Getting to grips with different types of auxiliary data:

Sarah Butt (City University London) and Kaisa Lahtinen (University of Liverpool)

26th May 2016

Tackling survey nonresponse: The role of geocoded auxiliary data

Getting to grips with different types of auxiliary data: Was it worth it?

Sarah Butt and Kaisa Lahtinen

26th May 2016

Tackling survey nonresponse: The role of geocoded auxiliary data

Outline

- ADDResponse data sources
- Challenges of using auxiliary data
- Points of Interest data
- Commercial data

Data sources

European Social Survey in UK

- ESS Round 6 fieldwork Sept 2012-Feb 2013.
- Carried out by Ipsos MORI
- Address-based sample using Postcode Address File (PAF)
- Sample of 4,520 addresses clustered in 226 postcode sectors
- 54% RR (38% refusals, 7% non-contacts)

002: prioritya

Small area data

- Geocoded data readily available
  - Census 2011 (ONS)
  - Crimes per 10,000 population (Home Office)
  - Benefit claimant rates (DWP)
  - Indices of multiple deprivation (DCLG)
  - School absences (DfE)
  - Electricity consumption (DECC)
  - OAC and census “hard to count” measure (ONS)

- Easily matched using National Postcode Lookup File (in theory…)
- Differences across four countries of UK

Commercial data

- Data purchased from two “value added resellers”
- Matched using exact address
- Consumer segmentation variables: ACORN, MOSAIC etc.
- Specific variables for e.g. length of residency, tenure, house price, age, employment status, children present, marital status
- Consumer preferences data -> too much missing

Points of interest data

- Ordnance Survey maintains record of 600+ “points of interest” in GB (not NI)
  - Astrologers to Zoos
- Using national grid coordinates, can locate ESS addresses in relation to POIs
  - Some POIs more prevalent in certain types of neighbourhood
  - Physical environment/amenities influence quality of life and behaviour
- OpenStreet Map as an alternative data source

Auxiliary data challenges

- No shortage of data (400+ variables, 20 different data sources)

BUT:

- Resource intensive
- Vagaries of geocodes
- Timing
- Too many variables?

CITY UNIVERSITY LONDON

THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

POI DATA

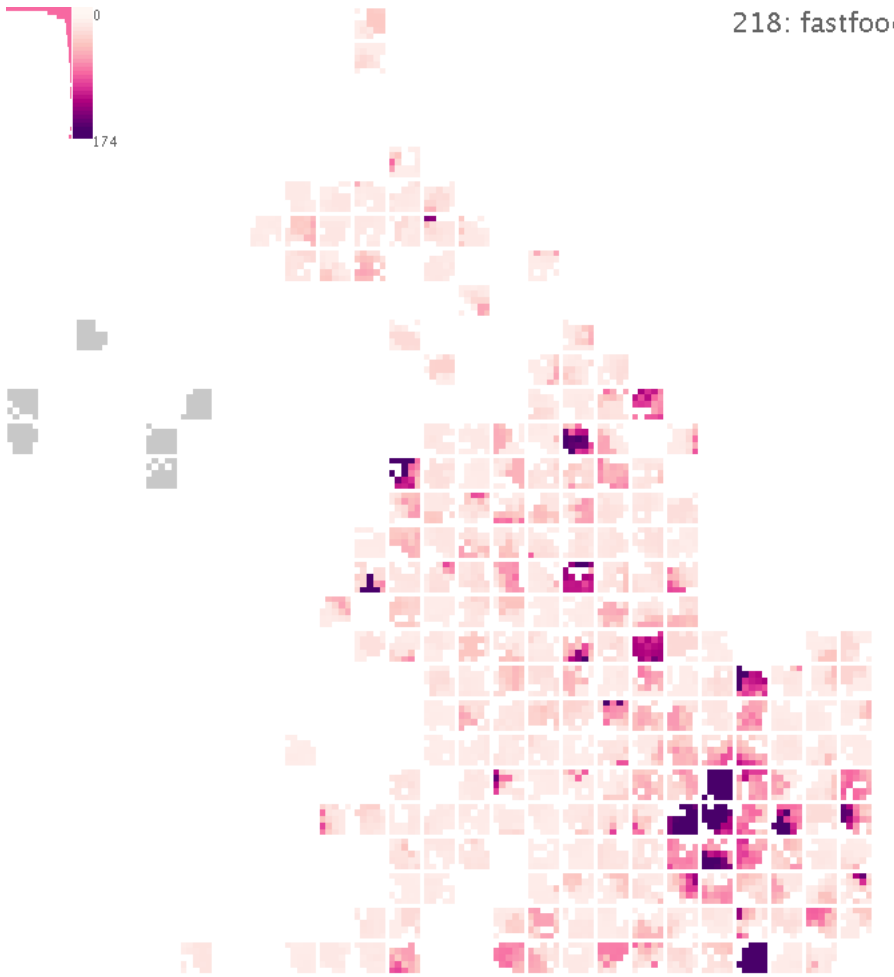
POI: Rationale

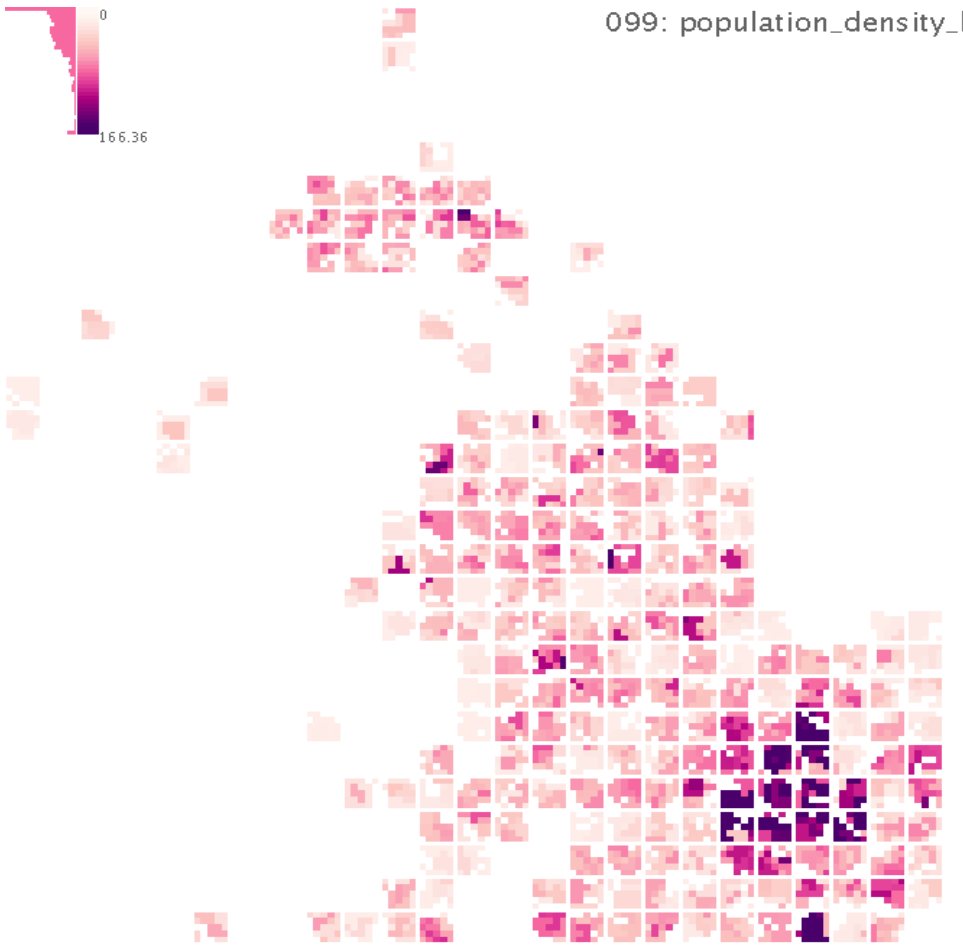
- Built environment can have a significant effect on behaviour and attitudes
  - Green space - > mental health/wellbeing, social trust, physical activity
  - Fast food outlets/alcohol vendors - > health
  - “Evening economy - > crime rates, perceptions of anti-social behaviour

- Possible influence on response behaviour and/or ESS survey variables?

Challenge to identify suitable POI variables

- 616 different POIs (4 million records)
- Different ways to construct POI-based variables
  - Counts, ratios, proximity (Euclidian vs. travel times)
  - Administrative areas vs. “buffer zones”
- Concentration of POIs highly skewed
- POIs concentrated in densely populated urban areas


0
174
218: fastfood_D


0
166.36
099: population_density_lsoa

Hypotheses

- Happier people more likely to respond to surveys (Groves et al, 1992)
  - POI measure of wellbeing: Distance (in m) to nearest recreational outdoor space
  - POI measure: Number of industrial sites within 1km radius
  - POI measure: Presence of waste plant within 1km radius
- Fear of crime/perception of anti-social behaviour reduces survey response (Groves and Couper, 1998)
  - POI measure: Number of pubs, bars and nightclubs within 800m ("evening economy")

Findings: Are POIs correlated with ESS variables?

- Bivariate (urban areas)
  - Wellbeing POIs and self-rated happiness - ns
  - "Evening economy" and social trust - ns
  - "Evening economy" and fear after dark - sig (+)

- Controlling for population density, IMD, crime rates
  - No significant relationships

Findings: Are POIs predictive of response propensity?

- Bivariate (urban areas)
  - "Evening economy" sig (-)

- Controlling for interviewer obs and small-area variables*
  - None of POI variables significant
  - No effect on model fit

* Barrier to entry, condition of property, % owner occupiers, % flats, % single <35, % retired, IMD, population density, violent crime rate

CITY UNIVERSITY LONDON

THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

COMMERICAL DATA

Commercial data: Rationale

- Source of household level data on all sample units

BUT

- Quality issues (West et al, 2015l; Pasek et al 2014; Sinibaldi et al, 2014)
- Resource implications

- Do commercial data add value over and above small area data?

There are quality issues…

- Commercial data from 2015 but survey conducted in 2012
  - 34% of addresses contain post-2012 movers
- Missing data
  - Company 1: 10%
  - Company 2: 20 -50%
- Differences between two commercial databases
  - N of adults matches in 54% of cases
  - Tenure matches in 75% of cases -> Company 2 underrepresents private renters
- Data modelling a black box - > often based on aggregate data

… but some evidence data may be useful

Missing addresses:
- More likely to be ineligible
- Less likely to be contacted

Geo-segmentation variables predict response behaviour

Contacted vs not	Cooperate vs not
Missing (-)	**Missing (-)**
Length of residence	Length of residence
Married couple (+)	Married couple
Children present	**Children present**
At least one person <30	At least one person <30
At least one person 65+ (+)	At least one person 65+ (+)
No one in work	**No one in work**
Owner occupier (+)	Owner occupier
High financial stress	High financial stress
Household income	Household income
Council tax band	Council tax band

Do commercial variables “add value”?

- Compare fit of three models
  - Model 1: Interviewer observations + census variables
  - Model 2a: Model 1 + geodemographic segmentation
  - Model 2b: Model 1 + individual commercial variables

- Responded vs. not
- ESS variables: Happiness, social trust, attitudes to immigration etc.

	P	Model 1	Model 2a	Model 2b
Responded vs not	(Pseudo) R2	0.013	0.015	0.014
	AIC	5679	5672	5675
	LRtest (Model 2 vs. Model 1)		36.96 *** (15 df)	26.38*** (11 df)

	P	Model 1	Model 2a	Model 2b
Happiness (0-10)	R2	0.05	0.07	0.08
	AIC	9093.5	9068.7	9057.3
	F test Model 2 vs Model 1)		3.55 *** (15, 225)	4.89 *** (11, 225)

	P	Model 1	Model 2a	Model 2b
Attitudes to immigration (0-30)	R2	0.04	0.08	0.07
	AIC	9539.9	9478.1	9500.5
	F test (Model 2 vs Model 1)		6.02 *** (15, 225)	5.06*** (11, 225)

Summary:

- Lots of auxiliary data available
  - Data matching not straightforward
  - Too many variables can be as much of a problem as too few
- Little value added
  - Failed to identify suitable POIs for nonresponse analysis
  - Commercial data potentially promising despite quality issues but gains in predictive power small

ADDResponse: Summary

- Lots of auxiliary data available
  - Data matching not straightforward
  - Too many variables can be as much of a problem as too few
- Struggle to find suitable auxiliary variables
  - Not looking in right place?
  - Or no systematic bias?
- Auxiliary data for nonresponse analysis/weighting not a priority for ESS
  - Focus efforts on improving collection of interviewer observations and reducing interviewer effects
- Other uses of auxiliary data?
  - Topic specific surveys
  - Identifying sub-populations
  - Local models

ADDResponse: Outputs

- (Hopefully) Dataset available via special license
- Info on all auxiliary data sources used
- Report on "lessons learned" from different data sources
- Summary of today's panel discussion
- Journal articles on nonresponse findings in preparation