Scottish Transport Applications Research

A spatial microsimulation approach to modelling capacity for active travel in Scotland

George Macklon, Matt Pearce – Sustrans Research and Monitoring Unit

Matthew Dallas, Ian Philips – University of Leeds

Table of contents

Summary

This paper sets out the steps we have taken in first simulating a spatial dataset and then developing a model to estimate the capacity of people to travel actively in different areas of Scotland. This model takes into account relevant characteristics of the population including age, gender and health.

Spatial microsimulation is a well-established technique that estimates what individuals are like in an area, based on aggregate statistics. Our approach is based on similar recent work conducted on English data. In Scotland our source of health data was the 2016 Scottish Health Survey and our spatial data source was the 2011 census which provides data at output level (geographical areas of around 125 households).

We identified attributes that are required to model an individual's capacity to travel actively and which are collected for the Scottish Health Survey, such as age, gender, whether the individual had a long-term limiting illness and whether he or she was economically active. The spatial microsimulation approach then allows us to combine these attributes with the spatial data to estimate the number of individuals in each category in each output area.

Our approach also incorporates data from the 2016 Scottish Household Survey to help simulate which individuals in our local populations have access to a bicycle – a critical factor in their ability to travel actively

Using the simulated population, our model calculates a maximum cycling and walking distance for each individual based on a number of factors: an estimate of their VO2max (the maximum amount of oxygen that an individual can make use of during exercise) which, when combined with information on weight and BMI produces an estimate of an individual's power output.

One of the other factors affecting the distance an individual can walk or cycle is the topography of the area in which they live. Our model takes into account the average slope of within five kilometres of the centre of each output area when calculating the cycling and walking speed for each individual.

We believe our model has produced a rich dataset with a variety of uses. For example, it provides a clear and easily communicable presentation of locations where individuals have relatively limited capacity to travel by active modes, both within and between areas. For example the mean distance that people have the capacity to cycle ranges from 10. 7 km in Inverclyde to 14.5 km in Angus.

Our modelling has also demonstrated that the common transport planning assumption that people can cycle 5 miles (8km) overestimates the capacity of the population to commute by active modes. We found that 21% of the individuals in our model did not have the capacity to cycle this far, while an even higher proportion had the capacity to cycle at least this far but did not own a bicycle. This highlights the risk of excluding substantial proportions of the population when relying on common assumptions.

There are a number of areas in which the methodology outlined in this paper could be improved and we intend to take these forward in the coming months. Nevertheless we are sharing our methodology at this stage to demonstrate the power of the approach and in the hope of inspiring further joint development in using it to model active travel behaviours in Scotland.

1 Introduction

This paper sets out the development of a spatial microsimulation approach to modelling the extent to which the Scottish population has the physical capacity to travel actively. This is quantified in the model by estimating the distance by which each member of the population can travel by bicycle or on foot.

Spatial microsimulation is a well-established method for simulating data on individuals in a population, including a spatial element (ie where they live or work) (e.g. O'Donoghue et al 2014, Tanton 2014). It is also known as population synthesis amongst transport modellers. This model is based on the work of Philips et al. (2018, 2017), which developed a similar approach for the population of England.

Because the simulation occurs at the level of individuals it is possible to describe variation in attributes within areas (ie the differences between individuals in a community) as well as between areas. This is useful because although we might say "an average person could cycle 5 miles", by assuming everybody is the same we risk excluding some people who cannot cycle that far, as well as underestimating the ability of some to travel longer distances. This might have policy implications for planners interested in transport equity or those seeking to maximise the energy and pollution reductions that could be achieved by active travel.

Policy makers can also use simulated datasets like this to ask "what if" questions as a way of understanding the effects of different policy options. For example asking – "what if everyone was given access to a bicycle?" – could be simulated by changing the bicycle ownership attribute in the model. This would provide an estimate of the change in capacity to travel actively if, say, an affordable bike share scheme were provided in a particular location. Insights such as this can be used to produce sophisticated models of behaviour and contribute towards the development of more impactful policy.

Understanding the capability to travel actively and how it varies within and between Scottish communities has important implications for policy and planning.

The paper is structured as follows:

- Section 2 explains the simulation process, including the selection of variables used
- Section 3 details the validation of the simulated population
- Section 4 explains the modelling of individuals' capacity to walk and cycle and presents the results
- Section 5 discusses the results and presents some maps of the data to highlight different ways the information can be communicated.

2 Spatial microsimulation

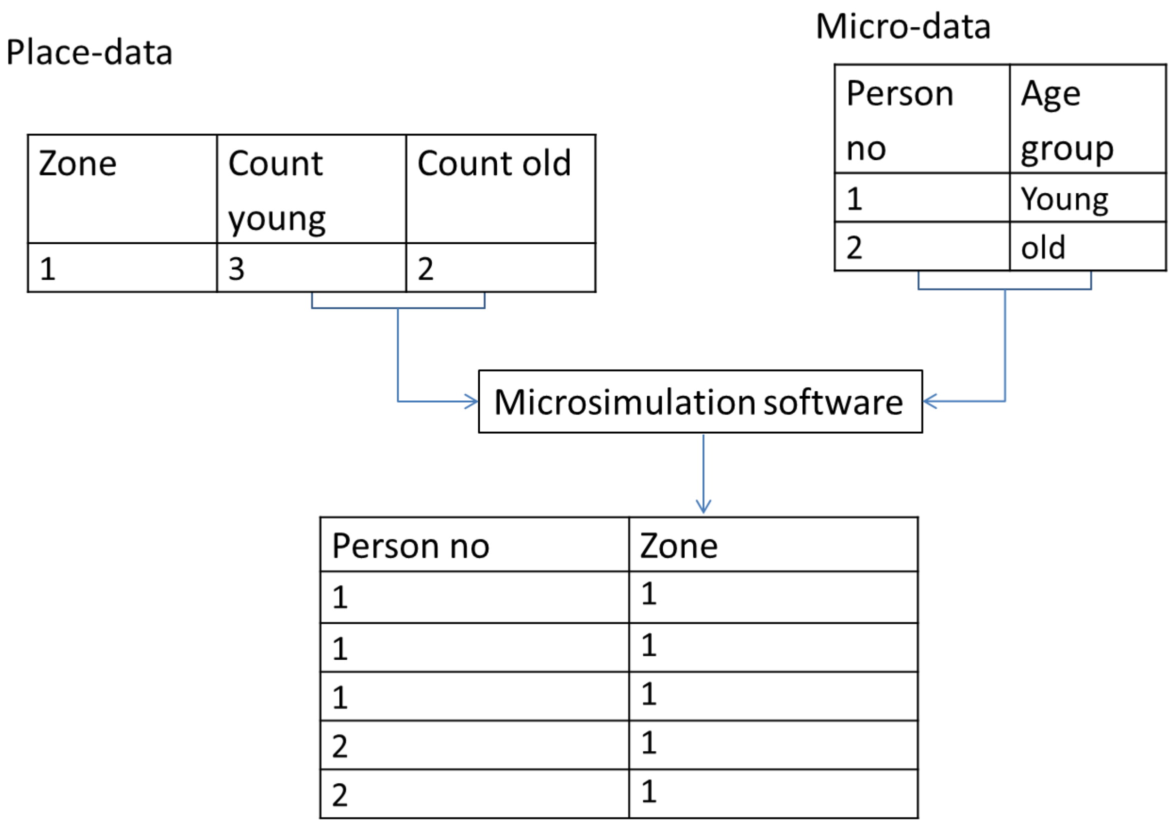
Spatial microsimulation does not require detailed confidential data about every member of the population; census data (or other ‘place-data’) and an anonymised survey of a sample of the population (the micro-data) are sufficient. This means that the approach is much more data efficient (not to say more feasible) than having to gather data about every member of the population.

Our micro-data source is the 2016 Scottish Health Survey and our place data source is the 2011 Census at Output Area (OA) level (the highest resolution available). An output area is a geographical unit that includes c.125 households. We can use the place-data to identify, for instance, the number of females of a specific age in an output area.

To link these datasets some attributes in the micro-data must also be found in the place-data. For example, the 2011 UK census contains age and gender as does the Scottish Health Survey. We call these attributes “constraints”.

Although there are many different microsimulation softwares and algorithms (a discussion of which is beyond the scope of this paper but see Tanton 2014 for review) the basic process of simulating the constrained attributes is shown in Figure 1. The constraint values allow individuals to be simulated and replicated to build a population for each area.

Figure 1 Basic simulation of constrained attributes



The model developed by Philips et al (2018) for estimating the capacity of individuals to walk and cycle requires information on various attributes of those individuals, primarily concerning their physical health. This type of data is not collected for the whole population in the census, but it is collected by the Scottish Health Survey (SHeS).

However, because health attributes have been shown in many studies[1] to be correlated with socio-demographic attributes which are collected by the census, spatial microsimulation can use this relationship to simulate the attributes which appear in the micro-data, but not the place-data. We call these “unconstrained attributes”.[2]

To do this, at least some of the variables in common between the micro-data and the place-data (the constraint attributes) must also be correlated with the unconstrained attributes.

The following sub-sections consider which constraint attributes should be used in this simulation.

2.1 Unconstrained variables

Using Philips et al (2017) as a guide, the following unconstrained attributes that we want to simulate for the whole population are identified in the SHeS[3]:

- Physical activity levels[4]
- Body Mass Index (BMI)
- Weight
- Age
- Gender

These are the attributes that are required for modelling individuals’ capacity to travel actively. Only age and gender are also found in the place dataset.

2.2 Potential constraint attributes

Using Philips et al (2017) again, possible constraint attributes common to both the place-data and the micro-data were identified. These are the variables that are used to combine the two datasets in the simulation – at least some of which need to be correlated with the unconstrained attributes.

- Age
- Gender
- Highest educational qualification
- Level of deprivation of household location
- Longstanding, activity limiting illness
- Level of economic activity

In some cases, there were multiple variations of a potential constraint variables in the two datasets (eg different ways of presenting an individual’s level of economic activity). In these cases the closest possible match between the two datasets was used, accounting for the similarity of the variable categories and the questions used to collect the data.

2.3 Exploring the possible constraint attributes

The simulation process requires at least some of the constraint attributes to be correlated with the unconstrained attributes in the micro-data. Although Philips’ simulation of the English population revealed relationships between the identified attributes (Philips et al, 2017), these relationships need to be re-explored in the Scottish micro-data before they form part of our new simulation.

There are two important factors to consider when exploring the relationship between the variables in the micro-data:

[1] Such as McArdle (2010)
[2] Based on these simulated attributes, we also model further variables for each individual and this is explained in Section 4.
[3] Note, Philips et al (2017) included VO2max as an unconstrained variable, but this information is not included in the 2016 SHeS dataset. Another unconstrained variable, bicycle availability, is considered in a later section, using a different data source.
[4] Using the metric of ‘average mins doing moderate to vigorous physical activity per week 10+ min’, as used in the NHS’s guidance on recommended levels of physical activity. Variable code “mintot10T” in SHeS 2016 dataset.

- The constraint attributes in the micro-data need to be comparable with the place-data. For instance, if the micro-data holds age data in discrete one year bands but the place data only has age data in discrete 5 year bands, a common age-band needs to be implemented. In this example, the micro-data would be adjusted to discrete 5 year bands as this is the lowest common age band possible.
- The simulation process has a ‘sweet spot’ where the combination of the constraint attributes provides a sufficiently detailed categorisation of individuals’ characteristics, but not so many that there are lots of ‘empty cells’ in the micro-data, where there are no individual cases with the given set of constraint characteristics. For instance, although in the above example, 5 year age bands may be the lowest common age band, this may result in too many empty cells in the micro-data, where there are no individuals that fall into certain age bands once the other constraint variables are also considered. In this case it may be more appropriate to use wider bands to reduce the number of possible categories.

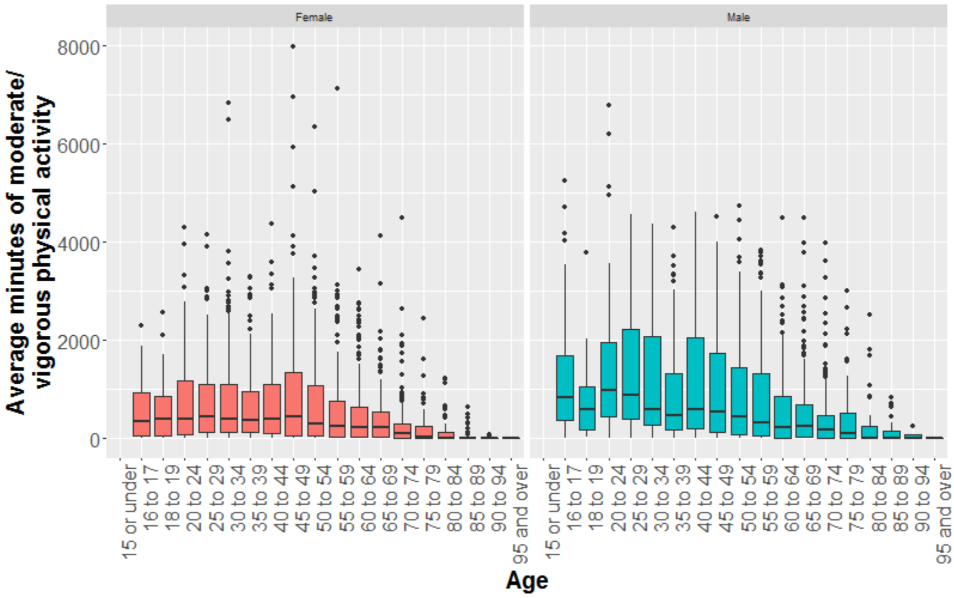
This sub-section of the paper explores the relationship between the possible constraint attributes and some of the unconstrained attributes in the micro-data (specifically BMI and the average minutes of moderate or vigorous activity per week). Common constraint categories between the micro-data and the population data are identified and the relationship between the constraint attribute and the unconstrained attributes is then explored. Unless otherwise specified, Kruskal-Wallis rank sum tests are used to quantify whether there is correlation between the attributes.

2.3.1 Age and gender

Census data for Scotland are available in five year age band (Census data table LC1117SC), with the exception of the age groups younger than 20 and older than 95. The SHeS provides age data as discrete single years. The SHeS data are therefore ‘binned’ into common five year age categories. The datasets share the same binary categories for the gender attribute.

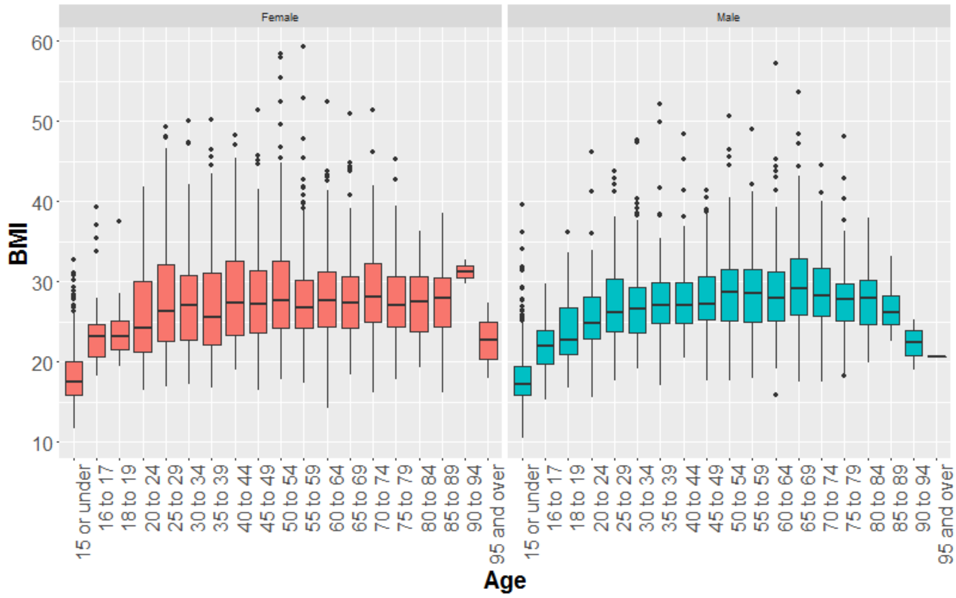
Figure 2 shows the relationship between age and gender and levels of physical activity. The data suggest that males typical undertake more physical activity than females, and that physical activity for both recorded genders levels decline from the age of 50. Kruskal-Wallis tests show that there are significant differences in levels of physical activity between genders (H(1) = 52.19, p = <0.001) and age groups (H(17) = 552.25, p = <0.001).

Figure 2 Physical activity levels by age and gender



There are also significant differences in BMI between age groups (H(18) = 2086.3, p <0.001), but not genders (H(1) = 0.09, p = 0.75). Figure 3 shows that generally there is a steady increase in BMI until around the age of 30 after which BMIs remain relatively flat until falling among the oldest individuals in the dataset.

Figure 3 BMI by age and gender



To reduce the number of categories in the constraint attribute, 10 year age bands are considered. The resulting analysis indicates that there are significant differences in levels of physical activity and BMI between the 10 year age bands (Kruskal-Wallis rank sum test, ($H(7) = 512.56$, $p < 0.001$) and ($H(7) = 2068$, $p < 0.001$ respectively)). The trends in the data are similar to those seen in the analysis using the smaller age bands (Figure 4 and Figure 5).

Figure 4 Physical activity by age and gender

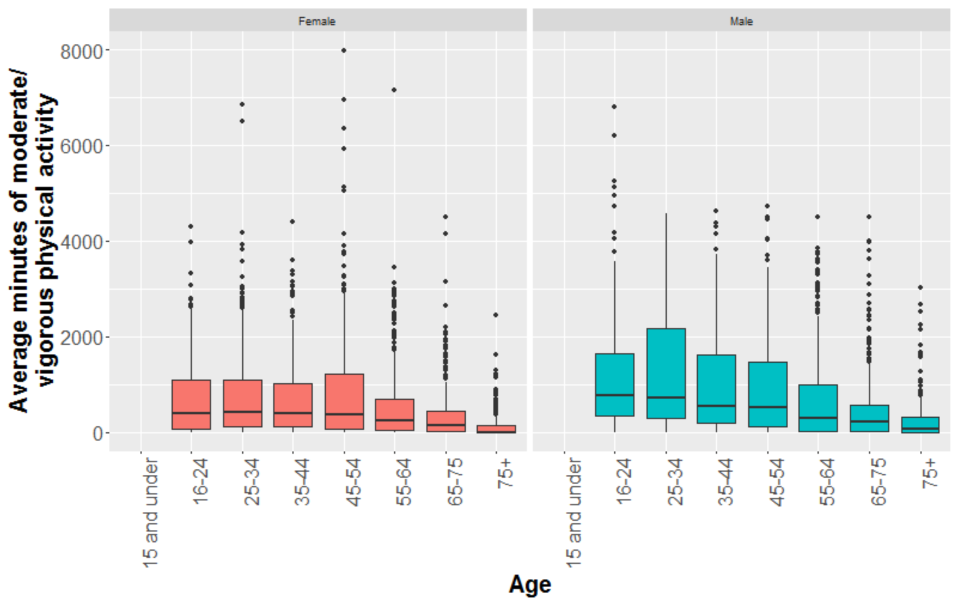
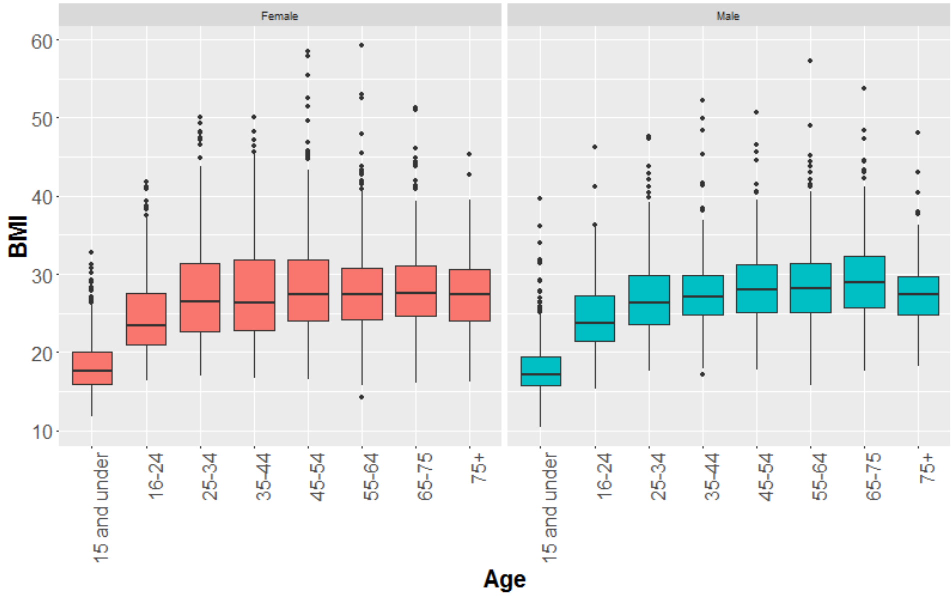


Figure 5 BMI by age and gender



2.3.2 Education

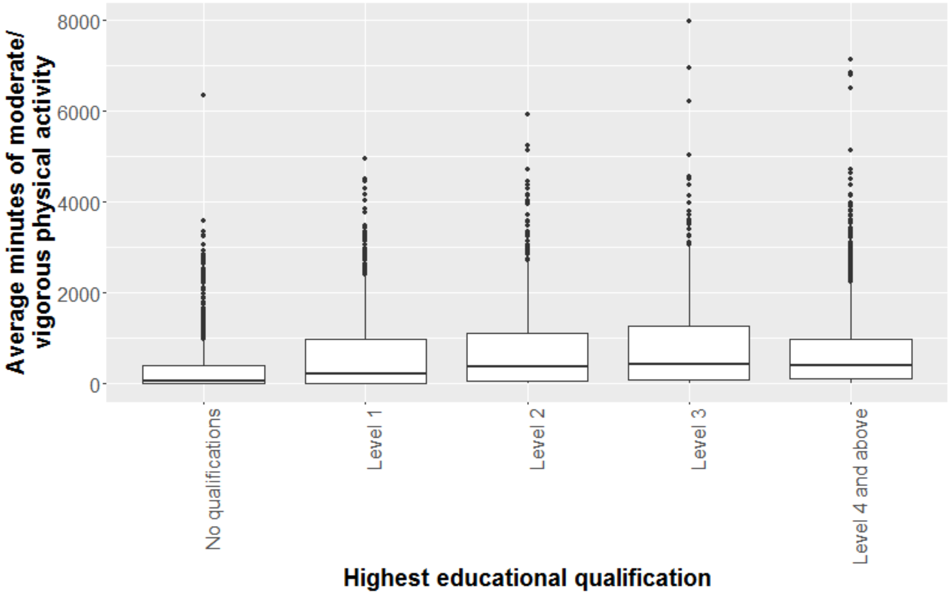
When compared with the population data (Census table HLQPS11), the SHeS variable 'hedqul08' contains one additional category ('other school level'). Using the census metadata, the 'Standard grade or equivalent' and 'Other school level' levels in the SHeS data can be combined to be comparable with the 'Level 1' category in the census data (Table 2-1).

Table 2-1 Education categories in micro-data and place-data

<table>
<tr><th>SHeS (2016) levels</th><th>Census 2011 levels</th></tr>
<tr><td>No qualifications</td><td>No qualifications</td></tr>
<tr><td>Standard grade or equivalent</td><td rowspan="2">Level 1
• O Grade, Standard Grade, Access 3 Cluster, Intermediate 1 or 2, GCSE, CSE, Senior Certificate or equivalent;
• GSVQ Foundation or Intermediate, SVQ level 1 or 2, SCOTVEC Module, City and Guilds Craft or equivalent;
• Other school qualifications not already mentioned (including foreign qualifications)</td></tr>
<tr><td>Other school level</td></tr>
<tr><td>Higher grade or equivalent</td><td>Level 2
• SCE Higher Grade, Higher, Advanced Higher, CSYS, A Level, AS Level, Advanced Senior Certificate or equivalent;
• GSVQ Advanced, SVQ level 3, ONC, OND, SCOTVEC National Diploma, City and Guilds Advanced Craft or equivalent</td></tr>
<tr><td>HNC/D or equivalent</td><td>Level 3
• HNC, HND, SVQ level 4 or equivalent;
• Other post-school but pre-Higher Education qualifications not already mentioned (including foreign qualifications)</td></tr>
<tr><td>Degree or higher</td><td>Level 4 and above
• Degree, Postgraduate qualifications, Masters, PhD, SVQ level 5 or equivalent;
• Professional qualifications (for example, teaching, nursing, accountancy);
• Other Higher Education qualifications not already mentioned (including foreign qualifications)</td></tr>
</table>

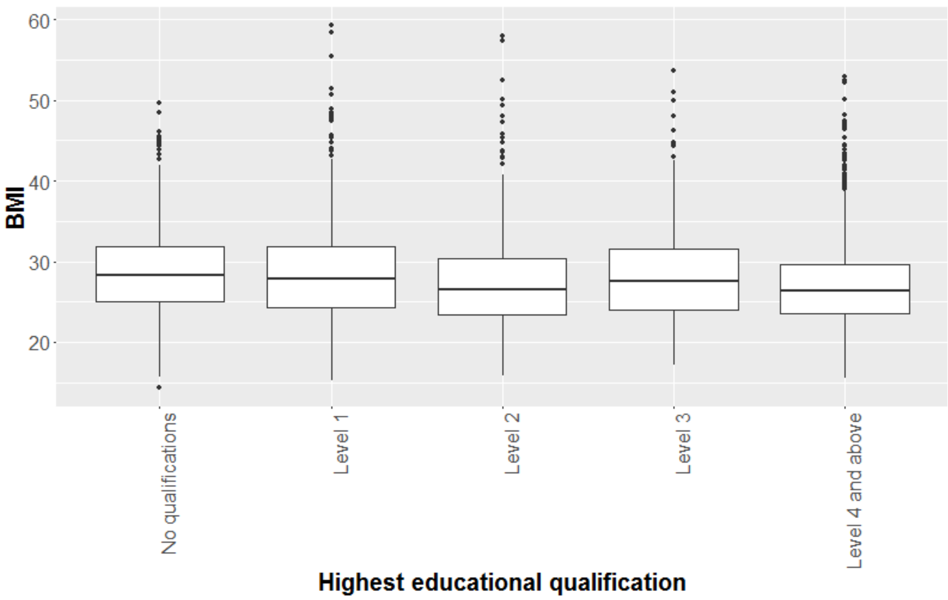
There are significant differences in levels of physical activity depending on the highest educational qualification obtained (Kruskal-Wallis rank sum test, $H(4) = 276.02$, $p < 0.001$). Figure 6 shows the relationship between highest educational qualification obtained and the average minutes of moderate/vigorous physical activity performed.

Figure 6 Physical activity by highest educational qualification



There are also significant differences in BMI depending on the highest educational qualification obtained (Kruskal-Wallis rank sum test, $H(4) = 61.68$, $p < 0.001$). Figure 7 suggests that the median BMI for those with no qualifications is slightly higher than that of those with qualifications, and that the median BMI for those with qualifications at level 2 or level 4 is lower than in the other categories.

Figure 7 BMI by highest educational qualification

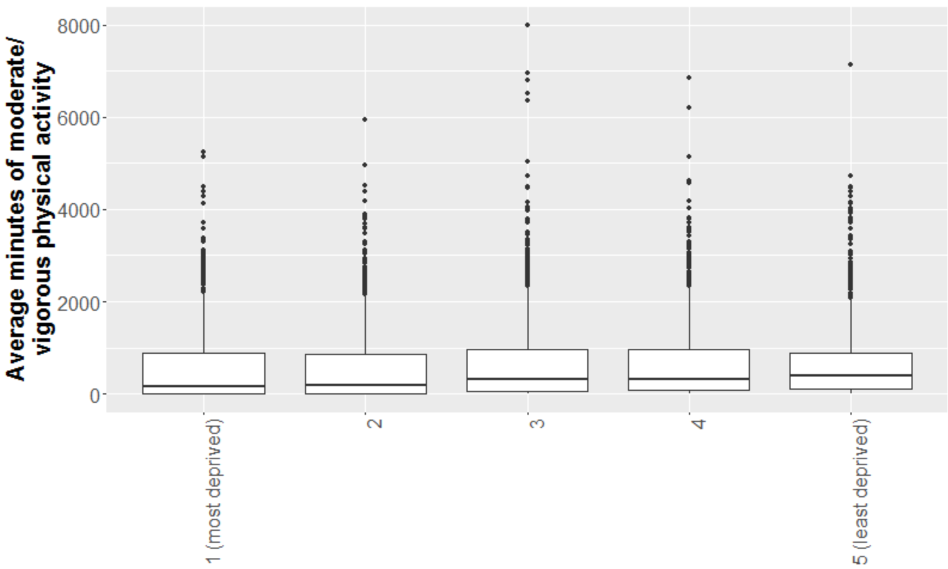


2.3.3 Scottish Index of Multiple Deprivation

The Scottish Index of Multiple Deprivation (SIMD) is a well-established metric for identifying the degree of deprivation of the area in which an individual lives. Quintiles are used as the common SIMD metric between the place-data and the micro-data.[5]

There are significant differences in levels of physical activity depending on the overall deprivation quintile of the data zone in which a person resides ($H(4) = 55.35$, $p < 0.001$). Figure 8 shows the relationship between these attributes. The median level of physical activity performed is higher for those in less deprived quintiles than those in more deprived quintiles.

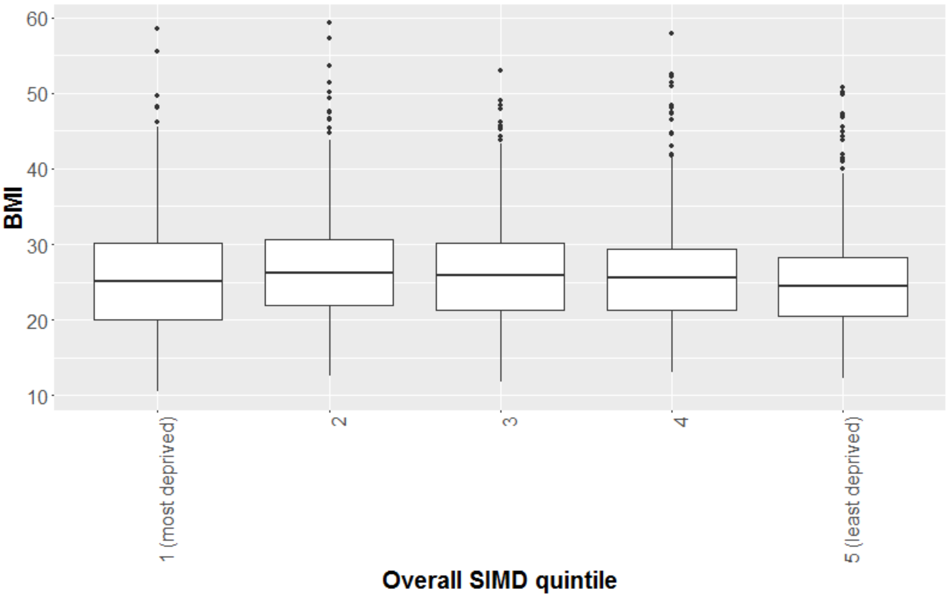
Figure 8 Physical activity by SIMD quintile



There are also significant differences in levels of BMI ($H(4) = 40.68$, $p < 0.001$) depending on the overall deprivation quintile of the data zone in which a person resides (Figure 9).

[5] Because the SIMD is only available at the level of Data Zones (DZ) (a geographical unit larger than an OA) the OAs in the population dataset are assigned to the SIMD quintile of the DZ in which they reside.

Figure 9 BMI by SIMD quintile



2.3.4 National Statistics Socio-Economic Classification

The National Statistics Socio-Economic Classification (NSSEC) is the official socio-economic classification used by the Office of National Statistics. It splits working age individuals into one of eight categories according to their employment characteristics. This classification is available in both the micro-data and the place-data. However, no individuals in the micro-data have been assigned to the lowest NSSEC category, "Never worked and long term unemployed".

There is a significant difference in both levels of physical activity (H(6) = 92.64, p <0.001) and BMI (H(6) = 31.93, p <0.001) between the different NSSEC categories. Figure 10 and Figure 11 show the data as boxplots.

Figure 10 Physical activity by NSSEC

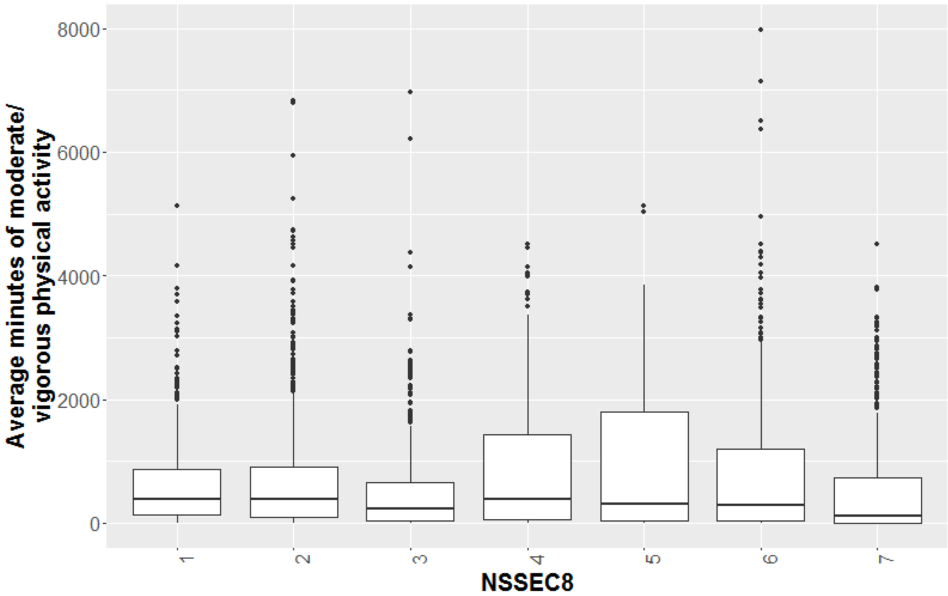
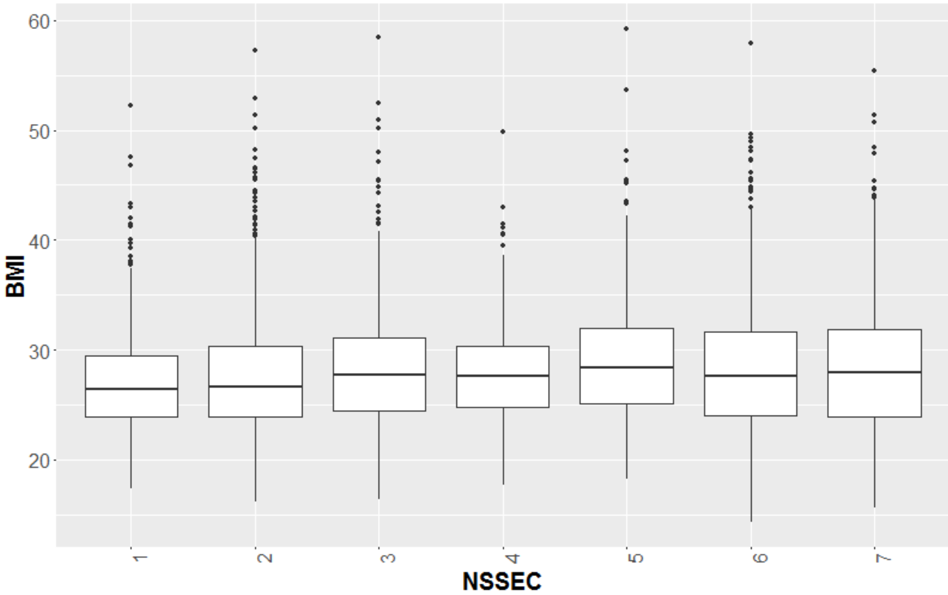


Figure 11 Physical activity by NSSEC



2.3.5 Limiting longstanding illness

Table 2-2 shows the difference between the micro-data and place-data in their categorisation of limiting longstanding illness. Combining these approaches results in a variable 'Limited activities due to longstanding illness' with two categories; 'Yes' and 'No'.

Table 2-2 Limiting longstanding illness categories in micro-data and population datasets

SHeS (2016) levels	Census 2011 levels
Limiting longstanding illness	Yes, limited a lot
	Yes, limited a little
Non limiting longstanding illness	No
No longstanding illness	

Figure 12 shows that those who are not limited in their activities due to longstanding illnesses perform a significantly higher average amount of moderate/vigorous physical activity compared to those that are limited (H(1) = 431.25, p <0.001).

Figure 12 Physical activity by limiting longstanding illness

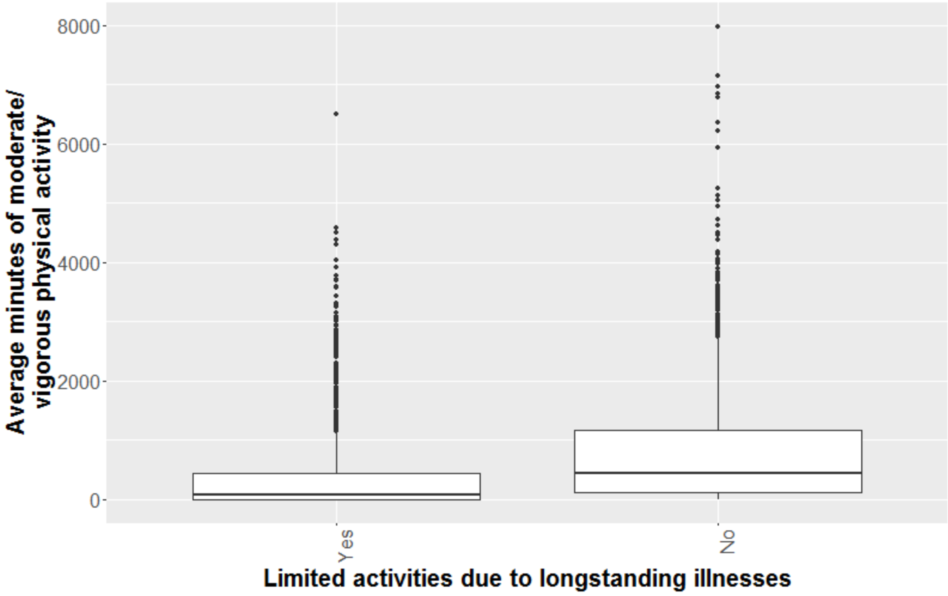
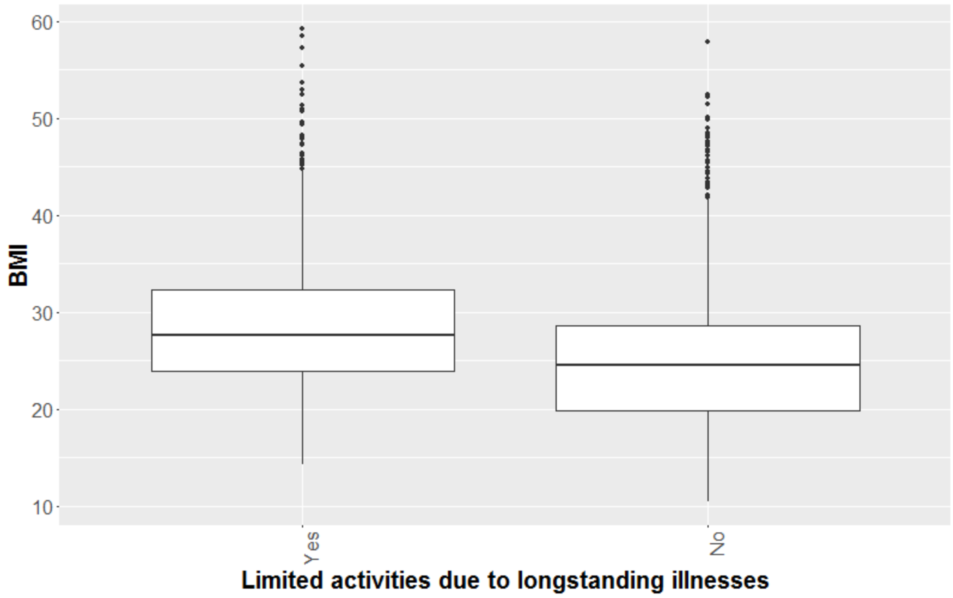


Figure 13 shows that those that are limited in their activities due to longstanding illnesses have a higher BMI than those that are not limited in their activities (H(1) = 257.85, p <0.001).

Figure 13 BMI by limiting longstanding illness



2.3.6 Economic activity

Table 2-3 shows the difference between the micro-data and place-data in their categorisation of economic activity. In addition there are a large number of categories for this constrained variable, which increases the chance of an ‘empty cell’ problem when producing the spatial simulation. Following the example of Philips (2017) we reduce the number of categories to four:

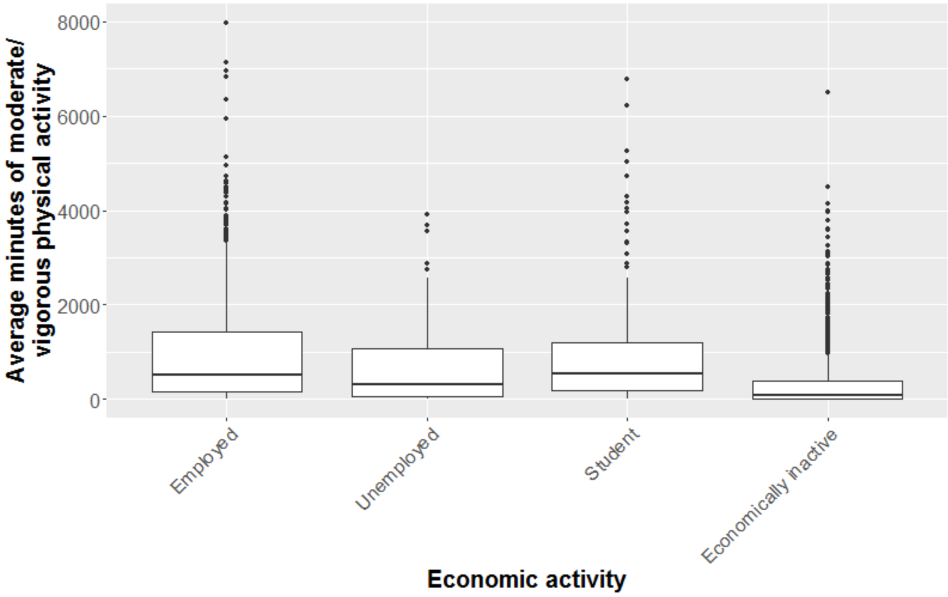
- Employed
- Unemployed
- Student
- Economically inactive

Table 2-3 Economic activity categories in micro-data and population datasets

SHeS (2016) levels	Census 2011 levels	After Philips (2017)
In paid employment, self-employed or on gov't training	Employee: Full-time	Employed
	Employee: Part-time	
	Self-employed with employees: Part-time	
	Self-employed with employees: Full-time	
	Self-employed without employees: Part-time	
	Self-employed without employees: Full-time	
Looking for/intending to look for paid work	Unemployed	Unemployed
In full-time education	Full-time student	Student
	Student	
Perm unable to work	Long-term sick or disabled	Economically inactive
Retired	Retired	
Looking after home/family	Looking after home or family	
Doing something else	Other	

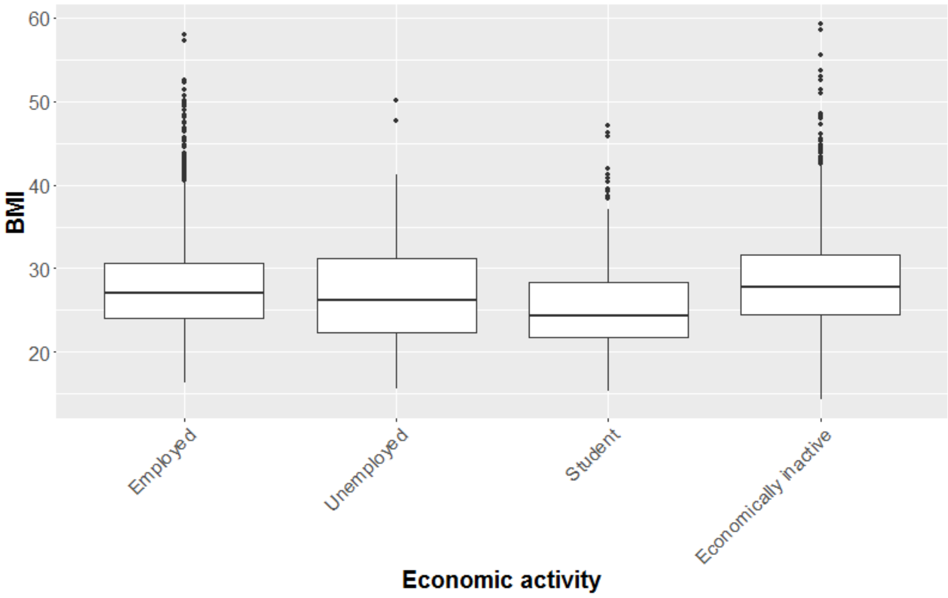
There are significant differences in levels of physical activity depending on economic activity ($H(3) = 670.93$, $p < 0.001$). Figure 14 shows that those that are unemployed or economically inactive typically have lower levels of moderate/vigorous physical activity than those that are in full-time education or employed.

Figure 14 Physical activity by economic activity



There are also significant differences in BMI depending on economic activity (H(3) = 58.12, p <0.001). Figure 15 shows that those that are in full-time education typically have lower BMIs compared to other economic activities. Those that are economically inactive have a higher median BMI than the other categories.

Figure 15 BMI by economic activity



2.4 Selecting constraint attributes for the microsimulation

The previous section identified the constraint attributes that could be used to combine the population dataset with the micro-data. However, when combined they categorise the dataset into c.7,700 distinct groups, including multiple ‘blank cells’ in the micro-data.

This means that the number of categories need to be reduced. This can be done in two ways:

- Removing one or more constraint attribute entirely
- Reducing the number of categories used for one or more constraint attribute

This section of the paper explains the process of identifying the final set of constraint attributes used in the simulation.

2.4.1 Scottish Index of Multiple Deprivation

The SIMD rank of a DataZone is calculated using a number of metrics, including the income, employment status, health and educations of its population. This constraint variable is removed from the simulation for two reasons; firstly, because the metric is evaluated at the DataZone population level, it may not be applicable at the micro level; and secondly, a number of the underlying metrics to the SIMD ranking are already included in some form as other potential constraint variables.

Removing the SIMD variable reduces the number of categories fivefold, but there are still a large number of empty cells in the micro-data.

2.4.2 Age bands

The 10 year age bands seen in Figure 4 above can be simplified to just three categories while retaining the essential relationship with the unconstrained variables. We use the following age bands in the simulation:

- Under 25
- 25 to 49
- 50+

However, even with this reduced level of categorisation, we are still subject to a high number of missing cells in the micro-data. This means that we will need to reduce the sophistication of our simulation in comparison to that used by Philips et al (2017). This is likely to be at least partially caused by the smaller sample size of the micro-data resulting in a higher number of ‘empty cells’.

2.4.3 National Statistics Socio-Economic Classification

The NSSEC variable is derived from an individual’s employment characteristics. This means that there is a strong correlation with the economic activity variable, so there is overlap between the two variables. In addition, there is a problem with the NSSEC variable in the micro-data as it is unclear why none of the SHeS respondents have been assigned to the ‘lowest’ category, and full time students are also excluded. The NSSEC variable is therefore eliminated as a potential constraint attribute.

2.4.4 Educational qualification

The five categories in the ‘highest educational qualification achieved’ attribute are too granular for the simulation. The data indicate that the relationship between highest level of education reached and an individual’s level of economic activity are strongly correlated. Removing the educational qualification attribute results in a categorisation of 48 distinct groups. There is one empty cell in the micro-data (male students aged over 50, with a longstanding limiting illness)

2.5 Selected constraint variables

Following the above elimination process, the constraint variables selected for the simulation are:

- Age and gender (6 categories)
- Economic activity (4 categories)
- Longstanding limiting illness (2 categories)

A total of 3,526 individuals are included in the micro-dataset[6]. The smallest category has 2 individuals, while the largest has over 400[7].

Table 2-4 Sample of the micro-data constraint variables

Unique ID[8]	Age/gender	Longstanding limiting illness	Economic activity
11111	25-49Female	Yes	Economically inactive
11112	50+Male	No	Employed
11113	25-49Female	No	Economically inactive
11114	16-24Female	Yes	Student
11115	50+Male	No	Employed
11116	25-49Female	Yes	Economically inactive
11117	25-49Male	No	Employed
11118	25-49Male	No	Employed
11119	25-49Female	No	Employed

We also apply these constraint attributes to the place-data to identify the number of individuals in each category in each output area.[9]

[6] After children under 16 are removed, along with any adult respondents who did not answer the relevant questions.
[7] See Appendix for full list of category sizes
[8] Falsified
[9] The Census table for economic activity only includes individuals aged 16-74. All individuals aged 75+ are assigned to the ‘economically inactive’ category.

Table 2-5 Sample of the categorised place-data (not all constraint attribute categories shown)

Output area ID	Number of 16 – 24 male	Number of 50+ female	Number of individuals with long-term limiting illness	Number of individuals who are employed	Number of individuals who are economically inactive
S00088956	6	23	18	113	31
S00088957	4	15	10	38	15
S00088958	3	5	3	41	12
S00088959	2	12	10	43	14
S00088960	4	14	3	37	13
S00088961	11	45	84	74	90
S00088962	3	15	6	39	15
S00088963	4	15	10	45	18
S00088964	4	17	10	53	12
S00088965	12	33	23	70	32

2.6 Applying the spatial microsimulation

Once the constraint attributes had been identified, the spatial microsimulation process was applied to simulate the population. Following the approach of Philips (2018, 2017), Flexible Modelling Framework (FMF) spatial microsimulation software (Harland, 2013) was used to apply the simulation process. An outline of the process is provided below, including an example using only age/gender and longstanding limiting illness as constraint attributes.

For each OA in the place data, the process randomly selects a sample of individuals from the micro-data the same size as the actual population of the OA[10]. The constraint attributes from this sample of the micro-data are aggregated and compared to the constraint attributes of the OA.

Table 2-6 shows an example of a random sample of micro-data records (n=10) to compare to an OA with a population of 10, while Table 2-7 shows the aggregated constrained variables for this sample compared to the OA.

Table 2-6 Example sample of micro-data records

Micro-data ID[11]	Age/gender	Longstanding limiting illness
11182	25-49Female	No
11132	50+Male	No
11182	25-49Female	No
11168	16-24Female	Yes
11165	16-24Male	No
11196	50+Female	No
11176	25-49Male	No
11205	50+Male	Yes
11195	25-49Male	No
11176	16-24Female	No

[10] It is possible for the same individual's record to be selected multiple times in a single sample.
[11] Falsified

Table 2-7 Aggregate constraint variables for micro-data sample and OA

Constraint variable	Age and gender						Long-term limiting illness	
Category	16 – 24 male	25 – 49 male	50+ male	16 – 24 female	25 – 49 female	50+ female	Yes	No
Number of individuals (sample)	1	2	2	2	2	1	2	8
Number of individuals (OA)	2	1	2	3	1	1	3	7

The aim of the microsimulation process is to generate a sample of the micro-data which matches the aggregate constraint attributes of the OA. The metric used to compare the sample and the OA is the Total Absolute Error (TAE), the sum of the number of differences between the two sets of data. In the example above, the TAE would be 6. The microsimulation process aims for a TAE of 0.

To do this it employs an iterative optimisation algorithm in which, if the TAE between the sample of micro-data (the simulated population) and population of the OA is greater than 0, a randomly selected individual from the sample of the micro-data is replaced by another randomly selected person from the micro-data. If this change lowers the TAE then it is accepted.[12]

Table 2-8 provides an example of a simulated population for which there is a TAE of 0 with the aggregate constraint variables of the example OA. Note how the simulated population can include the same individual more than once.

[12] There are various parameters that can be set, which determine how many iterations can be made and how the process handles increases in TAE following an iteration. The default parameters in the FMF software were applied in this case.

Table 2-8 Example simulated population with a TAE of 0

Micro-data ID	Age/gender	Longstanding limiting illness
11187	16-24Male	No
11137	16-24Male	No
11140	25-49Male	No
11132	50+Male	No
11192	50+Male	No
11209	16-24Female	Yes
11131	16-24Female	Yes
11131	16-24Female	Yes
11199	25-49Female	No
11206	50+Female	No

The micro-data ID can then be used to assign values for the unconstrained attributes to this simulated population.

Following the microsimulation process, simulated populations with a TAE of 0 had been created for all of the OAs in the population dataset.

2.7 Bicycle availability

In addition to the health of an individual, a major factor in their ability to travel actively is whether they have access to a bicycle. Before we can model the capacity of the simulated population to walk and cycle, we need to know which individuals have access to a bicycle.

The 2016 Scottish Household Survey (SHS) asked respondents "how many bicycles are available for use by adults in your household?" We can use these data to model which individuals in our simulated population have access to a bicycle. Similarly to the process of identifying constraint variables for the simulation process, we align the SHS gender, age and economic activity variables (as correlates of bike availability) with those in the SHeS micro-data and then calculate the weighted proportions of adults that have a bike available for use[13] for each category (Table 2-9).

Table 2-9 Weighted proportion of randomly selected adults from respondent households with a bike available to them by age/gender/economic activity category

Age/gender/economic activity	% that have an adult bike available to them
16-24 Female Economically inactive	8.2%
16-24 Female Employed	40.6%
16-24 Female Student	51.5%
16-24 Female Unemployed	33.6%
16-24 Male Economically inactive	23.9%
16-24 Male Employed	48.9%
16-24 Male Student	57.2%
16-24 Male Unemployed	51.9%
25-49 Female Economically inactive	30.2%
25-49 Female Employed	49.8%
25-49 Female Student	43.4%
25-49 Female Unemployed	23.6%
25-49 Male Economically inactive	21.6%
25-49 Male Employed	52.4%
25-49 Male Student	49.5%
25-49 Male Unemployed	31.6%
50+ Female Economically inactive	18.9%
50+ Female Employed	42.0%
50+ Female Student[14]	100.0%
50+ Female Unemployed	14.8%
50+ Male Economically inactive	24.1%
50+ Male Employed	46.6%
50+ Male Student	100.0%
50+ Male Unemployed	27.8%

These bike availability proportions are then used to randomly select the same proportion of individuals in each category of the simulated population who have access to a bicycle.

[13] Those with one or more bicycled available for use by adults in their household
[14] There weren't any adults that were members of the 50+ female student group, and so the equivalent male value has been applied.

3 Validation of the simulated data

Before using it in the model for estimating capacity to walk and cycle, it is important to validate the simulated population data to make sure it is a good representation of the actual population. The most straightforward approach is to calculate some national and local authority level statistics from the unconstrained attributes of the simulated population and compare these to the national and local authority level results from the SHeS.

It is worth noting that this comparison is imperfect as the SHeS is used to produce the simulated data, so there is a degree of circularity in the comparison. However, the microsimulation did not account for where the SHeS respondents came from, whereas the national and local results only use data from the relevant geography, and are also weighted to represent that population. In addition, there are not any other data sources which can be used for a more effective validation. This is a common issue in spatial microsimulation studies (Edwards and Tanton 2012)

3.1 National comparisons

Table 3-1 shows the national results from the SHeS compared to those calculated from the simulated population.

Table 3-1 Comparison of simulated population and national statistics for BMI and levels of physical exercise

Statistic	SHeS 2016		Simulated population		% difference from national figure
Mean BMI	Female:	27.84	Female:	27.46	-1.4%
	Male:	27.56	Male:	27.57	0.0%
	All:	27.70	All:	27.51	-0.7%
Proportion of population with a BMI >= 25	Female:	63.13%	Female:	60.06%	-4.9%
	Male:	67.38%	Male:	68.28%	1.3%
	All:	65.19%	All:	64.00%	-1.8%
Proportion of population with a BMI >= 30	Female:	30.21%	Female:	28.14%	-6.9%
	Male:	27.24%	Male:	26.91%	-1.2%
	All:	28.77%	All:	27.54%	-4.3%
Proportion of adults (16+) meeting recommended levels of physical activity	Female:	59.81%	Female:	64.23%	7.4%
	Male:	70.89%	Male:	73.54%	3.7%
	All:	65.14%	All:	68.69%	5.4%
Average hours of physical activity performed each week	Female:	9.9	Female:	11.0	11.1%
	Male:	15.1	Male:	14.7	-2.6%
	All:	12.4	All:	12.7	2.4%

We can see that the simulated population provides a fair approximation of the national level statistics, although the simulation appears to have been less accurate when modelling female BMI and levels of physical activity.

A similar comparison can be drawn for access to at least one bicycle, with the national data taken from the SHS 2016 (Table 3-2).

Table 3-2 Comparison of simulated population and national statistics for access to bicycles

	National figure	Simulated population*	% difference from national figure
Proportion of households with bicycles available for private use	33.8%	39.2%	16.0%

*This figure is not directly comparable to the national figure as it refers to individuals rather than households.

We can see that the simulation overestimates the number of people who have access to a bicycle at the national level. This could potentially be improved by undertaking a more robust selection of constraint variables.

3.2 Council comparisons

The plots on the following pages (Figure 16 to Figure 18) show a comparison between physical health metrics from the SHeS data and the simulated population at council (local authority) level. SHeS outputs are not produced for all councils if the sample size for that authority is too small.

The diagonal line on each of the charts represents a line of ‘perfect fit’, where the SHeS output matches the simulated population data exactly.

We can see that at the council level, the metrics for the simulated population are similar to those drawn from the SHeS data. However, there is not as much variation between the simulated councils as there is in the SHeS data, particularly for the BMI metrics. This is revealed by the points appearing horizontal relative to the line of perfect fit.

We can also see that, in respect of the proportion of the population achieving their recommended levels of physical activity, the simulation has overestimated the results for these councils in comparison to the SHeS data.

These charts again support the findings of the national level validation, that the simulated population provides a fair approximation of the micro-data, but that there is scope for improving the accuracy of the simulation.

It is worth noting however, particularly at the council level, that the SHeS data may be subject to sampling error, as the sample sizes for some councils is quite small. Thus the SHeS values are estimates of the ‘true’ value, as are the values derived from the simulated data.

Figure 16 Proportion of population with BMI > 25

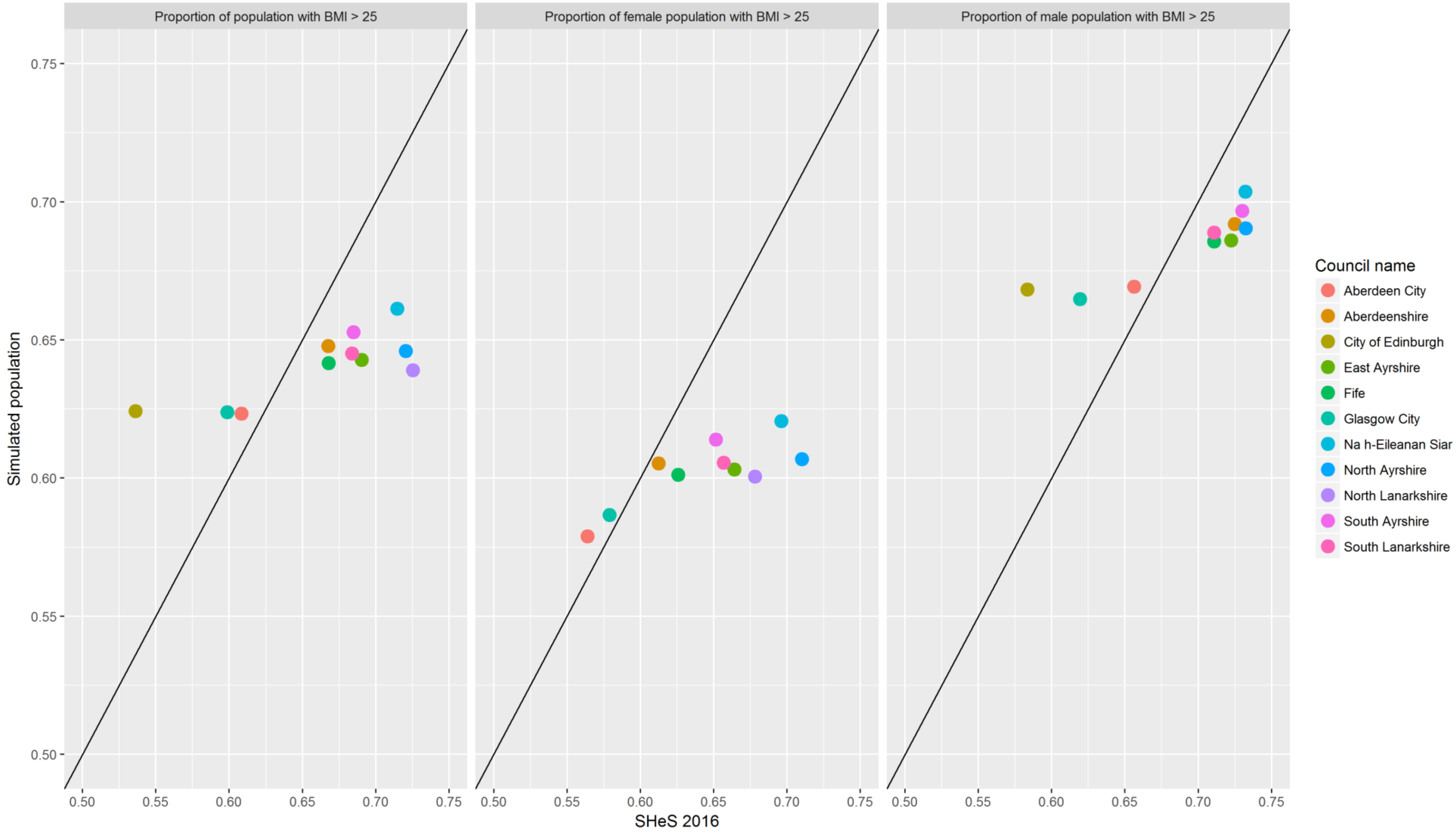


Figure 17 Proportion of population with BMI > 30

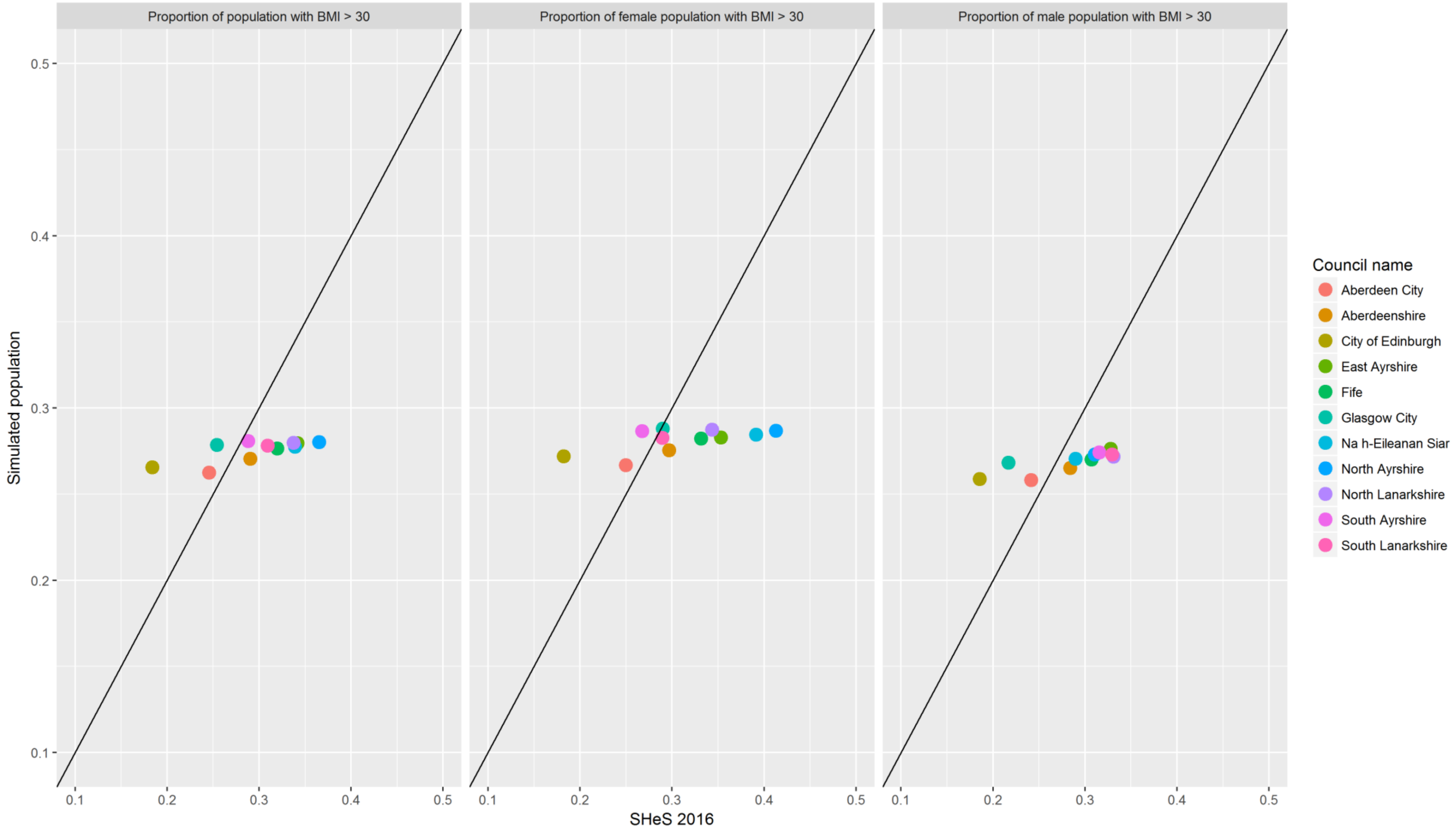
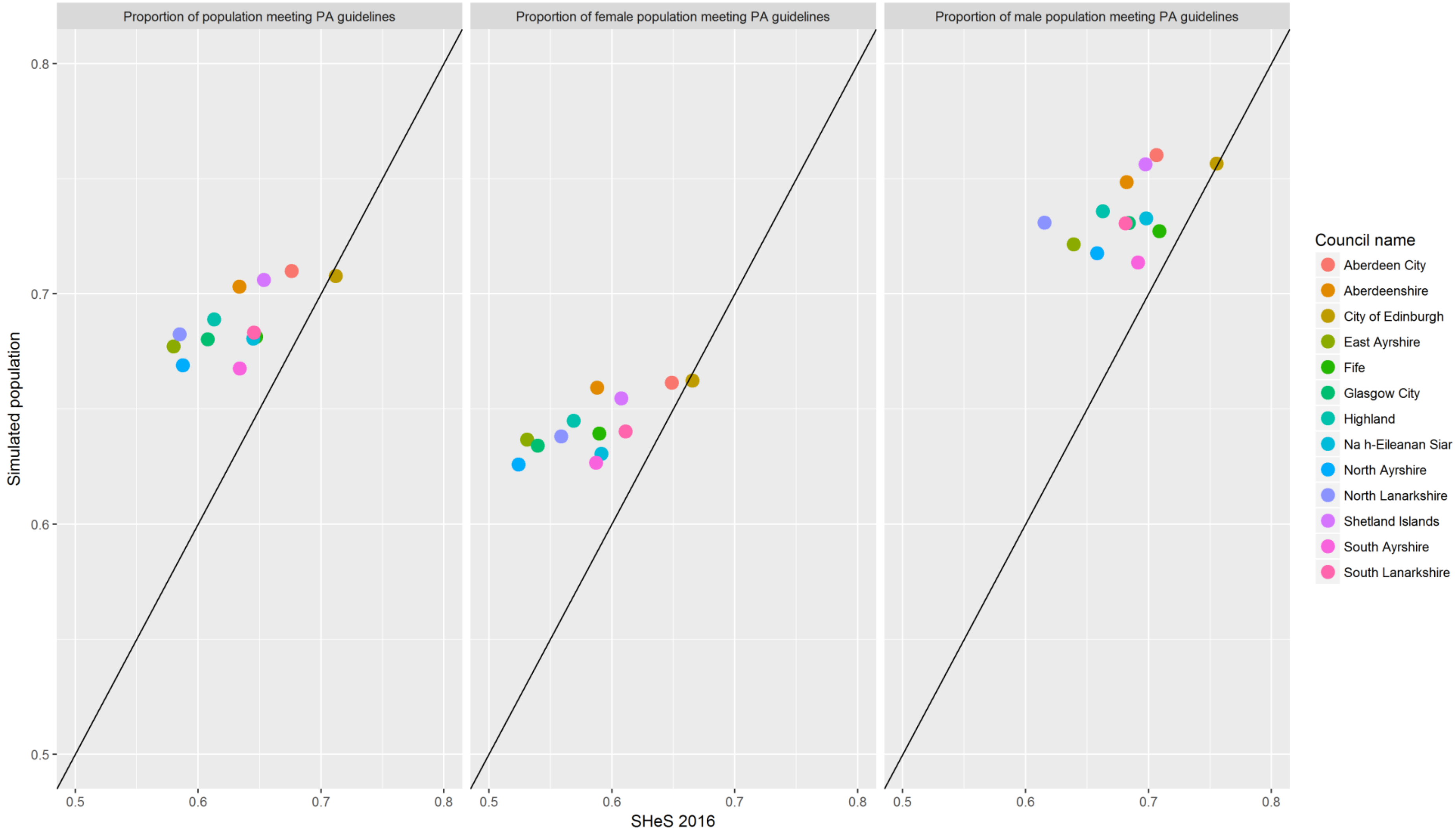


Figure 18 **Proportion of population meeting physical activity guidelines**



4 Modelling maximum cycling and walking distances for the simulated population

We can use the simulated data to model the extent to which each individual in the simulated population can walk and cycle. The stages of the modelling process can be broken down as follows:

- Estimating VO2max for each individual
- Estimating pedal power values for each individual
- Calculating an average slope value for each output area in the geography
- Calculating cycling and walking speed for each individual
- Calculating a maximum cycling and walking distance for each individual

This section of the paper explains each of these stages in turn.

4.1 Estimating VO2max

The model developed by Philips et al (2017) for estimating the distance someone can cycle takes the amount of pedal power that can be delivered by that individual as one of its inputs. This in turn is calculated from their VO2max, which is a measure of the maximum amount of oxygen that an individual can make use of during exercise. It is commonly used as a metric of an individual's aerobic fitness and can also be used to calculate a person's energy use and power output during exercise (McArdle, 2010).

However, our micro-data does not include a VO2max variable, so it is necessary to estimate this metric using the attributes that we do have in the micro-data. We use the regression model proposed by Weir et al (2006):

*VO2max = Constant(p) + (Age * Age(p)) + (Gender * Gender(p)) + (PASS score * PASS(p)) + (BMI * BMI(p))*

where (p) is the parameter associated with the respective variable.

4.1.1 Physical Activity Status Scale (PASS)

Before using this model a Physical Activity Status Scale (PASS) score also needs to be assigned to each person. PASS is an activity and exercise scale developed by NASA (Wier et al, 2006).

In Philips et al (2017), the number of days on which 30+ minutes were spent doing sport or vigorous activity in the past 4 weeks (VIG30SP in the Health survey for England dataset) was converted to a PASS score (Table 4-1).

Table 4-1 Conversion of Health survey for England variable 'VIG30SP' to PASS score

VIG30SP value	PASS score
0	1,2,3*
1-4	4
5-8	5
9-16	6
17-24	7
>24	8,9,10*

* Assigned with equal probability

The closest equivalent variable in our micro-data was the total number of days on which 30+ minutes were spent being active in the last 4 weeks (variable code ‘adtot10b’). The same conversion has been applied, although this attribute does not specify the intensity of the activity. This may result in an artificial inflation of the PASS score for some individuals, as they could be performing a high volume of low intensity exercise (eg walking the dog). Because the higher PASS scores assume a high degree of vigour, these individuals would receive an incorrectly high level PASS score. This could be developed in future iterations of this microsimulation.

A value for the adtotal10 variable was not available for two of the cases in the micro-data, but in both cases the value was set to 0 due to them having a value of 0 for the ‘Average minutes doing moderate and vigorous physical activity per week (10+ mins)’ variable (mintot10T).

4.1.2 VO2max

Once the PASS score had been calculated for each individual, Weir et al’s regression model (2006) was applied to each of the individuals in the simulated population. The regression parameters are shown in Table 4-2, and the error value was applied probabilistically after the VO2max values were calculated for each individual.

Table 4-2 Regression parameters for VO2max estimation from Wier et al (2006)

Attribute	Beta
Constant	57.402
Age	-0.372
Gender	8.596
PASS	1.396
BMI	-0.683
Error ml/kg/min (95%ci)	4.9

4.2 Estimating pedal power

Pedal power is the rate of useful work applied to moving the cranks to propel the bicycle. This is calculated using the VO2max, weight and BMI variables. We use the same method used by Philips et al (2018) to estimate pedal power, which accounts for an individual’s lactate threshold – below which physical activity can be sustained for long periods.

4.3 Average output area slope

One of the factors affecting the distance that an individual can walk or cycle is the topography of the area in which they live (Phillips et al, 2017). The hillier an area, the greater the effort required to travel any given distance.

To reflect this, the average slope of all roads within five kilometres of the centre of each output area was calculated. These values were derived using Ordnance Survey ‘Meridian’ roads data[15] and Ordnance Survey ‘Terrain 50’ open data 50 m digital elevation data[16] as follows:

[15] No longer available from Ordnance Survey
[16] Available from https://www.ordnancesurvey.co.uk/opendatadownload/products.html

1. Mosaic all tiles of raster digital elevation data
2. Extract the maximum elevation along a road segment
3. Extract the minimum elevation along a road segment
4. Calculate the slope of the road segment (maximum - the minimum) / length of segment
5. Create a 5km buffer around OA centroid
6. Use a spatial join to calculate the mean gradient of all segments within the buffer
7. Repeat 6 for all OA centroids

These values are then used in calculating the cycling and walking speed for each individual.

4.4 Calculating cycling and walking speed for each individual

4.4.1 Cycling

The pedal power, weight and average slope values for each individual are used for the calculation of their cycling speed. Uphill, downhill and flat terrain cycling speeds are calculated using algorithms taken from Wilson (2004). These values were initially expressed in meters per second, but then transformed to kilometres per hour.

4.4.2 Walking

A simple estimation of walking speed (which doesn't account for weight and is capped at 6km p/h) is taken from McArdle (2010):

*Walking speed (km p/h) = (((0.5 * VO2max) – 5) / 2.5)*

To account for the slope of the individual's output area, the rule set out by Naismith (1892) is used; for every metre ascended it takes six seconds longer to walk a given distance (compared to it being flat). This rule is applied to the walking speed value that has been calculated using the McArdle model.

4.5 Calculating a maximum cycling and walking distance for each individual

The final step in the modelling process is to estimate the distance that each individual can travel at their given speed. Following Philips et al (2017), it is assumed that each individual could travel at their given speed for a maximum of one hour. This is deemed to be the upper limit for the general population as the majority are sedentary and/or not regular cyclists, so walking or cycling for more than an hour at a time could lead to injury when repeated over several days.

Note that, because the effect of an individual's physical health on their ability to travel actively is already incorporated into the *speed* metric, it is not necessary to incorporate this into the *distance* metric as well, as this would add an un-evidenced bias into the calculation. We use the *distance* metric as it is a more easily understood value with which to communicate the results. It can also be used to explore the extent to which individuals can access their local services by active modes, in a way in which the *speed* metric does not allow.

4.5.1 Cycling

The distance value calculated was calculated by taking the weighted mean of the time taken to travel a single kilometre at the different uphill, downhill and flat terrain cycling speeds for each individual. The distance that the individual can travel in one hour is then derived from this figure.

The weights used for the uphill, downhill and flat terrain cycling speed values vary according to the average slope of the output area. These weightings are shown in Table 4-3. These are taken from Philips et al (2017).

Table 4-3 Weighting of uphill, downhill and flat terrain cycling speeds by output area slope

Output area slope	Uphill	Downhill	Flat terrain
≥ 4%	56.0%	44.0%	-
2% ≤ x < 4%	50.4%	39.6%	10.0%
<c2%	42.0%	33.0%	25.0%

4.5.2 Walking

As the estimated walking speed is calculated in kilometres per hour, it already reflects the maximum distance that the individual could travel in an hour.

5 Results and maps

It is now possible to examine the results of the modelling process to understand the capacity of the Scottish population to travel actively. Although this is quantified by the distance by which they can walk and cycle, in isolation this unit of measurement is not very meaningful. It becomes meaningful when making comparisons between different groups of individuals, as it reveals differences between them. It can also be used, as per Philips (2014), to provide insights into whether people can access jobs or services by active modes.

We first consider some summary statistics for the whole population to provide a national standard before looking into some of the local authorities in more detail. We then produce some maps to demonstrate possible presentations of the data.

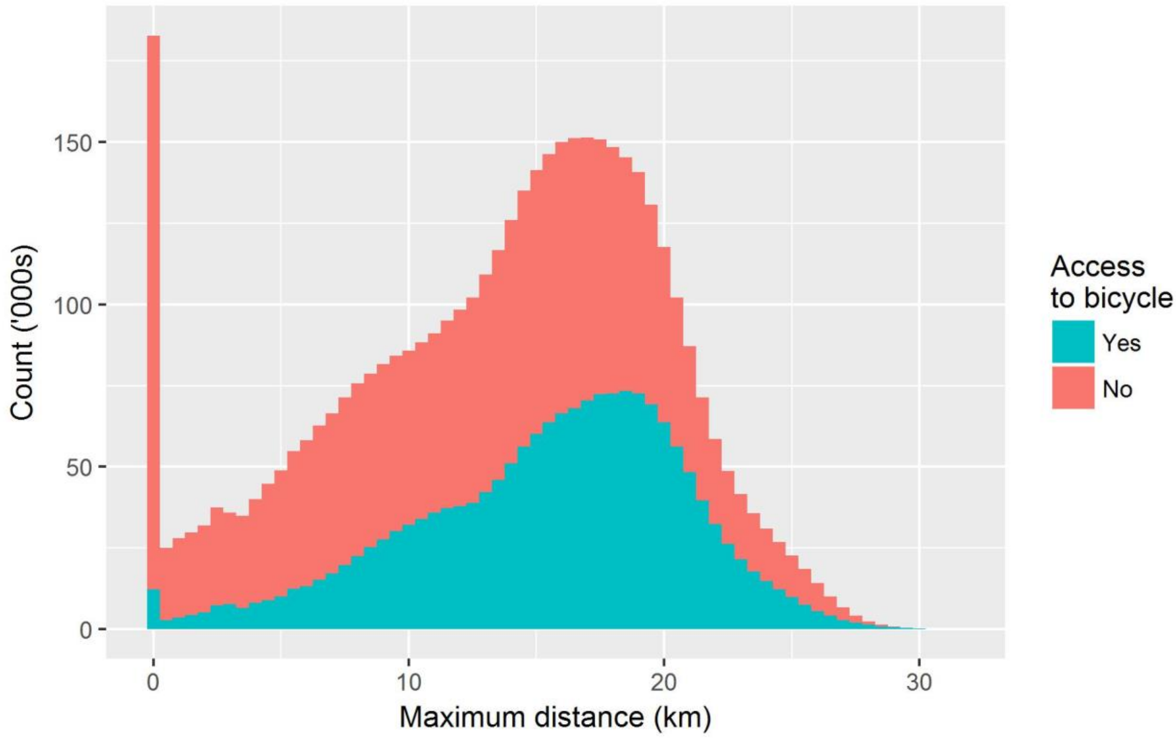
5.1 Summary national results

Table 5-1 shows the summary statistics for the distance that people could travel by bicycle (for everyone and for the subset of the population who currently has access to one), while Figure 19 shows a histogram of the same data, split by whether individuals have access to a bicycle or not. We can see that those who have access to a bicycle could, on average, cycle slightly further than those who do not currently have access to a bicycle (Welch Two Sample t-test, $p < 0.001$).

Table 5-1 Summary national results (cycling)

Statistic	Whole population	Those who can currently access a bicycle	Those who cannot currently access a bicycle
Minimum	0.06	0.06	0.06
1st quartile	9.29	12.19	7,41
Median	14.64	16.27	13.26
Mean	13.61	15.51	12.37
3rd quartile	18.33	19.29	17.42
Maximum	31.45	31.45	30.41

Figure 19 Histogram of maximum cycling distance



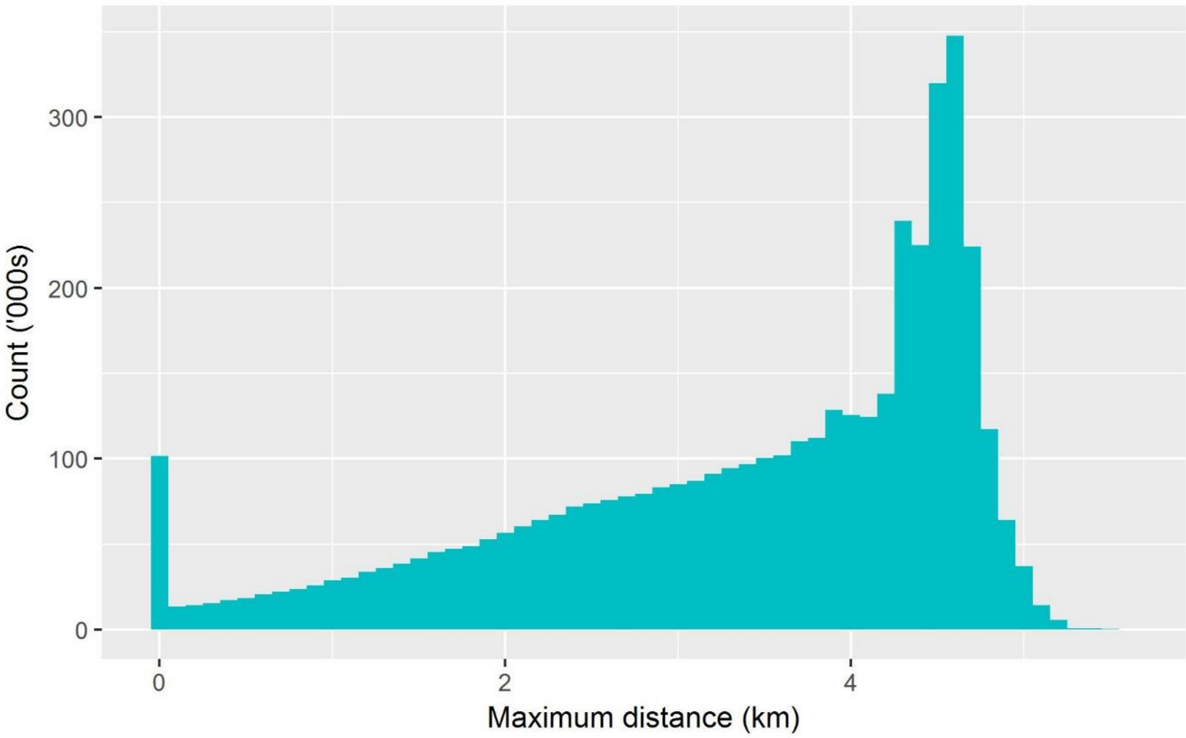
We can produce the same summary statistics and histogram for the distance that people can walk, although we don't have to differentiate the population according to those who can access a bicycle.

On average, the model indicates that people can walk around 3.5km in an hour. The slightly 'odd' distribution of the data may be indicative of the less sophisticated modelling approach taken to estimating walking distances in comparison to cycling distances.

Table 5-2 Summary national results (walking)

Statistic	Whole population
Minimum	0.00
1st quartile	2.58
Median	3.78
Mean	3.39
3rd quartile	4.47
Maximum	5.55

Figure 20 Histogram of maximum walking distance

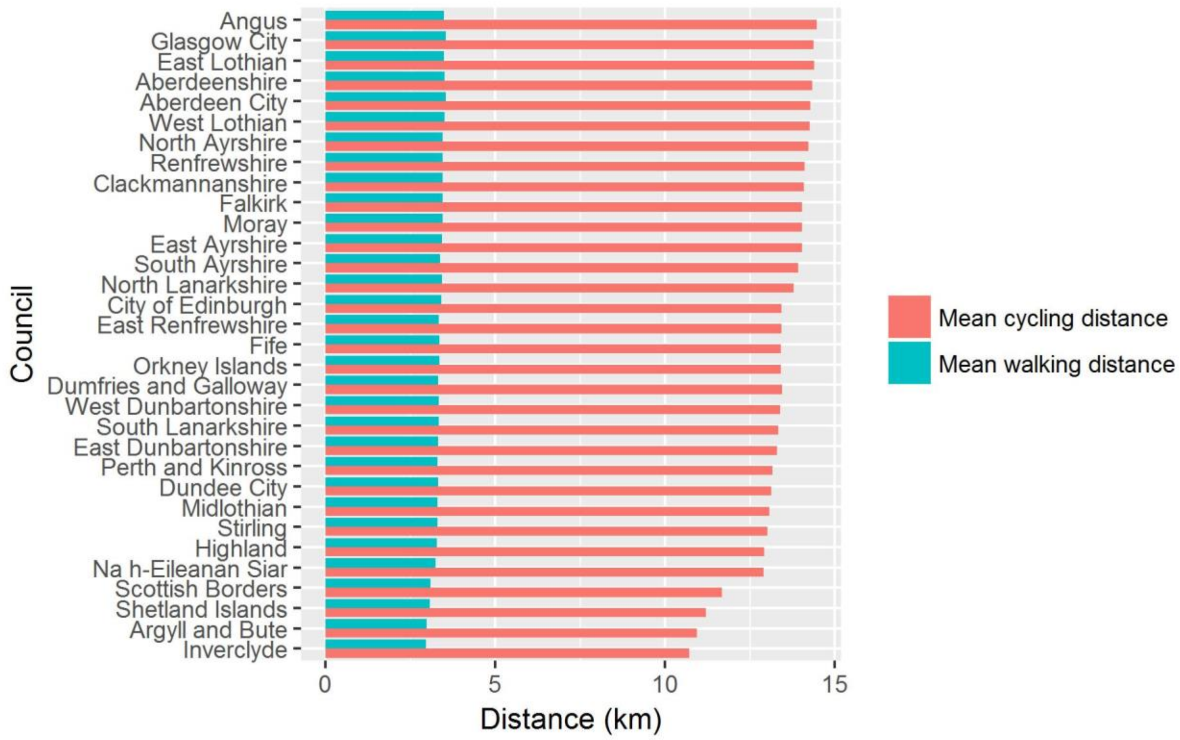


We can see in both Figure 19 and Figure 20 that there is a substantial number of individuals whose maximum cycling or walking distance is at or very close to zero. This is because the method for modelling both pedal power and walking distance can result in negative or zero values. In these instances the values are capped at zero, which in turn means there are a disproportionate number of simulated individuals with the same maximum walking and cycling distance.

5.2 Council results

There are 32 local authorities in Scotland, designated as Councils. The mean cycle distance for the adult population in each local authority ranges from 10.7km in Inverclyde to 14.5km Angus, while mean walking distances range from 3km in Inverclyde to 3.5km in Glasgow (Figure 21).

Figure 21 Mean walking and cycling distances by local authority



The maps in Figure 22 and Figure 23 show how the results are distributed across the country.

Figure 22 Mean cycling distances by local authority

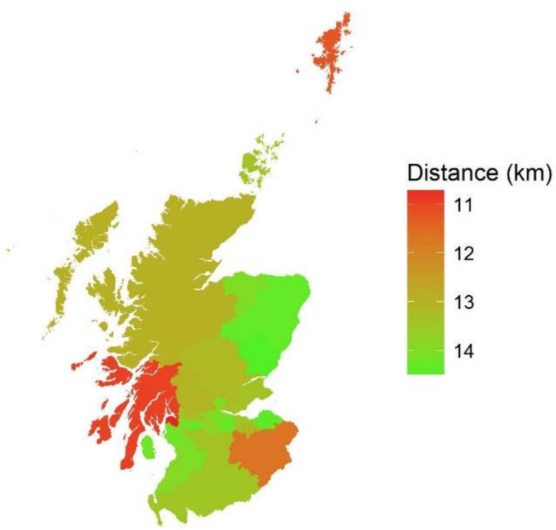
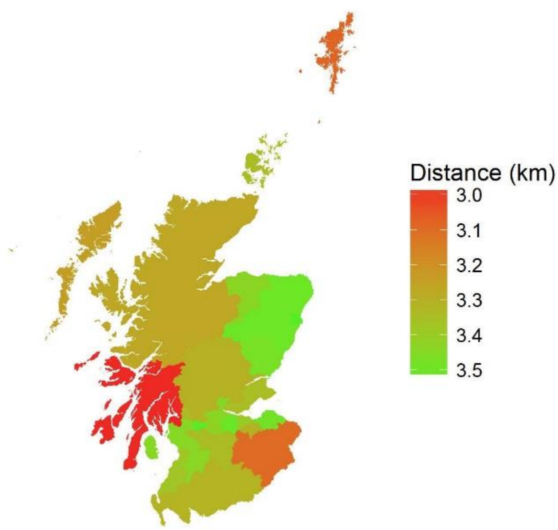


Figure 23 Mean walking distances by local authority



We can also visualise the results within each council area, which reveal how the variation in the mean cycling and walking distance by OA are distributed across each council. Examples are shown in the maps below (Figure 24 to Figure 27).

Figure 24 Mean cycling distances by OA (Dundee City)

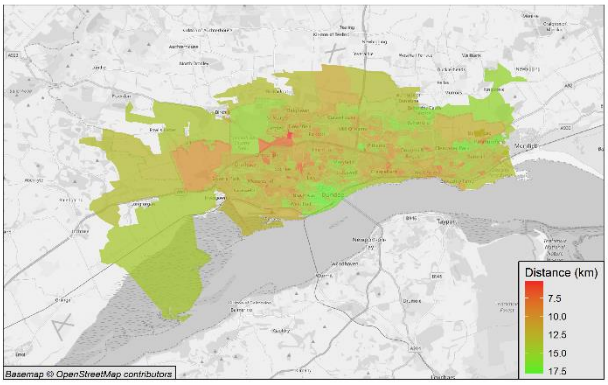


Figure 25 Mean cycling distances by OA (Aberdeen City)

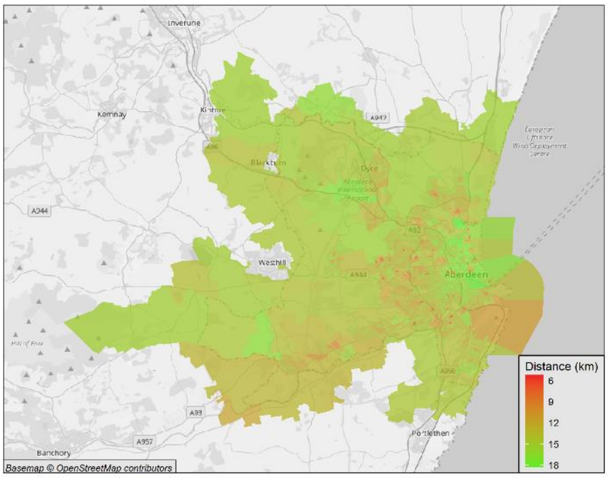


Figure 26 Mean walking distances by OA (Glasgow City)

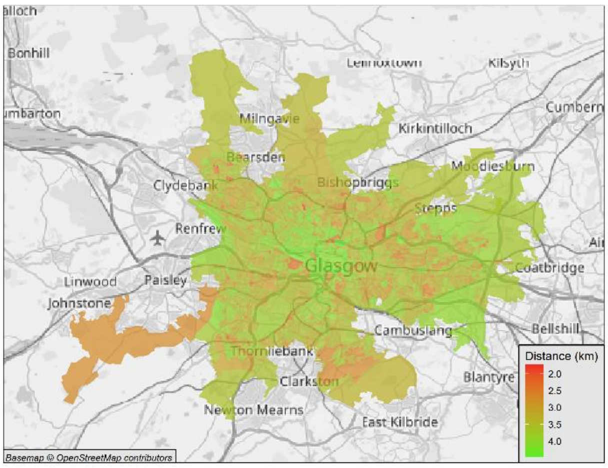
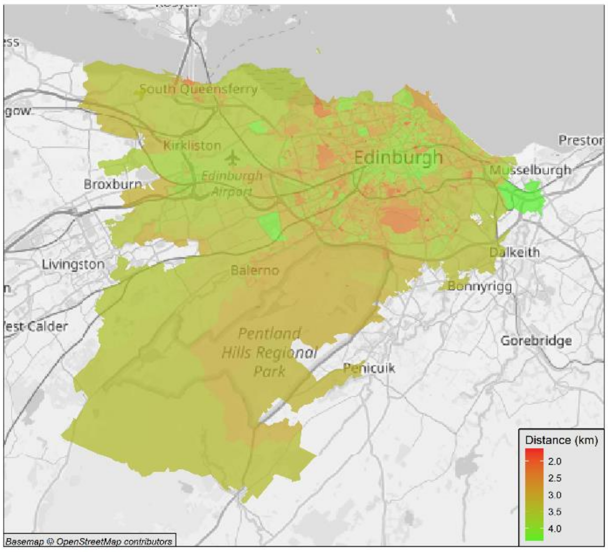


Figure 27 Mean walking distances by OA (City of Edinburgh)



The results can also be shown as intervals, such as quintiles in the examples below (Figure 28 and Figure 29).

Figure 28 Mean walking distance quintiles by OA (Perth and Kinross)

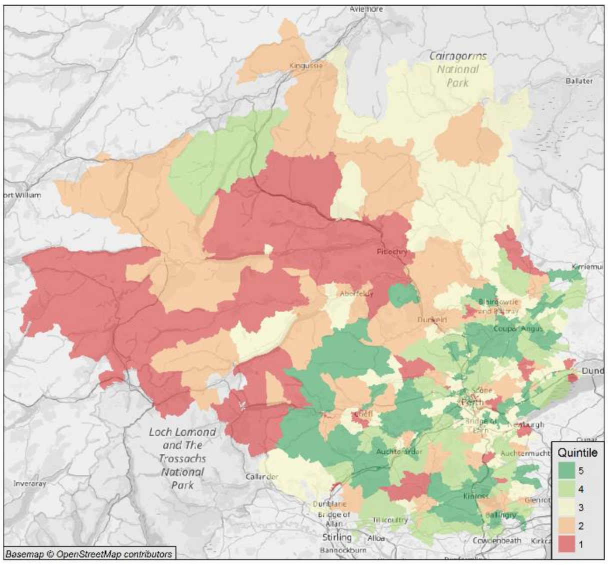
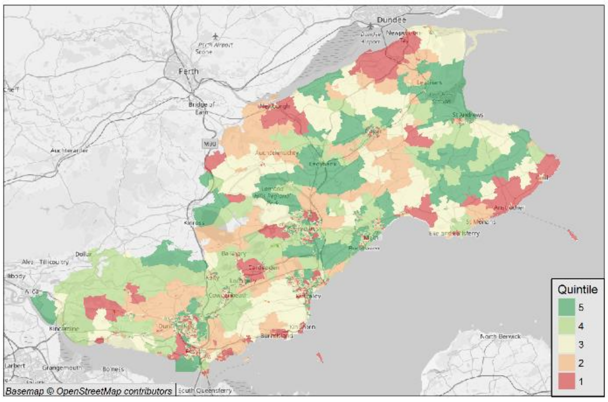


Figure 29 Mean cycling distances quintiles by OA (Fife)



6 Discussion and conclusion

6.1 Value of the simulated dataset

This paper has detailed the spatial microsimulation of the adult population of Scotland, producing a simulated dataset of health related attributes for every adult individual in the country. We have used this dataset to model the capacity of these individuals to travel by active modes, incorporating a spatial element to the modelling.

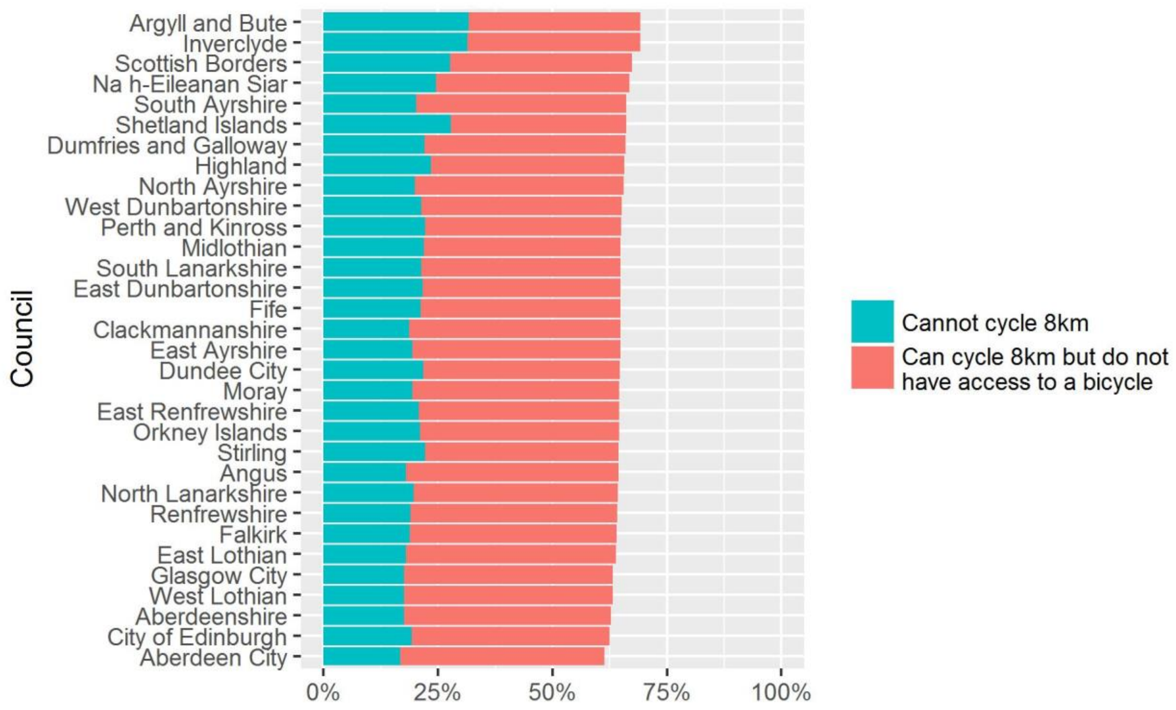
This process has produced a rich dataset with a variety of uses. Most immediately, it provides a clear and easily communicable presentation of locations where individuals have relatively limited capacity to travel by active modes. This has benefits for policy development and developing targeted interventions to improve this capacity.

It also illustrates where high capability exists and if compared to actual cycling / walking behaviours, would indicate the maximum potential for behavioural change and also possibly identify areas where some external factor was leading to a high degree of suppressed demand.

However, there are a wide range of possible uses for this dataset. For instance, in Philips (2014), the dataset was used to identify locations where individuals would be unable to travel to work by active modes in the event of a 'shock' event which rendered motorised transport unavailable.

This approach has also revealed that the assumption that everyone can cycle 5 miles (8 km – a commonly used assumption in transport planning, originating from Forester (1983)) overestimates capacity of the population to commute by active modes. On average 21% of the simulated individuals in each OA do not have the capacity to cycle 8 kilometres. In addition, a greater proportion of people can cycle over 8km, but don't have access to a bicycle (Figure 30).

Figure 30 Proportion of adults who cannot cycle 8km by local authority



6.2 Opportunities for further development

There are a number of areas where the methodology detailed in this paper could be improved. This is invariably the case with any modelling approach, but it is important that they are acknowledged. This simulation is not intended to be the last word in spatial micro-simulation for Scotland, but is meant to demonstrate the power of the approach and inspire further development in using it to model active travel behaviours.

The census data used to provide the spatial component of the simulation are several years out of date. This is a widespread problem for work that makes use of census data, but there is not currently any alternative data that could be used to provide this place-data. This does not affect the modelling of individual's capacity to travel actively – as the data used in this aspect of the model are drawn from up-to-date sources – but the spatial distribution of the results will be affected.

The correlation between the constraint and unconstrained variables is imperfect. Although all of the relationships examined show a statistically significant correlation at the 95% confidence level, there will inevitably be variance in the data. In addition, only a handful of variables have been included – other variables are likely to make a substantial contribution towards an individual's ability to travel actively. The impact of this is compounded by the compromises made in the selection of variables to enable the simulation approach to avoid the problem of 'empty cells' – in part caused by the relatively small sample of respondents in the micro-data.

6.3 Conclusion

This paper has set out a method for creating a simulated population dataset of health attributes incorporating a spatial aspect. It has used this data to model the capacity of the Scottish population to walk and cycle.

In doing so, the method presented and the synthetic population generated provides:

Insights into the capacity of every community in Scotland to travel actively

By incorporating a spatial element into the dataset and because the simulation occurs at the level of individuals we can use the dataset to explore how the health attributes of the population vary both between areas and within areas. This helps us to understand that, by relying on average values for different attributes (such as people's capacity to walk and cycle), we risk excluding or misrepresenting substantial numbers of individuals.

A framework for assessing the impacts of a scenarios and policies.

The method for modelling the population's capacity to walk and cycle includes a number of variables that could be affected by different policies. By changing the inputs to the modelling process we can use the dataset to ask "what if" questions about the impact of different policies. For instance, we could ask "what if a proportion of the population were provided with access to bicycles" or "what if everyone achieved their recommended level of physical activity". This can provide policy makers with valuable insights into the impacts of different policy decisions.

A basis for further behavioral models

Although the approach detailed in this paper models individual's capacity to walk and cycle, the health attributes that have been simulated could be used as the basis for a wide range of further behavioural models. In addition, the general approach of spatial microsimulation could be applied to a number of datasets to explore other facets of population behaviour.

7 References

Edwards, K.L., Tanton, R., 2012. Validation of Spatial Microsimulation Models, in: Tanton, R., Edwards, K. (Eds.), Spatial Microsimulation: A Reference Guide for Users. Springer Netherlands, Dordrecht, pp. 249–258.

Forester, J., 1983. Bicycle transportation. MIT Press, Cambridge, Mass.

McArdle, W. D. (2010). Exercise physiology: Nutrition, energy and human performance, (7th ed.). International ed.Philadelphia, P.A., London: Lippincott Williams & Wilkins.

Naismith, W., 1892. Untitled. Scott. Mt. Club J. 2, 135.

O'Donoghue, C., Morrissey, K., Lennon, J., 2014. Spatial Microsimulation Modelling: a Review of Applications and Methodological Choices. Int. J. Microsimulation 7, 26–75.

Philips, I., 2014. The potential role of walking and cycling to increase resilience of transport systems to future external shocks. University of Leeds, U.K., Leeds, UK

Philips, I., Clarke, G.P., Watling, D., 2017. A Fine Grained Hybrid Spatial Microsimulation Technique for Generating Detailed Synthetic Individuals from Multiple Data Sources: An Application To Walking And Cycling. Int. J. Microsimulation 10, 167–200.

Philips, I., Watling, D., Timms, P., 2018. Estimating Individual Physical Capability (IPC) to make journeys by bicycle. Int. J. Sustain. Transp. 12. https://doi.org/10.1080/15568318.2017.1368748

Tanton, R., 2014. A Review of Spatial Microsimulation Methods. NTERNATIONAL JOURNAL MICROSIMULATION 7, 4–25.

Wier, L.T., Jackson, A.S., Ayers, G.W., Arenare, B., 2006. Nonexercise Models for Estimating &OV0312;O2max with Waist Girth, Percent Fat, or BMI. Med. Sci. Sports Exerc. 38, 555–561. doi:10.1249/01.mss.0000193561.64152

Wilson, D.G., 2004. Bicycling science. MIT Press, Cambridge, Mass. [u.a.

Appendices

Appendix 1- Counts of unconstrained variable combinations in micro-data

Age/gender	Longstanding limiting illness	Economic activity	n
25-49Female	No	Employed	449
25-49Male	No	Employed	392
50+Female	Yes	Economically inactive	338
50+Female	No	Economically inactive	306
50+Male	No	Employed	286
50+Female	No	Employed	279
50+Male	Yes	Economically inactive	270
50+Male	No	Economically inactive	242
25-49Female	Yes	Employed	116
50+Female	Yes	Employed	108
50+Male	Yes	Employed	78
25-49Male	Yes	Employed	72
16-24Female	No	Student	60
16-24Male	No	Student	60
25-49Female	No	Economically inactive	59
25-49Female	Yes	Economically inactive	55
16-24Female	No	Employed	42
16-24Male	No	Employed	41
25-49Male	Yes	Economically inactive	32
25-49Female	No	Student	24
25-49Male	No	Student	19
25-49Female	No	Unemployed	18
16-24Female	Yes	Employed	17
50+Male	No	Unemployed	15
16-24Female	Yes	Student	14
25-49Male	Yes	Unemployed	13
16-24Male	Yes	Student	12
16-24Female	No	Economically inactive	11
25-49Male	No	Unemployed	11
25-49Female	Yes	Unemployed	9
25-49Female	Yes	Student	8
50+Female	Yes	Unemployed	8
16-24Male	No	Unemployed	7
50+Female	No	Unemployed	7
16-24Female	No	Unemployed	6
25-49Male	No	Economically inactive	6
16-24Female	Yes	Economically inactive	5
16-24Male	Yes	Employed	5
50+Female	No	Student	5
16-24Female	Yes	Unemployed	4
16-24Male	Yes	Economically inactive	4
16-24Male	No	Economically inactive	4
25-49Male	Yes	Student	4
50+Male	Yes	Unemployed	4
16-24Male	Yes	Unemployed	3
50+Female	Yes	Student	3
50+Male	No	Student	2
50+Male	Yes	Student	0