Population 24/7 – A method to account for daily population mobility in spatiotemporal population estimates

Contents

Key Messages of Paper

Purpose

We are assessing the feasibility of estimating the population of small areas by specific times of day (i.e., spatiotemporal population estimates that account for population mobility) using the Population 24/7 framework developed by Martin et al. (2015). Previous uses of this framework have relied exclusively on open data to develop the input files required. As such, we are evaluating the possibility of using sensitive, record-level data available within the ONS. Sensitive, record-level data may bring potential benefits in terms of increased temporal and spatial granularity, as well as the ability to produce spatiotemporal population counts for years and times not covered by existing open datasets.

This paper outlines the work done so far in developing origin, destination, and transit files used by the framework. These represent residential, educational, work-related, healthcare-related, and retail-based population movements, and include details of assumptions made. An initial evaluation of the output population estimates is compared to a population estimated created using an open-data methodology developed by Cockings et al. (2021). The paper then discusses the limitations in the methodology as it currently stands, as well as the issue of developing a strategy for validating the output population counts as accurate. Finally, possible validation methods are considered.

Recommendation

Initial feasibility work indicates that the Population 24/7 framework yields promising results. Further investigation and development into the Population 24/7 framework is recommended to:

- Validate and assess the quality of the spatiotemporal population estimates the method produces.
- Incorporate the use of anonymised and aggregated cellular phone mobility data.
- Scope work required to improve coverage of types and modes of population movement poorly handled by the method as e.g., leisure activities, rail travel, and non-vehicular transport
- Compare the method against alternatives under development in ONS e.g., agent-based modelling

However, the method requires a comprehensive repository of data that describe different types of population movements from a range of sources that must be continually maintained. The significant resource overhead required to curate these data and the variety in their timeliness and reliability may, based on current data and systems, make the method unsustainable for regular statistical production, and more suited to bespoke case studies.

Key Asks of MARP

The key asks of the panel for this paper are as follows:

- Are there any types/routes of population movement that we haven't included in our experimental outputs that the panel thinks should be a priority for inclusion in future work?
- The method does not currently include non-usual resident population groups like short-term migrants, tourists, and second-home owners, and including them will be challenging. If these groups cannot be included do the panel still see value in publishing day-time population statistics with this caveat?
- If the panel has any expertise in geospatial population methods, is there any work we have not referenced herein that we should be aware of? Or any potential for collaboration with academic research groups in the panel's institutions?
- Is our proposed approach to validation sufficient, or can the panel advise on any alternatives?
- Day-time population statistics will likely persistently have much larger error than our traditional population statistics. We will need to communicate this increased uncertainty clearly to users and prioritise the production of uncertainty measures. Does the panel consider any degree of uncertainty so wide as to recommend communicating limited use of the statistics to users?
- Private organisations like Google and Apple have access to rich mobility data and offer some population mobility services to users e.g., Google Popular Times. Assuming that privacy concerns over ONS having access to (likely anonymised and aggregated) device location data could be satisfactorily addressed, does the panel have a view on ONS developing our own population mobility methods and statistics vs. fostering collaboration with private organisations with more mature population mobility data collection and methods?
- There are a huge number of potential spatiotemporal scenarios (geospatial area and date-time combinations). We are considering options for producing and disseminating population mobility / day-time population statistics, and we expect the 'best' option to vary for different users. These could range from publishing a single generic 'day-time' population base, to publishing a wide range of spatiotemporal population estimates in different date-time scenarios to meet a variety of user needs. This could possibly involve making a tool available for users to specify a spatiotemporal scenario and download bespoke outputs. What other publication formats/routes/types does the panel think we should consider?

Executive Summary

- We are attempting to develop a methodology to create gridded population estimates representing the number of people in a particular location by time of day, taking into account population movements for work, education, healthcare, retail and other reasons.
- At present, population estimates for an area are typically obtained from Census counts of people at their primary residence, and Mid-Year Estimates that roll forward these population counts from the Census population for years without a Census. These estimates provide a good approximation of an area's night-time population but are less useful for estimating the population of an area at other moments in time, which would be valuable for public service planning, disaster response, and public health purposes.
- Following a review of potential methods of producing more temporally granular population estimates, we have identified the Population 24/7 framework, developed by Martin et al. (2015), as a potential avenue for investigation.
- This method has been previously used to produce population estimates for the whole of England for the year 2011 (Cockings, et al., 2021), and elsewhere for smaller areas, but these approaches have so far only been produced with access to openly-released data. We are thus looking at evaluating the potential improvements of using the framework with sensitive record-level data available within the Office for National Statistics.
- Based on the method described in Martin et al. (2015), and updated in Cockings et al. (2021), we develop a set of origin, destination, and background (transit) files covering the whole of England, using a reference year of 2019.
- These origins, background and transit files are not yet finalized, and there are numerous coverage issues, both in terms of data required for the currently implemented population movements and for new types of population movements not yet included, that we would like to improve in future work.
- We compare the population estimates created using this methodology to those created using the open data methodology developed by Cockings et al. (2021), for 8km by 8km areas around the centre of the city of Southampton, Hampshire, and Dartmoor, Devon, and find that in general, there are high correlations between our population estimates and those from Cockings et al. (2021). This correlation is highest when looking at night-time population estimates, which mainly represent the residential population of the area, with much less movement from origins to other locations. For population estimates at times of day where there is more population movement, the correlation drops slightly, particularly for student age groups which are more affected by changes to the underlying data preparation methodology, but it is not clear in what proportions these differences are due to differences in the reference year chosen (2011 in the open data method, 2019 in the new method), or changes in the assumptions made when developing destination profiles. Looking specifically at population cells that contain retail locations in the new methodology that are not accounted for in the old methodology, we generally see a higher percentage increase in the population estimates, the most noticeable being for the on-site population which has the lowest correlation between methods compared with the in-travel and resident populations estimates.

- While initial evaluation suggests that our population estimates generally concur with estimates produced using the open data methodology, it is not yet clear how to properly validate their accuracy against the actual number of people in the area. A number of validation strategies are currently being considered; each of these have limitations and would not, taken individually, represent a true validation of the methodology as a whole. These strategies include:
    - Comparison against similarly aggregated location counts derived from mobile phone data
    - Evaluation against other published daytime and night-time population counts, such as those from the EU Enact project (Batista e Silva, et al., 2017)
    - Comparison against footfall counts from mobile phone data and CCTV and Wi-Fi connection sources for specific areas where this data is available
    - Creation of a quality framework to individually examine the component data sources and assumptions used to produce the population estimates.

Population 24/7 - A method to account for daily population mobility in spatiotemporal population estimates

Introduction

Currently, estimates of population size in England & Wales are derived from Census counts of individuals at their usual residence, and from Mid-Year Estimates that roll forward these population counts for years without a Census. Although these provide a good approximation of a given area's night-time population, Census and Mid-Year Estimate counts are less useful for predicting day-time population counts; sub-populations residing within each geographical area move between areas over the course of the day between workplace, residential, educational, commercial and leisure destinations. These patterns vary by sub-population and can be cyclical, seasonal, and unpredictably disrupted by stochastic events. Demand for more granular population distribution estimates (for example, by time-of-day at low-level geographical breakdowns) is longstanding. Requirements are increasing in both the public and private sectors, specifically for the ability to nowcast and forecast population counts by time of day. There are numerous potential applications for these estimates in public health, service planning, transport planning, business or disaster response. For example, they could be used to understand how many people are in danger during potential flash flood events, inform decisions on investments in local areas and evaluate their impact on population movement, support decisions regarding required infrastructure changes, understand differences between term and out of term population movements (especially valuable for cities with multiple universities) or inform business decisions by serving as a basis to estimate potential footfall counts for certain areas of interest.

Background

As part of our research into the feasibility of producing granular spatiotemporal estimates for population mobility statistics we conducted a literature review and identified several approaches worthy of consideration.

The Centre for Ecology & Hydrology (CEH) have produced 1km gridded population estimates for both the workday and resident populations of the United Kingdom by combining population counts at census output area (OA) level with land use classification data (Reis, et al., 2017). This methodology produces reasonable estimates for daytime and night-time populations, but these do not properly account for primary, secondary, and higher education students (the workday population estimates produced by the ONS assign people under the age of 16 to their place of residence, whereas we would expect them to be at an educational establishment for most of the day). Similar daytime and night-time estimates have also been produced by the EU Enact project (Batista e Silva, et al., 2017), also taking into account student locations as well as tourism flows by month, and by the LandScan USA project (Bhaduri, Bright, Coleman, & Urban, 2007), which estimates daytime and night-time population for each census block in the United States of America, and then allocates each block's population to grid cells according to a cell likelihood coefficient derived from land use and estimated building occupancy data. The National Population Dataset is a commercially available tool developed by the Health and Safety Executive which can generate GB population estimates at high spatial resolution (including building locations) for five types of populations: residential, sensitive (childcare institutions, schools, colleges, hospitals, care homes, prisons), in transport (airports, London underground, railway stations, ports and ferries, A-roads and motorways), workplace and leisure (stadia). While it includes

an extensive number of population types compared to other methodologies, due to the non-modelling nature of the tool and the spatial/temporal scale incompatibilities between the available data sources, the outputs are limited to particular combinations of population type, time period, and spatial extent. Statistics New Zealand (StatsNZ) have developed a methodology to generate a mobility index using telecoms data which can be weighted to population totals to generate spatiotemporal population estimates (Data Ventures, 2021). StatsNZ have mobility data that covers most of their mobile phone market and have a close relationship with their telecoms suppliers, where they are able to implement bespoke data processing on the supplier side. Here in the UK, we have access to less complete data and a less mature relationship with suppliers that may make this approach challenging.

However, most of these methods assume that a person's day-time location is entirely predicated on their primary activity (i.e., a worker is going to be at their workplace for the entirety of the day, a student at their school/university, etc.), and so do not allow for population movements for other activities, which could have a significant effect on population numbers for certain locations at particular times, or transit to and from these locations. Additionally, shift- or night-working are not considered, or the effects that these could have on the workplace population of an area.

The main focus of our research is the Population 24/7 framework developed by Martin, Cockings & Leung (2015). The authors implemented the method in a piece of openly available but compiled software called SurfaceBuilder247 that cannot be used in ONS systems for security reasons. We have collaborated with the authors' group at the University of Southampton to develop a new Python implementation that is usable on ONS systems (and specifically, in the Data Access Platform (DAP)). This implementation is also openly available (University of Southampton GeoData Institute, 2022). This framework allows for the creation of rasterised population maps (i.e., presented in a gridded, cellular map format) for a given time of day, incorporating a variety of administrative and survey-based data. Populations are modelled in terms of centroids (easting and northing co-ordinates denoting the centre of a particular area of interest) for origins, which represent the residential population distribution of an area, divided into population sub-groups representing demographic breakdowns of interest, and for destinations, representing employment, education, healthcare, retail, leisure and other locations. Each destination centroid is then assigned a time profile related to that destination type, indicating the proportion of the location capacity expected to be present at or in transit through that destination at a given time. When creating a population estimate for a specific time, each destination centroid is queried in turn, and people are reallocated from origin centroids within defined catchment distances (specific to each destination). The populations assigned to each centroid after re-allocation are then assigned to grid cells around the centroid based on variable-kernel density estimation techniques (Martin D. J., 1989) according to the destination-specific extent radius defined by the model. These grid cells can theoretically be of any size, depending on the granularity of input data to the model. In-transit populations are distributed to grid locations around the destination centroid they are allocated to, based on a background map of weights representing the relative volume of traffic in that location for that time.

Because of its flexibility regarding the types of data that can be used to produce population estimates, the Population 24/7 framework can be applied to several use cases. Martin et al. (2015) demonstrate the framework by producing population maps for a number of times

covering the area around the city of Southampton, Hampshire. Smith et al. (2015) use the framework to evaluate the effects of a seasonally varying population on the risks of exposure to flooding danger, by modelling the flows of tourists into the area surrounding the town of St. Austell in Cornwall by month. Their model uses South West Tourism data to identify and populate tourist accommodations, and a variety of data sources to create time profiles for leisure and retail locations that tourists would visit. By intersecting the produced population estimates for a variety of times with rasterised flood risk models, they were able to estimate the number of people who would be at risk from flooding events to a higher degree of accuracy than would be possible with census-based estimates alone. Elsewhere, Alexis-Martin (2016) uses the framework to model the risks of a hypothetical radiation exposure event in the city of Exeter, Devon, by creating sets of age- and sex-specific time profiles and intersecting the results with a rasterised fallout model.

To date, all published applications of the Population 24/7 framework have relied entirely on openly published data. One purpose of this research is thus to investigate whether the introduction of sensitive data available to the ONS can lead to the production of population estimates with greater temporal and spatial granularity, and to identify methods to validate the accuracy of these produced population maps.

Data

Table 1 below shows the data used to create time profiles and location databases for each of the three layers (background, origins, and destinations) required as inputs to the SurfaceBuilder247 software, as described in Martin, Cockings & Leung (2015). Further details on the purpose of these layers and how they are constructed are given in Methods. The data sources used in the reference study, Cockings, et al. (2021), are listed along with the updated and additional data sources used in current work, including those we hope to include in the near future. For the current feasibility study, we have chosen the most recent data available up to the start of the COVID-19 pandemic (late 2019).

Table 1: A comprehensive list of all data sources used in Cockings, et al. (2021), and how these have been updated and added to in the current work. Data sources we hope to incorporate in the near future are given in grey italic.

Type	***Cockings, et al.** (2021)*	***This research***
Origins	Output Area (OA) centroids and communal establishments resident populations, Census 2011	OA centroids, Census 2011
		Prison populations from Ministry of Justice (MOJ), 2019
		MYE, 2019
		Care Quality Commission (CQC) Adult Social Care Capacity Tracker
Destinations		
Workplace	Workplace zones (WZ), Census 2011	Workplace zones (WZ), Census 2011
	United Kingdom Time Use Survey, 2014-2015 (time profiles)	*Time Use Survey (TUS), 2015 and 2019 (COVID sequence)*
		LFS, 2019
		Inter-Departmental Business Register (IDBR), 2019
Education	Higher Education Statistics Agency (HESA), 2006	HESA, 2019
	Get Information About Schools (GIAS)	*Get Information About Schools (GIAS)*
	Pupil attendance in schools, 2012 (time profiles)	
	Funding for 16 to 19 year olds in schools (time profiles), 2011	
	Participation in education, training and employment, 2011 (time profiles)	

	Schools, pupils and their characteristics, 2011	
	School and college performance measures, 2011	
		English School Census (ESC), 2019
		Welsh School Census (WSC), 2019
		UK Register for Learning Providers (UKRLP)
Healthcare	Hospital Episode Statistics (HES), 2011/2012	HES 2017 – Accident and Emergency (A&E) attendance, outpatient appointments and inpatient admissions
		Hospital address list, 2019
		A&E providers, 2017
		GP Episode Statistics *Private healthcare visitor numbers*
Tourism & Leisure	*No sources used*	*VisitBritain (VB) attractions annual visitor counts, up to 2020*
		Taking Part survey
Retail	*No sources used*	Geolytix supermarket locations, 2022
		MYE, 2019
		Retail sensor data 2012-2017 from CDRC (Consumer Data Research Centre)
General		National Statistics Postcode Lookup, November 2019
		Deimos/Mars telecoms data, 2019
Background	Ordnance Survey (OS)	*OS, 2019*
	Department for Transport (DfT) average annual daily flows (AADF)	*DfT AADF 2019*
	ONS 2015 Regions, Mean High Water	*Corine Land Use/Land Cover, 2019*
		EU Enact Population Grids, 2011

Work is ongoing to identify further sources of data, particularly for extension of public transport, immobile populations, healthcare, retail, tourism and leisure (including places of worship).

Q1 for panel

Are there any key datasets which come to mind that are missing from our plans?

Methods

The work described in this document builds on the framework described in Martin, Cockings & Leung (2015) and additional work from Cockings, et al. (2021) by updating data sources and incorporating sensitive (not publicly available) and/or new data within the existing framework. Our work focuses on a granularity of hourly population in grid cells of 200m x 200m (0.04km^2). For comparison, the median Output Area used in the 2011 census is 0.067km^2, and the median Workplace Zone is 0.3km^2.

Broadly, the methodology is split into three main layers.

1. Origins – This layer deals with allocating the base population to their home locations, ready for redistribution over several destinations dependent on day and time.
2. Destinations – This layer is made up of a variety of sectors, broadly broken down into: workplaces, education, healthcare, tourism and leisure, and retail. Each destination is associated with a time profile and a wide area dispersion string. The time profile determines how many people are present at the destination for any given day and time. Note that the time profile denotes that for an average day and is not date-specific, although any number of specific, continuous time profiles can be used in the model. In this iteration of the model, time profiles are constructed for an average day at hourly intervals. The 'wide area dispersion' (WAD) value determines a 'catchment area', or the radii from which the necessary population are drawn from their origins, and in what proportions. Further detail on the construction of WAD strings differs depending on the type of destination; in general, "population is reallocated from origin to destination containers by simple weighted allocation to nearest destinations or to meet known proportions of population traveling from successive distance bands (recorded in travel-to-work data)" (Martin, Cockings, & Leung, 2015).
3. Background – This layer covers the population who are in transit (i.e., present on the road network) at any given time. This methodology is currently unchanged from that described in Martin, Cockings & Leung (2015) and Cockings, et al. (2021). At present, this layer only incorporates the major road network (A-roads and motorways), but future work at the University of Southampton may extend the background layer to include public transport, such as rail links.

To obtain origin files for a given year, we use the produced Mid Year Estimates (MYEs) for that year, giving us estimates for the number of people of each age in each Output Area (OA). For each OA, the Ordnance Survey (OS) easting and northing co-ordinates (Ordnance Survey, 2022) of the population-weighted centroid – the mean location of the population that reside within the OA - is obtained. For each OA in the MYE file, we aggregate the counts of people at each age to pre-determined Age Bands (0-4, 5-9, 10-15, 16-17, 18-64, and over 65), as well as to an overall total. The 18-64 group is then split into two – 18-64 HE (i.e., students) and 18-64 NSTU (non-students), by using the HESA census for that year to obtain the count of the number of students with a term-time address in each OA, using that as the 18-64 HE count, and subtracting that count from the 18-64 count obtained using the MYE. The percentages that each demographic group make up of the total count is then computed.

Primary and secondary education destinations are geolocated by linking the English School Census to the UK Register for Learning Providers (UKRLP) to obtain the postcode of the

legal address, which can then be assigned Eastings and Northings co-ordinates using the National Statistics Postcode Lookup (NSPL). The radii for proportional population draw for each school can be calculated directly using student postcodes. Some of the codes used to identify schools can change over time, due to local government structures and establishment changes. These codes are updated in the UKRLP, but not in the School Census, so replacement of the old codes to new ones using a separate, maintained list is necessary before linkage can take place.

Primary campuses for higher education destinations can be geolocated in a similar way to those for primary and secondary education, by linking unique identifiers in HESA data with the UKRLP to obtain postcodes. The wide area distribution is again calculated using student term time postcodes. For current testing, an assumption has been made that all students are spread across a 200m radius from the geolocated primary campus easting/northing position. This assumption has been inherited from the methodology of the previous framework applications (Martin, Cockings, & Leung, 2015) (Cockings, et al., 2021) but is likely to be invalid in a number of cases; further discussion around this and other assumptions is provided below in the Discussion section of this paper.

Workplace locations are determined using the Inter-Departmental Business Register (IDBR), a register of all VAT and/or PAYE businesses (including public sector organizations and non-profit entities with paid employees). To obtain employee counts, we first identify the Workplace Zone (WZ) for each Local Unit (a distinct location operated by a business on the register) in the IDBR by linking to the NSPL on the address postcode. These units are then aggregated by WZ and by industry (obtained by extracting the first two digits of that local unit's standard industry classification code (SIC code) (Office for National Statistics, 2007)), and then further aggregated into industry clusters. We then sum the employee count for each local unit that is still active (that is, has a value of NULL in the field deathdate) to create one row for each WZ/Industry Cluster with a count of the number of people working in that industry grouping in that WZ. To derive Wide Area Distributions for each WZ, we assume that the average travel-to-work distances for each Workplace Zone has not significantly changed from the most recent census year, 2011 (the only pre-pandemic year for which we have this data at the required granularity), and calculate the Euclidean distance (as the crow flies) between the eastings and northings of each employee's workplace and home address postcode, aggregating by workplace address WZ, and calculating the proportion of employees with travel distances in predefined distance bands. We assume that travel to work distances do not differ for a WZ by industry, and have assumed that all workers are within the 18-64 age band and are not students, due to difficulty in getting accurate demographic breakdowns for the non-census benchmark year of 2019; in the Southampton methodology, the average proportion of the workforce between 18 and 65 was 95.3%, and only 8 WZs had a proportion less than 80%. Travelling and agency workers, as well as workers at temporary workplaces (e.g., short-term construction) are currently assigned to the location of the Local Unit for which they work, despite the fact that this is not likely to be the place at which they are usually present during their workday. There is also no adjustment made for workers who have a distinct workplace listed but choose to work from home or another location for a portion of their time. Whilst the home-working population is likely small and less likely to significantly impact the overall daytime population distribution for our reference date of 2019, adjustments for home working may cause substantial changes to population counts

when applying the model to 2020 onwards, and therefore further work will need to be done to include an adjustment for those partially working remotely for application to these timeframes.

In order to locate healthcare destinations, a list of NHS acute treatment units and a full list of NHS hospital addresses are linked to the NSPL to obtain eastings and northings co-ordinates for healthcare destinations. Hospital Episode Statistics (HES) deidentified data are used to obtain population estimates for hospitals; these include total inpatient, outpatient and A&E admissions for the UK. The set of HES records used for this analysis contains the patient's age and home census OA but no data on the location of the hospital they visited. As such, it has been assumed that each admission to A&E occurs at the acute treatment location closest to the admitted person's home OA (as the crow flies). Further work to update the criterion for closest location, based on distance via road links and incorporating travel barriers such as rivers, may be necessary, but has not been considered a priority at this time. Similarly, inpatient admissions and outpatient appointments occur at the closest hospital to the admitted person's home OA. To calculate the Wide Area Distribution for each destination, the travel distance for each patient assigned to it is calculated by computing the Euclidean distance between the eastings and northings of the closest hospital or acute treatment centre and the centre of the patient's home OA. The proportions of patients with travel distances in each travel band are then calculated.

For the above destinations, time profiles have been taken directly from Cockings, et al. (2021). Further work to update these time profiles is described in the future work section of this paper.

In addition to the destination categories included in the original Southampton case studies (Martin, Cockings, & Leung, 2015) (Cockings, et al., 2021), we have conducted some additional work to include retail destinations using Geolytix Retail Points data. The data provides Ordnance Survey Eastings and Northings for each retail point in the data set, as well as an indicator of how that easting and northing was obtained (typically by manual rooftop geocoding). There are a few cases where two stores appear in the same location; at present, these are dealt with by removing one of the seeming duplicates, to avoid potential double-counting of customers. To calculate the capacity of a store, we assume that the number of customers in a store at a given time is a scalar multiple of the number of employees working at that store at a given time; this assumption is solely common-sense-based, and we hope to update this with better retail data such as footfall cameras. To get the number of employees working at the store, Geolytix data is linked to the IDBR (described above in relation to workplaces). To get an estimate for the stores that cannot be matched in the IDBR, the capacities of the matched stores are averaged by retailer and size band and applied to each unmatched store. To calculate the customer demographics, we assume that the demographics are broadly similar to that of the residents of the Local Authority (LA) that each store is within the boundaries of. To obtain these, we use the NSPL to obtain the LA code for each store's postcode, and then link to MYEs for the chosen year to identify the proportions of each chosen age group and assign them to each retail point. The WAD for each retail point are created by assuming that, for each type of store (using the size band as a proxy), people will mostly go to the nearest store to their home. We can illustrate these boundaries using Voronoi polygons, which demarcate the area around a defined point (in this case, a retail location) for which that point is closer than any others. Thus, we can create a WAD string by

using the following steps for each retail size band: Create a set of Voronoi polygons for each retail point in that size band, clip the Voronoi polygons to the coastline of the country of interest, and calculate the distance between the retail point and each edge of the corresponding polygon, and keep the longest distance. The produced distance can then be used as the distance for a proportion of the population in the WAD string (in this methodology we have chosen 75%, with the remainder taken from beyond that).

Tourism and leisure destinations are not included in the original Southampton case study (Martin, Cockings, & Leung, 2015) (Cockings, et al., 2021), but are being considered as part of our ongoing and future work.

Five key test areas have been selected for testing. These are given in **Error! Reference source not found.** below. The test areas have been chosen to allow for specific validation of particular areas of the estimation process, especially those which are currently under significant development, including population in transit, retail and tourism/leisure activities. Each of the test areas is an 8km by 8km square, made up of 1600 grid cells, each measuring 200m.

Table 2: A list of test areas to be used for validation of the adapted Population24/7 framework and data inputs, with reasoning behind the choices and areas for specific validation.

Test area	***Category***	***Reason for selection***

Southampton (399 OAs fully or partially within test area; 111 WZs)	University city	• Large student population
		• Distinct seasonal differences in and out of term time
		• Comparable to previous case studies e.g., Martin, Cockings, & Leung (2015), Cockings, et al. (2021)
		• Not in current bounds, but test area could be expanded to include:
		  ○ unusual transport hubs e.g., cruise port, ferry terminals, airport
		  ○ a large retail and entertainment centre (West Quay, nearby Mayflower theatre)

Warwick (125 OAs; 111 WZs)	University city	• Unusual case where student housing may be outside of the test area, causing potential issues with student distribution
City of London (borough) (2211 OAs; 2377 WZs)	City location	• Dense population with lots of non-residential activity over a full 24-hour time window

		• Anticipated large impact of public transport usage on timestamped population • Comparable to other work being conducted in the Data Science Campus using Agent-Based Modelling
Dartmoor (7 OAs; 3 WZs)	Rural location	• Sparse population • Few major roads available for allocation of travelling population • Large immobile population due to prison
Blackpool (510 OAs; 150 WZs)	Tourism hub	• Significant seasonal population changes due to tourism

Further test areas of interest may include major railway hubs (for example, the area surrounding Birmingham New Street station) and large distribution centres (post office sorting centres, delivery processing warehouses). Additional data might also be necessary for date-specific estimates, for example, population mobility variation due to sporting events or festivals.

Q2 for panel Can you think of any types of population mobility relevant to spatiotemporal population estimation that aren't adequately captured by these areas, and can you suggest any areas that might better capture them?

In this paper, we focus on results from Southampton and Dartmoor test areas. For the purposes of this paper, we have chosen to compare the two implementations of the method (see Table 1 for the two methods' input data) by examining the outputs for both the Southampton and Dartmoor test areas, for three times – 02:00, 08:45 and 16:45, looking at both Term Time and Out of Term population movements. These times were chosen to account for various different types of population movements. At 02:00, we expect there to be minimal movements, and so the population estimates should primarily represent the residential population of the area, with some movements relating to night shift working and emergency healthcare. At 08:30, we would expect the majority of people in employment or education to be in transit to their respective day-time locations. At 16:45, we would expect a large proportion of working people to still be in their place of work, but that the majority of people in education to have left. Additional time periods to investigate were also considered, such as 14:00 to properly capture population movements to educational establishments, but were not available at the time of writing. Because the new methodology allows us to create healthcare datasets with different daily capacities for specific months, we have chosen to use the June healthcare destination files to accord with MYE reference dates. To assess the similarity of the two population maps quantitatively, we are using the Spearman's Rank correlation coefficient (cell population counts are not normally distributed) of the populations of each grid cell for the two data selection and preparation methodologies, excluding all cells

which contain fewer than one person in both population maps in order to reduce the effect of unpopulated cells that would be highly correlated with each other regardless of the differences in the two methods. We are interpreting the rank correlation as a measure of the similarity of relative cell population sizes between the two methods at a given location and time. We report p-values to adhere to normal standards but recommend that they are interpreted cautiously as the Population 247 method is effectively a deterministic simulation (parameterised by real data) and p-values may not be meaningful. As we are not intending to interpret p-values we have not attempted to adjust for multiple comparisons.

Results

At present, the biggest unanswered question for this research is how to best validate the population estimates produced. Further discussion of possible validation strategies considered follow in the Discussion section of this paper but have not yet been implemented due to time constraints and issues with data availability.

While we have not yet implemented a method to validate how accurate the estimates are, we were able to compare them with estimates produced using origin, destination and time profile files developed by Cockings et al. (2021), to ensure that the estimates produced from our method are sensible and similar to those produced in previous work, and identify areas where our estimates diverge from those previously produced, due to additional or updated sources of data. These estimates were developed to redistribute people based on population movements for workplace, education, and healthcare reasons, and are based on various sources of open data available for the year 2011. Because we are developing population estimates for the year 2019, we expect there to be (in some areas significant) differences between the estimates we produce and the Open Data Methodology due to changes in population over time, but we expect the two methodologies to broadly correlate overall.

Figure 1 below gives an example of the distribution of cell counts for the Southampton test area at 16:45 during term time. It can easily be seen from this example, and from Shapiro-Wilk normality tests, that although the distribution of cell counts across the test area appear similar for the previous and current framework implementations, they are not normally distributed (p<0.01 for both the previous and new implementations of the framework for this test area). Non-parametric assessment methods, such as Spearman's Rank correlation, have therefore been implemented for comparisons of the old and new methodology.

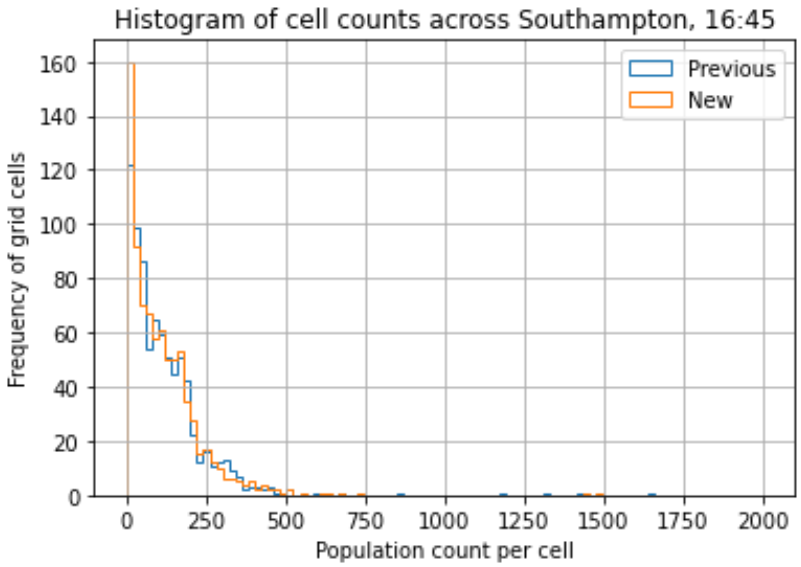

Figure 1: A histogram showing the distribution of cell counts for the Southampton test area at 16:45 during term time, for both the previous 2011 case study and our updated (new) use of the framework.

A Kolmogorov-Smirnov test on the above example (see Figure 1) fails to indicate a statistically significant difference between the distribution of cell counts from the old and new applications of the method ($D(1600) = 0.05$, $p = 0.25$).

Q3 for panel Can you think of a better way, other than Spearman's rank correlation, to assess areas of similarity and difference between the updated application and the previous case study usage of the Population 24/7 framework?

Figure 2 to Figure 9 in this section show population maps produced using the Population 24/7 methodology. Figure 2 shows how the population distributions differ according to reference time, and Figure 4 to Figure 9 show examples of how the population distribution output by the framework differs according to the data preparation methodology. Figure 3 shows how the population per cell in the Southampton test area changes between the old and new methodologies at varying times of day.

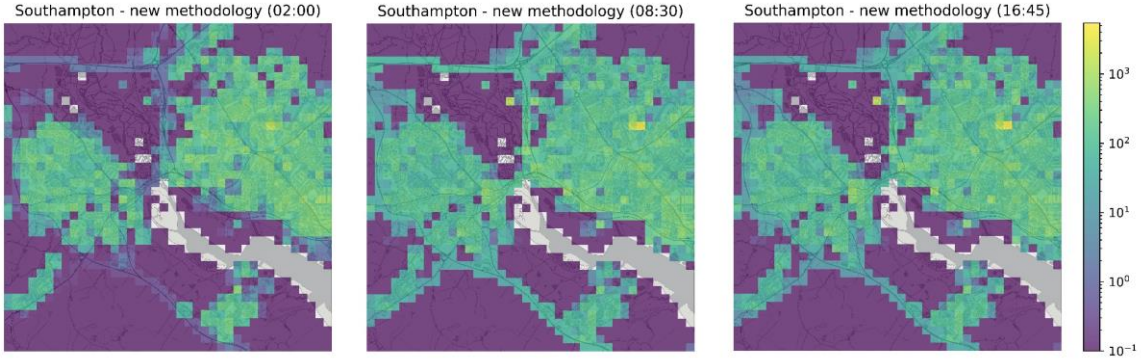

Figure 2: Maps showing the population changes over the three reference times for Southampton during term time, using a logarithmic colour scale. All cells with a value of less than 0.1 people have been clipped to 0.1 for this rendering.

In Figure 2, we can see that the motorways and other roads show the greatest difference across the time periods, while the population in built up areas differs less. In the top right corner of each map, we can see Southampton General Hospital as an area of higher population compared to the surrounding grid cells.

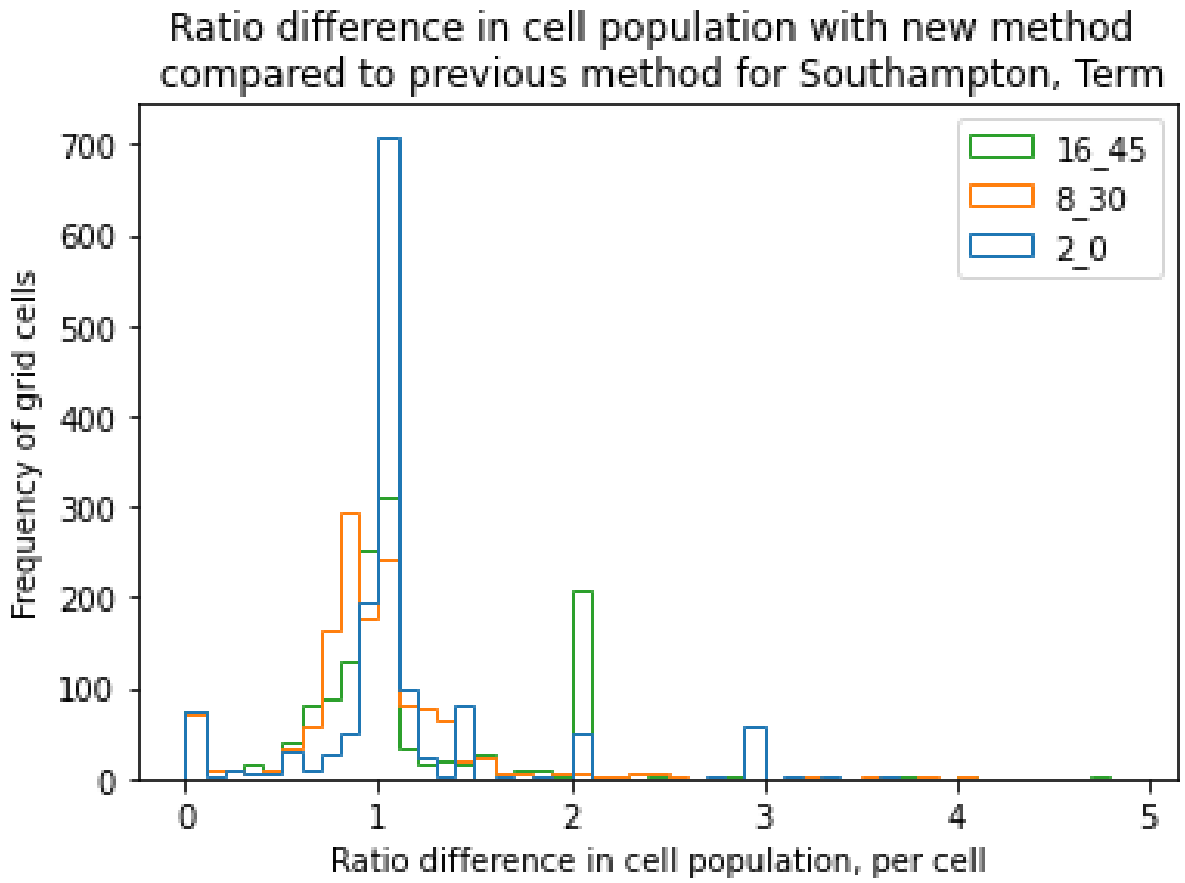

Figure 3: Histograms showing the ratio of the population in each cell of the Southampton test area for the new method compared to the previous method, for three reference times of day. For clarity, all cells with a ratio increase of five times or more have been grouped together in one bin to the far right of the plot.

Figure 3 shows that, at 2:00, the majority of cells in the Southampton test area have similar or slightly increased populations in the new application of the framework compared to the case study from 2011, in line with the increase of the population in England more generally shown in the MYEs between 2011 and 2019. There is a long upper tail indicating around 25% cells which have increased in population by a multiple of at least 1.25, and a short lower tail representing a smaller number of cells which have decreased in population in the 2019 methodology compared to the 2011 methodology. By 8:30, variability in the cell counts between 2011 and 2019 has increased compared to 02:00, with this variability becoming more pronounced at 16:45.

Southampton - 02:00

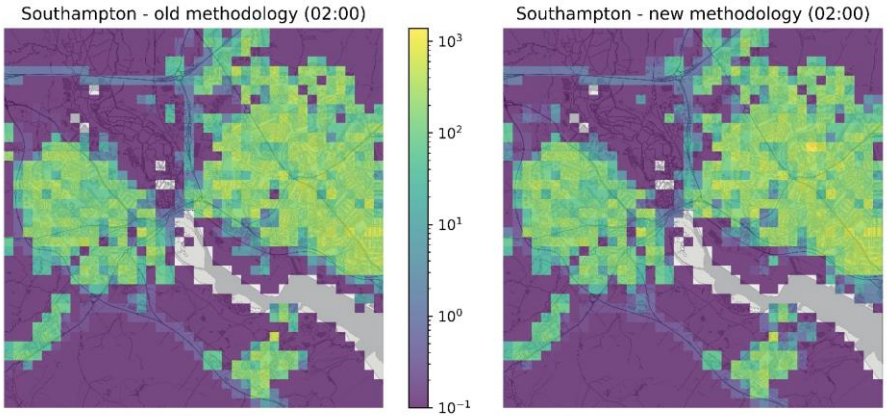

Figure 4: Population maps showing the overall population totals during term time for the Southampton Testing area at 2:00, using a logarithmic colour scale. All grid cells with a population of less than 0.1 people have been clipped to 0.1 for this rendering. Background image for this and other figures created using the Ordnance Survey Open Zoomstack, available under the Open Government Licence.

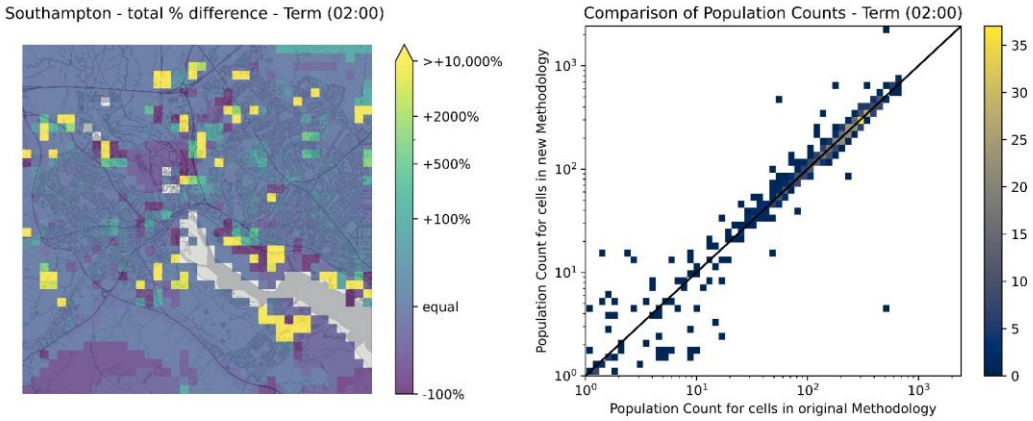

Figure 5: Right - A population map showing the percentage difference in population totals for the new methodology, compared to the previous case study, during term time for the Southampton Testing area at 2:00 using a logarithmic colour scale. All grid cells with a population of less than 0.1 people have been clipped to 0.1 for this rendering. Left – a colour map showing the pairwise comparison of the absolute population estimated to be present in each cell from the new methodology and the previous case study.

At 02:00, the population estimates for the open data method and the new method are very similar. For the Out of Term-Time populations, the total number of people within the area at that time has increased in the new method's results, from ~119,000 to ~124000, a 4% increase, and the overall correlation for the values of each grid cell between the two methodologies is 0.97 ($r(722) = 0.97$, $p<0.001$). There is a large difference in population values (~4000) for four cells centred on Southampton General Hospital, but this seems to be because of differences in the local dispersion parameter for that destination (the open data methodology spreads the population located at that origin over a larger area, resulting in fewer people in each cell). Breaking down these population maps by age group, we see very high correlations between the two maps for the age groups representing the under-16

population (for 0-4 year olds, $r(564) = 0.90$, $p<0.001$; for 5-9, $r(567) = 0.89$, $p<0.001$; for 10-15, $r(571) = 0.89$, $p<0.001$), as well as for the 18-64 age bracket ($r(719) = 0.97$, $p<0.001$) and the over-65 bracket ($r(582) = 0.91$, $p<0.001$), although there is a slightly lower correlation for the 16-17 age bracket ($r(532) = 0.81$, $p<0.001$), likely due to the removal of 16-17 year olds from the workplace movements in the new methodology. For the Term-Time populations, the overall correlation remains high, at 0.97 ($r(721) = 0.97$, $p<0.001$), though there is no correlation for the 18-64 student age bracket ($r(495) = 0.07$, $p=0.09$) – this is likely due to the changes in the way students are assigned to residential areas, the removal of students from workplace processing, and also possibly due to changes in the overall residential distributions of students over the intervening eight years, though more investigation is required.

From Figure 5, we can see that there are some cells which have increased dramatically in population between the previous and updated methodologies. Some of these differences can be accounted for by changes to local dispersion parameters; many of the cells with very large percentage increase are cells with almost zero allocated population in the previous methodology which lie close to the borders of cell groups with higher population and often have adjacent cells which have seen an almost total (close to 100%) drop in population between the two methodologies. Some other areas may see dramatic increases due to the addition of housing or other types of origins and destinations in the years between 2011 and 2019. Further investigation into the exact cause of the localised increases and decreases in cellular population at these times is needed in the future and is discussed in more detail later in the paper.

Southampton – 08:30

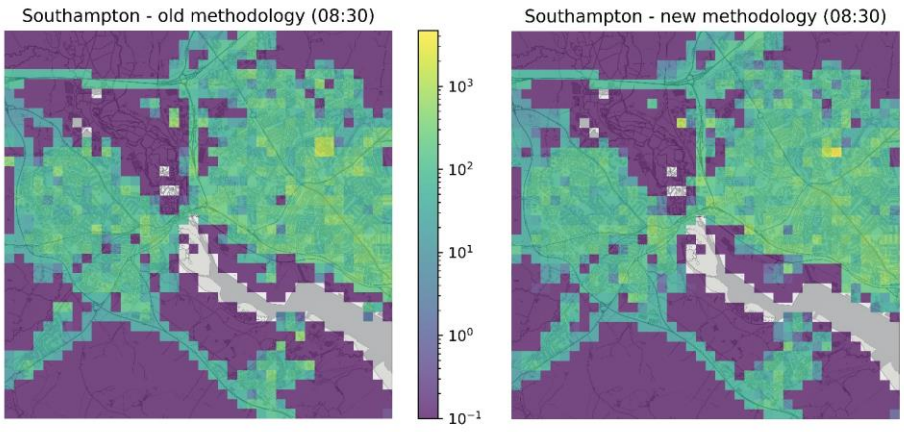

Figure 6: Population maps showing the overall population totals during term time for the Southampton Testing area at 8:30, using a logarithmic colour scale. All grid cells with a population of less than 0.1 people have been clipped to 0.1 for this rendering.

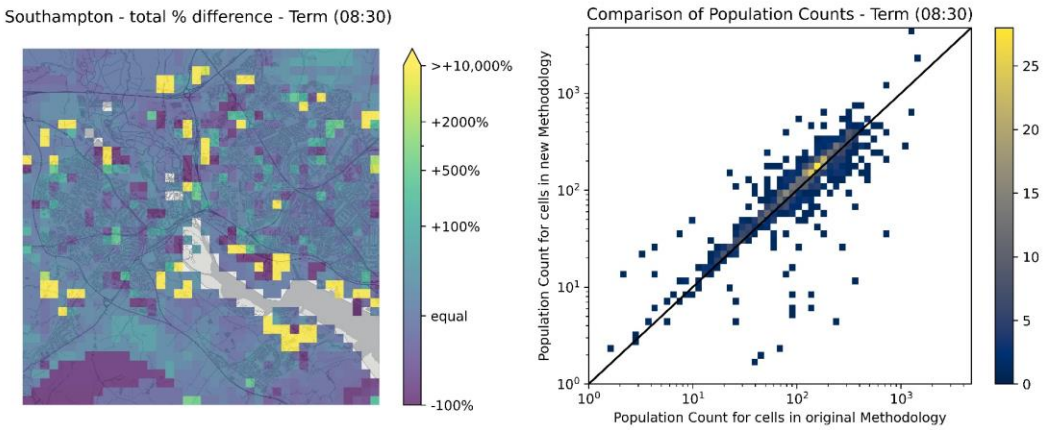

Figure 7: Right - A population map showing the percentage difference in population totals for the new methodology, compared to the previous case study, during term time for the Southampton Testing area at 8:30 using a logarithmic colour scale. All grid cells with a population of less than 0.1 people have been clipped to 0.1 for this rendering. Left – a colour map showing the pairwise comparison of the absolute population estimated to be present in each cell from the new methodology and the previous case study.

For the 08:30 population estimates, again we see broad similarities between the Open Data and the New Methodology. Overall, for the term time population, the cell-wise correlation of the two methodologies when looking at the total population out of Term-time is 0.87 ($r(844) = 0.87$, $p<0.001$), and a cursory evaluation of the produced population maps shows similarities in overall population distributions. Because we are using the same transit population background layers as in the Open Data Methodology, there is a very large correlation between the transit populations in both methodologies $r(333) = 0.84$, $p<0.001$), but even taking that into account by removing the in-transit population from both maps, the cell-wise correlation remains high at 0.85 ($r(742) = 0.85$, $p<0.001$), suggesting that this is not a major driver of the similarity of the results. Note that, even though the background layer is derived from the same data, the correlation of the in-transit population between the two models is not exactly 1; this is because the background layer weights a population total which is defined by other data sources, rather than estimating population counts directly. However, breaking down the maps by movement type, we see that while the population remaining at their origin locations is very high ($r(589) = 0.98$, $p<0.001$), the correlation between the populations at their day-time destinations is considerably lower ($r(426) = 0.13$, $p=0.009$). For the term time populations, the correlation between the two methods is 0.85 ($r(849) = 0.85$, $p<0.001$) overall, and 0.80 ($r(747) = 0.80$, $p<0.001$) excluding transit.

Southampton – 16:45

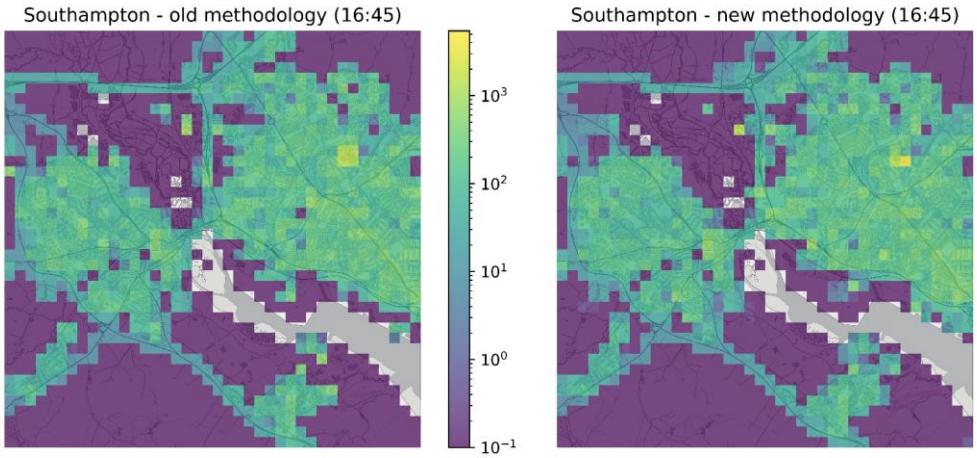

Figure 8: Population maps showing the overall population totals during term time for the Southampton Testing area at 16:45, using a logarithmic colour scale. All grid cells with a population of less than 0.1 people have been clipped to 0.1 for this rendering.

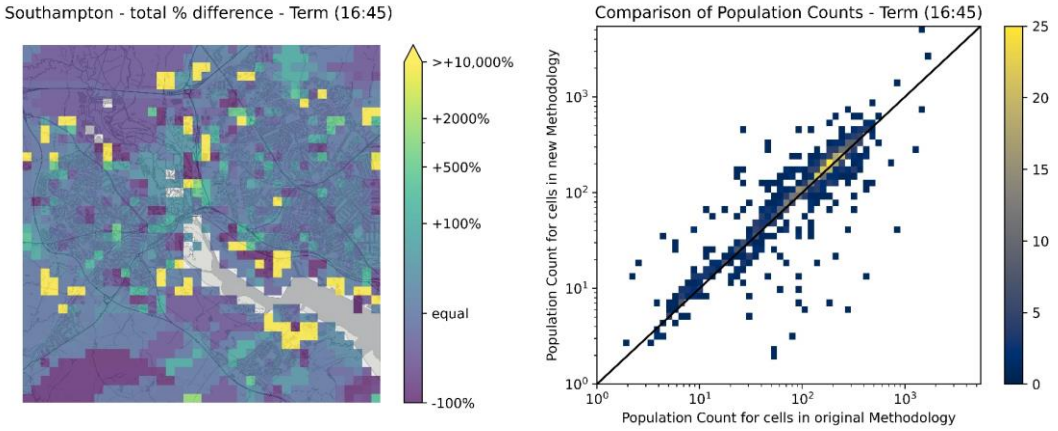

Figure 9: Right - A population map showing the percentage difference in population totals for the new methodology, compared to the previous case study, during term time for the Southampton Testing area at 16:45 using a logarithmic colour scale. All grid cells with a population of less than 0.1 people have been clipped to 0.1 for this rendering. Left – a colour map showing the pairwise comparison of the absolute population estimated to be present in each cell from the new methodology and the previous case study.

Again, for the 16:45 population estimates, there is a large correlation between the new and Open Data methodologies – 0.86 ($r(567) = 0.86$, $p<0.001$) for the Out-of-Term population maps, and 0.85 ($r(852) = 0.85$, $p<0.001$) for the Term Time populations, though with a weak correlation for the 16-17 age group ($r(587) = 0.18$, $p=0.004$) in the Term-time population maps, likely due to the removal of this age group from the workplace destination processing and the lack of data available for dedicated sixth forms in the education datasets.

Retail population estimates

The introduction of retail location data in the new methodology has made a noticeable difference to the results. The cells with retail destinations had, on average, 36.57% higher estimated population compared to the results obtained with the open data. The estimates for

cells without retail destinations increased by a much smaller degree (3.85%). In both cases, there was a strong correlation between the results obtained with both methodologies: r(697) = 0.81, p<0.001 for cells with retail destinations and r(24028) = 0.84, p<0.001 for cells without retail destinations. For cells containing retail destinations, the biggest differences between the methods have been observed for the on-site population (number of individuals who are estimated to have travelled to a certain non-residential location within the cell at a specific time) which, on average was 77.06% higher for the results obtained with the new method. In contrast, the on-site estimates for cells without retail destinations were 2.89% lower than the ones obtained with the open data methodology. In both instances, the correlations between methodologies were low to zero: for cells with retail: r(234) = 0.338 p<0.001 and without retail: r(4047) = 0.0002, p=0.99. The same pattern of results has been observed for the in-travel population estimate (but with a much higher correlation between methods, over 0.6) as well as for all age groups and times: increased total counts in cells with retail destinations with much smaller increases (or decreases) for locations without retail destinations (see Table 3for example).

Please note that although these correlations (or lack of) indicate that there may be substantial differences between the old and updated methodologies, more work needs to be done to investigate the causes of these discrepancies.

Table 3: Percentage increase in population summed over cells containing and not containing retail locations in the Southampton test area in our estimates compared to Southampton's methodology (Cockings, et al., 2021) for three different reference times.

Increase in total population of cells at given time	***02:00***	***08:30***	***16:45***
with retail (%)	56.74	29.00	27.55
without retail (%)	2.41	4.34	5.14

Dartmoor

Looking at the Dartmoor testing area, we again see very high correlations between the two methodologies (at 02:00, $r(13) = 0.99$, $p<0.001$; at 08:30, $r(19) = 0.72$, $p<0.001$; at 16:45, $r(10) = 0.65$, $p<0.001$) for the out-of-term population maps*)*. Interestingly, despite the fact that immobile populations are not marked as immobile yet in the new methodology, there is still a reasonable correlation for the working age population in this area (at 16:45, $r(19) = 0.56$, $p=0.009$), though this is likely due to the relative isolation of the prison and the lack of nearby destinations which could draw this population, and is not likely to hold true for other prison locations.

Analysis Conclusion

Overall, preliminary investigation of the new methodology as compared to the methodology used in Cockings et al. (2021) suggests that while there are minor discrepancies between the outputs of each methodology, the new methodology produces numbers that (taking into account population changes over the years between 2011 and 2019) are broadly sensible as estimates. Future work in this area will include further investigations into more granular comparisons of the two results by age group and by destination type. Additional work will include testing areas with different characteristics, for example, the City of London, an area with enormous differences between the workday and residential populations, and Blackpool, an area with significant seasonal tourist populations that are not yet incorporated into either

methodology. However, as we currently have no way of validating either of these methods against a benchmark population estimate, it is still unclear how accurate these estimates are, and whether any differences introduced using new data are for better or worse.

Discussion

Comparative results

There are a few key areas where our updated outputs differ from those used for the previous case study, which may go towards explaining some of the differences. The first of these is change to the baseline origin population; the population of the United Kingdom as a whole has grown between 2011 and 2019 and therefore, there is a larger starting population for redistribution in our new methodology compared to the previous iteration. Other changes to the population within the test area which are not already included in the origin information (e.g., travel in and out of the defined test area for visits, commuting or short-term migration) are not accounted for in either iteration of the model. A second is the change in the number and location of destinations (and origins) where new housing, leisure and retail spaces have been built, changed ownership or otherwise changed usage. A third difference may stem from changing accuracy or granularity due to the inclusion of sensitive data and updated assumptions. Additional work is necessary to unpick these differences and conduct more standardised comparisons to enable direct comparisons of, for example, the model outputs with and without sensitive data for the same time period.

Rank correlation has been used as a measure of change between various population counts in this paper; as discussed previously, it's important to note that a correlation coefficient of 1 is not expected in the majority of cases, but correlation coefficients do provide a standardised measure of the extent to which a variety of different components impact the results. Correlation coefficients are an easily interpretable and standardised way to compare the impact of software and input changes such as model differences and data differences, as well as population changes such as population growth or the impact of additional infrastructure. The rank correlations in this paper should therefore be interpreted as a tool to highlight areas of difference, and not directly as a measure of error or bias in the methodology itself.

Data caveats

There are currently several areas where additional data is needed to measure mobility of additional populations which are not currently included in the framework.

The origins data, from which people are drawn to fill destinations, does not currently cover predominantly immobile populations, such as incarcerated persons and those in specialist care facilities. Additional data sources are currently being investigated to add immobile populations to the framework. Temporary living quarters, such as hotels, shelters, and other types of short-term or impermanent living structures, for example, those used by transient communities, are also not included in the framework at present.

The background layer, accounting for the population in transit at any given time, currently only covers the population on major roads (primarily A-roads) and motorways. The background layer makes no account of the mode of transport; those using public transport (e.g., buses, trains), cycling, or walking will be distributed as in-transit to cells around the destination they are travelling to according to the cell weights derived from DfT average annual daily flow data used to create the background layer. No coverage is available for use of minor roads. This is likely to have a particularly noticeable effect in areas where use of public or non-motorised transport is particularly widespread (e.g., London and other large city centre locations), and in more rural areas where there are large distances separating major roads. Addition of the train network to the framework is being investigated by the

University of Southampton in parallel to the work being conducted within the ONS. Further work into modelling population mobility via other forms of transport is currently being done by the mobility squad in the Data Science Campus, using agent-based modelling methods and drawing on information from the National Travel Survey and telecoms data. We are in regular contact with these groups and may be able to incorporate elements of their work into the Population 24/7 framework in the future.

Whilst most publicly-funded primary and secondary education facilities have been included in processing, independent schools aren't accounted for using the present methodology, and variable time profiles for schools with boarding pupils have not been applied. Additionally, colleges and dedicated sixth-form facilities exclusively for pupils over the age of 16 are not listed on the English school census and so are not included in our processing. Although there are perhaps only a few independent schools and around five dedicated sixth forms and colleges in our Southampton test area, not including these facilities could lead to a large undercount of population present in the relevant cells in age bands between 4 and 18. One potential source of information which could be drawn from to fill most of these gaps is the Department for Education's Get Information About Schools (GIAS) service, which contains information on children's centres, academies, free schools, maintained schools, independent schools, further education colleges (further education and sixth form corporations, specialist designated colleges and special post 16 institutions) and higher education institutions. Much of the information in the UKRLP is drawn from GIAS. Additional information available from GIAS which is not currently available in the English/Welsh school census or the UKRLP could help to address issues arising from the current methodology with respect to teaching sites being different from the legal registration address (e.g., academy chains) and school address and status changes. Since GIAS contains information about schools in both England and Wales, the need for separate processing of the Welsh school census for schools in Wales would be reduced by using GIAS. However, use of GIAS has some drawbacks. GIAS does not readily surface historical student numbers, so it would be harder to create accurate estimates for anything other than the most recent year. Since no student address information is available in GIAS, it would not be possible to calculate school specific wide area distributions for schools based on GIAS alone.

At present, all higher education students are distributed evenly within a 200m radius centred on the primary campus. There are a few issues with this approach. Firstly, the presence of any secondary campuses or other teaching sites is not accounted for; whilst HESA does provide data on secondary campuses, this data is not currently available within ONS/DAP. In extreme cases this can cause coverage issues that span counties, for example the University of Exeter in Devon has a secondary campus in Cornwall. The assumption that campus areas fall roughly within a 200m radius is also unlikely to be accurate in many cases. Whilst HESA can provide information on the size of university estates, this is aggregated by provider and not by campus. In any case, acquisition of additional data on secondary campuses and estate sizes would not address issues with assigning students to locations in non-campus universities with buildings dispersed throughout a city or across a wide area (for example, Cambridge, many London universities). Further assumptions that time profiles for part-time students are proportional to those of full-time students may not be correct, but little data have been identified which can give insights into the exact study patterns of part-time students. Students

participating in distance learning, particularly since the pandemic, may also positively skew the estimated student population present on campus.

Estimation of time-varying healthcare populations is much less well-defined than those in education. The HES data we currently have access to contains the location of the census OA of residence of each patient but contains no information on which healthcare site or provider they visited. It has been assumed that each patient visited their closest hospital (in the case of inpatient and outpatient data) or A&E department (in the case of A&E admissions), but this assumption is unlikely to be accurate in many cases. For example, most patients with prebooked appointments will visit a hospital which provides specialist care in their specific ailment; this may not always be the closest hospital to the patient's home. In addition, the HES data does not contain any information relating to specialist clinics, general practices, private healthcare facilities or other healthcare locations. Additional data sources are needed to account for visits to non-hospital and non-NHS healthcare locations. Further work is planned to include GP Episode Statistics and locate data sources for measuring private hospital attendance. Additionally, it is hoped that access to a more disclosive cut of the HES data could be negotiated in the future, which would enable time profiles and magnitudes to be constructed more accurately for individual hospitals.

Workplace populations are estimated directly from the IDBR, using various methods of aggregation to apply appropriate time profiles. The key assumptions made in this approach are: travel to work distances do not differ for a WZ by industry, and have remained broadly similar to travel to work distances in the last census year, all workers are within the 18-64 age band and all workplaces within a given industry cluster follow the same time profile. In the methodology from Martin, Cockings, & Leung (2015), the average proportion of the workforce which lay outside of the 18 to 65 age band was only 4.7%. Since the reference year used by the Southampton paper of 2006 occurred before the introduction of the law requiring young people to remain in full-time education until the age of 18, (previously 16) (Education and Skills Act, 2008), the percentage of under 18s in the workforce is likely to have decreased further. The state pension age remained at 65 until 2018, and so the proportion of over-65s in the workplace is unlikely to have increased in our reference period of 2019 compared to the original work by Southampton; however, as state pension age increases, the proportion of over-65s is likely to increase and as such, this assumption that the vast majority of the workforce falls between 18 and 65 may need to be revisited in the future to include an additional component for estimating the number of over-65s in the workforce. The other key assumptions are difficult to test; following extensive research by the University of Southampton into the construction of time profiles for workplaces, it is likely that these assumptions are reliable. However, further validation of these assumptions may be necessary in the future. More generally, it is not yet clear that the IDBR accurately represents the distribution of the workforce at low spatial granularity – it is known that agency workers and travelling workers would be counted as working at the Local Unit they are registered to, despite the fact that they would not usually be working at that location, which would inflate the workforce population of those locations, and so more work may be needed to adjust the counts for these cells to take this into account.

Retail destinations are a new addition to the Population 24/7 framework and as such, data included in the modelling in this area is still relatively sparse and requires further development. Our assumption that customer numbers are proportional to the number of staff

working at a location at any given time may not be a good assumption for many retail destinations, for example, if shelf-stacking or online delivery preparation is done outside of usual opening hours. Proportions of staff to shoppers may vary depending on the retailer, and may ebb and flow through the day or with the seasons in a way that is not fully reflected by staff attendance. It is hoped that the addition of data from the UK Time Use Survey and sample Smart Sensor data (connections to Wi-Fi sensors for a single high-street retailer) can be used in the near future to reconstruct time profiles and more accurately estimate shopper magnitudes.

Tourism and leisure destinations are not currently accounted for in the Population 24/7 framework in any form. This is likely to lead to significant underestimates of daytime population in a number of urban and suburban areas where leisure parks and leisure venues are located, in addition to both urban and rural areas with high levels of tourism activity. There are several data sources which may be helpful in estimating tourism- and leisure-related activity, including open data on annual visitor numbers to attractions from VisitBritain and information on the regularity of specific leisure activities from the Taking Part survey, and work to incorporate leisure destinations is expected to commence soon. The Taking Part survey and its successor, the Participation Survey (available from 2021) are available with annual microdata via the UK Data Service and should be straightforward to obtain, but the full extent to which this data can be used is yet to be determined.

At present, the Population 24/7 framework can produce high granularity statistics, but since this relies entirely on the input data from a broad range of sources, timeliness of new inputs (and therefore outputs) is a concern. One possible option for increased timeliness of outputs is use of low-latency indications such as call data records, footfall and other mobility data; this approach has worked well for other statistics agencies such as StatsNZ but may not be as easy to apply to UK statistics. Additional discussion is provided later in this section.

Q3 for panel

Are there any obvious ways to address any gaps or assumptions we've made that aren't listed here?

Validation Strategy

Validation of the outputs is extremely challenging because we use the best available demographic data, mostly from official statistics, as inputs to the model to create population distribution estimates at high spatiotemporal resolutions which are not captured directly at a similar level of granularity by any other measurement system. Because there are no true values available to directly compare the model outputs with, most of the ideas discussed below will provide only partial indications of validity.

One method for qualitative evaluation of Population24/7 outputs involves sense-checking using specific test areas with known particularities; the test areas we've chosen for sense-checking are listed with their key characteristics in **Error! Reference source not found.**. Initial validation of this type is ongoing but is in the early stages; the work described in

Results around the Southampton and Dartmoor test areas is an initial example for how this kind of validation might be carried out. This kind of qualitative evaluation of test areas allows us to quickly identify any key areas where estimates may be inaccurate or where additional data may be needed. For example, one might logically expect a substantial population (predominantly tourists or those visiting for leisure purposes) to be present on the beach in Blackpool at midday on a weekend out of term in the summer, whereas the population in this same area can reasonably be assumed to be much reduced at midnight on a weekday in term time; if estimates produced for the area of Blackpool beach do not match our initial speculations, further investigation is warranted and a need for additional or more accurate data in this area is immediately highlighted.

Population24/7 relies on a wide range of input data sources, each collected at different time intervals, for specific purposes, possessing different characteristics, quality, and weaknesses that affect the validity of the model. As such, the second validation step we plan to take is to assess the quality of each input source on a range of criteria (e.g. timeliness, trustworthiness, completeness, purpose) and then to document the potential sources of noise resulting from the way we represent and adapt the data within the model (for example, currently, we assume that the number of customers in a store is a scalar multiple of the number of employees working at the store at that particular time, which is just an approximation and, as such, it introduces error in the model). Drawing on information resulting from these input source profiles and data representation assessments, we will identify key criteria to develop a scoring framework for the model that we can use to get an overall indication of expected accuracy.

Another validation exercise we intend to conduct in the short term is to compare our results with those published by the EU ENACT project which aims to produce day and night–time monthly population estimates for the entire EU. The ENACT model is less granular and uses fewer input sources to determine population dispersion (e.g., there is no account of healthcare or retail mobility, population in transit or of those working night shifts) and it has not been validated with UK data. To be able to compare the results, our outputs will have to first be aggregated and processed. These issues will limit the extent to which we can interpret the differences and to generalize the conclusions beyond the aggregated spatial, temporal and demographic units used, however, the comparison will serve as an opportunity to identify obvious inconsistencies and errors and to confirm that our outputs are within plausible ranges.

In the long term we are considering the possibility of acquiring footfall counts derived from high-frequency data e.g., CCTV, Wi-Fi sensors for several different small areas (e.g., high streets, shopping centres, hospitality venues, towns), and mobile phone mobility data with broad coverage. These data represent direct observations and are usually available for any time period, and thus would offer the best approximation possible for the true population distribution that we are trying to model. There are issues to consider in determining the validity of the footfall data itself (e.g. the quality of the methods used to extract the counts from each data source and the way it is integrated, the accuracy of the estimated demographic data, the actual temporal and spatial coverage, etc.), but if these data prove to be of sufficient quality, they will allow us to determine the model validity with the highest spatiotemporal accuracy possible (within the sampled areas).

The most promising source of high-frequency mobility data is telecoms data from mobile phone network providers initially acquired for COVID-19 research by Data Science Campus. These data record interactions with cell towers and have been used to derive hourly footfall counts down to the level of Middle Layer Super Output Areas (MSOAs). While there are questions around the coverage of the data due to incomplete mobile phone ownership in the population and incomplete market share coverage for the suppliers working with ONS, these data could be used to derive indexed mobility trends. The data are weighted by the suppliers using ONS population statistics to make them more representative of the population, but it is unclear whether these data could be used to determine population level over time to a reasonable quality. The suppliers also apply bespoke data engineering processes to the data that we have no control over (due to acquisition of the data through a number of intermediate organisations), for example to determine a user's usual residence based on overnight phone location across a period of time. This could pose definitional problems where the suppliers' operational definitions do not meet our statistical needs, and could threaten the stability of the data if these processes change.

Discussions with our colleagues in StatsNZ have revealed that it is possible to use mobile phone data to produce population maps in a timely manner with great accuracy. StatsNZ have very high coverage of the mobile network and a close relationship with suppliers such that they can fully control and understand the methodology used to, e.g., weight and aggregate their data. Unfortunately, due to complexities in the agreements via which the data are made available within the ONS and concerns from mobile providers around public perception of the sharing of personal data, we are unable to foster that same two-way with providers as those managed by StatsNZ. Further issues arise from the market share dispersion of a larger number of major mobile networks in the UK which isn't the case in New Zealand. These factors make use of mobile phone data as a reliable indicator of magnitude unrealistic in our case. However, this mobility data may still provide a good signal of movement trends and provide a more timely indicator of population mobility than any of our current data sources. Mobile phone data could therefore either be used as an input data source, for validation of our current models, or a combination of the two; for example, mobility data from the area around a single destination (such as a single shopping centre) could be used to create a general time profile for other destinations with similar characteristics, and validated using the mobility data for those other destinations.

Some key validation measures which we are considering implementing in the future include comparing correlations and trends from mobile usage data, using data from other sources such as the Google Community Mobility Reports and comparisons of our results with those from other synthetic models (e.g., the Data Science Campus Agent Based Modelling project for travel to work statistics). Further validation in the form of user engagement with, for example, local authorities, could be considered in the future; however, at present, the feasibility study (benchmarked at 2019) is not considered timely enough for this to be productive.

Q4 for panel

What are your opinions on our planned validation strategy? What other methods of validation do you think are worth pursuing?

Practical limitations

In addition to the technical limitations discussed above, there are several operational limitations which need to be considered when using the Population24/7 framework within the ONS.

To incorporate sensitive data held within the office, all processing must be carried out within the Data Access Platform (DAP). The SurfaceBuilder247 software, which the University of Southampton have written to implement the framework, was originally written in Visual Basic and was incompatible with ONS systems. The software has since been rewritten in Python and an early part of ONS's work has been to test this new software implementation and import it into DAP. This process is now complete, however, any future changes to the SurfaceBuilder software will not be automatically updated and need to be pulled to the working area via a ServiceDesk request from the software user. Processing in DAP presents additional issues, such as a lack of mature geospatial tools within the platform and difficulties in processing non-standard data formats (e.g., SPSS).

Preparation of data for input to the SurfaceBuilder247 software is a time-consuming task which, due to the complexity and range of data sources used, cannot be easily automated. This is further exacerbated by the necessity to import and export all data through a third-party ingestion team, which, whilst necessary, introduces additional delays to the processing pipeline, especially in the cases where data need to be processed partially outside of and partially within DAP (e.g., where file formats are non-standard and need to be converted before ingestion or where, due to platform limitations, geospatial processing needs to be carried out using tools not available in DAP). In the cases where data need to be manipulated in more than one location (i.e., both inside and outside of DAP), updates to processing when new data become available may be particularly slow and difficult to apply.

Due to the nature of the framework, several small datasets are used which aren't commonly part of larger ONS outputs. A large proportion of these data aren't georeferenced (i.e., don't contain geospatial codes for analysis), and prioritisation for georeferencing via the address index matching service (AIMS) is low. This introduces a significant overhead in manually geolocating origins and destinations for input to the SurfaceBuilder software.

Taking these constraints into consideration, Population 24/7 may not be an appropriate method for the production of regular population mobility statistics but, in its current form, may be more suited to bespoke case studies with particular spatiotemporal population research questions. The degree to which high-frequency and low-latency mobility data (e.g., mobile phone data) could adjust for spatiotemporal coverage issues may determine whether the method is practical if any such adjustment could lessen the need for such a wide array of other data.

For production of this document, the data identification, acquisition and pre-processing stage has taken approximately 6 months with a rough allocation of 1.4 full-time equivalent hours. Although work could be done to automate parts of this process to produce regular mobility statistics, it is difficult to estimate the resource which would be necessary on an ongoing basis. The bulk of the work which would still need to be done on a regular basis would involve constant research into new data sources and quality assurance on more timely iterations of data sources already included. Modification of these new data and quality

assurance of updated iterations of existing data could require significant resource which is very difficult to estimate until the particulars of the work have been identified.

Accuracy vs Impact: Data Inclusion
There are a vast number of types of travel which aren't currently included in the framework. When adding additional data sources to the framework, consideration needs to be given to the size of that data source and the impact on total population movement that the data will make. For example, it may be possible to conduct a significant amount of work to integrate a small flow of individuals (for example, obtaining volunteer numbers by time of day for a small local charity shop), but inclusion of this data is unlikely to have any noticeable impact on model outputs. A decision needs to be made about the minimum size/capacity of an origin/destination to warrant inclusion; ultimately, this decision will need to be made in response to specific user needs and the purpose of the statistics, but may have a big impact on the specific locations and data that are included/excluded and processed as part of the outputs, and on the amount of resource required to pre-process and maintain the additional data.

Reliance on Census
Much of the data used in the current implementation of the Population 24/7 framework is drawn directly from the Census. In the absence of another census, replacements would need to be identified for the following data, currently sourced from Census and Mid-Year Estimates:

- General resident population, broken down by age, at a geographical granularity equivalent to that currently provided (output area)
- Communal establishments resident populations
- Geographical information or maps to match new inputs for resident populations
- Geographical information on work zones matching the Labour Force Survey and Inter-Departmental Business Register outputs

Accuracy in these estimates, especially those for resident populations, is of paramount importance. Since the current framework uses these resident population estimates as a base for redistribution, the validity of the entire framework will be compromised and granularity restricted without a sufficiently reliable and accurate source for resident numbers.

Presentation and communication of error
'Day-time' spatiotemporal population statistics are likely to persistently have higher error than our traditional population statistics. There are too many unpredictable and stochastic factors that determine small-area population presence at any given time of day to estimate population accurately without high-coverage real-time data (which are not currently available to us). This means that when developing spatiotemporal population estimates we need to consider how to best present these imperfect statistics and communicate their error to users. For example, the degree of imprecision or bias in the statistics may mean it is inappropriate to publish granular population counts, and instead a relative indexed measure of population presence may be preferable.

Our work has not yet begun considering this in detail but should be led by discussions with users on requirements for 'day-time' population statistics.

Conclusion

Given the highly flexible and granular nature of the Population 24/7 framework and associated SurfaceBuilder247 software, the Population 24/7 framework was determined to be worthy of further investigation as a method for determining highly granular spatiotemporal estimates of population mobility. Initial feasibility work indicates that the Population 24/7 framework yields promising results, producing sensible estimates at a granularity of hourly windows for a grid of 200m square cells. Further investigation and development into the Population 24/7 framework is recommended, in order to validate the outputs and establish an accurate and reliable method for highly granular estimates of population mobility, using SurfaceBuilder247 software.

The key strengths of the Population 24/7 framework are its flexibility for inclusion of a range of data sources and capability of producing highly granular estimates of population mobility in both spatial and temporal domains. The main weaknesses of the current framework centres on the significant resource necessary for creating and maintaining the inputs to the dataset, and the dependence on data derived from the Census that may not exist in the transformed statistical system. Other indirect weaknesses in the application of the framework by ONS centre around availability of data; population mobility estimates from the software depend on the data input and therefore current estimates are not very timely and have several associated caveats and gaps in areas such as tourism. This isn't an intrinsic weakness of the framework but of the inputs to the framework itself, and can be remedied by introducing more data to the model to plug any gaps in coverage, including more timely indicators (for example, telecomms data). The lack of benchmark figures to assess the quality of estimates is also a general problem for any attempt to produce population mobility statistics.

Future Steps

Possibilities for future work on this project are extensive; in the coming months, we hope to:

- Plan and implement a strategy for model validation, as discussed above. (We see this as the highest priority.)
- Continue communications with Data Science Campus around potential overlap and progress on agent-based modelling methods for travel-to-work estimates
- Incorporate mobile phone mobility data as a source of generic mobility information i.e., not related to any specific type of activity
- Update the background layer with more recent traffic data. This step is in progress and traffic counts data from 2019 has been acquired and processed. However, this processing has been temporarily paused due to difficulties in computing resource related to geospatial processing.
- Enhance estimates for healthcare destinations using additional data sources, such as GP Episode Statistics.
- Augment origins information by adding data for immobile populations.
- Extend work into leisure and tourism destinations, including research into new sources like the Time Use and Taking Part surveys.
- Reconstruct and update time profiles for most destinations (retail, leisure, workplace) using Time Use Surveys.
- Extend and improve upon time profiles and magnitude estimation for retail data, using Smart Sensor wi-fi connection data for shopping centres.

- Conduct further analysis and comparisons of data between SurfaceBuilder runs using the University of Southampton's benchmarking run (Cockings, et al., 2021) and runs using sensitive and updated data from ONS (using 2019 as a reference year).

References

Alexis-Martin, B. (2016). RADPOP: A New Modelling Framework for Radiation Protection. *PhD Thesis*.

Batista e Silva, F., Rosina, K., Schiavina, M., Marín Herrera, M., Freire, S., Craglia, M., & Lavalle, C. (2017). Spatiotemporal mapping of population in Europe: The ENACT project in a nutshell. *57th European Regional Science Association (ERSA) Congress*. Retrieved from https://publications.jrc.ec.europa.eu/repository/handle/JRC107109

Bhaduri, B., Bright, E., Coleman, P., & Urban, M. L. (2007). LandScan USA: a high-resolution geospatial and temporal modeling approach for population distribution and dynamics. *GeoJournal, 63*, 103-117. Retrieved from https://link.springer.com/article/10.1007/s10708-007-9105-9

Cockings, S., Martin, D., Harfoot, A., Branson, J., Campbell-Sutton, A., & Gubbins, G. (2021). Population 24/7 Near Real Time: Data Library, Sample Outputs and Batch Files for England, 2011. [data collection]. *UK Data Service*. doi:DOI: 10.5255/UKDA-SN-853950

Data Ventures. (2021). *Mobility Index* . Retrieved 2021, from Github Code Repository: https://github.com/dataventuresnz/mobility-index

(2008). *Education and Skills Act.* Retrieved 10 05, 2022, from https://www.legislation.gov.uk/ukpga/2008/25/contents

Martin, D. J. (1989). Mapping Population Data from Zone Centroid Locations. *Transactions of the Institute of British Geographers*, 90-97.

Martin, D., Cockings, S., & Leung, S. (2015). Developing a Flexible Framework for Spatiotemporal Population Modeling. *Annals of the Association of American Geographers, 105*(4), 754-772. doi:10.1080/00045608.2015.1022089

Office for National Statistics. (2007). *UK SIC 2007*. Retrieved from https://www.ons.gov.uk/methodology/classificationsandstandards/ukstandardindustrialclassificationofeconomicactivities/uksic2007

Ordnance Survey. (2022). *Beginner's Guide to Grid References*. Retrieved from https://getoutside.ordnancesurvey.co.uk/guides/beginners-guide-to-grid-references/

Reis, S., Liska, T., Steinle, S., Carnell, E., Leaver, D., Roberts, E., . . . Dragosits, U. (2017). UK gridded population 2011 based on Census 2011 and Land Cover Map 2015. NERC Environmental Information Data Centre. Retrieved from https://doi.org/10.5285/0995e94d-6d42-40c1-8ed4-5090d82471e1

Smith, A., Newing, A., Quinn, N., Martin, D., Cockings, S., & Neal, J. (2015). Assessing the Impact of Seasonal Population Fluctuation on Regional Flood Risk Management. *ISPRS Int. J. Geo-Inf., 4*(3), 1118-1141. doi:10.3390/ijgi4031118