**1) Preparing the Census datasets**

**1.1 Synthetic population of commuters, including LSOA-level OD pairs [see Appendix Figure 1]**

1. Age\*sex\*mode\*LSOA origin-destination (OD) pair is known for all commuters. Add home and work MSOA to this. Convert this OD-pair-level dataset to an individual-level dataset.
2. For each MSOA-level OD pair, some data is available on car ownership[[1]](#footnote-1):
   1. Selected MSOA OD pairs: number of car owners by mode (only for OD pairs containing 10+ commuters).
   2. All MSOA OD pairs: number of car owners in total.

Use this data to probabilistically assign car ownership to individual commuters, such that the total number of people owning no car in each MSOA OD pair \* mode combination is correct. For OD pairs where the number of car owners by mode is not available, probabilistically assign car ownership such that the total number of people owning no car in each OD pair is correct.

The probabilistic assignment is done as a function of home region, age, sex, and mode. The probabilities used were calculated by combining the two 5% individual-level samples from the Census, and running logistic regression models predicting “owning no car” as a function of sex, age, and mode, stratified by region.[[2]](#footnote-2)

1. For each MSOA-level OD pair, some data is available on ethnicity:
   1. Selected MSOA OD pairs: number of non-white individuals by mode (only for OD pairs containing 5+ white commuters and 5+ non-white commuters).
   2. All MSOA OD pairs: number of non-white individuals in total.

Use this data to probabilistically assign ethnicity to individual commuters, such that the total number of non-white individuals in each MSOA OD pair \* mode combination is correct. For OD pairs where ethnicity by mode is not available, probabilistically assign ethnicity such that the total number of non-white individuals in each OD pair is correct.

The probabilistic assignment is done as a function of home region, age, sex, mode, and car ownership. The probabilities used were calculated by combining the two 5% individual-level samples that have been made available from the Census. We pooled these two datasets together to increase power, although note that because the samples are overlapping they will double count 1 in 40 of the individuals included. In this combined dataset, we ran logistic regression models predicting “being non-white as a function of sex, age, mode and car ownership, stratified by region.

1. Merge in distance between home and work, defined in two ways:
   1. Euclidean distance.
   2. Fastest-route cycling distance, for OD pairs with distance <20km. *[Ideally one would re-route all OD pairs to estimate distance based on the actual mode used - but I don’t think we will have capacity for this. I think we can live with it as a limitation – this variable is only currently used when trying to match commuters by the distance they travel to work, which in any case is done in fairly crude bands.]*

**1.2 Synthetic population of non-commuters, including home LSOA [see Appendix Figure 2]**

1. Age\*sex\*economic activity\*home LSOA is known for all adults.[[3]](#footnote-3) Exclude commuters, leaving those who are unemployed or economically inactive. Add home MSOA to this. Convert this LSOA-level dataset to an individual-level dataset.
2. For each home LSOA, data is available on car ownership\*economic activity. Use this data to probabilistically assign car ownership to individuals, such that the total number of people owning no car in each LSOA\*economic activity combination is correct. The probabilistic assignment is done as a function of home region, age, sex, and economic activity (unemployed versus economically inactive). The probabilities used were calculated by combining the two 5% individual-level samples from the Census, and running logistic regression models predicting “owning no car” as a function of sex, age, and economic activity, stratified by region.
3. For each home LSOA, data is available on ethnicity\*economic activity\*age (<50 vs 50+). Use this data to probabilistically assign ethnicity to individuals, such that the total number of non-white individuals in each LSOA\*economic activity\*age combination is correct. The probabilistic assignment is done as a function of home region, age, sex, economic activity, and car ownership. The probabilities used were calculated by combining the two 5% individual-level samples from the Census, and running logistic regression models predicting “being non-white” as a function of sex, age, economic activity, and car ownership, stratified by region.

**1.3 Merging to create synthetic population of all adults**

1. Append the synthetic populations of commuters and non-commuters, to generate a total population of 42,989,620 adults living in England.
2. Use home LSOA to merge in data on Index of Multiple Deprivation 2015 income decile, urban/rural status, and area sparsity. Also merge in the hilliness of the home LSOA (in twentieths); plus the proportion of people in each LSOA who travel to work, in the 2011 Census, by a) bicycle, b) bus, and c) tube/train (in twelfths).
3. The variables in this dataset are:
   1. For all individuals: home LSOA/MSOA/LA/region, sex, age, economic activity, car ownership, ethnicity, income deprivation, urban/rural status, area sparsity, hilliness of the home area, and the local prevalence of commuting by bicycle, bus and tube/train.
   2. For commuters: work LSOA/MSOA, commute mode, commute distance.

**1.4 adding in postcode**

* Noise is estimated in Metahit at the level of postcode. We therefore assigned a postcode to each individual in our synthetic population. This allows us to capture better the variation across an LSOA in noise levels, and also provides a means of routing between LSOAs with more granularity.
* In the 2011 Census, there were 1.3 million postcodes in England containing at least one usual resident in the Census, i.e. an average of 46 (range 5 to 291) for each of the 34,753 2011 LSOAs in England. For each individual in our synthetic population, we randomly assigned one of the postcodes belonging to their LSOA. This was done with the probability proportional to the total number of inhabitants in that postcode in the 2011 Census.
* Note that the boundary geography of postcodes does not align with the output area/lower super output area Census geography. Instead the National Statistics Postcode Directory assigns each postcode to an output area, and thereby to an LSOA, based on where the centroid of that postcode falls. (Postcode centroids are calculated fall within the building of the matched address that is closest to the mean location of all addresses in postcode.) Thus a few individuals in some postcodes will live outside the geographical boundaries of the LSOA to which they are assigned.

**2) Preparing the NTS datasets**

1. Restrict NTS to participants living in England aged 16+, in the years 2010-2016. Choose 2010 as the earliest date, as older data is more affected by secular trends since the

year 2000, e.g. falling trip rates and increasing levels of driving among older people.

1. Prepare the NTS dataset to be a trip-level dataset, expanding up short walks on the final day. Among commuters, calculate the median commute distance for all past-week commute trips.
2. Have DfT merge in anonymized data on the proportion of people in each LSOA who travel to work, in the 2011 Census, by a) bicycle, b) bus, and c) tube/train. This data is defined as fourths.
3. We may want to redo the creation of the synthetic population, adding in the NTS data for 2017, when that is released in autumn 2018.

**3) Preparing the ALS datasets**

1. Use all 4 available ALS surveys, covering November 2015 – November 2018. Res trict ALS to participants aged 16+. ALS is already restricted to participants living in England.
2. Process questions on past-week walking and cycling time. In addition, estimate time spent in other nonoccupational physical activity by taking the total weekly duration of nonoccupational physical activity reported and subtracting walking and cycling from this.

**4) Merging the Census, NTS and ALS datasets at the individual level; plus generating frequency weights for Census**

**4.1 Merging NTS with the Census**

1. Seek to match each individual in the Census to a randomly-selected individual in NTS who shares a defined set of matching criteria. Steadily relax the set of criteria in successive matches, such that eventually all individuals in the Census find an NTS match. Only allow matches where the number of NTS individuals for the category in question is at least 20, to avoid having results for subsets skewed by continually resampling from the same small number of NTS participants.
2. The matches used are outlined in Table 2.

Table 1: Matching criteria used when matching Census with NTS

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Match 1 | Match 2 | Match 3 | Match 4 |
| Home region | 9 categories | 9 categories | - | - |
| Urban/rural status | Binary | Binary | Binary | - |
| Sex | Binary | Binary | Binary | - |
| Age category | 6 categories | 6 categories | 6 categories | 5 categories |
| Commuters/non-commuter | Binary | Binary | Binary | Binary |
| Car ownership (any/none) | Binary | Binary | Binary | - |
| Ethnicity (white/non-white) | Binary | Binary | Binary | - |
| Commute main mode among commuters | 9 categories | 5 categories | 5 categories | 5 categories |
| Commute distance among commuters | 11 categories | 6 categories | - | - |
| LSOA % cycle commute among non-commuters | 3 categories | 3 categories | 3 categories | Binary |
| LSOA % bus commute among non-commuters | 3 categories | 3 categories | 3 categories | Binary |
| LSOA % train commute among non-commuters | 3 categories | 3 categories | 3 categories | Binary |
|  |  |  |  |  |
| % Census individuals who can be matched | 27.6% | 39.9% | 87.4% | 100.0% |
| % Census individuals using this match | 27.6% | 12.3% | 47.5% | 12.6% |

1. Variables that NTS adds on top of what is available in the Census are:
   1. Past-week trip level travel diary (see Section 5).
   2. Potentially also some individual variables, e.g. number of children in the household.

**4.2 Merging ALS with the Census-NTS composite**

1. Seek to match each individual in the Census to a randomly-selected individual in ALS who shares a defined set of matching criteria. These matching criteria include both demographic variables from the Census and merged walking and cycling variables from NTS. As before, steadily relax the set of criteria in successive matches, such that eventually all individuals in the Census find an ALS match. Again, only allow matches where the number of ALS individuals for the category in question is at least 20.
2. The matches used are outlined in Table 2.

Table 2: Matching criteria used when matching the Census-NTS composite with ALS

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Match 1 | Match 2 | Match 3 | Match 4 |
| Home local authority | 326 categories | 326 categories | - | - |
| Home region | 9 categories | 9 categories | 9 categories | 9 categories |
| Sex | Binary | Binary | Binary | Binary |
| Age category | 6 categories | 4 categories | 4 categories | 4 categories |
| Ethnicity (white/non-white) | Binary | Binary | Binary | - |
| Usually cycle weekly (yes/no)[[4]](#footnote-4) | Binary | Binary | Binary | Binary |
| Past-week time spent walking for transport[[5]](#footnote-5) | 4 categories | 4 categories | 3 categories | 3 categories |
|  |  |  |  |  |
| % Census individuals who can be matched | 67.8% | 78.5% | 99.9% | 100% |
| % Census individuals using this match | 67.8% | 10.6% | 21.4% | 0.1% |

1. The key variable that ALS adds on top of what is available in the Census is average weekly minutes in sports/recreational activities (excluding walking and cycling).
2. We also considered adding in past-week duration of cycling for recreation and duration walking for recreation. NTS technically only includes walking and cycling for transport, plus some walking for recreation. But in practice we believe there is considerable double counting/overestimation from using these variables. Moreover, even without adding in this additional physical activity, the distribution of physical activity in our synthetic population was slightly higher than that recorded in 6 large representative studies from European countries that contributed to our meta-analysis examining the health impacts of physical activity. As such, assigning any additional physical activity to our synthetic population would have made the results of the meta-analysis less applicable. In addition, note that our use of RTS scaling rates should mean that we are capturing most cycling (including recreational), except what is happening fully off road.

**4.3 Defining frequency weights to update the size and structure of the Census population**

1. Generate frequency weights to adjust for changes in the size and demographic structure of the English population since 2011. Generate these frequency weights as a function of home local authority\*age\*sex, using mid-year ONS population estimates from 2018. For example, in 2011 the estimated number of women age 16-24 living in Cambridge was 12365, whereas in 2018 this had grown to 13662. We therefore generated a frequency weight of 13662/12365=1.10 to women age 16-24 living in Cambridge in our Census population.

**4.4 Defining scaling weights to scale up NTS trip volumes to match RTS**

1. In the synthetic population (after applying the population scaling weights described in section 4.3), we calculated the total volume of travel (annual millions of kilometres) by bicycle; car or taxi; motorcycle; and van/lorry. We only counted car, motorcycle and van/lorry trips where the person was the driver.[[6]](#footnote-6)
2. We compared this with the total volume of travel estimated by DFT in their 2015 Road Travel Statistics (2015 is the most recent year with data available to us for all modes). These Road Travel Statistics are calculated through a rolling series of 12-hour manual counts of vehicles, supplemented by 10,000 automatic traffic counters. Data are collected on all motorways, all A roads and a representative sample of minor roads.
3. For buses, we used the DfT 2015/16 estimates of total passenger distance travelled on local buses. This data comes from an annual survey of over 500 bus operators.
4. Our synthetic population only includes individuals aged 16+. For cycling and bus travel (the only modes where a child could count, since children cannot be drivers), we calculated the proportion of all distance accounted for by adults age 16+. This was 92% nationally for cycling (range 89%-94% across GOR) and 83% for bus travel (range 80%-87% across GOR). We factored this into our comparisons, i.e. first scaling down the relevant RTS and bus passenger data by the proportion of travel done by adults, and then making the comparison with our synthetic population.
5. NTS does not distinguish between van and lorry travel. We assume that light vans are the mode used for all ‘household vehicle’ van/lorry travel in NTS (household vehicle travel accounts for 78% of all van/lorry trips). For the purposes of our modelling, we further assume that all ‘non-household vehicle’ van/lorry trips are also by van – i.e. that lorry travel is not captured in NTS. As a sensitivity analysis, we instead assumed that amongst ‘non-household vehicle’ van/lorry trips, A) a random 50% and B) 100% are done by lorries.
6. Nationally, the volume of travel estimated by RTS was higher than in NTS as follows:
   * 1.12 times higher for cycling
   * 1.23 times higher for car driving & taxis
   * 1.82 times higher for motorcycle driving
   * 1.02 times higher for local bus travel
   * 7.12 times higher for light van driving and infinitely higher for lorry driving. Alternatively, 8 times higher for vans and 24 times higher for lorries in sensitivity analysis A in which 50% of non-household vehicle van/lorry travel was by lorries; and 9 times higher for vans and 12 times higher for lorries in sensitivity analysis B in which 50% of non-household vehicle van/lorry travel was by lorries.
7. It is likely that these discrepancies reflect a mixture of some trips not being reported in NTS plus some types of people choosing not to complete the travel diary – e.g. couriers or delivery drivers who make so many trips that they would not want to complete travel survey. We therefore assume that RTS and the DfT bus passenger data are the more reliable source. Moreover, regardless of whether RTS is the more reliable source, it is the source that feeds into the NAEI model that underpins our air pollution and noise data – so is the better source to use in terms of creating comparability across the model.
8. In applying these scaling weights, we assume that:

* Trips are missing at random, i.e. the existing pool of trips can be used to approximate the missing trips. This is unlikely to address the problem of couriers/delivery drivers being disproportionately missing, but we do not see how we can reasonably add this class of person.
* Weights applied to travel as a driver could also be applied to travel as a passenger by the same mode.
* The weights for local bus travel can also be applied to long distance express buses and other public transport (in practice, these other modes are not a major focus of metahit, reducing the impact of this assumption).
* Walking is captured correctly in NTS + ALS, i.e. we do no scaling for walking. This is inconsistent in terms of the relative amount of travel reported in NTS, but means that the NTS walking estimates match better onto the physical activity dose response data used elsewhere in metahit.
* Lorry travel cannot be reliably captured in NTS; this is true even in the sensitivity analysis in which we assume that some or all of non-household van/lorry travel is by lorries. Therefore, lorry travel will need to be added separately, rather than via awaiting approach. It is arguable whether the same should also be true for van travel, but since we think the current method is acceptable since a) van travel is more reliably identifiable in NTS (via ‘household’ vehicles) and b) the underreporting of an travel is not as marked as for lorry travel.

1. Based on all these assumptions, we apply scaling weights to the existing NTS ‘trip weight (excluding household weight)’ variable. This scales up all NTS trips, such that the total volume of travel generated is correct. The trip weights are calculated as described above, except with stratification by city region to give the values shown in the table below. For buses, it was not possible to distinguish across regions. For cycling we also used the national scalar as we were less confident for cycling in the accuracy of RTS data at the city region level. Using the national scalar effectively gives more weight to 2011 Census data on proportion of adults cycling to work, as this is one of the matching criteria used to create the synthetic population.
2. The scaling weights below are used in our scenarios. When creating denominator data for analysing injury risks in stats19 data from 2005-2017, we adjust these scaling weights to map onto the relevant year between 2005-2017 *[this was planned, may or may not have been implemented by Robin the injuries analysis]*. For bicycles and motorcycles, RTS data was only available up to 2015, so we use the 2015 weight for 2016 and 2017 analyses.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| City region | Cycle weight | Car and taxi weight | Motorcycle weight | Van weight | Bus + other public transport weight |
| Bristol | 1.08 | 1.17 | 1.81 | 5.36 | 1.04 |
| Greater Manchester | 1.08 | 1.11 | 1.35 | 7.14 | 1.04 |
| Leeds | 1.08 | 1.06 | 1.47 | 6.81 | 1.04 |
| Liverpool | 1.08 | 1.26 | 1.67 | 8.04 | 1.04 |
| London | 1.08 | 0.94 | 1.72 | 5.89 | 1.04 |
| North East | 1.08 | 1.09 | 1.11 | 6.21 | 1.04 |
| Nottingham | 1.08 | 1.13 | 2.04 | 6.58 | 1.04 |
| Sheffield | 1.08 | 1.23 | 2.06 | 6.58 | 1.04 |
| West Midlands | 1.08 | 1.17 | 1.57 | 8.09 | 1.04 |

**4.5 estimating background air pollution exposure**

* In the synthetic population, I have added measures of background air pollution:
  1. in the home LSOA (everyone)
  2. in the work LSOA (commuters with an identified workplace)
  3. in the home local authority (everyone)
* Each of these three measures exist for both background PM2.5 and background NO2, using the 2016 values from the UKIAM model.
* We created an individual-level estimate of baseline air pollution exposure, assuming that people spend:
  + 12 hours \* 7 days at home.
  + those with a workplace spend 8 hours \* 5 days at the workplace.
  + and the rest of the non-travelling time is spent in the local authority.

**5) Merging in trip data**

1. Stages 1-4 have created an individual-level dataset containing the 42,989,620 adults in England in the 2011 Census. These individuals have been matched to individuals in NTS, and so can also be merged with a trip-level NTS dataset, to create a trip-level dataset for England.
2. This creates issues around dataset size: the average number of past-trips per individual in NTS is around 17, meaning that the dataset of 43 million individuals would become over 700 million trips. I have created trip-level datasets at the local authority level. These could readily be aggregated to regions or to the national level depending on whether it is more computer-efficient to run the same analysis 326 times in series or once in a very large dataset.
3. Another issue that I have currently left unresolved in the data is what to do when the home LSOA and main workplace LSOA, given in the Census, are not consistent with the travel diary data from NTS. This is bound to be the case for at least some trips. For example, perhaps the home LSOA and work LSOA are 4.2 km apart Euclidean distance, but the travel diary reports a commute trip of distance 3 km. This could happen a) because we are matching imperfectly on commute distance, as it is defined in categories for the first two matches and not used at all for the final two matches; or b) because people make their commute trips to different workplaces, or from different homes, on different days. Options could include:
   1. Do we believe the Census, thereby imputing the origin and destination of all home-work commutes? In that case, do we overwrite the distance and duration data provided in the NTS with alternative plausible values when the two are incompatible?
   2. Do we believe NTS, thereby losing the ability to routes trips along the road, since we’re back to not knowing where they are going. NB in this case we would still be able to consider the home location as known, which would be helpful in terms of e.g. exposure to air pollution /noise. Also, we could use the Census for a sensitivity analysis of “how much difference does it make if we know the routes”.
   3. Do we do some sort of hybrid, e.g. we simultaneously have NTS distance/duration values paired with Census origins and destinations, even though the two are not fully compatible? The problem with this is then our results are not internally consistent, e.g. a scenario might see a 10km/person decrease in car use based on NTS, but an 8km/person decrease based on Census. But possibly in such a case we could do some sort of scaling, e.g. we believe the NTS value of 10 km/person is correct, and so we take the spatial distribution of effects estimated by the Census and scale that up by 10/8. This appeals to me to some extent, but is potentially fiddly and hard to explain.

In practice we are using option b in Metahit, but could consider improving in future work.

**Appendix**

Figure : Schematic representation of the process used to create the synthetic population of commuters in the Census

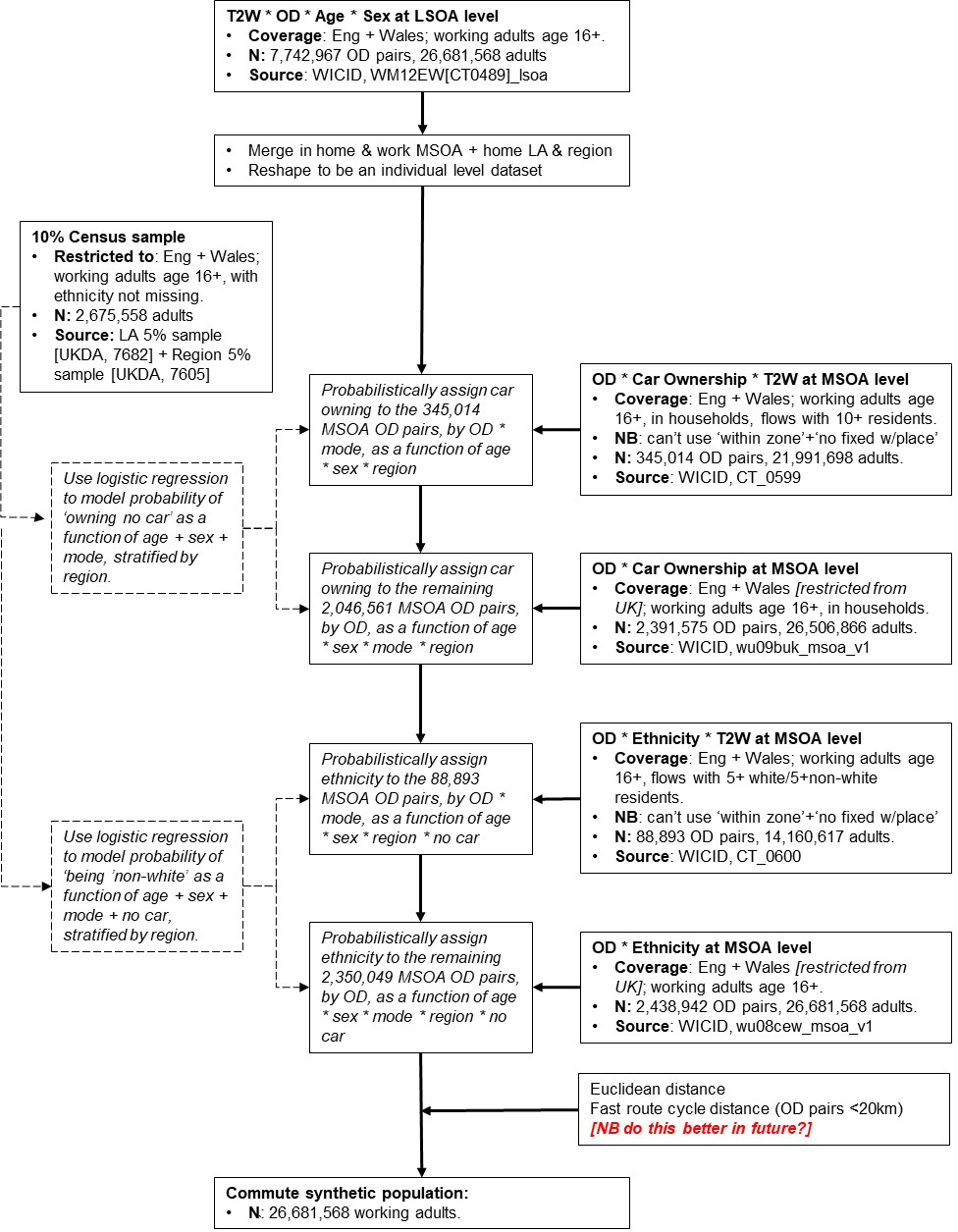
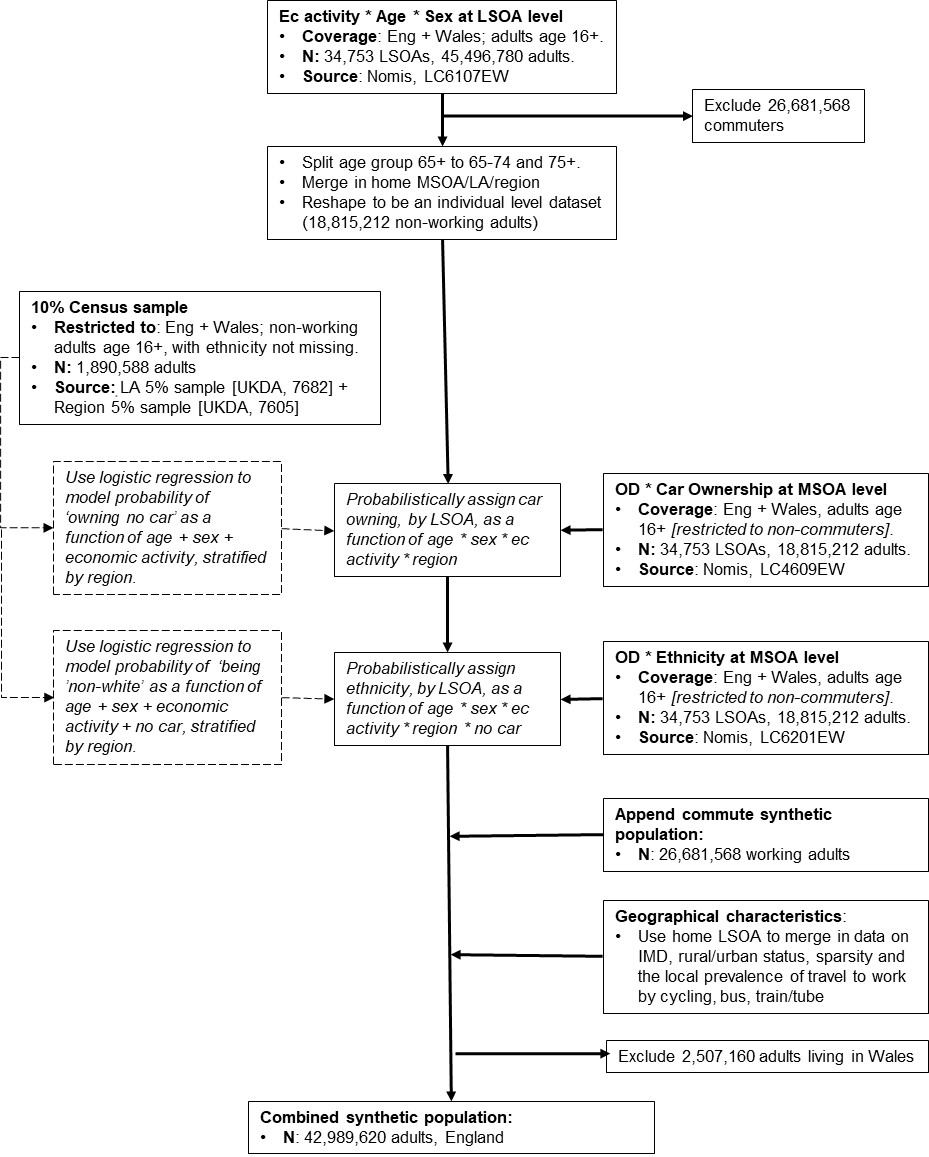


Figure 2: Schematic representation of the process used to create the synthetic population of non-commuters in the Census, and then merge in the commuters



1. Around 1% of commuters do not live in households so were not eligible to be asked this question – here and below, they are treated as having no car. [↑](#footnote-ref-1)
2. For example, among commuters living in the North East, the individual-level Census samples estimate that the probability of owning no car is 26% for female bicycle commuters age 35-49, and 52% for male bus commuters age 25-34. Thus, in any MSOA OD pair in the North East, it would be twice as likely for a male bus commuter age 25-34 to be assigned the status of not owning a car than for a female bicycle commuter age 35-49. [↑](#footnote-ref-2)
3. The highest age group is 65+, but this can be divided into 65-74 and 75+ by triangulating with a more detailed age\*sex\*home LSOA breakdown and then subtracting the number of commuters, to give the number of non-commuters in each age group. The small number of unemployed people age 65+ are all assumed to be 65-74. [↑](#footnote-ref-3)
4. Calculated in NTS as the proportion of people who say they usually cycle at least weekly, based on survey questions about usual travel patterns. Calculated in ALS as the proportion of people who report cycling on at least four days in the past 28 days. [↑](#footnote-ref-4)
5. Calculated on NTS as the total duration of past-week walking time in the one-week travel diary, using stage-level data. This duration is divided into four categories: no walking (69% of NTS sample), <1 hour (9%), 1-2.49 hours (11%), 2.5+ hours (11%). Calculated in ALS as the average weekly duration of walking for transport. NB the average amount of walking for transport reported in ALS is higher than that reported in NTS – 1.6 hours/week vs 1.2 hours in NTS. This discrepancy may partly reflect the fact that short walking trips (less than 1 mile) are only collected in NTS on the final day of data collection, and so many participants may do some short walking during the week that is not captured. It also seems likely that ALS overestimates past-week walking (and physical activity in general). Instead of matching on absolute walking time, therefore, we divided the ALS data into four categories that approximate the frequency distribution observed in NTS. Specifically, define categories: no walking (54% of ALS sample), <2 hours (22%), 2-3.99 hours (11%), 4+ hours (13%). [↑](#footnote-ref-5)
6. For taxis the comparison between the datasets is not exact, but as taxis are much rarer than car driving (accounting for 0.7% of distance in NTS as opposed to 55% for car driving) this is not expected to have any great effect on the results. [↑](#footnote-ref-6)