Division of Computing Science and Mathematics
Faculty of Natural Sciences
University of Stirling

Creating a Synthetic UK Population

Matthew Senseman

Dissertation submitted in partial fulfilment for the degree of
Master of Science in Mathematics and Data Science

September 2023

Abstract

The United Kingdom's census produces a large number of data tables that represent characteristics of both individuals and households in the United Kingdom (i.e., age, income, etc.), but the time and resources can be costly to examine all of these outputted data tables. There is also the issue that the United Kingdom census is taken every ten years, therefore the population makeup of the UK can change significantly in this time gap. Also of note, the United Kingdom census data is anonymized, meaning we do not know the true or exact characteristics that make up these households and individuals. These issues associated with census data tables could be solved by several different methods to improve census data usability and understanding. In this paper, the solution to these issues is by developing a synthetic population of the UK using this census data.

Through aggregating the different 2011 census data tables together for research usability, it is possible to then reverse-engineer the United Kingdom census data and explore a synthetic population that was created of the United Kingdom. Just as the census tables contain many detailing characteristics of the population, the newly made synthetic model also contains the individual and household characteristics, but on a singular data table for each of the 8480 Middle Super Output Areas (MSOAs). Simpler access and usage of census style data produces metrics that are easily accessed to look practically identical to the real United Kingdom population make-up. Once the desired household and individual characteristics are added to the synthetic data table, the simulated population can be queried to obtain the demographic makeup of the United Kingdom population without sifting through the numerous amounts of real census data tables. Through cautious and thorough programming, taking in all possible dependencies that characteristics might have with one another, a successful synthetic population can be produced from the original United Kingdom census data tables.

Because of the sheer number of data tables that are outputted by the United Kingdom census and a limited amount of research time, characteristics were chosen for households and individuals that gave a solid basis for a population. The tables used in this research project are Household Size, Household Composition, Age/Sex of Individuals, and Employment Status of Individuals by Age/Sex. Through a probability based iterative "bottom up" approach, the synthetic data frame was filled in using the following steps: input the name of one MSOA, initialise a spreadsheet, with each row corresponding to one household, apply a household size to each row, apply a household composition to each row (based on the household size), distribute individuals into households based on household size, composition, and existing individuals age/sex, and finally apply employment status to these individuals. This results in the probable household makeup of the inputted MSOA and an accurate synthetic United Kingdom population.

Based on metrics such as average ages of parents in United Kingdom households, average age difference between partners in households, and average household size it was possible to determine if the developed synthetic United Kingdom population was accurate. By comparing the previously listed averages provided by the UK's Office for National Statistics (ONS) with those same metrics extracted from the created synthetic population, there seemed to be little to no difference between them. Also of note, the exact number of households and individuals are added to each MSOA's created data frame, with the only difference being rounding errors that lead to a small difference in the number of household compositions. Overall, this model for developing a synthetic United Kingdom population was successful at its implemented scale and may be useful for researchers that need to use multiple census data tables but do not have the time or resources to effectively aggregate them.

Attestation

I understand the nature of plagiarism, and I am aware of the University's policy on this.

I certify that this dissertation reports original work by me during my University project except for the following:

- N/A

Signature *Matthew Senseman* **Date** *08/09/23*

Acknowledgements

I would like to thank my supervisor, Dr. Anthony O'Hare for guiding me through this research and supporting my methodology that was developed throughout the research process.

Table of Contents

List of Figures

1 Introduction

This dissertation introduces a new basic method for advancing synthetic population modelling in the United Kingdom, emphasizing the importance of building an understanding of both census data and synthetic data concepts. Numerous contemporary methodologies do already exist in the field, but the method presented in this paper narrows its scope to the UK using 2011 census data. The central methodology to this research is a probability based iterative bottom-up model for generating a synthetic population of the United Kingdom. The anticipated outcomes of this approach aim to efficiently aggregate census data tables, leading to enhanced accessibility, improved research applicability, and heightened overall precision of a synthetic UK population.

1.1 Background and Context

The United Kingdom's census data can be difficult to parse, as the answers pulled from the filled-out census forms submitted by the population are picked apart and placed into anonymized data tables. In order to effectively use the tables, one may need to perform a deep dive into the data to fully understand the syntax and formatting before use in research. In the case of this research, it is also useful to understand why and how synthetic data sets are created and used. Census data and synthetic data generation pair very well together, allowing for powerful algorithms to be developed that ingest the census data and output valuable synthetic data sets that could be used in many fields such as disease modelling or even used to train artificial intelligence models. There is a clear case for further development in the field of human data sets such as the census for the creation of synthetic data sets.

1.1.1 Census Definitions

Every 10 years the census is taken in the United Kingdom by the Office for National Statistics (ONS) in England and Wales, then the National Records of Scotland and Northern Ireland Statistics and Research Agency also contribute to census data in their respective regions [14]. It is important to note that the census is compulsory by law, so the data gathered can be as accurate to the whole UK population as possible. For this project, the UK's 2011 census was used to create the synthetic population and it is important to note that questions and formatting can change between census taking years. According to the ONS website, the 2011 census contained "56 questions on the questionnaire: 14 about the household and its accommodation and 42 for each member of the household" [15]. Only four questions were used in this research project, so there are many more household or individual characteristics available like health, national identity, or religion. Some of the data taken in the census is also based on the date that the census is taken and submitted by the population, for example any statistic that uses age is based on the individual's age at the day of the census, which was on March 27th in 2011. All of the data collected is then anonymized and used only for statistical publications by the ONS, not even other government departments have access to the de-anonymized data sets. It is clear that the UK census is a highly useful tool for assimilating accurate and comprehensive data about the country's population. The resulting data tables are essential for effective governance and resource allocation in the UK's communities by allowing governments or even businesses have accurate information to aid their populations.

1.1.2 Synthetic Data

Synthetic data generation is found in many research areas that overlap with the tools or skills used in data science, making it an ideal technique to be used on large and complex census data sets. It is the intended purpose that a synthetic data generation method creates an artificial data set that mimics the characteristics of a real data set, maintaining accuracy and privacy of

the data. Synthetic data can be generated using several different algorithms and statistical methods to resemble the distributions and relationships found in the true data [16]. There are generally a few methods to generate synthetic data that varies by type of data and desired outcomes. Statistical methods, such as the one used in this project, involve using statistical distributions to generate data points that resemble the distribution of the original data. This seemed to be the easiest to implement method for these census data sets. There are also more complex models called "Generative Models" that utilize machine learning to detect and learn the underlying data distribution from the real data and generate new data points that closely match a distribution [17]. The generated synthetic data should be evaluated to ensure its quality and trueness of fit to the original data. By comparing statistical metrics such as the mean or median between the original and synthetic data to ensure they match up with high accuracy.

1.2 Scope and Objectives

There exist numerous techniques in the field of synthetic population modelling that vary in scope and objectives of the research, this trend continues with the method introduced in this paper. The overall goal of this project is to investigate and develop a different approach to modelling a synthetic United Kingdom population by utilizing census data. With this goal in mind, the scope and objectives of the research project can be set and accomplished with accurate and impactful results.

The scope of this research has been limited to the modelling of a synthetic United Kingdom population using only UK census data from 2011. Much of the focus in the project will be on being able to effectively aggregate census data tables into a singular table outputted by a code base with high accuracy results. There also exists several limitations of the problem approach that narrow the way which the results of the research are outputted. These come as a result of the disposable computational power of the machine this code base was developed on, as well as the time availability to efficiently and accurately include more data tables from the UK census. Due to these limitations the scope of the project is reduced to only include certain data tables from the census, so the resulting synthetic population is also limited in information contained in it. Also of note, the use of this research method as a means of creating a synthetic population are geographically limited to the United Kingdom. Therefore, any further use of the research or observations about this research can only be used in the context of the United Kingdom's population and census data. Even further reducing the scope of this project is that the method only uses data tables from the UK 2011 census. Some of the syntax contained in the data, and therefore the coding, is unique to the 2011 census tables so this is critical to take into account if the project is to be utilized or updated any further. The creation of a synthetic United Kingdom population from census data was developed by considering the following project objectives:

1. Aggregate the 2011 United Kingdom census data tables for further research usability in the field of synthetic population modelling.

2. Reduce access time for those users and researchers that want to use multiple UK census data tables without having to access them one-by-one.

3. Produce an accurate synthetic United Kingdom population that contains useful household and individual level characteristics.

1.3 Achievements

It is necessary to unbiasedly evaluate the success and completion of the research as set by the objectives in the section above. If the overarching project goal and all objectives are entirely completed with accuracy to a satisfactory level, then this research project can then be deemed

as successful. At the level of this research the achievements are not necessarily ground-breaking, but it is desired that the research is seen as a success and merits further use of the methods as presently described within this paper.

Objective 1, which can be considered the primary motive for this research, can be deemed as successful through evaluation of some chosen accuracy metrics found in the *Results* section. The chosen census data tables were extracted from the UK data service and successfully loaded into the synthetic population and aggregated together. There was no significant loss of data during this process, therefore the resulting synthetic population that was created through the data aggregation can now be queried by researchers that wish to use the data for their research in synthetic UK populations. More data tables can be aggregated as described in the *Further Work* section, but the current model was successful in aggregating the originally desired tables for research use.

As a somewhat dependent result of objective 1, the results from objective 2 can also be considered successful. Through the success of aggregating the census data tables into one, researchers using this method only need to query the synthetic data table versus all census tables one-by-one. Users can ask for the characteristics included in the synthetic data at the same time such as age and sex of individuals in a household, this prevents them from having to download the tables and extract these characteristics themselves. Therefore, the time it takes to go through the process of getting multiple information fields from the census data is reduced significantly through the use of this research method.

It can also be said that objective 3 for this project was also successful as the metrics in the *Results* section indicate. The synthetic population that was created though this method was measured and evaluated to find discrepancies between the census data tables and the synthetic data. This resulted in little to no significant difference between the numbers or household and individual characteristics in the true census tables with the synthetic census tables.

Through the successful completion of all three objectives set forth for the project, it can be said that the creation of a synthetic United Kingdom population using 2011 census data was achieved with satisfactory results.

1.4 Overview of Dissertation

The project paper has been written as five sections; two sections as introductory and concluding sections and three more technical sections between them. These sections of the paper and their structure are outlined below:

1 – *Introduction*: The introduction section of this paper provides the reader with a background of the subject and problem, then details the necessity of the research that was carried out. It also includes the problem scope and objectives of the research so the reader has knowledge on what is trying to be accomplished with the research. Then the introduction provides a concise summary of the achievements completed throughout the development of this research project.

2 – *State-of-the-Art*: This section gives the reader context on problems in the research field addressed in this paper. It gives depth to the current findings on the topic(s) related to this project's research. Much of the section describes other research methods that have been released on the field of synthetic population modelling around the world. The section also includes a review of the current technological tools that have been developed for use in the research field.

3 - *Methodology*: This section provides the methods, detailed steps, and the technology tools used for the research. The inclusion of the methodology section is to allow for the reader to

see the step-by-step process that was carried out to accomplish the research and allow for a degree of reproducibility. Background on the datasets used and how the research was accomplished are displayed in order of use to allow the research process to be more easily followed.

4 – *Results and Discussion*: The results of the method and research performed are presented in this section at the household and individual levels respectively. Much of these results are accompanied by visual indicators such as graphs as a means to display the results accuracy. These results are then discussed in depth to present both the success of the method and the flaws that it may have.

5 - *Conclusion*: The conclusion section of this paper demonstrates a summary and brief evaluation of the processes and of the research that was carried out. It also presents potential further work that could be done utilizing or improving upon the research method written in this paper.

2 State-of-The-Art

The field of synthetic population modelling is a fascinating field that blends computer science, statistics, and social sciences to create simulated populations that mimic real-world behaviours and characteristics. Created synthetic populations can then be used for various simulations and analyses where using actual data might be computationally expensive or ethically sensitive. This technology in the field has developed into applications that can be used in research fields such as urban planning, disease spread analysis, transportation modelling, and artificial intelligence or machine learning. Researchers that want to develop a synthetic population typically start with real-world data sources such as census data, surveys, and other relevant datasets to their field. These data sources provide basic information about demographics, household structures, commuting patterns, and other attributes that can be manipulated into a synthetic population.

2.1 Review of Current Works

The field of synthetic population modelling contains numerous works that have been effective in their respective scope and objectives. [1] creates a synthetic population dataset for individuals in Great Britain. The dataset is based on a few sources including the United Kingdom Census and a few outside survey datasets. It contains detailed attributes for each individual, such as age, sex, ethnicity, education level, employment status, health status, and socio-economic characteristics. The model of the population is at the Lower Super Output Areas (LSOAs) census data level in comparison with the MSOAs used in this dissertation paper's method. The authors argue that the resulting synthetic population dataset can be used to estimate a wide range of health and socio-economic outcomes at the LSOA level, with example results for distinct sub-population groups. Much of the method used in the authors' research is successful in the completion of its objectives, but there are some small issues that are present. The method is based on a model of the population at a single point in time. This means that it cannot be used to estimate changes in the population over time. The method is also sensitive to the quality of the data used to create the model. If the data is inaccurate or incomplete, this can lead to errors in the estimates produced by the model. But all in all, these types of issues are present in many of the current synthetic population creation methods.

At another national level scope, the creation of a synthetic population using Canadian census data created a synthetic population of the country [2]. It uses multiple sources of data from the Canadian census to generate a synthetic individual population that can be assigned into households, which are then assigned a household type. The method can then use population projection data provided by the Canadian government to project growth in their synthetic model and compare it to the government provided projections. Validation on the synthetic data characteristics was conducted by resolution level, i.e., city or national level, which were taken from 2021 census characteristics data tables. A correlation was found between the synthetic population and the census population for the majority of census characteristics used. Although the authors do note that the synthetic population is less reliable in areas where there have been significant changes in land use since the 2016 census data was taken. They also found that the synthetic population is less reliable for categories that represent a small proportion of the population, such as households of 5 persons or more and one-parent-families. But overall, the authors show that the synthetic population is a valuable tool for understanding and projecting the population of Canada.

One other application of synthetic population modelling applies the concept to the field of transportation and movement of peoples within a modelled population. The paper [20] proposes a replicable and reproducible process for synthesizing travel demand for Paris and Île-de-France. The process in the research is using open-source data that is available to the

public in France, and it also utilizes open-source software. The process generates daily activity patterns for each person in the synthetic population, which are generated using a statistical matching approach algorithm. Then the locations of activities in the synthetic population are assigned using a space–time prism concept. The performance of the proposed method is evaluated by comparing it to a reference travel demand data set. The results of this comparison show that the proposed method is able to generate a synthetic travel demand data set that is highly comparable to the reference data set. The authors argue that the process is the first entirely reproducible travel demand data set that can be used and altered at any time by any researcher. They also argue that the process can be used as a blueprint for other environments. However, it is noted that this method could also be improved by incorporating more data sources and making the assumptions the model utilizes more transparent to the user or reader.

2.2 Existing Tools and Packages

There exist a few tools in synthetic population modelling that can aid researchers in producing a reliable code base for developing the models. One tool that generates synthetic populations for use in urban planning and policy analysis is PopGen [18]. PopGen starts with individual-level data and probabilistically reconstructs a synthetic population dataset. This synthesized dataset enables researchers to use it in urban planning and other fields that require accurate and fine-grained population information. However, the accuracy of synthetic populations from PopGen depend on source data, which might limit its applicability in regions or domains with limited data resources.

Another existing method present today is SynthPop, an R package for generating synthetic populations [19]. The package utilizes statistical matching techniques to ensure that the distributions of variables in the synthetic population closely resemble those of the actual data. One of the strengths of the SynthPop is its ability to handle complex data structures, including categorical and continuous variables, as well as relationships between variables. But the package to be effectively used requires a solid understanding of statistical concepts and the specific methods implemented in the package. This complexity could be a barrier for users who are less familiar with these concepts.

One other programming package that is useful in this topic is the HumanLeague package available for use in R. The development background of the project, found in [21], introduces a novel approach to population synthesis, a technique intended for use in urban planning and transportation modeling to generate realistic synthetic populations. The key idea of the paper is to employ the developed quasirandom sampling techniques, specifically Sobol sequences, to improve the process of assigning attributes to synthetic individuals within a population. These quasirandom sequences provide better coverage of the sample space, reducing the likelihood of clumping or overrepresentation of specific subpopulations. A set of statistically significant metrics of the sampled population are outputted to determine the accuarcy of the generated population. But similar to other programming packages mentioned above, this method struggles with the need for high-quality data table inputs and user knowledge of programming systems or highly technical syntax.

3 Methodology

The methodology for this project was driven primarily by the goals set out in Section 1.2 but was also driven by trial and error towards initial stages. As more data tables from the census were incorporated, it became easier to understand the structure of the data better and the requirements needed in the code in order to effectively utilize the census data and avoid data loss of pertinent information. Much of this method was not pre-planned and was carried out on a table-by-table basis, dependent on the exact structure of the census data table and the dependency on existing structures within the synthetic population. The methodology was also developed through ease of usability, meaning the structure of the code should be more easily accessible or readable. By using Jupyter Notebooks and Python's Pandas package extensively, this allows for the code base to be edited by users of the research method to suit their needs. It also allows for the addition of new characteristics into the codebase by simply downloading new data into the code and following a similar methodology to what already exists as detailed in this section.

3.1 Overview of Methodology

By using the following steps, a probability based iterative "bottom up" approach (starting with the larger scope households and ending with smaller scope individuals) to synthetic population modelling of the United Kingdom population utilizing 2011 census data was developed.

1. Identify and download the desired United Kingdom census data table, containing a characteristic for UK households or individuals.

2. Import the downloaded data table into the Jupyter Notebook using the Pandas library, keeping only the necessary columns.

3. Import and input the name of one MSOA (found in any of the census data tables) to initialise the creation of a synthetic MSOA

4. Give each row a unique Household ID (i.e., HH123) and apply a household size to each row.

5. Apply a household composition to each row, based on the size of the household and the remaining rows without a composition assignment.

6. Distribute individuals into households based on household size, composition, and existing individuals' age/sex.

7. Assign employment status to individuals in the MSOA, dependent on age/sex of an individual.

Note that if or when further data tables are added, therefore adding new characteristics to individuals and households in the population, then a slightly altered or extended methodology is needed, but the basic process remains the same. This is explored in more detail in Section 4.3

3.1.1 Technology Toolkit

Jupyter Notebook was identified as the simplest and best option for maintaining and editing the code base the method relies on. Python is well integrated into the notebook(s) and the visualizations and tables that are outputted into the notebook allow for testing to be easily validated for accuracy.

Python was chosen as it is familiar to those in data science research and industry and contains a number of useful packages when dealing with tabular data. The included packages used in the project are:

- *Pandas:* To create data frames that held all information for the project.
- *Itertools:* For use of the dictionary flattening function 'chain.'
- *Random:* Used as the basis for defining and selecting from probability distributions.
- *Collections:* For use of the list to dictionary function 'Counter.'
- *Math:* For use of the 'ceiling' and 'floor' functions when rounding numbers.
- *NumPy:* For use in the results section when making calculations.

Hardware was a limiting factor for analysis of project success (discussed in depth in 4.3) but the current format of this method can still be performed under these hardware specifications:

- *Processor:* 2.7 GHz Quad-Core Intel Core i7
- *Graphics:* Intel Iris Plus Graphics 655 1536 MB
- *Memory:* 8 GB 2133 MHz LPDDR3

For further work to be performed using this method, including running more than one MSOA at a time, it is suggested that the hardware be significantly more upgraded, or a cluster computing network be utilized.

3.2 Exploration and Preparation of Data

The United Kingdom census data website and tables are complex and contain repetitive or unusable information for use in this method. To reduce an overload of information into the final results of the synthetic population, one must pick and choose which data tables from the census will explain the desired population traits best. There are a set of characteristics that every household and individual require for usability by practically every research field that utilizes synthetic populations. The goal when exploring and prepping the data tables in this method is to ensure some improved form of simplicity of the information being extracted from the tables. When the data tables are aggregated together into a single table, their information is able to be separated apart again to get back the original data tables.

Figure 1. Representation of how the data tables were used.

When exploring all of the census data sets one can see that there are four columns at the beginning of the sheet containing information about the type of output area. In the case of the data used in this research project, all data tables used and manipulated were at the Middle Super Output Layer (MSOA) level. The corresponding first four columns of the census data tables therefore contain the geographic label, MSOA name, and output area type of each of the 8480 MSOAs in England, Scotland, and Wales. For this research project, only the first column containing the name of each MSOA was utilized in the code, although the geographic label code in the second column could prove useful in future work that wanted to create geographic maps and plots of the data. The remaining columns in each sheet contained varying levels/numbers of a characteristic of individuals or a household in an MSOA. Some characteristics or data tables become arbitrary for use when it is anonymized and released for

use. These arbitrary characteristics should not be used in this model as they would be difficult to accurately assign due to a lack of probabilistic certainty. Instead, a set of basis characteristics are needed for the model to be used at any capacity. These basis characteristics include household size, household composition, individual's age and sex, as well as individual's employment status. All 4 of these characteristics form a solid basis to be able to explore a population's make-up and allow for this population to be manipulated for further research purposes and improvement of the method itself.

3.2.1 Household Size Data

Because this method is described as a probability based iterative bottom-up approach, the first data table to be synthesized and added to the method's synthetic population data frame should be the size of households in a given MSOA. The data table provided by the UK Data Service, Household size 2011, is a cleaned and well-presented data set that gives the information for each MSOA. It contains more columns than necessary for use, so only the following columns were used in this project's method:

GEO_LABEL
Household size : Total\ Household size - Unit : Household spaces
Household size : 1 person in household - Unit : Household spaces
Household size : 2 people in household - Unit : Household spaces
Household size : 3 people in household - Unit : Household spaces
Household size : 4 people in household - Unit : Household spaces
Household size : 5 people in household - Unit : Household spaces
Household size : 6 people in household - Unit : Household spaces
Household size : 7 people in household - Unit : Household spaces
Household size : 8 or more people in household - Unit : Household spaces

Figure 2. Columns used from the Household size 2011 data table.

The column GEO_LABEL corresponds to the MSOA's name (i.e., 'City of London 001'), and the remaining columns provide integer numbers of households that correspond to how many households contain the given number of people. So, a 1-person household is clearly only one person, but a household size of 8 or more can contain 8 people (related or unrelated) or more. The "or more" part of this column, which is an ambiguous description, can cause issues with this method. This is not a definitive size of household, a household could contain anywhere from eight people to 100s of people, this issue can lead to some discrepancies between the real make-up of households in the MSOA and the synthetic households created from this data. This issue will be addressed more in-depth in Section 3.3.3, but for now these "8 or more" households are treated the same as the other sizes.

By knowing the sizes of households in an MSOA, it can become clearer as to how much interaction potential a household has based off how many people are moving in and out of the household. But the household size characteristic is also useful intrinsically for this method, as it provides a basis for other characteristics to build dependencies upon it. It would be impossible to find how many individuals should be placed in a synthetic household without this data table's information.

3.2.2 Household Composition Data

Even further than just household size, the households of each MSOA require more in-depth descriptions to reduce the uncertainty in their usability. An important characteristic of households is their composition, or what type of people reside in each household. This not only builds upon household size but also creates another basis for how to assign individuals into households more accurately for research usability. The UK Data Service takes in the census answers and processes the responses about types of individuals in the household and produces the Household composition 2011 data set. The resulting columns in this data table, as it contains many variables, can be seen in detail in Appendix 1 Figure 12.

There exist several reasons behind the inclusion of the household composition for the purposes of this research method. One being the addition of more information to a household, therefore making it easier for the assignment of individuals into these households. Composition provides some details of age and sex of individuals that could potentially exist within these households, so when it comes to selecting individuals to be placed in these households it can be known with a higher certainty if they belong in a certain composition. For example, it is known with absolute certainty that "Household composition : One person household\ Aged 65 and over - Unit : Households" will contain a person over 65 by themselves. But some of the composition types introduce a degree of ambiguity into the method. Primarily, if a composition has "2 or more dependents" in the household then it cannot be certain if the household has two or three or more dependents and leads to some conflict when assigning the composition to a given household size. This ambiguity is addressed further in Section 3.3.2.

3.2.3 Individuals by Age and Sex

Individuals encompass many detailing characteristics that are asked by the UK's census and processed into data tables by the UK Data Service. However, individuals within a population need the very basic characteristics of age and sex in order to be useful in a research capacity. Of course, the census asks for the age and sex of individuals in the household, but because of the UK census data becoming processed and anonymized, the resulting data tables can only list the number totals of each age and sex in the MSOA. In this "bottom up" approach to a synthetic population, the households are created first as the bottom or base of the model, and then the individuals are created and distributed. This means a probabilistic certainty can be introduced for whether or not an individual with a given age and sex can exist in a household. The ages range from "under 1 year old" to "85 or over" and the number of each age and sex vary by MSOA. Some areas have a lot of younger adults and college aged persons; some have a higher average age and contain many persons over the age of 65. Note that a table representing this data is not present in the appendix because its structure in the used census table contains labels as simple as: "Age : Age 68 - Sex : Males - Unit : Persons" with 250 in an MSOA.

3.2.4 Employment Status by Age and Sex

While some individual characteristics have little to no usability in this algorithm, such as "Skills in Welsh", the employment status of individuals by age and sex is a simple data table to implement and add usable information for further research. Looking at Appendix 1 Figure 13, the individuals are broken up into age groups and sex with employment status part-time or full-time. Initially, the UK Data Service data table for this characteristic included exact number of hours worked, this was omitted in favour of either the full-time or part-time descriptor as a way to reduce complexity and time in creating the synthetic data from this table. In the case of census data, full-time working is defined as working 31 hours or more per week, and part-time working is defined as working 30 hours or less per week [3]. The inclusion of the employment

status data is to allow users of this method to be able to infer the movement of individuals to and from their houses to some degree of certainty.

3.3 Problem Approach

This method incorporates a “bottom up” approach to synthetic population modelling, the bottom being the households in the population and the top being individuals with characteristics. By starting with the creation of households it can be easier to work our way towards a finer resolution of population with the assignment of individuals. Essentially the base of the method is the larger in scope households, built upon these households are the smaller in scope individuals, then even further reduced in scope the attributes of individuals are added.

3.3.1 Probability Methods Used

The inclusion of a select few probability sampling methods is necessary due to the anonymous and ambiguous nature of the census data being used in this research method. By incorporating the sampling of a true population, a synthetic population can be developed to look as close to the true population as possible. When comparing the distributions of household and individual characteristics they should appear identical as the selected characteristics add up to the total population. Where issues arrive is whether or not the combination of these characteristics represent a real household with real individuals in the chosen MSOA. But it is not possible to know if the synthetic households exist as real households in the MSOA, that is where the method relies on probability distribution sampling to create a set of the “most likely” household make-ups.

Almost all parts of this synthetic population method utilize the very basics of probability by a simple random probability distribution sampling method [6]. This sampling method selects a sample, in these methods case a household or individual, from the larger population of households or individuals. By selecting a sample this way, every individual or household in the population has a known chance of being chosen based upon its proportional existence in the population. For example, if 4 out of 30 households are of the size 6 persons, then the simple random probability of being sampled would be 4 in 30. This random sampling method ensures that the sample is representative of the entire population, making the results more reliable and generalizable. As more samples are taken from a population, the closer the mean of the sampled population approaches the mean of the true population. In this research method, by taking a household or individual characteristic and applying it to a household, a sample of the whole population is taken and given to the synthetic population. As more samples are taken from the census data tables, the most likely household is created based on a simple random probability distribution sampling.

Even further utilizing the simple random probability distribution sampling in this research method is a more biased version, judgement sampling. By adding a degree of bias to a probability distribution the uncertainty in assignment was reduced and unwanted households or individual characteristics were removed from the equation. Judgement sampling can be defined as “the researcher using their judgement in selecting the units from the population for study based on the population’s parameters” [7]. This probability distribution method was used extensively when defining the distributions associated with selecting individuals by age and sex from the population. Through defining an age range and only taking certain sexes in some compositions, judgement sampling was introduced to the model. If an individual likely would not exist in a certain household composition, they would be rejected in favour of an individual that could exist. By using prior knowledge of the distribution, a more favourable set of households were created that avoided unlikely household make-ups and ensured that the synthetic population was able to represent the real population.

3.3.2 Household Creation

By first developing a set of potential households in an MSOA, the building blocks of a synthetic household were established for reference in all other aspects of the synthetic population. The total number of households is referenced, and a blank data table is created using Python Pandas library containing the same number of rows as there are households in the given MSOA. For further usability purposes, each of these households are given an identifier i.e., HH25 so the information contained in this row equivalent of HH25 has the ability to pulled from the data table easily. Then, from the household size census data table, a size or number of persons is applied to each of the households. The application of a size to each household is random and does not depend on any prior information as there is none available yet in the synthetic population, size is the first characteristic and defines the initial distribution for individuals.

While size of a household is important in and of itself, as detailed in Section 3.2.1, household composition brings the idea of a dynamic and varying household to fruition. The issue with assigning this new characteristic to households is determining if the size of the household is a realistic match for a household composition. By looking at Figure 12 in Appendix 1, there are several household compositions that can be seen as a "guaranteed" match for some household sizes. Of course, a composition that is a one-person household age 65 + or one person household age 0-64 are immediately be assigned to all available one person households. Any of the one family households with no children are a couple living alone, so these compositions are deemed guaranteed two person households. Similarly, some of the guaranteed three person households are assigned compositions such as a married couple with one dependent child. This removes a significant number of available compositions available for assignment to households from the total pool, but the remaining compositions do not have a guaranteed size.

When looking at these household compositions available from the census data there exists a degree of ambiguity amongst them. For example, a household composition that contains "two or more dependent children" is very ambiguous as it cannot be said with uncertainty that these compositions can be assigned to a definitive household size. This is where an iterative probability distribution sampling method of the remaining households comes into use.

$$(H_x / H_{tot}) * Comp$$

The number of empty household slots of size *X* remaining (H_x) are determined and then divided by the total empty slots remaining (H_{tot}), giving ratio that is then multiplied by the total number of a given household composition (Comp) that could realistically exist at H_x. This process was applied to every household size and household composition and the resulting synthetic households all contain a household composition characteristic that is appropriate given the household's size.

3.3.3 Individual Creation

While having descriptive households are useful on their own, a comprehensive synthetic population method requires an even finer scope on population that includes individuals themselves. Because this population is built upon households, a random distribution of individuals cannot be applied to the households as it would lead to a significant number of household compositions not making any logical sense. But the issue with the data set is because of its anonymity, it also cannot be determined the exact makeup of the households. So, there is no certainty with which individuals in the MSOA can be placed in their exact household with their exact family members. Just as the assignment of household composition,

the assignment of individuals samples from probability distributions of the population of individuals.

The first step for adding individuals into the synthetic population is initializing columns to be filled. Eight columns were created, one for each possible individual slot and all being labelled as empty by adding a “0” into each. The eighth individual column also serves as the “8 or more” individual slot so it may not be representative of just a single person, rather a group of individuals if the household size is greater than eight. This method of assigning individuals goes column by column, starting with individual one being the “reference person” of the household. Several other parts of the method utilize a reference person for creation of a synthetic population as detailed in Section 3.2, but those parts of the method tend to use a data set that gives the age and sex of a reference person guaranteed to be in a household. The method used in this research does not have access to a 2011 UK census data table that contains reference person information, therefore it relies on the first person assigned to a household as its reference person. Dependent on the composition type, the first person assigned to each household tended to be of dependent age 0-18 for household with at least one dependent. Other household types such as those without dependents were assigned an individual picked from a distribution of ages 18 or greater, and households that only contained an individual 65 or older were assigned an individual in that range. Now the entirety of individual column one is filled, and a reference person is in place for households of size greater than one.

A second individual is placed into the household based on the composition of the household, age, and sex of the household reference person in slot individual one. For example, in a household consisting of a married couple and one dependent child, the household reference person in slot one will be the dependent. Then, individual two is assigned as one of the parents of the dependent, selecting from a distribution of ages centred around 30 years older than the initial dependent [4]. Individual three is then assigned as the partner of individual two (if that makes sense to the household composition), selecting from both an age and sex distribution of individuals centred around a three-year age difference from their partner [5]. This continues up to individual eight, depending on the household’s composition and household’s size, with compositions such as “two or more dependents” being assigned more dependent age children than adults.

One issue encountered during this process occurs with the ambiguous definition of household size when a household is assigned the 8 *or more* size to it. This method handles this ambiguity by assigning a number of these households to be “communal spaces” such as student dormitories, elderly retirement facilities, or some other form of communal housing. By determining the remining individuals that have yet to be assigned, and the number of 8 or more sized households, a percentage of the remaining individuals are assigned to each 8 or more household. In some MSOAs there will be very few individuals remaining to assign to these slots, while others have many. For example, MSOAs in Oxford will likely have several student dormitories that account for 8 or more sized households. But after this discrepancy in the data is processed, all possible individuals have been distributed into households so that the total number of individuals in the true population match the total created in the synthetic population.

3.3.3.1 Employment Status

The United Kingdom’s census contains many questions that after processing, provide useful data sets detailing characteristics of individuals in a household. Only some though may be more useful for inclusion in this research method than others, such as the employment status of individuals in a given MSOA. This research method randomly selects a household, ensures that at least one individual in the household that meets the age and sex requirements, then gives that individual an employment status. Through this process, each household may contain

individuals with a part-time or full-time employment status based on the available numbers from the census data table. Those who are not assigned part-time or full-time employment status are taken to be not of employment age, unemployed, or retired. By including employment status, further research done using this method may be able to determine which individuals are at risk for a higher number of interactions. Other characteristics such as this one could be added to the method as long as the data available is able to correspond to existing individuals and households without a high uncertainty.

4 Results and Discussion

This section asserts measures of success of the synthetic population in comparison with the true population. These results are on a limited basis as without more computing power or time it is costly to compare the entirety of the synthetic UK population with the entirety of the true UK population. Instead, the results of this synthetic population research method are based on 10 MSOAs as well as demonstrating the success in solving the objectives set forth in Section 1.2. Much of the results contain a straightforward comparison between synthetic and true MSOA counts, while other results are measured by comparing means of the synthetic population to the equivalent relevant statistics provided by the Office of National Statistics. While this is not the best method for testing accuracy of a model, it is what was on hand at the time this research was carried out. A plethora of visualizations are included to represent to the reader how well this method works for producing a synthetic population that appears as close to the true population as possible. Because of this method's bottom-up approach to synthetic population creation, each level of the population from household to individual must be evaluated independently. The evaluation of success of the synthetic population must start at measuring how successful the creation of synthetic households was, then the success of individual assignment is evaluated because of its dependency on the success of household creation.

Mentioned in the sections below, the synthetic MSOAs tested gave accurate results when being compared with the true MSOAs. Below is an example (using MSOA 'Barking and Dagenham 012') of how the accuracy of this method was tested when comparing numbers of synthetic characteristics with the census stable numbers:

Difference in total households: 0

Household size : 1 person in household - Unit : Household spaces 0

Household size : 2 people in household - Unit : Household spaces 0

Household size : 3 people in household - Unit : Household spaces 0

Household size : 4 people in household - Unit : Household spaces 0

Household size : 5 people in household - Unit : Household spaces 0

Household size : 6 people in household - Unit : Household spaces 0

Household size : 7 people in household - Unit : Household spaces 0

Household size : 8 or more people in household - Unit : Household spaces 0

Difference by household comp: [0 0 0 0 0 0 1 0.0 0.0 0.0 0.0 0 0 0 1 0 0 1 0 1 1 0 -5]

Difference in total household comps: 0.0

Difference in individuals by age/sex:

[0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0 0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0

0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0

0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0

0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0]

Difference in total individuals: 0.0

Difference in persons by employed status:

[0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0

0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0

0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0]

Difference in total employed: 0.0

Results such as those shown above are consistently accurate throughout each MSOA and at both the household and individual levels of characteristics present in the synthetic data. The total numbers of all characteristics included came to a *zero* difference between real and synthetic data sets. But as seen in the results above, the only synthetic numbers that tend to differentiate from the census numbers are when household compositions are generated and rounded to whole numbers.

4.1 Household Level Results

The success of this project depends on how accurately the households that make up the MSOA were created, as this model is a bottom-up approach and households are the base element. Evaluating the model on a step-by-step basis allows for higher accuracy the further the model reduces in scope to individual characteristics level. The first aspect tested for accuracy is the creation of households; the number of households in the census data table compared to the number of households in the synthetic MSOA. Because this method begins by extracting the number of households in each MSOA from the census data, then creating a synthetic MSOA with that many rows, there should be no difference in these numbers. As observed when testing 10 MSOAs (see Appendix 1, Figure 14), the number of real households in each MSOA is equivalent to the number of rows and therefore number of households in the created synthetic MSOA. This is likely the simplest evaluation that occurs when reviewing the effectiveness of this synthetic population method but is also critical as to not develop a synthetic population on a poorly designed starting off point.

Household size is the next building block of the synthetic population, and by ensuring that the number of households of a given size are equivalent in both real and synthetic MSOAs then the foundation become even more solidified. Not so dissimilar from total number of households, evaluating the number of each household size utilizes the extraction of the original number of household sizes in the MSOA and distributing them randomly into rows of the synthetic MSOA. Due to a lack of dependency on other characteristics of the population, the household size data is easily transferred into the synthetic population with no data loss. When comparing the number of household size present I the synthetic MSOA with the true numbers from the census, there is no difference between them. But more information can be extracted and compared for household sizes, such as the average household size in each synthetic MSOA.

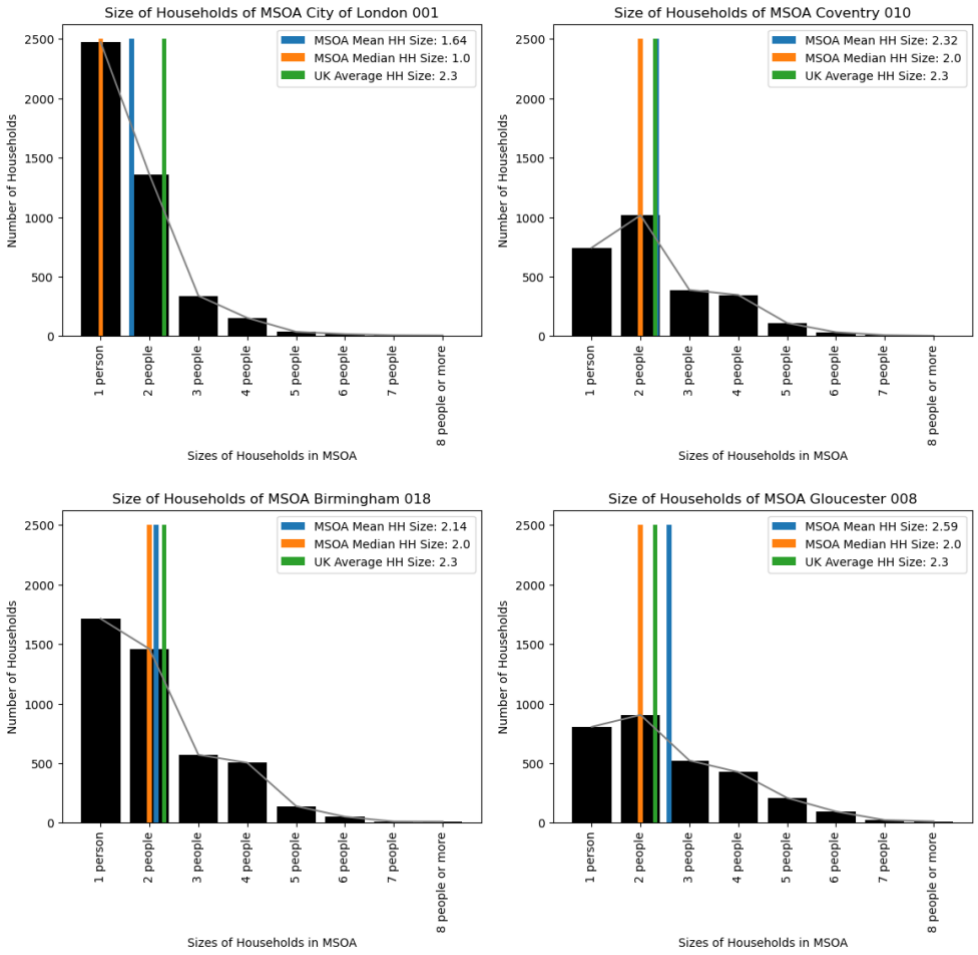

Figure 3. A comparison between average household sizes.

Figure 3 above demonstrates how easily the information about the synthetic households can be extracted and used to convey results. While the MSOA City of London 001 in the upper left may not be an average kind of MSOA (highly skewed towards one person households), if the average was taken over all synthetic MSOAs then an approach towards the ONS provided average of 2.3 persons per household would be observed [8]. This is a preliminary indicator of a successful generation of households in the synthetic MSOA up to this point, but note that further confirmation of these results is needed as discussed in the *Further Works* section. Having accurate household size for each household row allows this characteristic to be used as a dependent column with certainty if further use of household size data is necessary.

The last characteristic that is added to the creation of households in the synthetic population is household composition. Assigning a household composition to each synthetic household is largely probability based and also contains a dependency on household size, which leads to a small degree of difference between true and synthetic counts. The assignment of household compositions to households in the synthetic MSOA is based off of the methods describe in Sections 3.3.1 and 3.3.2 as well as the equation shown in Section 3.3.2. Because this produces an integer number of rows to assign a household composition to, rounding errors are made such as giving a household composition one too many rows or one too few.

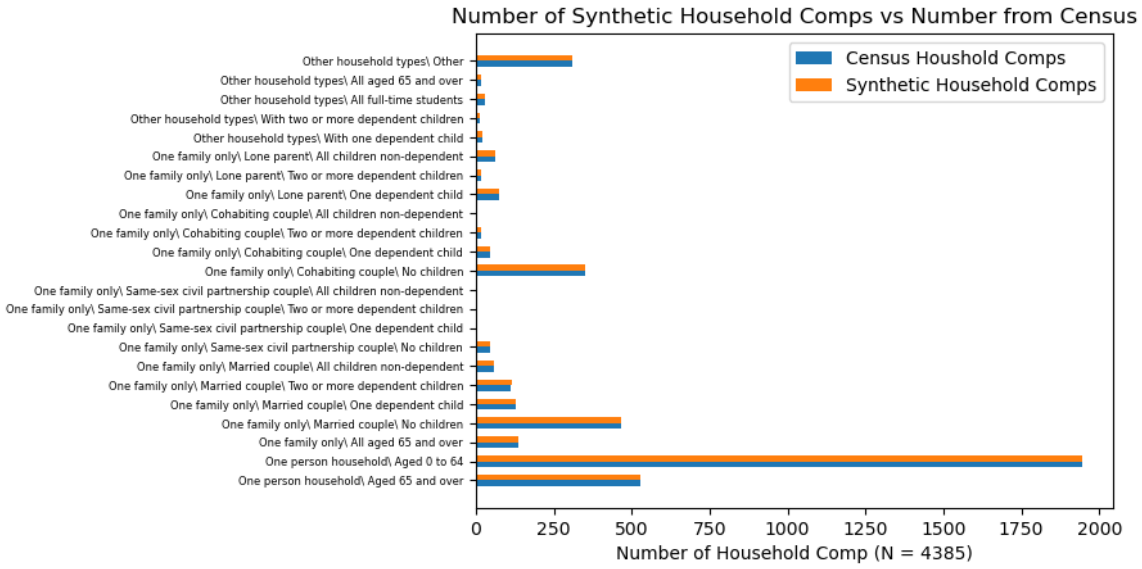

Figure 4. City of London comparison of real vs synthetic household compositions.

Of the 10 MSOAs (see Appendix 1, Figure 14) examined for this results section most did not vary by more than just a few counts of household compositions. In Figure 4 above it is nearly impossible to notice any differences between the counts of household compositions in the real and synthetic populations. Also of note, rounding errors do not affect the total number of household compositions assigned to the synthetic household rows. Even if a household composition has one less count in the MSOA, it is corrected by another household composition having one extra count, therefore the total number of household compositions assigned is equivalent to the total number of households in the real census MSOA. Because of this discrepancy, this has a small effect on how individuals are distributed into households due to this method's creation of a dependency between individual assignment and household composition. That being said, it is an insignificant effect compared to the total size of an MSOA and does not diminish the overall accuracy of the synthetic population creation significantly to cause concern over this method's integrity or usability. The assignment of the household composition to a household row in the synthetic population effectively matches the composition with a realistic household size and provides the synthetic population with a vital characteristic that can be used for further research on the topic.

4.2 Individual Level Results

A synthetic population can be used at just a household level to some extent, but the addition of individuals into the synthetic population increases the usability of the method exponentially. By analysing the success of individual creation in the synthetic population, several metrics are extracted and used to determine the effectiveness of tis research's methodology. Before these metrics are extracted, the individuals have to be found in the UK census data, manipulated for readability and usability, and then assigned to households based on prior information existing on the household. Just as with household data sets, the number of each individual by age and sex are taken directly from the census data table and assigned one-by-one into households, leading to no difference in the number of individuals in the real MSOA and the synthetic MSOA.

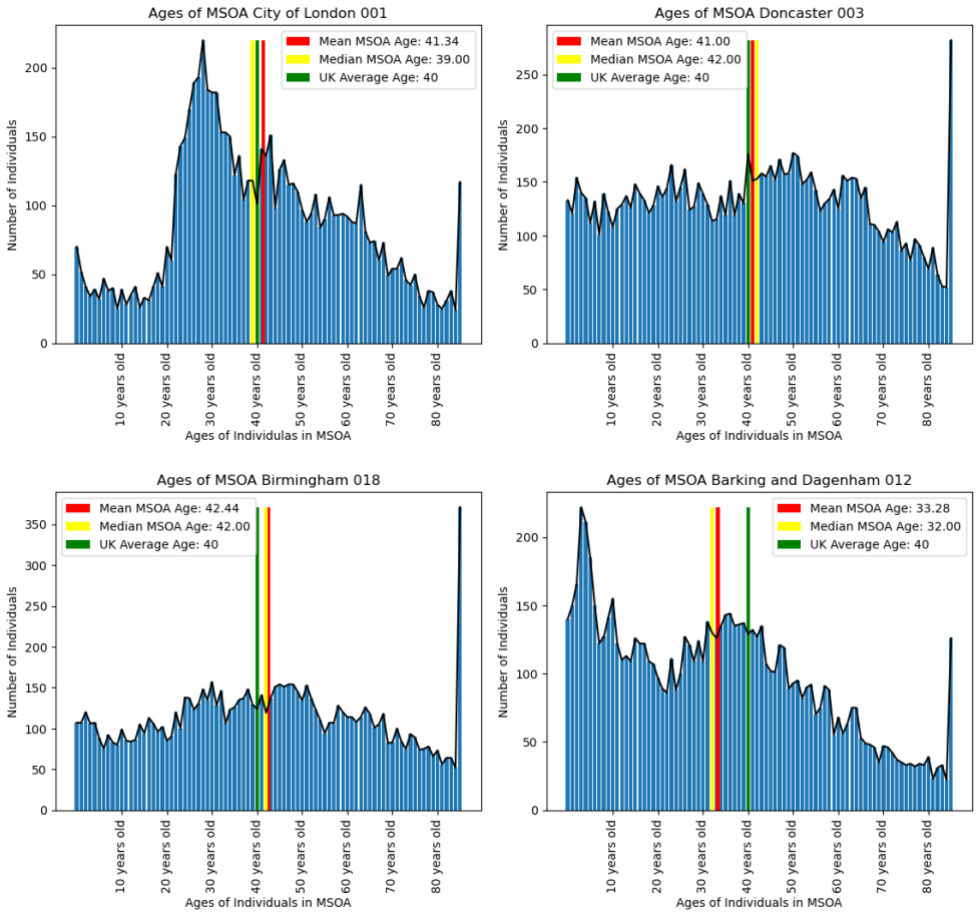

Figure 5. Number of Individuals by Age in MSOAs.

Figure 5 above displays the synthetic populations individuals by age, note that this population is also an exact representation of the real individuals ages in as taken from the census data table. It can also be seen that MSOA City of London 001 represents the UK population as whole quite well, considering the average age of the UK and the average age of the MSOA are close to equivalent. On the other hand, an MSOA like Barking and Dagenham 012 contains more younger persons, bringing down the average age. But once all the MSOAs are ran through this model, average statistics compiled and averaged for the whole UK, future users will see that the synthetic population is highly similar to the true population of the UK based on these statistics. City of London 001 can also be directly compared with real statistics because the ONS does output a statistic for the median age for the City of London, which is 37 years old [22], this number matches up exactly with the synthetic City of London 001 median age in the figure above. With more computing power, all MSOAs can be ran through this algorithm and the average age over all the MSOAs would more than likely be approach the UK average individual age of 40 years old [9].

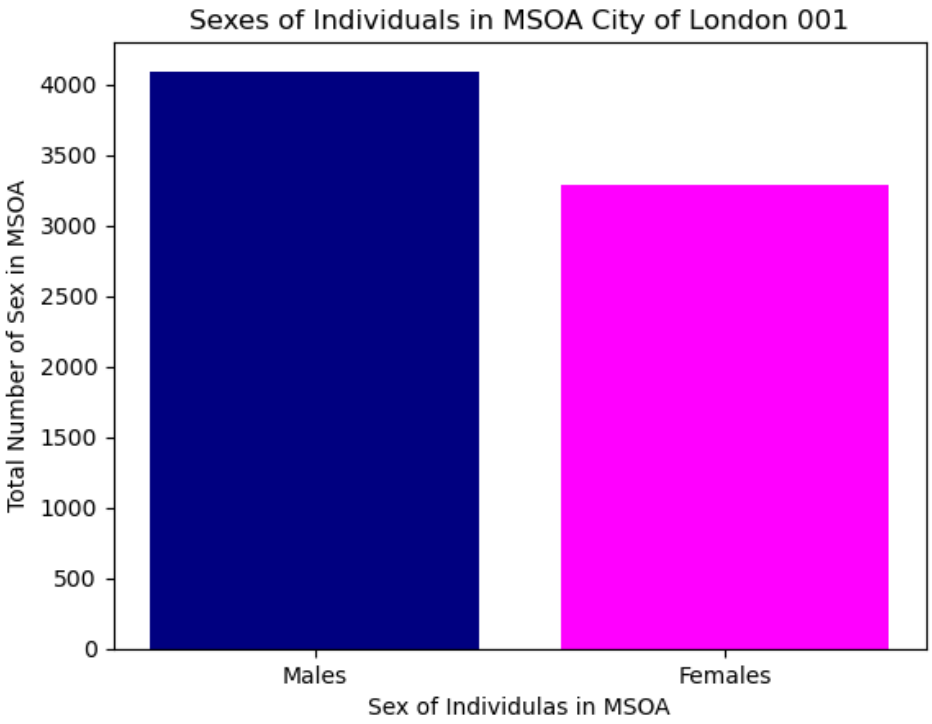

Figure 6. MSOA City of London 001 Sex Counts

Another quick check on the success of the individual assignment to the synthetic MSOA households is a measure of sexes of individuals present in the households. In Figure 7 above it can be seen that males outnumber females in the City of London MSOA 001, but this is not representative of all UK census MSOAs. The true ratio of females to males is 51:49 [10] but that is not the case above. This is no cause for concern as one MSOA cannot represent the whole population, so if all synthetic MSOAs were queried then the 51:49 ratio would more than likely be present. Just as with Figure 5, the figure above is a simple way of displaying the accuracy of this synthetic population method as the number of females and males taken from the 2011 census data table match up exactly with the synthetic population.

Although this method is successful with incorporating individuals by age and sex into a synthetic population with no deviation from the original data tables, there are more metrics that can be derived. It is important that the sum of characteristics of individuals in a household allows for statistics to be extracted that describe their households with individuals inside them. By proving that the individuals were placed into households with a higher degree of certainty, this method can be proven effective as an accurate representation of the United Kingdom's population. A sense of the synthetic population's make-up can be developed by looking at statistics such as average dependent age, average parent ages, and the age difference between parents and their children.

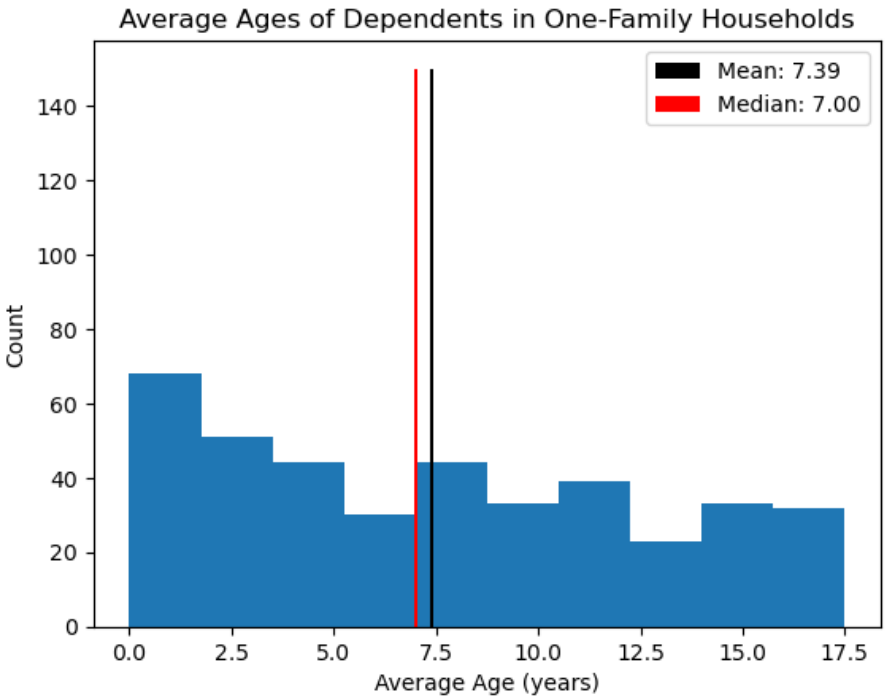

Figure 7. MSOA City of London 001 Dependent Ages

Figure 7 includes the ages of all dependents, ages 0 up to 18, in a given MSOA. Statistics such as this prove this method's usefulness as a means to extract metadata on the synthetic population. If statistics such as average dependent age hold true for all MSOAs, then this method is proven useful for researchers to be able to extract data on a population without having to aggregate the census data table themselves.

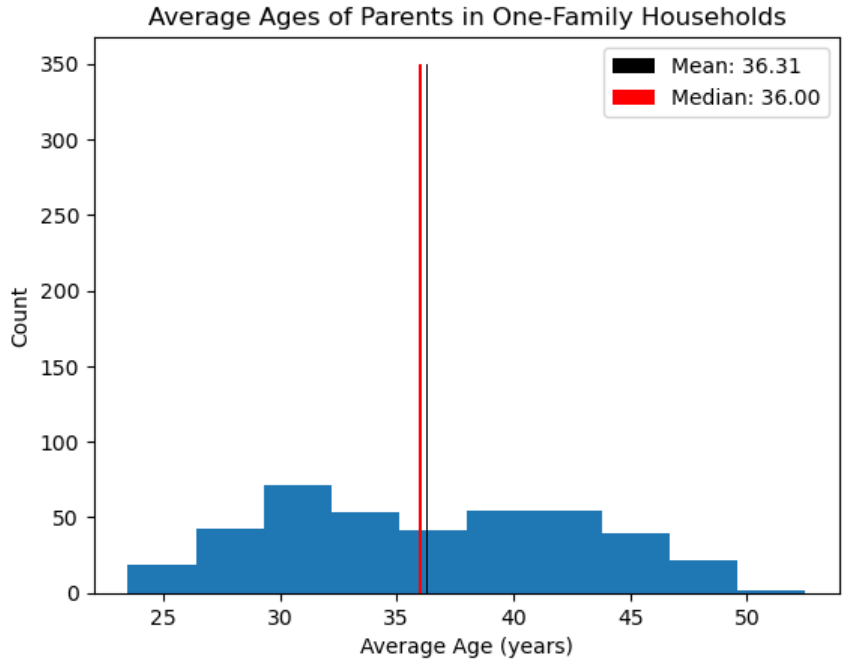

Figure 8. MSOA City of London 001 Parent Ages

Similar to dependent ages, parent ages in the given MSOA are extracted to give the user an idea of the synthetic population make-up. The synthetic MSOA is than queried reasonably easy for any statistics similar to these and taken as the true census population age statistics. Although, the addition of more computing power to calculate these age statistics for the entirety of a synthetic UK population would be more useful and easier to confirm as statistics for the UK population as whole are more widely available. But this research's analysis of 10 MSOAs (see Appendix 1, Figure 14) demonstrated the methods reliability to a certain degree, especially as more descriptive statistics of the MSOAs are extracted.

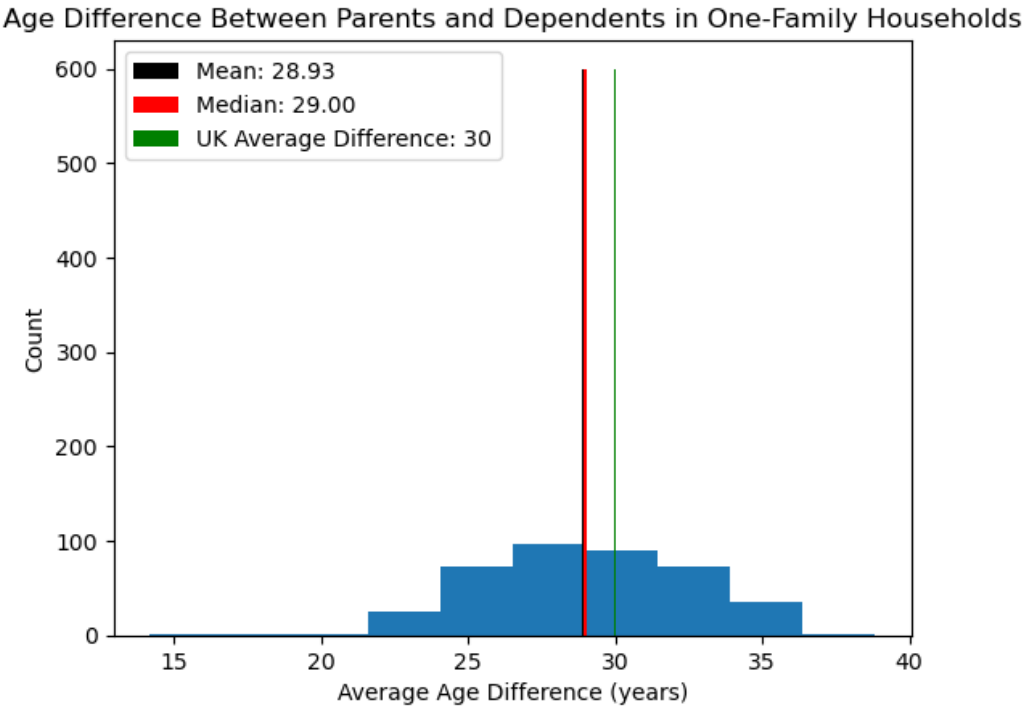

Figure 9. MSOA City of London 001 Difference in Age between Dependents/Parents

One metric that was used as a measure of success of the synthetic population creation and the success of individual household assignment is the difference between dependents and their parents in an MSOAs household. As seen in Figure 9, the average age difference for MSOA City of London 001 differed only slightly from the UK's average [4]. The results of the other 9 MSOAs tested produced similar results to this one. But just 10 synthetic MSOA's results cannot be a representative of the synthetic population's results as a whole, more MSOAs will need to be analysed for confirmation of this method. But the success of the 10 synthetic MSOAs created demonstrates this methods success in assigning individuals into households that correctly fit their age and sex with existing members of the household. If there was a more significant difference between the 10 synthetic MSOA averages and the true UK averages, then this synthetic population method would need to be re-evaluated and tuned to find a more accurate method.

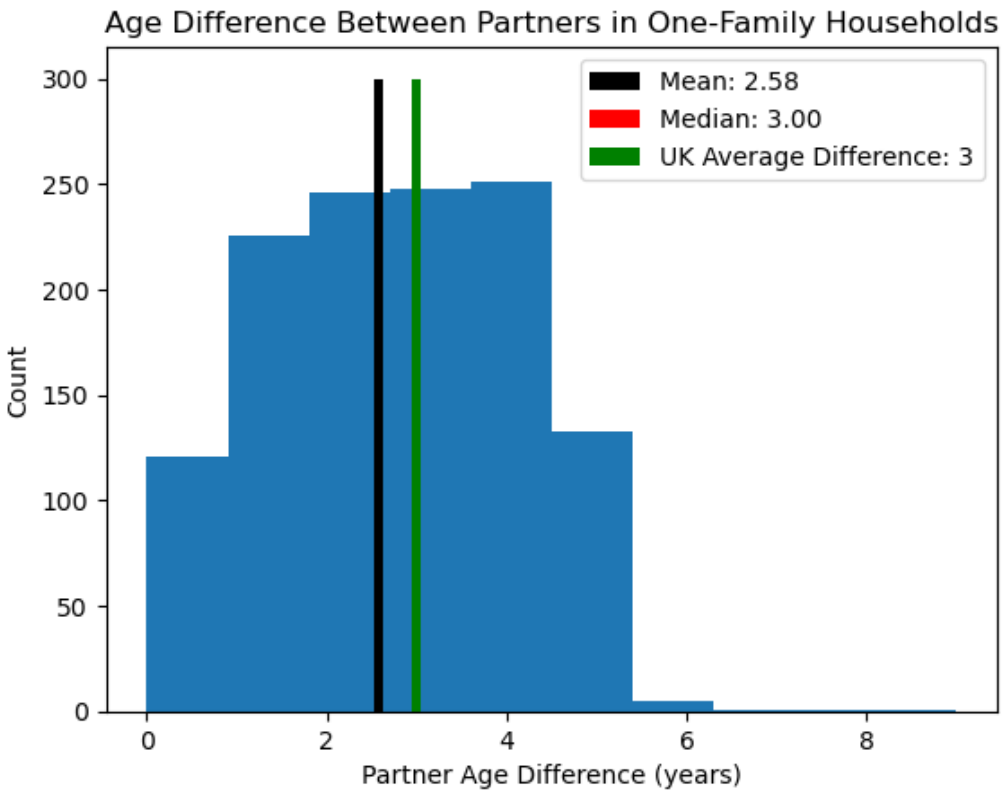

Figure 10. MSOA City of London 001 Partner Age Differences

The final metric that was extracted from the synthetic MSOAs was the differences in age between partners of all types in the synthetic households. When assigning a partner to a household, a distribution was taken starting at $\pm$ 3 years the prior existing partner's age. If a suitable partner of age and sex for the existing partner did not exist in the pre-defined range, then one year was added to each end of the distribution. This methodology allowed for most households to fall around a pre-determined average of 3-year age difference between partners in households [5]. A background and reasoning for the use of this probability distribution in this way is found in Section 3.3.1. Because of this assignment method, the synthetic population shows success when gathering metrics such as the one shown in Figure 10. Overall, the creation and assignment of individual to the synthetic population can be seen as successful in this research's scope based upon the lack of difference across all aspects of individual creation seen in Figures 6 through 11.

4.2.1 Employment Status Results

Adding characteristics to individuals in a synthetic MSOA can be difficult because of a lack of reasonable existing dependencies and the anonymity value of the characteristic's census data table. When adding the synthetic individual's employment status there was only some information present, so a probabilistic model was implemented as detailed in Section 3.3.1.

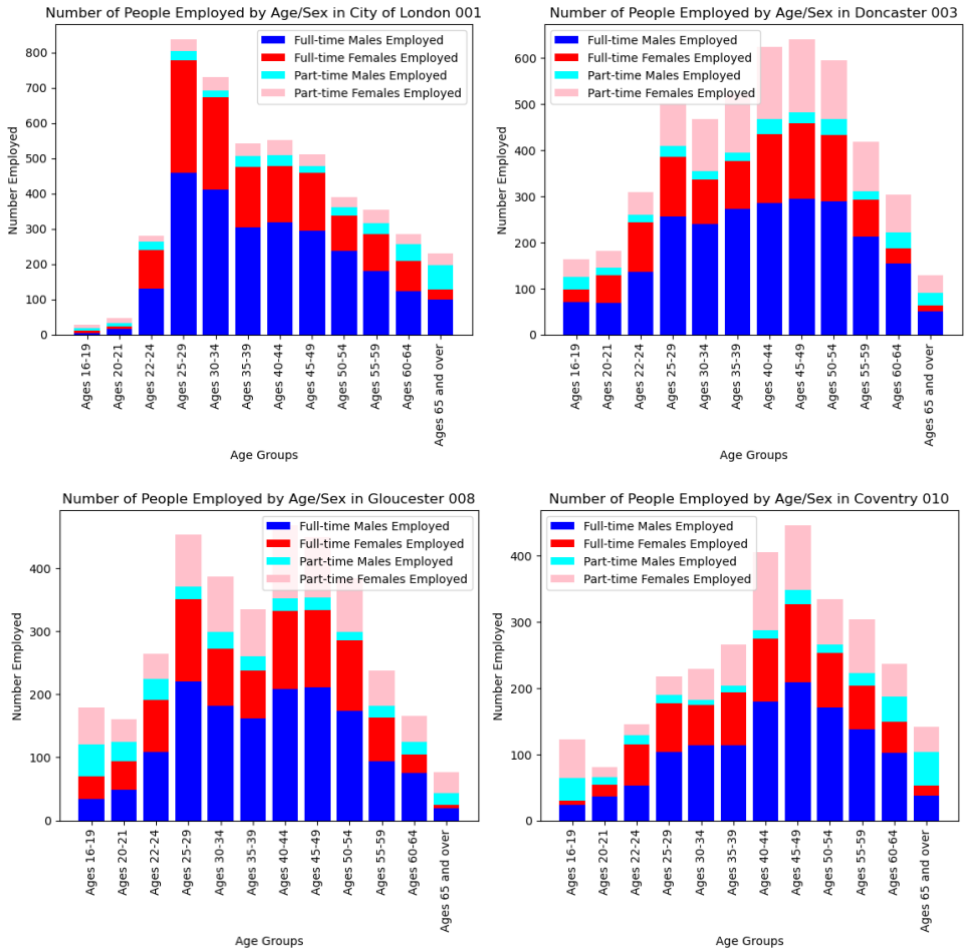

Figure 11. MSOA Employment Status Comparisons

The process of assigning an employment status to individuals looks accurate if the entire MSOA is looked at but may look more inaccurate when seen at the household level. If an individual inside a household fell into one of the age groups shown in Appendix 1 Figure 13, they had a chance at being assigned a full-time or part-time employment status. Therefore, some household may contain a husband working part-time and a wife working part-time with their child also working part-time. But because the census data is anonymized, there is no way at knowing if a household in the MSOA would contain this exact make-up. Issues seen here with employment status assignment and others within this method are addressed further in the discussion section below.

4.3 Discussion

Much of the results above show the effectiveness of this method for its intended purposes outlined in Section 1.2. The creation and assignment of all household and individual characteristics to the synthetic population differs only slightly in comparison with the true UK population make-up. In the methodology it can be seen how easy aggregation of the census data tables into a single synthetic table is, and the results prove the methodology holds up. When looking at some of the figures present in the section above, it can also be seen that metric present in the synthetic data can be extracted at a faster speed than going through the census data tables or in some cases it is not possible with a singular census data table. Overall, the synthetic UK population presented in this research method can be used as a successful and accurate representation of the UK population as long as the shortfalls of this method are kept

in mind. But of course, the method is not perfect and a few issues with both the methods used, and results displayed need to be addressed further.

4.3.1 Anonymity Issues

One of the most difficult to assess issues within this synthetic population creation method is the anonymity factor of the census tables. As discussed in Section 1.1.1, the census data is collected on one form, then the census office separates the data entered into the census form into the tables that are used in this method. Through this process they anonymize the data making it impossible to know the exact characteristics that make-up any given household in an MSOA. And this anonymity of the household and individual characteristics will remain so for one hundred years after the census data is taken [11]. Much of the validation of this method is done on a characteristic-to-characteristic basis, meaning only one field of data can be evaluated at a time and the synthetic population characteristics as a whole would require de-anonymized data sets. Therefore, this research method cannot determine the true accuracy of the synthetic households and the individuals inside them. For example, if a household with a married couple with ages 33/36 and one dependent age 4 is created in the synthetic MSOA, there is no way of knowing if this exact household with these exact individuals exists in the true MSOA. This method utilizes probability, so the most likely households are created based off the census data tables. But it is unlikely the exact set of households in the synthetic MSOA match up with the real MSOA. However, the synthetic households are in fact created to be close to existing households, so they can still be used for further research purposes after being fully evaluated for UK wide accuracy.

4.3.2 Computing Power

Another issue found within this data and being able to analyse the effectiveness of this research method is the sheer size of the UK census and the tables it creates. The machine that this code was created and ran on is not powerful enough to analyse the entirety of these data sets and produce valuable results of the method. That is why it is recommended that for those who wish to utilize this method to have access to a higher-level computational machine or set of machines. The Office for National Statistics outputs a plethora of metrics and statistics on the 2011 census data, many of these describe the UK population as a whole. Metrics such as employment rate [12], percentage of population in cohabitation versus marriage and living arrangements [13] are only comparable when looking at the entirety of the United Kingdom's population. The results section in this paper uses a select few whole population level metrics on the MSOA level for results comparisons as these can still be used to some effect at a smaller scale. But many available statistics are not able to be scaled down to the single MSOA level, so an implementation of this method on the whole population is necessary for a full validation of the methods success.

4.3.3 Complex Coding

Some sections of code for this synthetic population creation are somewhat simple for those familiar with the census data tables and their formats. But there are other sections of the code base that are far more complex and took quite a bit of time to implement successfully. The complexity of this code can partially be attributed to the anonymity factor of the data as discussed in Section 4.3.1, but this method also utilizes dependencies between characteristics quite heavily. For example, the assignment of individuals is based on age, sex, household composition, and existing household members age or sex. These dependencies can even create upwards of six if/else statements in the code that need to be evaluated in order to assign an individual to the most likely household. If further research and/or code was to be implemented into this codebase, then the complexity would increase exponentially as the

number of characteristics increases. Because of this code implementation complexity, the usability of the code for further research is diminished. While the current existing code has been evaluated to be quite usable currently, one would need to become familiar with both census data table structure and highly dependent coding practices to be able to advance this method even further.

5 Conclusion

5.1 Summary

The research presented in this dissertation presents a supportive case for further development in the field of synthetic population modelling. Understanding the background of census data and synthetic data is necessary to contribute to this progress in the field, especially with all of the current methods in the field that are presented in this paper. The method included in this dissertation is presented as a probability based iterative bottom-up approach to creation of a synthetic United Kingdom population. Results of this method are intended to successfully complete the goal of aggregating census data tables for faster access times, research usability, and overall accuracy.

A key factor in the methodology presented in this research is working effectively with the syntax and format of census data, and ensuring different tables are able to mesh together efficiently. The initial approach involves selecting specific data tables with household or individual characteristics from the census and using them to generate a synthetic population that maintains these characteristics of the real data. Household size, household composition, individual's age and sex, and employment status of individuals are the characteristics included in the current methodology. This method utilizes probability-based techniques such as simple random sampling and judgmental sampling to assign characteristics to households and individuals. These probability techniques are intended to create a synthetic population that represents the real population's demographic traits. But the method encounters challenges with dealing in ambiguous data and being able to avoid the creation of unrealistic households.

The results of this method are evaluated at different levels of the population created, the household and individual levels. The household level characteristics, size and composition, were evaluated by comparing the number of households with each size or composition in the synthetic population with the census numbers. Household size counts were equivalent, however household compositions slightly varied due to rounding errors but still showed a high level of similarity between the synthetic and true population. Individuals were assessed by comparing age/sex distributions, as well as other metrics such as average dependent ages and parent ages, between the synthetic and full UK population. The synthetic population accurately reflected these UK wide metrics, proving individual assignment was likely successful. Employment status of individuals was then included, but evaluation of the assignment method's accuracy was hard to determine due to anonymity of the data set.

It is also necessary to acknowledge that this research method has limitations in approach and results. The anonymization of census data makes it nearly impossible to validate the method's accuracy on a household-to-household basis. It is also important to note that higher computational power is needed for a full-scale implementation of the research method. The method also contains highly syntaxial code due to many dependencies between characteristics, making the method less user-friendly for further research. But overall, more research in the field and more eyes on this research method could lead to improvements in accuracy and elimination of these limitations present in the method.

5.2 Evaluation

The overarching goal of this research project was to build a model to create a synthetic United Kingdom population using only UK census data from 2011. There were objectives under this goal that are an integral part of the goal's completion:

1. Aggregate the 2011 United Kingdom census data tables for further research usability in the field of synthetic population modelling.

2. Reduce access time for those users and researchers that want to use multiple UK census data tables without having to access them one-by-one.

3. Produce an accurate synthetic United Kingdom population that contains useful household and individual level characteristics.

The first objective can be considered the most important one of the three because it is the basis for the other two. It was completed by adding the desired household and individual characteristics to one data frame that can be queried for statistics, as detailed in Section 3. This method to aggregate tables was then evaluated for accuracy in Section 4, which resulted in a success. Tables were able to be put together on the same data frame and then desired research metrics such as MSOA ages can be pulled from the synthetic data and look highly similar to the true data used from the census.

Objective two is a result of the success of objective one, therefore with the success of objective one, objective two was also a success. By following the methodology in Section 3 of putting multiple census tables together that contain basic information of a UK population, researchers will not need to look at the tables individually. The resulting synthetic data table contains a combination of four census characteristics in combination, so researchers can instead look at the synthetic population for multiple characteristics instead of looking one-by-one.

The trickiest objective was number three, as there are only a few metrics that can be used to measure the success of an accurate synthetic population. By comparing UK average statistics provided by the ONS such as average age, average household size, and average age difference between partners, a measure of success for the synthetic population method was created. The 10 MSOAs tested for results were close to UK averages in these used metrics. Although further testing on all MSOAs using more computation power is vital to ensure that this method works on all MSOAs and can be used as a UK wide synthetic population model.

Overall, the goal and objective were completed with a satisfactory accuracy and usability for future research in the field of synthetic population modelling.

5.3 Future Work

Overall, there has been a good degree of satisfaction with the results of this research as it accomplished all of the objectives effectively with few discrepancies with the true census data. But there still does exist these discrepancies that need to be addresses as this synthetic United Kingdom population method is not perfect. Perhaps the most glaring limitation of the way it was decided to code the method is when there was the introduction of some of personal bias into the probabilities used to assign households and individuals. One example of this bias that was included was during the assignment of individuals into their households based on ages of others existing in the household already. It was necessary to define the age ranges for each individual to be assigned dependent on the household composition (i.e., for households with a dependent child set the age range to be 0-18), and then the algorithm would base the first parent's age off of being 35-35 years older than the dependent. Furthermore, the initial age range to assign the 2nd partner to a household would be $\pm$ 5 years from the initial partner's age. By predefining these age ranges, and therefore pre-defining the probability of assignment to a household, I incorporated my bias into the probability essentially producing a form of sampling bias. However, this seemed to be a vitally necessary bias as without it, the creation of synthetic households that made little to no logical sense would exist. One other limitation of this methodology is that it is intensive to implement as adding a new feature (new census data table) to the MSOA's data frame can produce many dependencies on the other features, both household and individual. While taking into account all of the dependencies, a new feature can be added and it will make the model more accurate, but it takes a significant amount of coding

complexity and time to implement all these dependencies. On the other side of complexity, some data tables from the census would require little to no effort to implement into the model but are unlikely to be accurate when distributed as a characteristic to households in the MSOA due to a *lack* of dependencies. In summation, this model can become complex to implement very quickly when new features are added, but not all data tables from the United Kingdom's census can be inputted into the model and have an accurate and usable affect.

Because this synthetic population model was developed for usability by researchers and population data enthusiasts, there is still a plethora of work that can still be applied to the model and enhanced by the model. It is suggested that by following the same basic methodology that was used to add features to the synthetic population that the addition of even more features could be useful, such as general health, country of birth, etc. This would lead to a more accurate model of the population and even further this project's goal of aggregating multiple census data tables into a single synthetic population table. Another way to develop this project further would be to edit the code to be able to get the output results of as many MSOAs at a time as the user wanted, because currently the model can only output a single MSOA's results. Still, this current model can be used in some capacity in other research projects that have the necessity for the synthetic United Kingdom's population. Research sectors such as epidemiology or demography could particularly use the method so they would not have to aggregate census data themselves and can be confident that they are using a population that is highly similar to the real population. This research method is just a jumping off point for UK synthetic population modelling as it is a simple model with plenty of room for improvement. Further work is definitively needed for it to become more easily usable for those unfamiliar with the syntax and coding style of the method presented in this research paper.

Based off the 10 MSOAs (see Appendix 1, Figure 14) that were simulated (as the computer used could only handle so many), all results that are set forth in Section 4.2 for were up to satisfactory levels. All metrics that were used to measure the accuracy of the results were well within an acceptable range to consider the synthetic population a good representation of the true population. The number of households, household sizes, individuals, and number of employment statuses were all exactly the same as the true census' numbers. Household composition did vary slightly as rounding errors occurred when creating probabilities for composition for assignment to household rows, but these tended not to vary more than ± 5 of a type of household composition. This variation did not affect the total number of household compositions, meaning all households still did get assigned a composition which was absolutely necessary for assigning individuals. There exist other ways to approach this particular research problem, so it is important to acknowledge that this method has both flaws and strengths as demonstrated in the sections above.

There has already been a great deal of achievement shown in the synthetic population modelling field from small-scale to large-scale populations around the globe as described in Section 2.1. Wu, G., Heppenstall, A., Meier, P. *et al.* [1] performs a similar method to this paper's when looking at Lower Super Output Areas (LSOAs) in the UK to find health and socio-economic outcomes based of the 2011 United Kingdom census data tables. They successfully created a synthetic UK population at a small-scale using census health related data and some other sources of health data, which is similar to this project but has a narrower scope on applications. One other project in the field that I found was an agent-based synthetic population of Canada [2] using Canadian census household and individual data. They were very successful with their methods, as they successfully produced a synthetic population of Canada and allowed for projection of the Canadian population up to the year 2042. If there was more time for research and more resources as well, then this is the level of project that would be ideal to develop into eventually. Although, it is difficult to compare this paper's method with

others when they are set out to perform different tasks and are different levels of expertise and resources. This research project sets a decent standard for its level, but future work on it could bring it up to the level of the existing projects in the field aforementioned.

References

[1] Wu, G., Heppenstall, A., Meier, P. *et al.* A synthetic population dataset for estimating small area health and socio-economic outcomes in Great Britain. *Sci Data* **9**, 19 (2022). https://doi.org/10.1038/s41597-022-01124-9

[2] Prédhumeau, M., Manley, E. A synthetic population for agent-based modelling in Canada. *Sci Data* **10**, 148 (2023). https://doi.org/10.1038/s41597-023-02030-4

[3] "Age by Hours worked by Sex 2011 - UK Data Service CKAN," *statistics.ukdataservice.ac.uk*. https://statistics.ukdataservice.ac.uk/dataset/age-hours-worked-sex-2011 (accessed Aug. 13, 2023).

[4] "Births by parents' characteristics in England and Wales - Office for National Statistics," *www.ons.gov.uk*. https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/livebirths/bulletins/birthsbyparentscharacteristicsinenglandandwales/2015#:~:text=2.- (accessed Aug. 15, 2023).

[5] M. Bhrolcháin, "The age difference at marriage in England and Wales: a century of patterns and trends National Statistics." Available: https://eprints.soton.ac.uk/34801/1/PT120AgeDifference.pdf

[6] "Sampling Distribution | Simply Psychology," *www.simplypsychology.org*. https://www.simplypsychology.org/sampling-distribution.html

[7] "Judgement Sampling | School of Hospitality, Food & Tourism Management," *www.uoguelph.ca*. https://www.uoguelph.ca/hftm/judgement-sampling

[8] "2011 Census - Office for National Statistics," *www.ons.gov.uk*. https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationes-mates/bulletins/keystatisticsandquickstatisticsforlocalauthoritiesintheunitedkingdom/2013-10-11#:~:text=Key%20figures (accessed Aug. 17, 2023).

[9] Gov.uk, "Age groups," *Service.gov.uk*, Aug. 22, 2018. https://www.ethnicity-facts-figures.service.gov.uk/uk-population-by-ethnicity/demographics/age-groups/latest

[10] Office for National Statistics, "Population and household estimates, England and Wales - Office for National Statistics," *www.ons.gov.uk*, Jun. 28, 2022. https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/bulletins/populationandhouseholdestimatesenglandandwales/census2021

[11] "Your confidentiality - Office for National Statistics," *www.ons.gov.uk*. https://www.ons.gov.uk/census/aboutcensus/yourconfidentiality

[12] "2011 Census - Office for National Statistics," *www.ons.gov.uk*. https://www.ons.gov.uk/employmentandlabourmarket/peoplenotinwork/unemployment/articles/2011census/2013-11-29#:~:text=There%20were%2025.7%20million%20people (accessed Aug. 19, 2023).

[13] "People's living arrangements in England and Wales - Office for National Statistics," *www.ons.gov.uk*. https://www.ons.gov.uk/peoplepopulationandcommunity/householdcharacteristics/homeinternetandsocialmediausage/articles/livingarrangementsofpeopleinenglandandwales/census2021

[14] "About the census - Office for National Statistics," *www.ons.gov.uk*, Jul. 19, 2022. https://www.ons.gov.uk/census/aboutcensus/aboutthecensus

[15] "Questionnaires, delivery, completion and return - Office for National Statistics," *www.ons.gov.uk*. https://www.ons.gov.uk/census/2011census/howourcensusworks/howwetookthe2011ce

nsus/howwecollectedtheinformation/questionnairesdeliverycompletionandreturn#:~:text=2011%20Census%20questions&text=Questions%20included%20those%20about%20work (accessed Aug. 25, 2023).

[16] J.-F. Rajotte, R. Bergen, D. L. Buckeridge, K. El Emam, R. Ng, and E. Strome, "Synthetic data as an enabler for machine learning applications in medicine," *iScience*, vol. 25, no. 11, p. 105331, Nov. 2022, doi: https://doi.org/10.1016/j.isci.2022.105331.

[17] B. N. Jacobsen, "Machine learning and the politics of synthetic data," *Big Data & Society*, vol. 10, no. 1, p. 205395172211453, Jan. 2023, doi: https://doi.org/10.1177/20539517221145372.

[18] "PopGen," *MARG - Mobility Analytics Research Group*. https://www.mobilityanalytics.org/popgen.html#:~:text=PopGen%20is%20capable%20of%20synthesizing (accessed Aug. 28, 2023).

[19] Nowok, B., G.M. Raab & C. Dibben (2016), synthpop: Bespoke creation of synthetic data in R. *Journal of Statistical Software*, 74:1-26; DOI:10.18637/jss.v074.i11. Available at: https://www.jstatsoft.org/article/view/v074i11

[20] S. Hörl and M. Balac, "Synthetic population and travel demand for Paris and Île-de-France based on open and publicly available data," *Transportation Research Part C: Emerging Technologies*, vol. 130, p. 103291, Sep. 2021, doi: https://doi.org/10.1016/j.trc.2021.103291.

[21] A. Smith, R. Lovelace, and M. Birkin, "Population Synthesis with Quasirandom Integer Sampling," *Journal of Artificial Societies and Social Simulation*, vol. 20, no. 4, p. 14, 2017, Accessed: Sep. 04, 2023. [Online]. Available: https://jasss.soc.surrey.ac.uk/20/4/14.html

[22] "How life has changed in City of London: Census 2021," *sveltekit-prerender*. https://www.ons.gov.uk/visualisations/censusareachanges/E09000001

Appendix 1 – Extra Tables

GEO_LABEL
Household composition : One person household\ Aged 65 and over - Unit : Households
Household composition : One person household\ Aged 0 to 64 - Unit : Households
Household composition : One family only\ All aged 65 and over - Unit : Households
Household composition : One family only\ Married couple\ No children - Unit : Households
Household composition : One family only\ Married couple\ One dependent child - Unit : Households
Household composition : One family only\ Married couple\ Two or more dependent children - Unit : Households
Household composition : One family only\ Married couple\ All children non-dependent - Unit : Households
Household composition : One family only\ Same-sex civil partnership couple\ No children - Unit : Households
Household composition : One family only\ Same-sex civil partnership couple\ One dependent child - Unit : Households
Household composition : One family only\ Same-sex civil partnership couple\ Two or more dependent children - Unit : Households
Household composition : One family only\ Same-sex civil partnership couple\ All children non-dependent - Unit : Households
Household composition : One family only\ Cohabiting couple\ No children - Unit : Households
Household composition : One family only\ Cohabiting couple\ One dependent child - Unit : Households
Household composition : One family only\ Cohabiting couple\ Two or more dependent children - Unit : Households
Household composition : One family only\ Cohabiting couple\ All children non-dependent - Unit : Households
Household composition : One family only\ Lone parent\ One dependent child - Unit : Households
Household composition : One family only\ Lone parent\ Two or more dependent children - Unit : Households
Household composition : One family only\ Lone parent\ All children non-dependent - Unit : Households
Household composition : Other household types\ With one dependent child - Unit : Households
Household composition : Other household types\ With two or more dependent children - Unit : Households
Household composition : Other household types\ All full-time students - Unit : Households
Household composition : Other household types\ All aged 65 and over - Unit : Households
Household composition : Other household types\ Other - Unit : Households

Figure 12. Columns used from the Household composition 2011 data table.

Full-time Male	Full-time Female	Part-time Male	Part-time Female
Ages 16-19	Ages 16-19	Ages 16-19	Ages 16-19
Ages 20-21	Ages 20-21	Ages 20-21	Ages 20-21
Ages 22-24	Ages 22-24	Ages 22-24	Ages 22-24
Ages 25-29	Ages 25-29	Ages 25-29	Ages 25-29
Ages 30-34	Ages 30-34	Ages 30-34	Ages 30-34
Ages 35-39	Ages 35-39	Ages 35-39	Ages 35-39

Ages 40-44	Ages 40-44	Ages 40-44	Ages 40-44
Ages 45-49	Ages 45-49	Ages 45-49	Ages 45-49
Ages 50-54	Ages 50-54	Ages 50-54	Ages 50-54
Ages 55-59	Ages 55-59	Ages 55-59	Ages 55-59
Ages 60-64	Ages 60-64	Ages 60-64	Ages 60-64
Ages 65+	Ages 65+	Ages 65+	Ages 65+

Figure 13. Employment Status by Age and Sex

MSOA Name	Total Households	Total Individuals
City of London 001	4385	7375
Barking and Dagenham 012	3125	8402
Greenwich 027	2502	5970
Manchester 042	2949	6594
Trafford 014	3396	7850
Doncaster 003	4876	11277
Newcastle upon Tyne 005	4214	9924
Birmingham 018	4459	9716
Coventry 010	2648	6281
Gloucester 008	3000	7817

Figure 14. The 10 MSOAs tested in this research.

Appendix 2 – User guide

While I have developed the algorithm to be used and edited by others, the current code repository is somewhat messy as it contains a lot of notes and print statements that I used for research. But you can still visit my GitHub to get the data files used and the code used at:

https://github.com/MatthewSenseman/MatthewSenseman

1. Download the IPython Notebook: code_uncleaned.ipynb
2. In the first code cell you will see the names of which census data files are necessary to download also from my GitHub.
3. Then you can simply run the notebook in your browser, or you may choose to download it as a .py file and run it in your computers terminal command line.
4. The program will prompt you for a path to where you have the census data files saved at, make sure you enter the correct path otherwise you will need to restart the program and try again.
5. The code will ask you to enter an MSOA name (e.g., City of London 001), which are found in any of the census data files first column, so just choose one and it will output a data frame for that inputted MSOA.
6. You may also choose to download the csv file for the synthetic MSOA data frame. When prompted by the program (near the end), enter Yes or yes to download the csv, any other input will not download it.
7. Because this method only applies to one MSOA at a time, it is difficult to apply it to all MSOAs in the United Kingdom. You will either need a computer with a large amount of RAM/computing power or a cluster computing network to use it for all MSOAs.

Appendix 3 – Installation guide

The only installations that are needed is to have Python downloaded with the respective packages noted in Section 3.1.1 and have them all updated, as well as Anaconda which contains the Jupyter Notebook system. These two applications can be installed using their official guides here:

Anaconda: https://docs.anaconda.com/free/anaconda/install/index.html

Python: https://www.python.org/downloads/

You may also need to download the census files used in the code, these are named in the Jupyter notebook in Appendix 2 and can be downloaded here:

https://statistics.ukdataservice.ac.uk/dataset/?q=Census+2011&vocab_Year=2011&res_format=CSV&organization=office-national-statistics-national-records-scotland-northern-ireland-statistics-and-research&sort=score+desc%2C+metadata_modified+desc