Automated zone design for data collection: designing workload areas for censuses and surveys

Samantha Cockings, Andrew Harfoot, David Martin
Geography and Environment, University of Southampton

Census Research User Conference, RSS, London
22 June 2017
Outline

- Designing collection geographies (workload areas)
  - Concepts, methods
- Automated zone design for collection geographies
  - New functionality (AZTool)
- Examples
  - Census Coverage Survey
  - Labour Force Survey
- Conclusions, observations
Designing collection geographies

• Censuses and surveys usually involve enumerators/interviewers visiting households/addresses (even for 2021 Census)

• How do we design areas which are as efficient and effective as possible?
  – statistical quality of outputs
  – cost of field operations

• Concept of ‘workload areas’
Workload areas

- Equalise workloads across workload areas as much as possible
- Usually requires consideration of some combination of:
  - Number of addresses/hhds to be enumerated/interviewed
  - Weighted level of difficulty to access/enumerate address (HtC?)
  - Amount of time required at each address (often linked to difficulty)
  - Travel-time (time taken to reach each address): function of distance, speed and connectivity
- Census v surveys (sampling design)
Automated zone design for data collection?

- Automated zone design (AZTool) used successfully for design of output geographies
  - Output Areas, Super Output Areas, Workplace Zones, Small Areas etc.
  - Based on contiguity (adjacency)
- Can these methods be enhanced to aid the design of data collection geographies (i.e. workload areas)?
  - Requires new functionality:
    - Connectivity
    - New metrics for workload (e.g. interviewer effort, travel time)
Automated zone design

- Iterative combination & re-combination of sets of input zones (building blocks) to create output zones (tracts) which optimise pre-specified design criteria.
AZTool – new functionality:
Connectivity mode – building blocks
AZTool – new functionality:
Connectivity mode – connectivity matrix
AZTool – new functionality: Connectivity mode – building blocks
AZTool – new functionality:
Connectivity mode – connectivity matrix
AZTool – new functionality: Connectivity mode – impedance
AZTool – new functionality

• Connectivity matrix
• Network compactness metric
  – Minimise the sum of the weights of the minimum spanning trees connecting all BBs within each tract
  – i.e. minimise the time taken to travel to all BBs within tracts and choose the solution with the lowest overall travel time for all tracts
• New workload metrics for interviewer effort
  – E.g. number of households * HtC weighting
Example 1: 2021 Census Coverage Survey
2011 Census Coverage Survey (CCS)

- Post- census survey designed to measure census coverage
- Sampling design independent of census
- 2011 CCS design
  - Stratified two-stage sample:
    - Select OAs by LAD & Hard-to-Count (HtC) index
    - Select c.50% of postcodes from selected OAs
    - Interviewers identify households within their area
- Cost model (interviewer costs + travel costs)
- Straight line distances used to calculate travel costs
2021 Census Coverage Survey (CCS) research

- Collaboration: ONS Methodology, ONS Geography, UoS
- Evaluate different sampling designs
  - Statistical quality
  - Operational costs
- Operational costs
  - Use AZTool to design workload areas for specified sampling designs
  - Evaluate relative operational costs of sampling designs, incorporating improved estimates of travel-time and interviewer effort
- NB Lists of postcodes/addresses rather than areas per se
### 2021 CCS – sampling designs for evaluation

<table>
<thead>
<tr>
<th>Sampling design (scenario)</th>
<th>Primary Sampling Unit (PSU)</th>
<th>Secondary Sampling Unit (SSU)</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>OA</td>
<td>postcode</td>
<td>2-stage clustered sample with postcodes as SSUs, i.e., as per 2011</td>
</tr>
<tr>
<td>2</td>
<td>OA</td>
<td>address</td>
<td>2-stage clustered sample with addresses as SSUs, so less clustered than in 2011</td>
</tr>
<tr>
<td>3</td>
<td>LSOA</td>
<td>postcode</td>
<td>2-stage clustered sample with postcodes as SSUs but LSOAs as PSUs</td>
</tr>
<tr>
<td>4</td>
<td>LSOA</td>
<td>address</td>
<td>2-stage clustered sample with addresses as SSUs and LSOAs as PSUs</td>
</tr>
<tr>
<td>5</td>
<td>-</td>
<td>postcode</td>
<td>1-stage clustered sample with postcodes as sampling unit</td>
</tr>
<tr>
<td>6</td>
<td>-</td>
<td>address</td>
<td>1-stage unclustered sample of addresses (or households)</td>
</tr>
</tbody>
</table>
Designing workload areas for the 2021 CCS

- **Study areas**
  - South West Wales & Powys, Manchester, South East Hampshire

- **Data**
  - 2011 Census: postcode/address, HtC category, number of households
  - OS MasterMap Integrated Transport Network (ITN) - roads

- 400 replicates per scenario (+ varying sampling fractions)
Designing workload areas for the 2021 CCS

• Methods
  – Interviewer effort per postcode/address, based on number of households weighted by HtC category
  – Connectivity matrix
  – Network compactness score

• Design criteria
  – Minimise variation in interviewer effort between workload areas
  – Minimise travel time across all workload areas
2021 CCS:
Illustrative sample points (postcodes)
2021 CCS:
Road network (ITN)
2021 CCS: Connectivity matrix & network compactness
2021 CCS:
Illustrative tracts (Workload areas)
2021 CCS: Next steps

• Finish all runs for all scenarios for all study areas

• Evaluate relative costs of scenarios
  – Further develop cost models
    • Interviewing costs (e.g. hours per interviewer, regional pay rate, number of interviewers required)
    • Travel costs (e.g. number of journeys, travel time, pay rate, mileage, mileage rate)
  – Evaluate scenarios

• If suitable, implement for CCS following 2019 Census Rehearsal?
Example 2: Labour Force Survey reinstition
Labour Force Survey (LFS) restinting

- LFS: systematic sample of addresses ordered by postcode
- Stints / Half stints used to allocate workloads, within Interviewer Areas and Field Management Areas
- Last restinting in 2003
- Issues: changes in response rates, increasing proportion of telephone interviews, no more interviews for 75+ years, changes in population/hhd distribution, various ‘boosts’
- Need for new restinting to re- equalise workloads and hopefully improve statistical quality of outputs
Labour Force Survey (LFS) restinting

- Collaboration: LFS, ONS Geography, UoS
- Aims
  - Develop methods to redesign LFS hierarchy using AZTool
    - (Re-)equalise workloads
    - Incorporate more sophisticated measure of travel-time
  - Pilot study areas: Glasgow, East Midlands
- NB Areas important here, so hybrid connectivity/contiguity mode
LFS hierarchy

GOR5s

Interviewer areas (IAs)

Stints

Half stints

BBs: Postcodes

Field Management Areas (FMAs)

Input: Constraining geography

Input: Building block

Output: Workload area
Labour Force Survey (LFS) restinting

- Workload (postcodes) = Delivery Point count * Country weighting * Urban/Rural weighting
- Target workload c. 17400 per stint (adapted depending on total workload within GOR5)
- Workload per stint equalised within GOR5 but may vary slightly between GOR5s depending on total workload within the GOR5
LFS restinting process (AZTool)

For each GOR5 region

- Postcodes
  - Merge
  - Stints
    - Split
      - Half stints
    - Merge
  - Merge
    - IAs

Process (AZTool)

Input/output dataset
LFS (Glasgow):
Postcode centroids
LFS (Glasgow):
Postcode Thiessen polygons
LFS (Glasgow): Postcode polygons re-aligned to roads, rivers and railways
LFS (Glasgow): Road network
LFS (Glasgow):
Hybrid contiguity/connectivity matrix
LFS (Glasgow):
Stints

N = 26
Target workload = 17006
Hybrid mode
Network compactness
LFS (Glasgow): Half stints

N = 52
Target workload = 8503
Hybrid mode
Network compactness
LFS (Glasgow): Stints
LFS (Glasgow):
Interviewer Areas

N = 2
Target (dummy var) = 13
Contiguity mode
Shape = Perimeter²/Area
LFS – next steps

• Evaluate outputs from pilot areas (with LFS/field operations)
• If suitable, roll-out to entire country
• Integrate into new waves of LFS
Conclusions, observations …

- Automated zone design methods can be employed to design collection geographies (workload areas in this context)

- Enhanced methods/ functionality:
  - Connectivity
  - Hybrid contiguity/ connectivity
  - Network compactness (travel-time)
  - [Accessibility]

- NB Not attempting to prescribe best routes or deploy actual resources (“travelling salesman problem”): network routing algorithms still better for this; rather, aiding area-based management of resources

Could methods also be used to design for statistical quality?
Acknowledgements

• Funding
  – University of Southampton Faculty Enterprise Fund grant (2013-15)
  – ESRC Impact Acceleration Account grant (2016-17)

• Teams
  – ONS Geography (Andy Tait, Donna Viney)
  – ONS Methodology (Adriana Castaldo, Owen Abbott, Marie Cruddas)
  – Labour Force Survey (Dean Fletcher)

• Data
  – ITN: OS MasterMap® Integrated Transport Network Layer [GML2 geospatial data], Scale 1:1250,
    Tiles: GB, Updated: 1 April 2016, Ordnance Survey (GB), Using: EDINA Digimap Ordnance Survey
  – Postcodes: Contains Ordnance Survey data © Crown copyright and database right 2011;
    Contains National Statistics data © Crown copyright and database right 2011;
    Contains Royal Mail data © Royal Mail copyright and database right [2011]
Contacts

Email: s.cockings@soton.ac.uk
Twitter: @SamCGeog

AZTool
Email: aztool@geodata.soton.ac.uk
Web: https://www.geodata.soton.ac.uk/software/AZTool/