MSMCity Report

Introduction

This report outlines a solution to extrapolate crowdsourced data from the Catch! mobile application.

A microsimulation model has been built for the case study area of Newcastle-upon-Tyne. The model represents the entire population of Newcastle at an individual level. Each individual has a set of associated attributes (age, sex, economic status). Travel characteristics have been attributed to individuals in the synthetic population by linking them to individuals in the app user database.

The process for building the microsimulation model has been automated in R so that the outputs are reproducible for other Local Authority areas.

This work will be taken forward by Andrew Smith as of February 2017.

Data

The following Census tables were extracted from [Nomis](https://www.nomisweb.co.uk/census/2011/data_finder) for the Local Authority area of Newcastle-upon-Tyne.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Table code | Table name | Geographic level | Variables | Format |
| DC1117EW | Sex by single year of age | MSOA | Age: select all categories  Sex: Males and Females | Database (.tsv)  Do not include area codes |
| DC6107EW | Economic activity by sex by age | MSOA | Age: select all categories  Economic Activity: select all categories  Sex: Males and Females | Database (.tsv)  Do not include area codes |
| WF01BEW | Location of usual residence and place of work | MSOA | Place Of Work:  MSOA = All  At home, offshore, not fixed or outside UK = All  Countries = All  Usual Residence:  MSOA = Some (select Local Authority area) | Database (.tsv)  Do not include area codes |

Use the table finder to select the appropriate table and click on the ‘Query data’ button on the left hand side of the page to choose the required geography and variables.

To extract data for MSOAs within a particular Local Authority area:

* Under ‘Make selections’ click ‘Geography’
* Select ‘Some’ from the drop down menu next to ‘2011 super output areas – mid layer’
* Select the Local Authority area from the drop down menu that appears
* Click ‘tick all’

Mobile app data was supplied by TravelAi. Two data sets were provided: a database of GPS tracks and a database of app users. Extensive work was carried out by Nikée Groot to clean the GPS data. Home and work locations of app users were inferred from the GPS data. The app user database contained a unique identifier for each device and a set of user attributes including gender, year of birth and weight.

The 2011 Output Area Classification (OAC) was sourced from the [Office for National Statistics](http://webarchive.nationalarchives.gov.uk/20160105160709/http:/www.ons.gov.uk/ons/guide-method/geography/products/area-classifications/ns-area-classifications/ns-2011-area-classifications/datasets/index.html). This classifies each OA in England by socio-economic group.

An OAC to MSOA lookup was required which was sourced from the [Office for National Statistics](http://webarchive.nationalarchives.gov.uk/20160105160709/http:/www.ons.gov.uk/ons/guide-method/geography/products/census/lookup/2011/index.html) under ‘Exact-fit 2011 OA lookup files’ and ‘2011 OAs to 2011 lower layer super output areas (LSOAs), middle layer super output areas (MSOA) and LADs’.

MSOA boundary data in shape file format was sourced from the [UK Data Service](https://census.ukdataservice.ac.uk/get-data/boundary-data) using the Boundary Data Selector tool. The selected target geography was English Middle Layer Super Output Areas, 2011.

Methodology

The microsimulation approach involved the creation of a synthetic population of individuals for the chosen Local Authority area. Each individual was assigned a set of attributes including gender, age, economic status and workplace. Individuals in the synthetic population were then linked to app users to attribute travel characteristics.

The following section outlines what each R script does (NB. the scripts should be run in this order).

Population.R

* Reads in population by age and gender file (DC1117EW Census table. NB. File name will need to be updated in script)
* Removes records containing totals
* Creates a new column containing the numerical age
* Creates a new column containing an age band code (0=0-16, 1=16-24, 2=25-34, 3=35-44, 4=45-54, 5=55-64, 6=65+). The breaks for the age bands can be adjusted if necessary.
* Creates a new column containing gender code (1=male, 0=female)
* Summarises the data by MSOA, gender and age band (sum\_people table)
* Calculates cumulative frequency based on the number people in each age and gender band

EconomicActivity.R

* Reads in population by gender, age and economic status file (DC6107EW Census table. NB. File name will need to be updated in script)
* Removes records containing totals
* Creates an additional column containing an age band code (1=16-24, 2=25-34, 3=35-44, 4=45-54, 5=55-64, 6=65+). The age bands correspond with those in the Population script.
* Creates a new column containing gender code (1=male, 0=female)
* Creates a new column containing economic status as a numeric code
* Summaries the data by MSOA, gender, age band and economic status (sum\_econ\_act table)
* Crosstabs MSOA, gender and age by economic status (econ\_act\_crosstab table)
* Calculates percentage of people economically active by MSOA, gender and age band
* Calculates percentage of people in employment by MSOA, gender and age band

Workplace.R

* Reads in origin-destination commuter flows (WF01BEW Census table. NB. File name will need to be updated in script)
* Sums the number of people residing in each origin MSOA and adds the totals as a new column (originSums column in workplaces table)
* Calculates the percentage of people in each origin (residential MSOA) that commute to each destination (workplace MSOA)
* Calculates the cumulative frequency for each origin (residential MSOA)

OAC.R

* Reads in OAC csv file
* Filters data to relevant Local Authority area
* Reads in OA to MSOA lookup file
* Appends MSOA from lookup file to OAC data (joins on OA code)
* Sums the number of OAs within each MSOA and adds the totals as a new column (oacGroupSums column in oacGroup table)
* Calculates the percentage of OAs within each MSOA that are in each OAC group
* Calculates the cumulative frequency for each MSOA

AppUsers.R

* Reads in app data
* Reads in app users data
* Merges files based on device ID
* Creates a new column containing a gender code (1=male, 0=female)
* Calculates age based on year of birth and creates a new column containing an age band code (which corresponds with the age bands in the Population script)
* Reads in app users inferred home and work location (supplied by Nikée Groot)
* Reads in MSOA boundary shape file
* Uses point-in-polygon analysis to append the MSOA of the app users home and work locations (appUsers table)

MicroSim.R

* Creates an empty data frame
* Loops through the aggregated population table (sum\_people) and creates a record for each individual with the corresponding attributes (individuals2 table)
* Assigns an economically active flag (1=economically active, 0=economically inactive) to each individual by comparing a random number to the economically active percentage for the corresponding MSOA, gender and age band. Assigns zero to people aged under 16.
* Assigns an in employment flag (1=employed, 0=unemployed) to each individual by comparing a random number to the employed percentage for the corresponding MSOA, gender and age band. Assigns zero to people aged under 16.
* Assigns a workplace to each individual by comparing a random number to the cumulative percentage for the corresponding MSOA. People aged under 16 not assigned a workplace.
* Assigns an area type (OAC Group) to each individuals home location.
* Assigns an individual in the app user’s database to each individual in the synthetic population. Creates a table of app users in the corresponding gender and age categories as the individual in the synthetic population. Selects a random app user from within the table of possible matches and returns the agentID of the app user.
* Writes out the results to a .csv file (output.csv)
* NB. Due to the low number of app users in Newcastle the process currently links any individual in the app user database to the synthetic population of Newcastle. As the number of app users increases this could be adjusted to link individuals based on home and/or work location as well as gender and age band. As the data analysis progresses additional travel characteristics could be appended to the synthetic population.

Outputs

The microsimulation model contains the following variables.

|  |  |  |
| --- | --- | --- |
| Variable | Type | Description |
| ID | Integer | Unique identifier |
| MSOA | Character | Residential MSOA |
| num\_sex | Numeric | Sex (1=male, 0=female) |
| age\_band | Numeric | Age band (0=0-16, 1=16-24, 2=25-34, 3=35-44, 4=45-54, 5=55-64, 6=65+) |
| econ\_active | Numeric | Flag (1=economically active, 0= economically inactive) |
| in\_employment | Numeric | Flag (1=in employment, 0=unemployed) |
| work\_MSOA | Character | Workplace MSOA |
| area\_type | Character | OAC Group |
| agent | Integer | App user agent ID |

Next Steps

This report has outlined a methodology for generating a synthetic population of individuals from Census tables for any Local Authority area.

It has also outlined a process for linking a synthetic population of individuals to crowd-sourced mobile app data.

Andrew Smith will be building on the work outlined in this report to generate a synthetic population that can be used as the basis for an agent-based model.