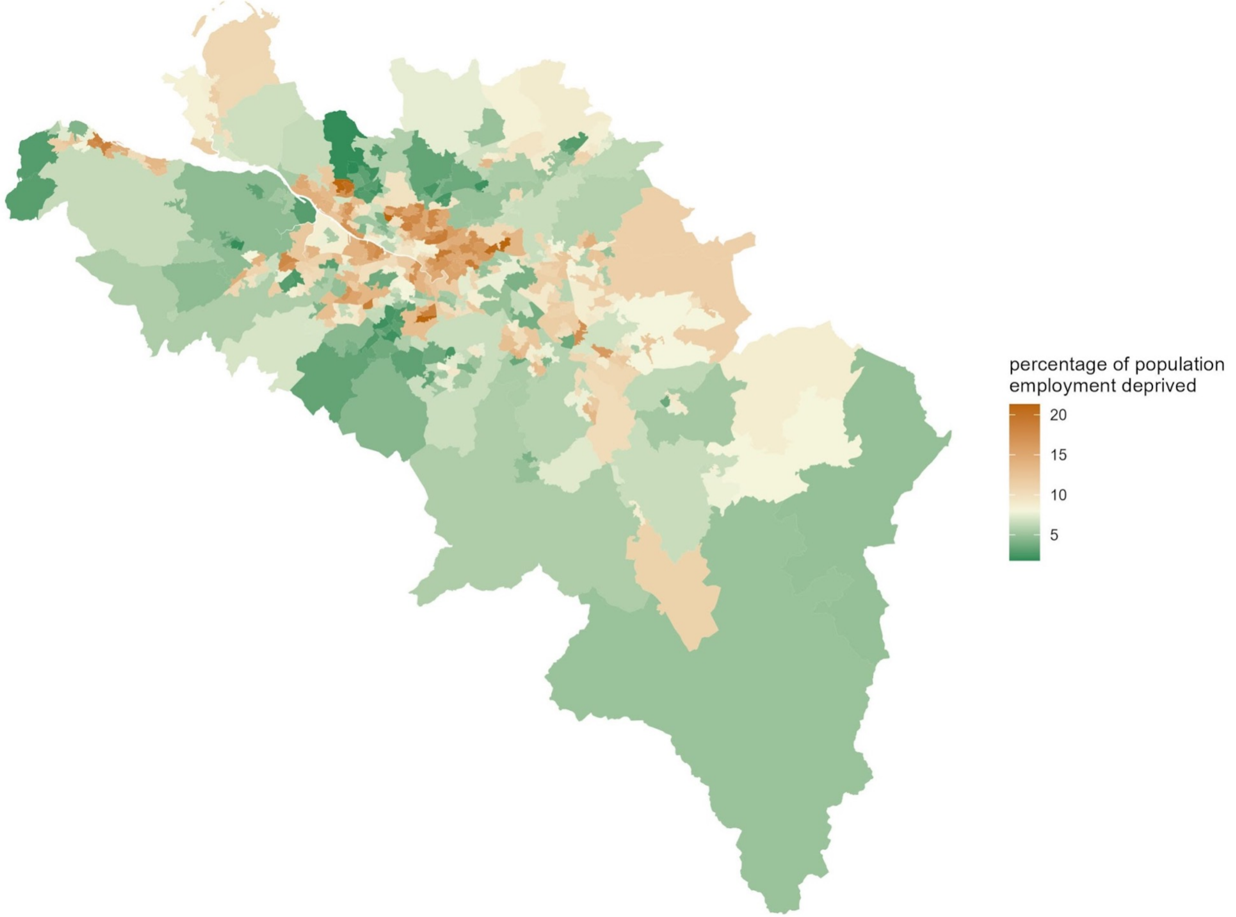
Synthetic population data

An example of utility: allows for spatially detailed analysis

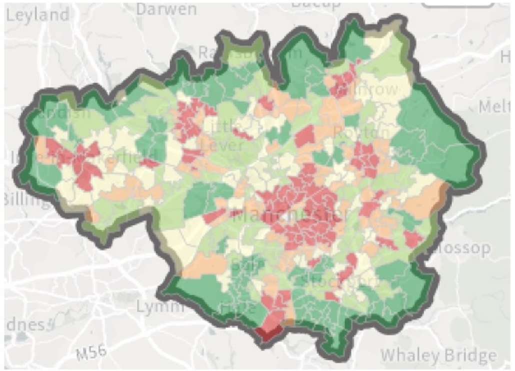


Mental health at CA level

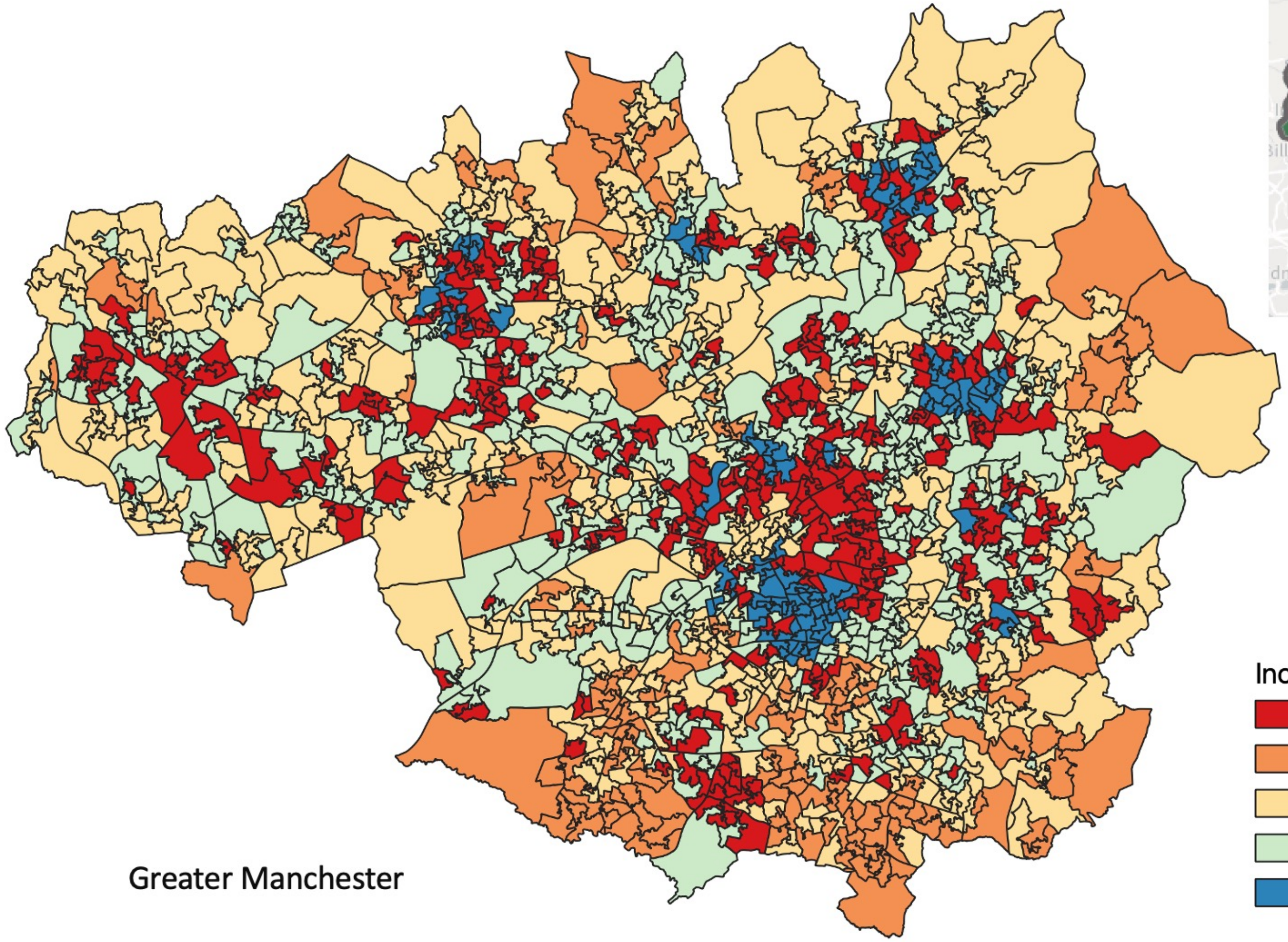
Employment Deprivation



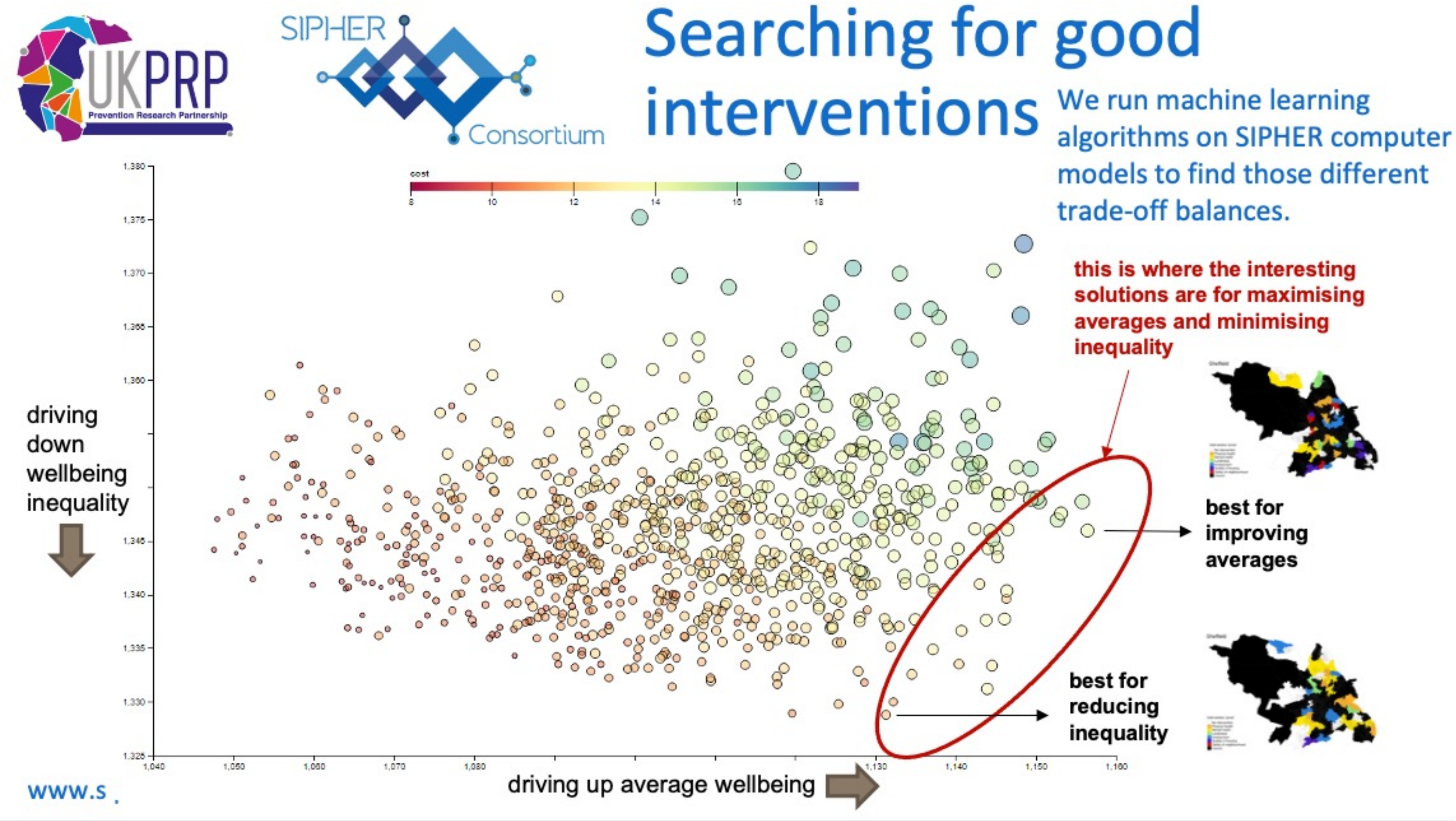
An example of utility: calculating metrics and indexes



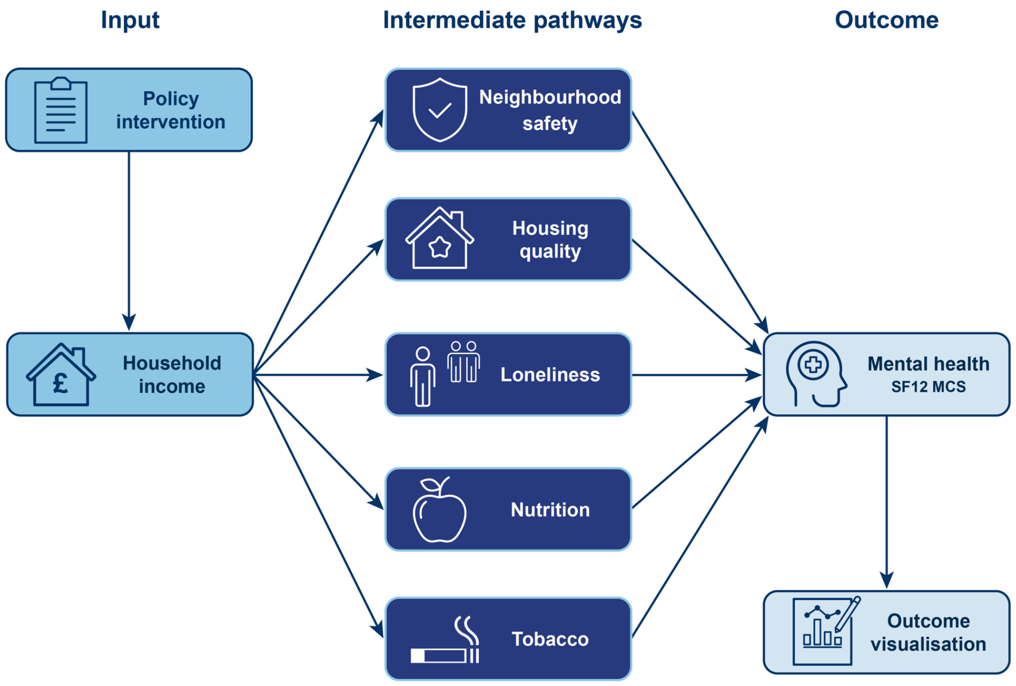
Healthy life expectancy


Greater Manchester

An example of utility: spatial optimisation

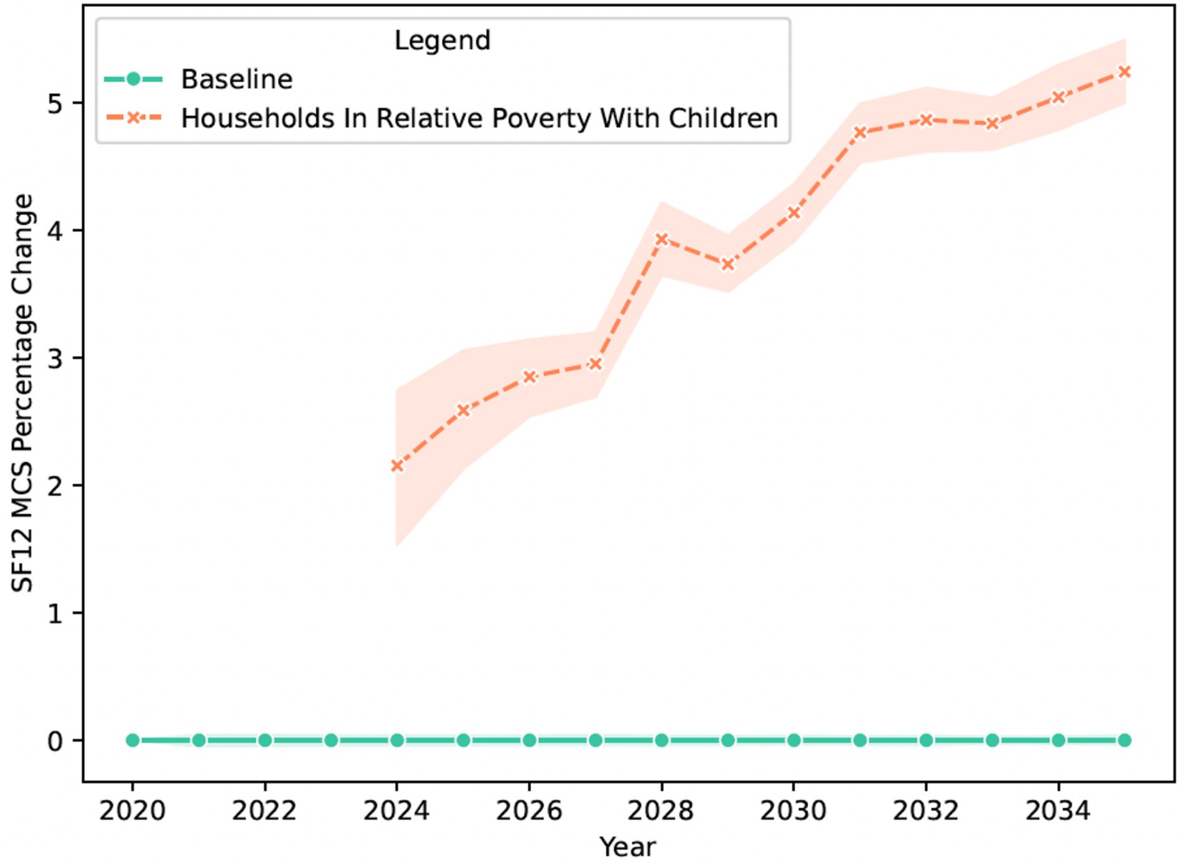


As an input to other (dynamic) models

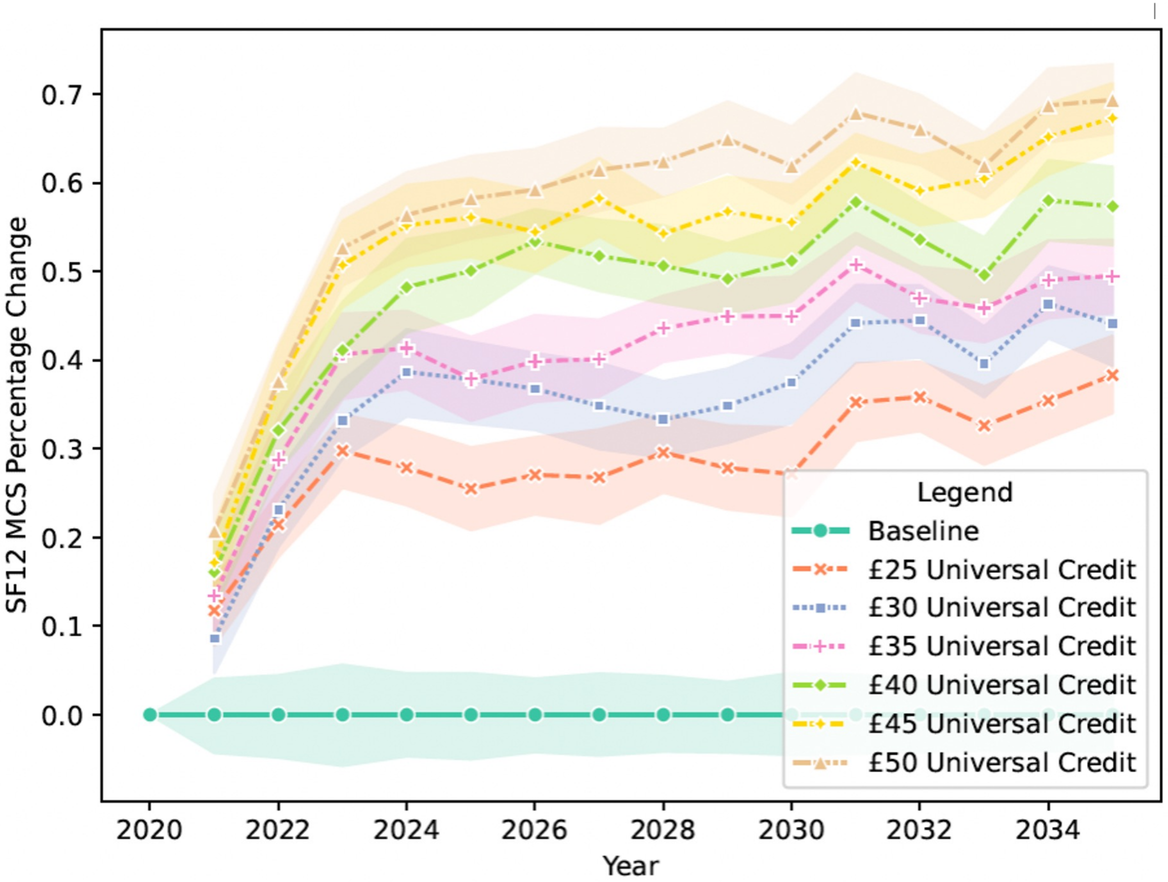


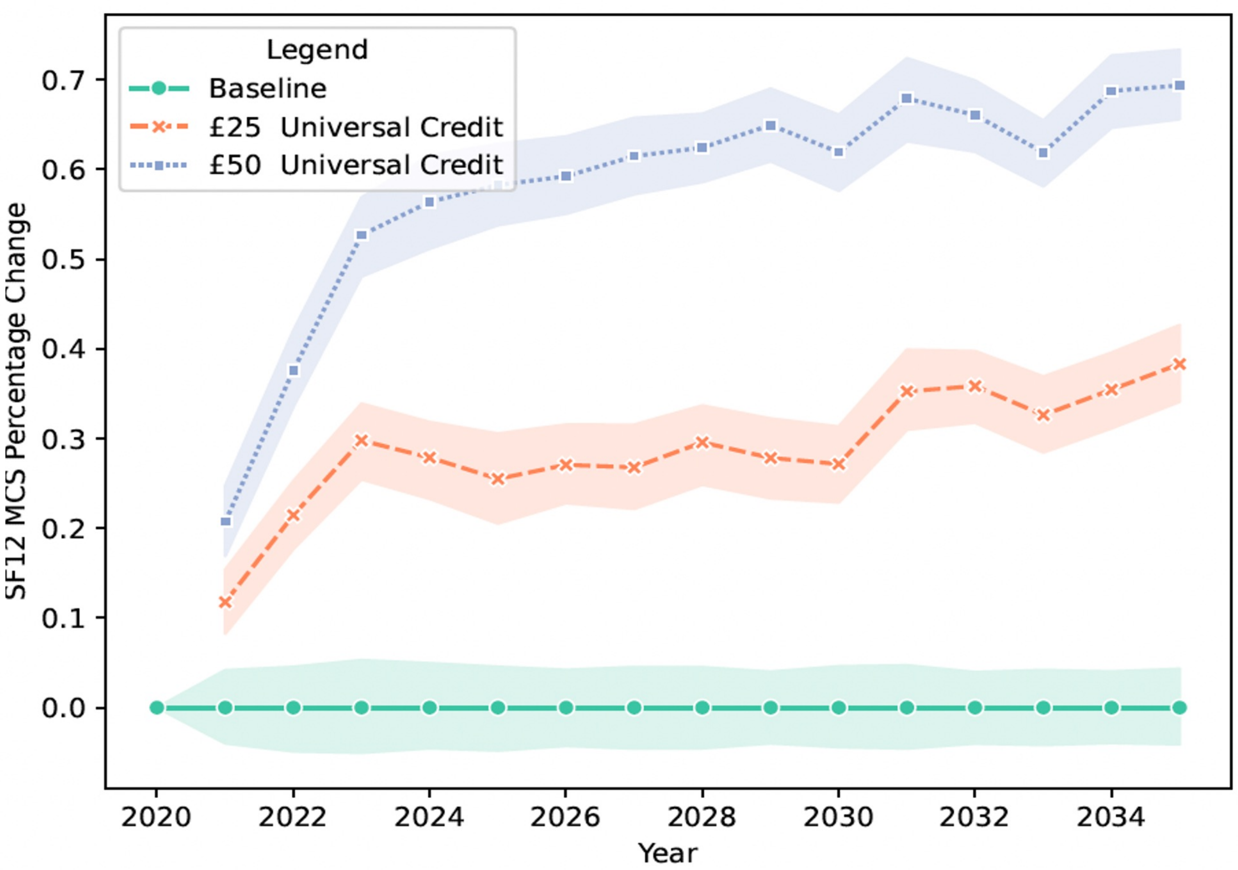
Graphs showing the SF-12 MCS improvement if the Relative Poverty target is met by 2030

- Fewer than 10% of children living in families in relative poverty
- Cost: Initially costs £405m per month (£900 per head in the relevant group).



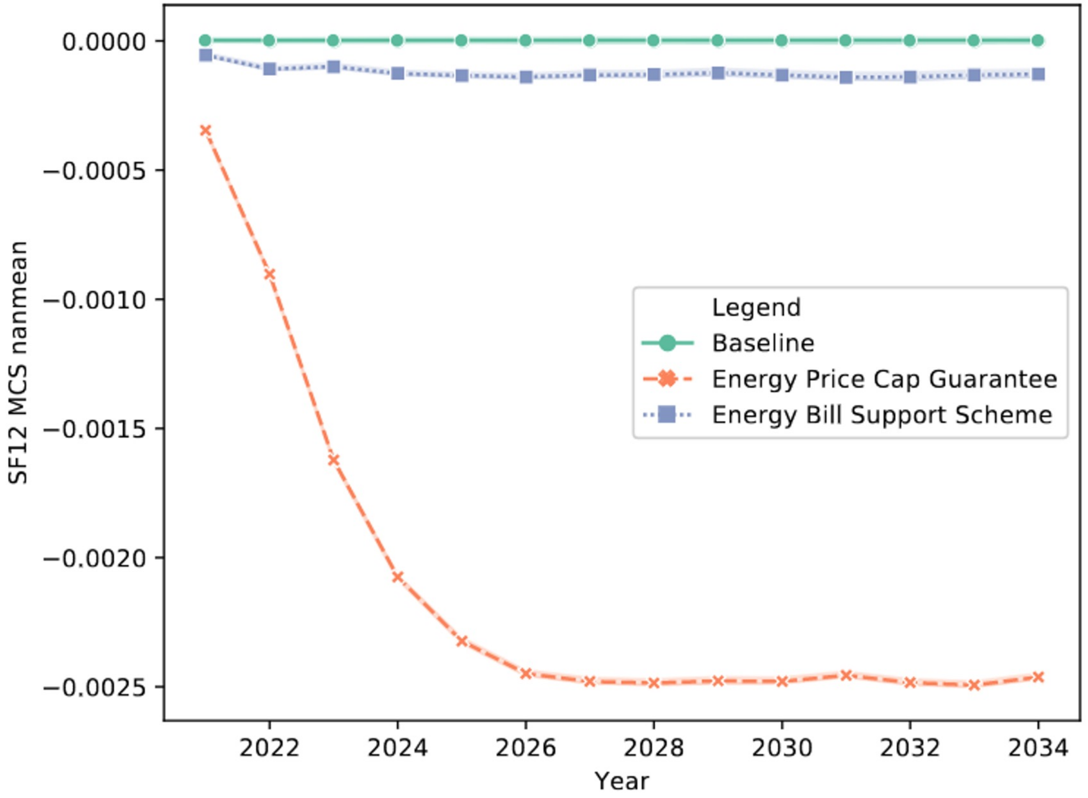
The SF-12 MCS improvement gained in the 16+ population from the Scottish Child Payments: Universal Credit group

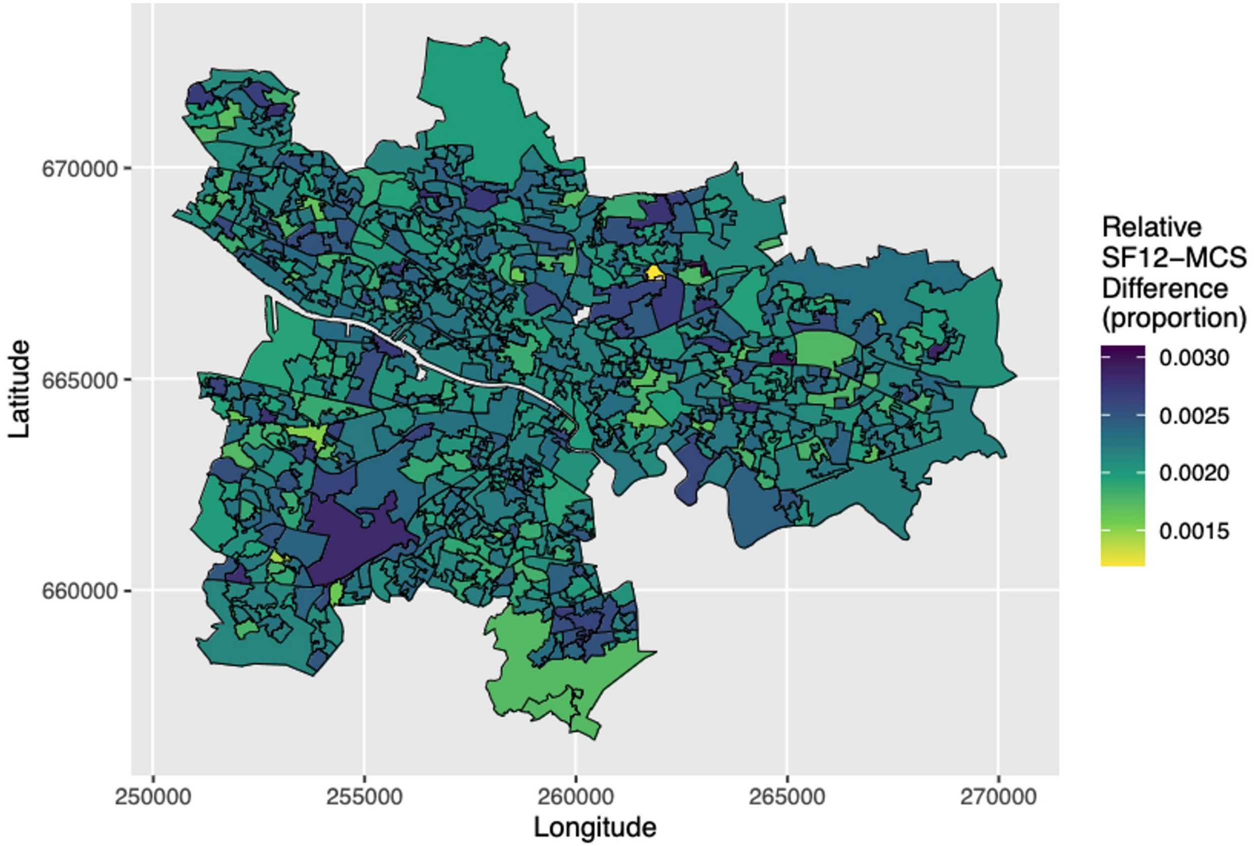




Interventions Cost £13.8m and £27.6m per month respectively in 2022/2023

Assessing the spatial distribution of model results: Energy price cap





DZs in Glasgow that benefited most from support

How to access the data

If you want the data now

1. Contact me and I will share the lookup file (individual ID -> LSOA/DZ code)
2. Register with the UK Data Service: https://ukdataservice.ac.uk/
3. Agree to conditions then download Understanding Society data
4. Merge lookup with US data (I can supply R code)

If you can wait a couple of weeks…

A data deposit will be available via the UKDS: https://ukdataservice.ac.uk/

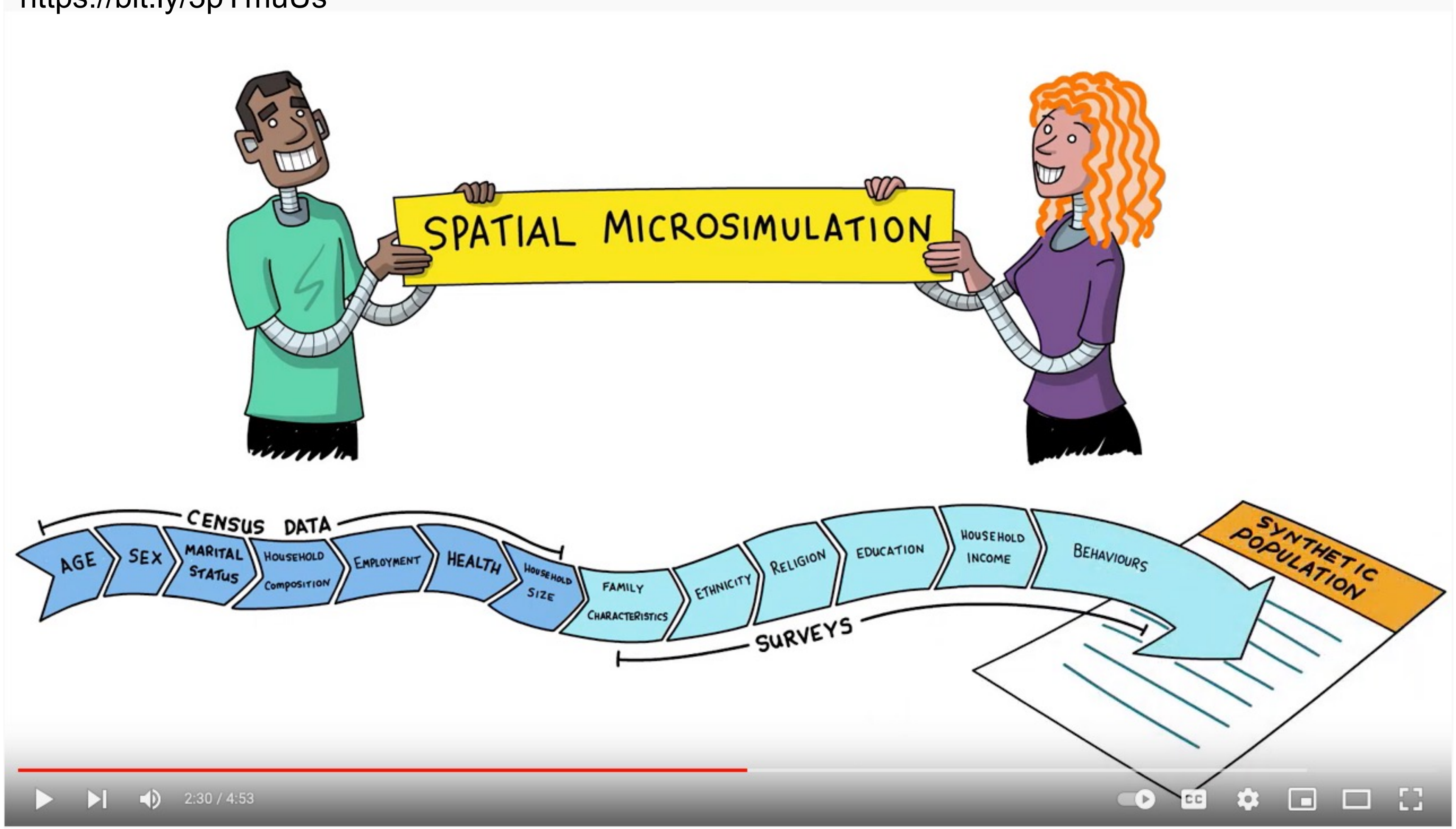
Complete with technical user manual

This deposit is undergoing final review by partners at Understanding Society and UKDS

Further information

https://www.nature.com/articles/s41597-022-01124-9

https://bit.ly/3pYmuUs



www.nature.com/scientificdata

scientific **data**

Check for updates

OPEN

DATA DESCRIPTOR

A synthetic population dataset for estimating small area health and socio-economic outcomes in Great Britain

Guoqiang Wu[1✉], Alison Heppenstall[1,2], Petra Meier[3], Robin Purshouse[4] & Nik Lomax[1,2]

In order to understand the health outcomes for distinct sub-groups of the population or across different geographies, it is advantageous to be able to build bespoke groupings from individual level data. Individuals possess distinct characteristics, exhibit distinct behaviours and accumulate their own unique history of exposure or experiences. However, in most disciplines, not least public health, there is a lack of individual level data available outside of secure settings, especially covering large portions of the population. This paper provides detail on the creation of a synthetic micro dataset for individuals in Great Britain who have detailed attributes which can be used to model a wide range of health and other outcomes. These attributes are constructed from a range of sources including the United Kingdom Census, survey and administrative datasets. It provides a rationale for the need for this synthetic population, discusses methods for creating this dataset and provides some example results of different attribute distributions for distinct sub-population groups and over different geographical areas.

Background & Summary

One of the central issues that researchers and policy makers face when modelling outcomes in a public health context is access to spatially representative individual-level data. Access to this data would enable researchers to examine bespoke spatial and sub-group effects of interventions and policy scenarios, thereby assessing their equability and implications within a wider policy making context. However, access to such individual level data are understandably restricted, owing to their sensitive nature. This presents a major barrier to the development of models that can inform spatially relevant interventions in a timely fashion. One way of dealing with this is the creation of synthetic data that are representative of the relationships contained within the real population.

A well established method for creating such synthetic datasets is microsimulation. In brief, microsimulation uses attribute-rich individual-level sample data to estimate the characteristics of a larger population[1,2]. An extension of this approach that explicitly accounts for spatial distributions is often termed spatial microsimulation[3]. In both microsimulation and spatial microsimulation, the resulting synthetic population dataset can be used to simulate impacts of interventions or evaluation of policy changes at an individual level which can then be aggregated over population sub-groups or geographies to calculate the overall impact of the policy scenario[4].

Typically, a synthetic population generated using microsimulation has a census or other large scale coverage survey as its backbone. Depending on the focus of the research agenda being addressed, this base population can be further enriched from other data sources. There are numerous examples of this approach being successfully applied to answer key policy questions which have a spatial dimension. These include the assessment of consumer expenditure patterns[5], estimating local area infrastructure demand[6] and health care planning in relation to the spatial distribution of morbidities[7].

Normally, the micro component of microsimulation represents units such as individuals, households or firms, which are simulated via a process of assigning attributes to those microunits from other data sources[2].

[1]Leeds Institute for Data Analytics and School of Geography, University of Leeds, Woodhouse Lane, Leeds, West Yorkshire, LS2 9JT, UK. [2]Alan Turing Institute for Data Science & AI, The British Library, London, NW1 2DB, UK. [3]MRC/CSO Social and Public Health Sciences Unit, University of Glasgow, Berkeley Square, 99 Berkeley Street, Glasgow, G3 7HR, UK. [4]Department of Automatic Control and Systems Engineering, University of Sheffield, Portobello Street, Sheffield, S1 3JD, UK. ✉e-mail: g.wu@leeds.ac.uk

SIPHER Consortium

Policy Partners

Sheffield City Council

GMCA GREATER MANCHESTER COMBINED AUTHORITY

Public Health Scotland

Scottish Government Riaghaltas na h-Alba

Academic Partners

University of Glasgow

University of Sheffield

University of Strathclyde Glasgow

UNIVERSITY OF LEEDS

Newcastle University

MANCHESTER 1824 The University of Manchester

University of Nottingham UK | CHINA | MALAYSIA

The SIPHER Consortium is funded by:

The SIPHER Consortium is supported by the UK Prevention Research Partnership (Grant MR/S037578/2), which is funded by the British Heart Foundation, Cancer Research UK, Chief Scientist Office of the Scottish Government Health and Social Care Directorates, Engineering and Physical Sciences Research Council, Economic and Social Research Council, Health and Social Care Research and Development Division (Welsh Government), Medical Research Council, National Institute for Health Research, Natural Environment Research Council, Public Health Agency (Northern Ireland), The Health Foundation and Wellcome.

SIPHER Consortium

THANK YOU FOR LISTENING

Find out more

Website: www.sipher.ac.uk

Email: sipher@Glasgow.ac.uk

Twitter: @SipherC