


Special Issue

An Open-Source Model for Projecting Small Area Demographic and Land-Use Change

Nik Lomax^{1,2} , Andrew P. Smith², Luke Archer², Alistair Ford³, James Virgo³

¹School of Geography, University of Leeds, Leeds, UK, ²Leeds Institute for Data Analytics, University of Leeds, Leeds, UK, ³School of Engineering, Cassie Building, Newcastle University, Newcastle upon Tyne, UK

The size, composition, and spatial distribution of both people and households have a substantial impact on the demand for and development and delivery of infrastructure required to support the population. Infrastructure encompasses a wide range of domains including energy, transport, and water, each of which has its own set of spatial catchments at differing scales. Demographic projections are required to assess potential future demand; however, official projections are usually not provided at a high level of spatial resolution required for infrastructure planning. Furthermore, generating bespoke demographic projections, often incorporating a range of scenarios of possible future demographic change is a specialist, resource intensive job and as such is often missing from infrastructure development projects. In this paper we make the case that such demographic projections should be at the heart of infrastructure planning and present a set of open-source models which can be used to undertake this demographic projection work, thus providing the tools needed to fill the identified gap. We make use of a case study for the United Kingdom to exemplify how a range of scenarios can be assessed using our model.

Introduction

Population and household estimates and projections are necessary to make informed decisions about infrastructure investments, from building new rail lines to broadband provision. The location and the composition of the population dictates current needs and demands, while projected changes in the population will dictate future needs. There is also a feedback, whereby any changes to the current infrastructure might influence demographic decisions, driving future population change. For example, developing new transport infrastructure might make an area more accessible, and thus attractive, but this might also drive-up property prices which would change

Correspondence: Nik Lomax, School of Geography, University of Leeds, Woodhouse Lane, Leeds, LS2 9JT, UK
e-mail: n.m.lomax@leeds.ac.uk

Submitted: 8 October 2021; Revised version accepted: 24 January 2022

doi: 10.1111/gean.12320

© 2022 The Authors. *Geographical Analysis* published by Wiley Periodicals LLC on behalf of The Ohio State University.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

the socio-economic composition of the area. We argue therefore that the estimation, modeling, and projection of demographic demand is an essential but often overlooked component of any project which is focused on the delivery of infrastructure projects or the assessment of current and future infrastructure needs.

A key issue in many countries is the lack of fine-scale and detailed population data available to help inform infrastructure plans. National Statistical Agencies (NSAs) around the world produce population estimates and projections; however, projections in particular are not routinely available at fine spatial scale. For example, in England and Wales, the Office for National Statistics (ONS) produce small area population estimates, usually released for the previous year, to a fine-scale Lower Super Output Area (LSOA) level (statistical areas which contain on average 1,500 people). However, population projections are only produced at a coarse Local Authority (LA) scale and the same is true for household projections (LAs denote the boundaries for local government and vary in population size, from around 2,500 to over 1.5 million people).

Additionally, there is a lack of flexibility in that the NSAs release pre-tabulated outputs, with models not generally available to be adapted to the specific needs of the infrastructure modeling team. This means that to assess different growth scenarios requires a bespoke demographic modeling solution. That demographic projections are needed at a local level for all kinds of applications is emphasized by Diamond, Tesfaghiorghis, and Joshi (1990), who make the point that many users will use official projections as a base and supplement these with additional information to deliver the detail they need.

Developing the models and providing the data needed for effective demand estimation and projection is the gap that the work outlined in this paper fills. We describe a comprehensive framework of open-source models for the projection of people and of households at fine spatial scale for the whole of the United Kingdom, which can be flexibly adapted to incorporate a wide range of infrastructure planning scenarios. These high-resolution projections from the demographic models are then coupled with a land-use model which allows us to better understand the spatial requirements and implications of the scenarios. Our models sit at the heart of a broader ecosystem called the National Infrastructure Systems Model (NISMOD) (Hall, Tran and Hickford 2016), with demographic data driving infrastructure demand models across a range of different sectors including transport, digital communications, energy, solid waste, and water supply. The modeling teams working on each of these different infrastructure sectors form the Infrastructure Transitions Research Consortium (ITRC), focused on creating a joined-up “system of systems” model, providing tools and results which are helping to inform and shape the future of infrastructure development in the United Kingdom.

This paper sets out the methodology and rationale for the demographic models used by the ITRC and fulfils the following objectives: (1) to demonstrate that demographic estimates are an essential input to infrastructure demand models, (2) to provide an overview of a comprehensive modeling framework that can be used by other modeling teams to produce demographic estimates and projections linked to land-use outputs, and (3) to demonstrate how that framework can be used to explore a range of spatial development scenarios. This latter objective is fulfilled by providing a case study of development in the Cambridge, Milton Keynes, and Oxford development “Arc” in the United Kingdom, but the model is adaptable to other areas where sufficient data exist. The model can also be used to estimate demographic change across a range of other sectors outside of infrastructure where high resolution and bespoke data are required.

Literature review

In the introduction, we outlined the need for detailed and customisable demographic estimates and projections when planning for infrastructure delivery. Here we deal briefly with water, transport, and energy to demonstrate the utility of such data where they can provide detailed disaggregation in terms of geographical scale and demographic attributes. We also discuss methods and approaches used for small area projection and summarize the theoretical underpinning of the land-use model used to allocate the demographic projection data.

Demographic variation in demand

Water consumption varies by the socio demographic attributes of households, as discussed by Willis et al. (2013) in an Australian context, while Shandas and Parandvash (2010) make the link between socio-economic variables and water demand in a case study of Oregon. Custom forecasts of households were developed by Rees, Clark, and Nawaz (2020) for the purpose of forecasting domestic water demand in London and the Thames Valley. They argue for the necessity of having good forecasts given that consumption is dependent on the number and type of individuals within households. They forecast population, households, dwellings, and then assess household and per-capita water consumption. For wastewater, Schlor, Hake, and Kuckshinrichs (2009) use three case studies to investigate sustainable development in the context of German population aging and decline. The challenge of “shrinking cities” when planning for water (and wastewater) infrastructure is covered by Faust, Abraham, and McElmurry (2016), Faust, Mannering, and Abraham (2016) who demonstrate that it is not only population growth which causes planning and delivery challenges, the reduced funding, and reduction in demand associated with population decline causes its own set of problems.

Energy consumption differs by geography and household socio-economic composition (Druckman and Jackson 2008). Population size is a key input into an energy demand model developed for New Zealand by Mohamed and Bodger (2005), while population size is a determinant in predicting electricity consumption in Hong Kong by Fung and Rao Tummala (1993). At the aggregate level, energy consumption is found to be impacted by population size in a study covering a number of European countries (York 2007). However, they note that at the national level there is a high degree of elasticity and that as the population gets older, energy consumption increases. These findings which expand on the simple relationship between population size and consumption point toward a need for micro-level analysis. A model built for the assessment of local level domestic energy consumption is outlined by Cheng and Steemers (2011), which takes into account occupancy based on employment status and find that at LA level there was a relationship between energy consumption, household income, and socio-economic status. The demographic variables and features of the built environment are used alongside domestic electricity consumption data to create clusters of small areas (LSOA) within London Boroughs. The domestic gas consumption of these clusters is assessed, with reference to cluster characteristics, for example high consumption clusters contain a high proportion of detached houses, and low consumption clusters have lower household income and higher unemployment rates.

Metz (2012) discusses how per capita demand for travel has plateaued in many developing countries across all modes except air travel, with the result that increasing demand is driven by population growth and population aging. The spatial distribution of this population growth is important, given that patterns of mobility differ between new development on

greenfield land and development in urban centers. Metz (2012) reports substantial decline in car use in London as rail and cycling infrastructure have been developed. Demographic change sits alongside, economic change, global fossil-fuel costs, and climate change in a transport demand model produced by Blainey and Preston (2019) which results in 504 different possible future scenarios. Transport is also an example of an infrastructure where the relationship is two-way, whereby development can drive demographic change at a local level. Debrezion, Pels, and Rietveld (2011) demonstrate that proximity to railway stations and the quality of provision (measured as the most frequented station even if this were not the closest) have a positive impact on house price in the Netherlands. In an appraisal of the English rental market, Clark and Lomax (2018) find that shorter distances to railway or underground stations have the effect of increasing rental price.

Small area demographic projection methods

The above infrastructure specific examples demonstrate the need for detailed demographic estimates and projections, which take into account both the spatial disaggregation and composition of the population. There are, however, some issues to overcome in the provision and use of these demographic estimates and projections. Firstly, demographic projections at a detailed spatial resolution are not readily available from official providers. This is partly due to a second key issue, which is the complexity of undertaking such projections at fine spatial scale. Third is that forecast uncertainty is larger for areas with small populations than those with larger populations (Cameron and Poot 2011) meaning considerable work is required to properly calibrate the models and communicate this uncertainty.

Cameron and Cochrane (2017) identify four broad categories of small-area projection methods, providing an assessment of the advantages and disadvantages of each. The first are “naïve” methods based on extrapolation or the allocation of a share of headline growth to small areas. The second is the widely used cohort component model. Third are statistical models of population change and fourth is a group of techniques which fall within the category of urban growth modeling. The models presented in this paper draw strength from each of these categories, so we consider the broad context here and return to how each have informed our models in the discussion section of the paper. A subsequent review of small area projection methods over the past decade undertaken by Wilson et al. (2021) identifies similar broad headings, with the addition of microsimulation and machine-learning methods. We utilize microsimulation, but not machine learning in our models.

The naïve models include extrapolation of previous, observed trends in population change and the allocation of these extrapolated results to small areas, based on previously observed distributions (growth share models). The key criticism leveled at these approaches by Cameron and Cochrane (2017) is that they lack a strong theoretical basis given their deterministic reliance on past trends, especially when used for areas that exhibit less predictable growth. Nonetheless, they have been found to perform well in terms of overall accuracy (Smith, Tayman, and Swanson 2013), often better than more complex methods such as cohort component models at a small area level (Smith and Tayman 2003) and have the distinct advantage of relatively low data requirements (Wilson et al. 2021) so can be applied in a broad range of contexts.

An increased level of complexity is offered by cohort component models, routinely used to produce population projections at national and regional scale (Rees et al. 2017; Lomax, Wohland and Rees 2020). These models age the population and then deal with each of the

demographic components (births, deaths, migration) separately by applying rates of change, disaggregated by demographic group. It is a cohort component model which provides the headline constraints for our model. Comparatively, these models have substantial data requirements for each of the demographic components, and while this is seldom a problem at national and regional scale, these requirements are often too great to effectively use cohort component models for small area projections (Swanson, Schlottmann, and Schmidt 2010; Wilson, 2015).

Wilson et al. (2021) refer to as simplified cohort-component methods, namely cohort change ratios, of which the Hamilton-Perry method (Hamilton and Perry 1962) is widely used, e.g. by Swanson, Schlottmann, and Schmidt (2010) to project census tract populations in the United States. These methods are less data intensive than cohort component models, requiring as input age and sex population counts from two time points, usually derived from census data, from which the ratio change is calculated and then projected forward. However, in an assessment of methods used to project local government area population in New South Wales, Wilson (2016) concludes that an unconstrained Hamilton-Perry model is the least accurate in a comparison of five different models, the Hamilton-Perry plus four cohort component models which deal with the migration component in different ways. The best performing model in terms of lowest error is found to be a constrained bi-regional cohort component model.

While statistical models for demographic projection are attractive because they offer the opportunity to include contextual variables which might be relevant at small area scale, the reviews offered by both Wilson et al. (2021) and Cameron and Cochrane (2017) point out that that regression-based models generally do not perform any better in terms of forecast accuracy than simple extrapolative methods. This has been demonstrated by Chi (2009) in a study comparing a regression model (including variables capturing small area characteristics) with simple extrapolative approaches, and Chi and Voss (2011) who additionally incorporated variables capturing the characteristics of neighboring areas. In both cases, the extrapolative methods outperformed the regression methods.

Microsimulation methods that deal with the demographic projection of individuals (or synthetic representations of individuals) rather than population groups/cohorts are becoming more prevalent in the small area projection domain. The strengths of taking a micro over a macro approach for population projection are addressed by Van Imhoff and Post (1998), namely that microsimulation allows for the inclusion of a large number of individual attributes (which impact on demographic behavior) and are capable of producing richer output than macro models in the form of a database of individuals. Examples of implementation include a model for Ireland (Ballas, Clarke, and Wiemers 2005), for Britain (Ballas et al. 2005), and for the London borough of Tower Hamlets (Lomax and Smith 2017). Further advantages of micromodels identified by Van Imhoff and Post (1998) are that they are better at dealing with interaction effects between variables, and between individuals; however, it could be argued that if these are requirements of a model then they are more appropriately dealt with within an Agent Based framework (e.g. see Wu and Birkin 2012). A recurring criticism of using microsimulation approaches for demographic projection is that they are extremely data intensive (Wilson et al. 2021) and by extension they are difficult to calibrate and validate (Ballas et al. 2005).

It is not unusual to combine methods to produce a more robust set of outputs. For example Kanaroglou et al. (2009) combine the Rogers multiregional population projection model

(Rogers 2008) at the regional (municipality) level and an aggregated spatial multinomial logit model to better account for migration at small area (census tract) level for projections in Ontario. Similarly, there is considerable strength to be gained from producing a population projection using one method and then allocating that population to small areas using another model. Land-use models are often used to undertake that allocation, for example Cameron and Cochrane (2017) use population projections (the output of a cohort component model) at the Territorial Authority level for the Waikato region of New Zealand and allocate these to the smaller scale Area Unit level using a Cellular Automata (CA) land-use model. In their work, the CA apportions land in to four-hectare grid cells based on zoning constraints, suitability, accessibility, and the composition of neighboring cells to identify suitable locations for the population to be allocated. Similar approaches have been used by Tayman (1996) and Tayman and Swanson (1996). This is broadly similar to the approach we take in this paper, so further context the following section provides more information on the theoretical underpinnings of land-use models.

The development of land-use models

Population change is often closely linked to land-use change, and both are often driven by changes to transport infrastructure. Wegener (2021) reviewed models of land use and transport, demonstrating how changes of land use are driven by human activities. Briassoulis (2000) categorized models of land-use change by their underlying modeling tradition, be that statistical/econometric, spatial interaction, optimization, or integrated. An in-depth review of land-use modeling approaches was offered by van Schroyen et al. (2011), highlighting the wide-spread adoption of Cellular Automata (CA) in land-use applications.

There is a long history of the use of CA modeling of land-use change and urban development. Couclelis (1985) introduced the concept of cellular worlds to simulate dynamic geographical processes. Cecchini (1996) and Clarke, Hoppen, and Gaydos (1997) applied these concepts in practical implementations of CA simulations of urban land-use change. These concepts were further developed by Engelen, Geertman and Smits (1999) in a decision-support framework to explore land-use change using CA approaches, which were built-upon by White, Uljee, and Engelen (2012) to demonstrate the spatial interlinkages between population change and land-use change. Tong and Feng (2020) offer an up-to-date review of CA approaches to modeling urban growth.

The identification of suitable locations for land-use change arising from population growth through the spatial Multi-Criteria Evaluation (MCE) is an approach first suggested by Carver (1991). Malczewski (2004) further demonstrated how this approach could be used to identify suitability for land-use changes. The coupling of CA and MCE approaches to simulate land-use change has been previously described by Ford et al. (2019). By mapping land-use change at fine spatial scale, and making assumptions about the density of development, it is possible to simulate the distribution of population (and thus infrastructure demand) at a sub-zonal level.

Summary

This review has outlined how demand for infrastructure varies by geography and demographic attributes to demonstrate that estimates and projections are needed to take this into account when planning for delivery. It has also discussed how, while there are not necessarily readily available official data at fine spatial scale, there is a long-standing interest in producing small area projections,

and that the methods used to develop these are varied and ever evolving. The link to land-use models as discussed is important for understanding the spatial allocation of projection outputs sub-zonally.

Material and methods

The demographic model structure comprises a number of standalone packages which form the interlinked workflow, from the download and cleaning of data from a range of sources, through the creation of baseline synthetic population and household datasets, projection of the baseline, custom scenario generation, and data output. The packages and processes which form the workflow are outlined in Fig. 1 and we explain these in the next sub-sections. The ethos of the model from the outset has been to create an ecosystem of self-contained open-source packages that are well documented and produce reproducible results so that the models can be run by anyone. This means that the models are flexible and adaptable, rather than just providing data and outputs, and the project provides the tools to produce those data and users are then free to undertake any analysis they like. Links to the Github repositories are available below, where further detailed documentation and model code (published under an open MIT licence) is available to download and run.

Broadly there are six main phases to the workflow outlined in Fig. 1: Phase 1 involves downloading, cleaning, and making consistent population, household, and projection data for the United Kingdom; Phase 2 involves the creation of synthetic individual level population and household datasets which can be used as a baseline input for projection; Phase 3 involves the simulation and projection of populations and households over time at a fine spatial resolution; Phase 4 involves the creation of custom scenario constraints for projection; Phase 5 deals with the output and storage of demographic results; while Phase 6 allocates these demographic outputs to 1 hectare cells within the zone. The modular structure is designed in such a way as to allow for components to be updated, adapted, and even substituted entirely depending on user requirements.

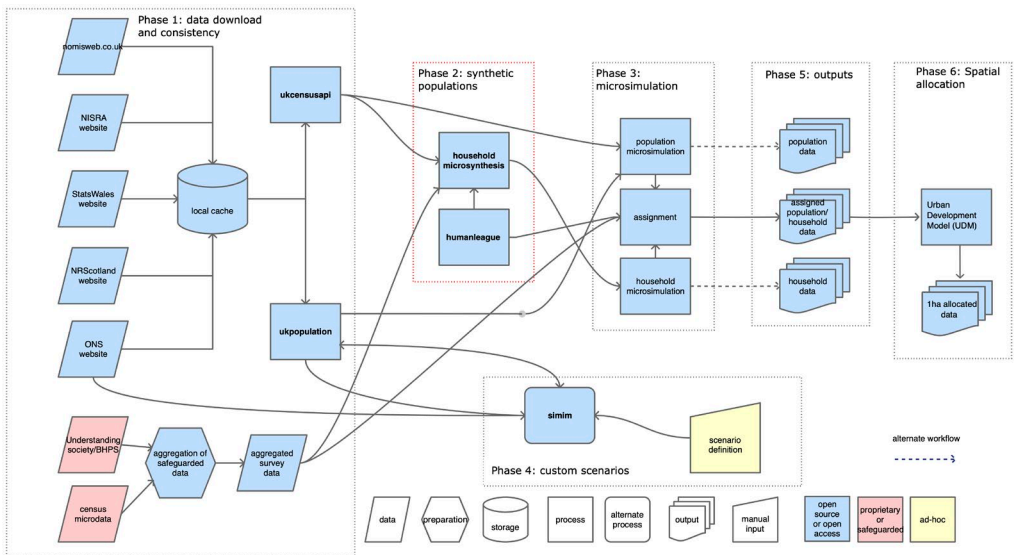


Figure 1. A schematic diagram of the different software packages and how they fit together.

Phase 1: Download, cleaning, and making consistent data (1a) and creating projection constraints (1b)

Demographic data for the United Kingdom is increasingly being made available from a range of sources, thanks largely to improvement in computing, storage, and the publication of application programming interfaces (APIs), which facilitate the automation of data download when developing software. There still exist, however, substantial inconsistencies in the format and accessibility of datasets produced by the NSAs which are responsible for data for different parts of the United Kingdom: The ONS for England and Wales, the National Records for Scotland (NRS), Statistics Wales and the Northern Ireland Statistics, and Research Agency (NISRA). These inconsistencies have been identified previously, see for example Lomax, Norman and Rees (2013) who construct a consistent set of internal migration estimates from data derived from the NSAs. In the past, pre-processing of demographic data was done manually, largely due to the difficulties of producing a “one size fits all” tool for dealing with inconsistent data and partly because the reproducibility of work in the social sciences is a relatively new, but very welcome, phenomenon (Hardwicke et al. 2020).

We discuss two packages in this section which occupy the right-hand side of Fig. 1: *UKCensusAPI* extracts and makes consistent census and mid-year population estimate data from a range of sources and *UKPopulation* extracts household estimate data and projection data (and produces constraining projection results). There is a third process which involves the aggregation of 2011 Census microdata and processing of survey data which has been implemented to produce a dataset used in the assignment of people to households.

Access, query, and make consistent baseline census data using the UKCensusAPI package

As a baseline, the estimates and projections need a consistent dataset containing counts of people and households at small area scale. This baseline is constructed from 2011 Census data, which are extracted by the *UKCensusAPI* package (Smith 2017). This package provides both a Python and an R wrapper around the API provided via Nomis (the web-based service provided by ONS for accessing labor market and census data), as well as the NRS and NISRA websites. *UKCensusAPI* allows for the querying of table metadata and auto-generating customized Python and R query code for future use. The automated cache downloads data modify the geography of queries and add descriptive information to tables (from metadata). For more information on the *UKCensusAPI* package on Github see <https://github.com/virgesmith/UKCensusAPI>. *UKCensusAPI* feeds data directly to the packages *ukpopulation* and *household microsynthesis*.

Access, query, and make consistent the official U.K. population projections using the ukpopulation package

The next step in the process is to download and make consistent the official projection data for the United Kingdom. These headline projections are needed because they form the constraints for small area projections. Having a model which is constrained to official estimates provides reassurance to users that the overall totals are in line with what they would expect to see from other sources. We do however offer an opportunity to create custom constraints which are independent from the official projections (Phase 4, discussed at section 2.4 and demonstrated later in the case study). Constraints are produced at LA scale in the first instance.

The *ukpopulation* package (Smith and Russell 2018) downloads and harmonizes the projection data from different sources. Projections here are comprised of the National Population

Projections (NPP), Sub National Population Projections (SNPP), and the Sub National Household Projections (SNHP). The other dataset used for constraining small area population estimates are the Mid-Year Estimates (MYE).

For population projection data, while the data sources are disparate, the basic attributes are consistent: a count of population by age, sex, and LA area. In creating the constraints for population data, the steps are to:

- use MYE data up to 2018, disaggregated by age and sex.
- then use SNPP data up to 2041, disaggregated by age and sex.
- After 2041, extrapolate the SNPP using NPP data and age-sex structure where the NPP horizon is 100 years into the future.

The extrapolation of the SNPP using NPP trends is done independently for each age and sex in order to try to capture the age-sex structure and trends in the original population. Aggregation only takes place on the extrapolated age-sex specific values. This means that the trends shown by SNPP geographies with different age-sex structures will differ.

This methodology can be more formally explained by the following equation for the aggregate SNPP $S(g,y)$ for a given geography (g) and year (y):

$$S(g,y) = \sum_a \sum_s S(a,s,g,\bar{y}) \frac{N(a,s,y,c(g))}{N(a,s,\bar{y},c(g))}$$

where N is the NPP, a is age, s is sex, \bar{y} is a reference year (typically the final year in the SNPP data), and $c(g)$ represents a mapping from a SNPP geography (LA) to a NPP (country).

This method can also be used to create variants from the data. Fig. 2 shows the high (hhh) principal (ppp) and low (lll) projections for Newcastle LA. These are created by taking the national variant (published in the ONS data), weighting by the age and sex structure in Newcastle (relative to the principal projection) and plotting the result.

The extrapolation methodology above can equally be applied to synthesizing SNPP variants from SNPP principal and NPP variant data. The equivalent expression to the above is:

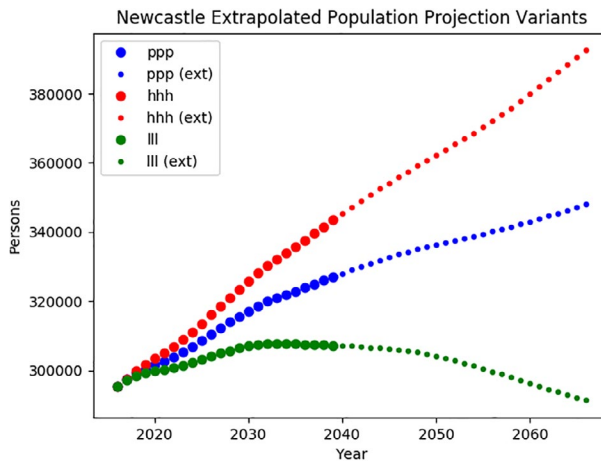


Figure 2. Example of the headline variants for Newcastle local authority district.

$$S_v(g, y) = \sum_a \sum_s S_0(a, s, g, y) \frac{N_v(a, s, y, c(g))}{N_0(a, s, y, c(g))}$$

where the subscripts v and 0 refer to the variant and the principal projections, respectively. In Fig. 2, the large points represent the SNPP data disaggregated by the NPP variant, and the smaller points represent the extrapolated data variants. Note that while these NPP data are available to the year 2118 and the extrapolation could continue to this point; however, we stop in 2065 for the purpose of this example. Anyone using these data will need to consider that uncertainty increases in line with the time horizon and that this method would need augmenting if a user wishes to project beyond the data points available in the NPP.

Household projection data are less consistent than the population projection data. Each country's NSA provides projections of households, disaggregated by household type. However, there is not enough consistency between the definitions to produce a unified classification, so we instead use the definitions "as is" when producing the LA level constraints. This means that the categories used in the different datasets are maintained in the data used for constraint. No extrapolation, nor application of a national projection variant, is currently undertaken on these datasets. A detailed summary of the different methods and subsequently different outputs produced for each of the constituent nations of the United Kingdom can be found in ONS (2020). The definition disparities are summarized in Table 1.

In summary, *ukpopulation* provides the constraints at LA scale for the population estimates, population projections, and household projections produced by our model. Documentation and code for *ukpopulation* can be found at <https://github.com/nismod/ukpopulation>.

Preserving individual and household relationships in the simulation

There is a further process in Phase 1 of the model. This involves the aggregation of census microdata into a relationship matrix, which allows us to preserve the relationship between household reference person (HRP) and other household members, disaggregated by age, sex, and ethnicity. We also use data from the Understanding Society survey (University Of Essex 2020) to assess the relationship between household size and number of bedrooms needed for the baseline projections. Because these datasets are safeguarded/licenced, we do not provide a process for the automated extraction of this matrix. This is one of the few manual steps in the workflow.

Creating synthetic microdata

Once data have been downloaded and cached, we can turn our attention to creating the baseline microdata needed as an input to the projection models. Seen in the Phase 2 box of Fig. 1, this involves creating a synthetic, individual level population of households using the *household microsynthesis* package and a synthetic, individual level population of individuals using the *humanleague* package (Smith 2018).

Generating synthetic household data using the household microsynthesis package

household microsynthesis is a Python package for creating household populations from census data, including communal and unoccupied residences. *household microsynthesis* combines census data (fed from the *UKCensusAPI* package) on occupied private households, communal residences, and unoccupied dwellings to generate a synthetic population of households classified across a number of categories. The synthetic population is consistent with the census aggregates

Table 1. The Categories of Household Type Produced as Part of the Sub-National Household Projections in Each Country of the United Kingdom

	England (ONS)	Wales (StatsWales)	Scotland (NRS)	Northern Ireland (NISRA)
Lowest common denominator				
Single person	“One person households: Female” “One person households: Male”	“1 person”	“1 adult: female” “1 adult: male”	“One adult households”
Adults and children	“Households with one dependent child”	“2 person (1 adult, 1 child)”	“1 adult 1+ children”	“One adult households with children”
	“Households with two dependent children”	“3 person (2 adults, 1 child)”	“2+ adults 1+ children”	“Other households with children”
	“Households with three or more dependent children”	“3 person (1 adult, 2 children)”		
		“4 person (2+ adults, 1+ children)”		
		“4 person (1 adult, 3 children)”		
		“5+ person (2+ adults, 1+ children)”		
		“5+ person (1 adult, 4+ children)”		
Adults only	“Other households with two or more adults”	“2 person (No children)”	“2 adults”	“Two adults without children”
		“3 person (No children)”	“3+ adults”	“Other households without children”
		“4 person (No children)”		
		“5+ person (No children)”		

at the specified geographical resolution and can be simulated from Output Area (OA) level upwards. The output data consists of a csv file containing the synthetic population where each row represents a single household. The code and documentation for *household microsynthesis* can be found at https://github.com/nismod/household_microsynth.

Generating synthetic population data using the humanleague package

humanleague (Smith 2018) is a Python and an R package for creating synthetic populations from marginal and seed data using microsimulation. The package is implemented in C++ for performance. Three microsimulation methods are offered to undertake the population synthesis within the package: the deterministic reweighting method of Iterative Proportional Fitting (IPF) (Lomax and Norman 2016), a probabilistic resampling method called Quasirandom Integer Sampling (QIS) (Smith, Lovelace, and Birkin 2017), and a hybrid approach called Quasirandom Integer Sampling of IPF (QISI). A worked example of the output can be found in Lomax and Smith (2017). Also see <https://github.com/virgesmith/humanleague>.

Projections

Using the synthetic data as an input, the processes outlined under Phase 3 in Fig. 1 are concerned with projecting the population and households forward through time. Households and individuals are handled separately, with an *assignment* algorithm which places people into households. Broadly speaking, the processes used in the main part of the workflow are static, in so much as they are time independent and reliant on the headline constraints calculated at LAD level to adjust a small area seed in each year that the projection is run. The projection of population, households, and the assignment is described in the documentation for the *microsimulation* package <https://github.com/nismod/microsimulation>.

Small area projections

Population projections are handled within the *microsimulation* package. A seed population of individuals at Middle Super Output Area (MSOA) level (which contain on average is fed through from the *humanleague* package and is adjusted to fit the local authority level constraints taken from either the *ukpopulation* package described above, or alternatively a custom projection variant derived from the *SIMIM* package described later. The simulated population is estimated by single year of age, sex, and ethnicity.

Household projections require as an input the synthetic household dataset generated by the *household microsynthesis* package. Households persist according to a survival probability and new households are created randomly to match the local authority level household constraints generated by *ukpopulation*.

The adjustment of the microdata can be undertaken using two different methods. The first is iterative proportional fitting. The second is quasi-random integer sampling. IPF is much faster so is used where a large number of areas need to be simulated. Quasi-random integer sampling is much slower so should be used if only a small number of areas need to be simulated or High Performance Computing (HPC) is available. The model can be run in parallel processing to speed up data generation for multiple areas at once.

Household assignment algorithm

The relationship between individuals and households is essential for (1) providing consistency between the two projections and (2) understanding the composition of households, which have

different infrastructure demands. Individuals are grouped into households by randomly sampling the synthetic population to form distributions defined by census microdata. This preserves the relationship between age, sex, and ethnicity of the HRP and the age, sex, and ethnicity of other household members. This helps to avoid nonsensical or unlikely household combinations such as children who are only fractionally younger than a parent. The effect is to largely preserve the distribution of household structures seen in the 2011 census but aligned with the household and population projections which are consistent with the projection constraint.

Of the household structures defined in the census, all contain one household reference person, and some categories are more precise about the number and status of the occupants. For example, single-occupant households must contain a single adult; single-parent households of size 3 must contain one adult and two children. Conversely, multiple occupant households containing 4+ occupants are less well defined.

Assignment means linking rows in two tables: the household table is given an additional column that refers to an entry in the person table, this is the HRP. The person table is given a column containing a household ID. Once assignment is complete, every person will be associated with a household, and every household will be associated with a HRP. The approach taken by the algorithm is to get the specific structures assigned first. Once a household is filled, it is marked as such and no more people can be assigned to it. The algorithm loops over the MSOAs in the LA, assigning people to households in the following order:

- First HRP, as this is the key link between people and households. We rely on distributions from census microdata that link the HRP characteristics with those of other members of the household.
- partners of HRPs are then sampled for the relevant households.
- children are then sampled.
- multi-person households are constructed.
- communal establishments are constructed.

At this point many households will be fully assigned, but there will generally be unassigned adults and children in the population. They are assigned to those households that are not already full.

Creating custom growth scenarios

There is often a need to create custom headline constraints that reflect changes which might be anticipated at local level, for example the development of new housing, new jobs, or transport infrastructure. These custom constraints can replace those produced by the *UKPopulation* package. While any custom constraint could be introduced to the model (see Phase 4 in Fig. 1 for detail of where this fits in), we have created a scenario generation tool called *SIMIM*, a spatial interaction model (SIM) which takes into account user assumptions about the development of housing, transport connectivity, and the strength of the local economy in a specific area.

Internal migration is one of the most difficult demographic processes to model given that migration has an impact at both origin, where people leave, and destination, where people arrive. This complexity is further extended when there are changes to the infrastructure at either origin or destination. The provision of new housing, roads, or job creation can make an area more attractive, increasing the inflow of internal migration to that area as people move to take advantage of improved infrastructure. Conversely, these people move from other areas which are relatively less attractive, which has implications for the origin area.

There is a rich history of modeling internal migration, taking into account the relative attractiveness of different areas using SIMs. The idea that areas that are closer to one another are more attractive than those which are further away from a human migration perspective was developed by Zipf (1946) as “gravity models.” Subsequent work by Wilson (1971) introduced constraints at origin, destination, or both (termed doubly constrained models) which meant that limits could be imposed on the total outflow or inflow. A useful guide for the implementation of SIMs can be found in Dennett (2018). In their SMILE model of the Irish population, Ballas et al. (2005) note that the treatment of internal migration could be better handled using a SIM.

Our model *SIMIM* builds on these well-established principles and utilizes the SpInt module in the Python spatial analysis library (PySAL) (Oshan 2016). SpInt provides the functionality to produce both constrained and unconstrained models. We introduce the option to alter the attractiveness of areas based on the number of new jobs created, the local distribution of regional Gross Value added (GVA) (as a proxy for relative economic strength), and number of new homes being built.

In order to generate scenarios of change, all of these factors can be varied in each year of the projection to be run. These can be set manually, but we have also produced a web-based user interface which allows for changes to an area to be set using this interface. This model interface was developed in collaboration with the Data Analytic Facility for National Infrastructure (DAFNI) and the interface is hosted on the platform (Hall 2019). An account of the DAFNI implementation of *SIMIM* can be found in Lomax and Smith (2020) but by way of example:

- Within a region of the United Kingdom, each LA area has a baseline projection (supplied via the *UKPopulation* package).
- A user can use the *SIMIM* interface to increase the supply of housing and jobs within an area to reflect a planning scenario. This can vary for each year of the projection.
- The user can also increase the GVA of that area (relative to other areas)—this is a proxy for economic growth and based on prior expertise, or is the output from some other econometric model (as is the case in the example presented later in this paper). This can also be varied in each year.
- The user can then choose the type of SIM to run: unconstrained, origin-, destination-, or doubly-constrained as well as the type of distance metric to be used.
- *SIMIM* is run for the first year of the projection, which redistributes the LA population based on the relative attractiveness of that area and distance from other areas. Some areas (e.g. those with new houses and jobs) will gain population while others will lose population (e.g. those without any new development) to those areas. There is a distance decay effect, whereby areas that are closer are more likely to lose population to the more attractive areas than those which are further away.
- The new headline (LA level) projection for the given year replaces the *UKPopulation* constraint
- *SIMIM* then runs for the next year of the projection, taking in to account user assumptions for that time period.

In each year, the total population can be adjusted to match the totals reported in the SNPP/NPP extrapolated data, so *SIMIM* effectively deals with the redistribution of population

within the country. The documentation for SIMIM can be found at <https://github.com/nismo/simim>.

Allocation of demographic outputs within zones

The final step of the workflow, outlined at Phase 6 of Fig. 1, involves calculating the spatial distribution of the demographic scenarios at a sub-zonal scale. To do this, we employ a model called the *Urban Development Model (UDM)*. *UDM* takes the outputs described at Phase 5 as input, alongside constraints including current population density and land availability. *UDM* uses a spatial multi-criteria evaluation approach coupled to a cellular automata model to calculate the likely spatial locations of demographic outputs within a given MSOA zone. The projected population increases are used to estimate future land requirements, based on a set of density assumptions, and these requirements are then mapped to their most likely spatial locations according to a set of suitability drivers (such as proximity to transport infrastructure) and planning constraints (such as protected land). The output is a one-hectare raster grid of urban development at a given future timestep arising from the population growth in a given zone. The model is described in detail in Ford et al. (2019) and the open-source code can be found at <https://github.com/geospatialncl/OpenUDM>.

Summary of methodological contribution

The packages and processes described in this section have been designed to fill a gap by providing a framework for producing small area population estimates and results in a flexible and reproducible way. In summary, the contribution of the framework is five-fold: (1) to streamline the download and processing of data needed for demographic estimates; (2) to generate baseline synthetic estimates of people *and* of households; (3) to project both of these units forward through time and assign people to households in order to retain a relationship; (4) provide the tools for flexibly generating custom scenarios of small area composition which drive demographic change; and (5) to assign these outputs to sub-zone spatial units. This project will continue to evolve, so the above provides a snapshot of current capability. Further iterations will follow the guiding principles of providing open-source and reproducible models and workflows.

Results and discussion: A case study of the Arc development corridor

We apply our models to the Oxford-Milton Keynes-Cambridge “Arc” region of the United Kingdom (Valler, Jonas, and Robinson 2020) to demonstrate their utility. The Arc has been identified as an area for potential urban and economic development (National Infrastructure Commission 2017), and population change scenarios produced by *SIMIM* have already informed analysis of demand for 5G mobile phone infrastructure in the Arc (Oughton and Russell 2020). The impacts of the different growth scenarios on road transport, energy, water, urban drainage, urban form and green infrastructure have been addressed by the ITRC and are reported in Hickford, Russel and Hall (2020). Here we summarize how the headline (LA-scale) constraints of potential Arc growth scenarios are translated to small area projections of households and people in order to inform high resolution estimates of demand for infrastructure. We assign the demographic outputs to 1 hectare grid squares and assess the levels of urban development required and resulting population densities.

The development scenarios

We make use of four different development scenarios for the Arc, which are outlined in Hickford et al. (2020, p. 24). In brief these comprise:

1. **Baseline**, which takes the average annual number of new dwellings completed between 2007 and 2017 (14,500 per annum for the entire Arc region) and carries this forward for each year. There is no new transport infrastructure developed for the Arc and no allowance for development on Greenbelt land.
2. **Unplanned development**, assumes a higher number of new dwellings per annum (19,000) and that laissez-fair planning policy allows for development driven by market forces: dwellings are concentrated closer to a newly developed road expressway and rail line running east to west through the Arc region. This scenario allows for some limited development on Greenbelt land.
3. **Expansion of existing settlements**, whereby a larger number of new dwelling completions (30,000 per annum) are split between the existing towns and cities within the Arc and both expressway and east-west rail are developed. Construction is allowed on Greenbelt land.
4. **New Settlements**, whereby there would be major growth (30,000 dwellings per annum) in five new towns within the Arc region. Both rail and road infrastructure are developed and some limited building on Greenbelt land is allowed.

These scenarios are translated to LA level assumptions which reflect annual dwelling completions, the distribution of regional Gross Value Added (GVA) (which represents the relative economic attractiveness of each LA, calculated separately using an input-output approach), and the number of new jobs in each year. These LA level constraints are used by *SIMIM* to redistribute population based on the relative attractiveness of each area, and these new totals form the basis for the spatially disaggregated results presented in the next section. It should be noted that *SIMIM* is run nationally, i.e. population can be gained or lost from areas outside of the Arc region, but we limit our results to the impact on the Arc region only. With reference to Fig. 1, the scenario constraints replace the official constraints generated by *UKPopulation*. The dataset containing the parameters used by *SIMIM* can be found in Russell (2019).

Results: Sub-LA population distributions

Fig. 3 demonstrates that when the headline constraints are translated to MSOA level outputs, there are clear differences in the spatial distribution of population for each of the scenarios. Each map displays the percentage increase/decrease in total population in 2050 compared with 2018 data.

Under the Baseline scenario, all MSOA populations grow and while there is a fairly uniform distribution of growth, higher percentage change occurs in the area between Oxford and Milton Keynes. Population growth under the new settlements scenario is highest in those MSOAs which constitute or are adjacent to those new towns, both because of the additional new housing and the increase in GVA and employment opportunities. There is population loss from Cambridge under the New Settlements scenario. Unplanned growth results in more pronounced population growth in the southern and eastern areas of the Arc when compared with Baseline. Under the Expansion of existing settlements scenario, population growth is highest in the existing urban centers and

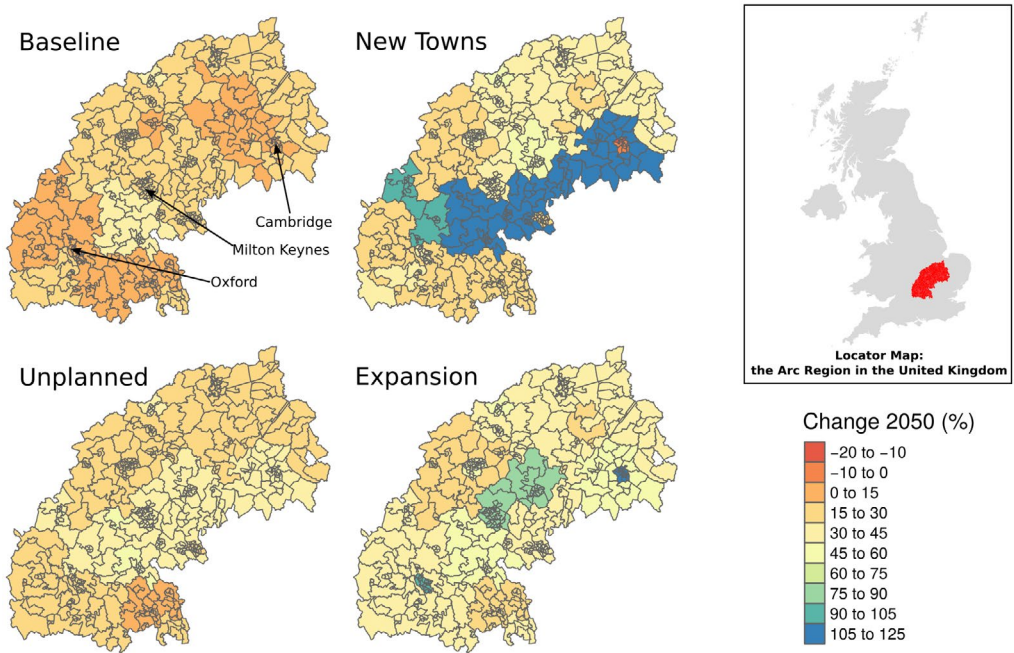


Figure 3. Percent population change at MSOA level in 2050 compared with 2018 under four scenarios.

areas surrounding, not surprising given the focus on providing more housing and jobs in these areas.

Under these four scenarios, we can see that the spatial distribution of population change differs substantially. This in itself is useful for the planning of infrastructure and services in that it provides information about the level of demand which might need to be provisioned for. For some applications, however, a higher resolution, sub-zonal indication of potential population distribution is needed. For example, estimation of increased energy demand or additional pressure on drainage systems requires indications of potential settlement patterns at the highest resolution. We may also need an idea of how realistic or sustainable the scenario is in terms of the land available to house the projected population. To provide this additional detail, the MSOA level scenarios are passed to *UDM*.

Results: Allocation of population using UDM

UDM was used to simulate 1ha (100 × 100 m) resolution development patterns for the four MSOA-level demographic scenarios. By using a set of spatial attractors (such as proximity to transport networks, proximity to existing urban development, and natural capital scores) and constraints (nature reserves, water bodies, open green space, etc.), the most likely locations of land development are mapped. Development takes place at the observed density of people/ha in each MSOA to retain the local character of each geographical zone.

From Fig. 4, it can be seen that the different population growth rates combined with different planning policies for each scenario result in contrasting locations and amounts of development in each case. The new settlement scenario concentrates growth along the proposed new east-west

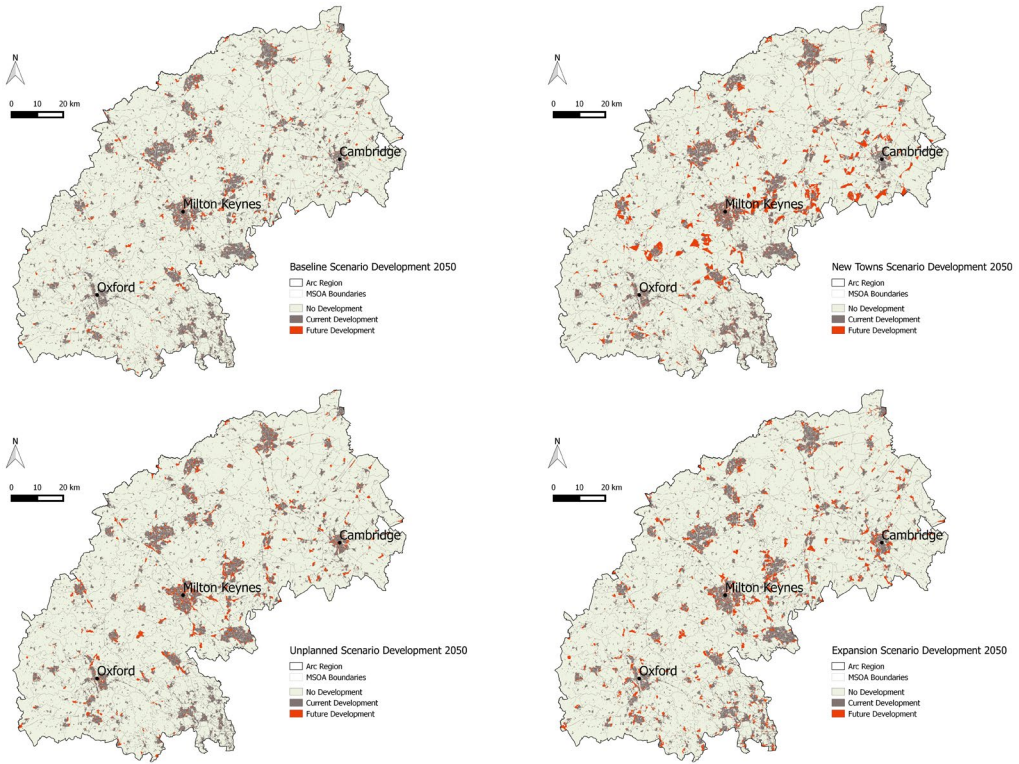


Figure 4. Land-use development patterns arising from population increases at MSOA scale for the Oxford-Cambridge Arc region, showing future development in red.

rail corridor in the Oxford-Cambridge Arc, showing the close links between land-use change and transport infrastructure development.

Fig. 5 shows a larger-scale depiction of the Expansion scenario development pattern for the city of Cambridge. The areas of development can be seen to agglomerate around existing urban areas, shaped by constraints and attractors in *UDM*. MSOA boundaries are shown in the figure, demonstrating the improvement in spatial resolution that can be gained by running the MSOA-scale population projections through this model stage.

The outputs from *UDM* can be used to provide assessments of the impacts of population changes on factors such as available land, loss of greenspace, and changes to population density (and thus urban character). In some cases, *UDM* can also provide a check on the realism of the population projections by taking into account available land and planning constraints. If all projected population cannot be accommodated in the available land for new development in an MSOA zone, the model will report back an overflow and the required increase in population density that would be required to accommodate the projected population. This gives a useful feedback on the areas where there is the most pressure on land or where the projected population increases may be unrealistic. Fig. 6 shows the required population densities for each MSOA for each of the four scenarios.

It can be seen from Fig. 6 that in some cases the increased population density is very high, with almost an additional 100 people/ha required in some MSOAs in the Expansion scenario in

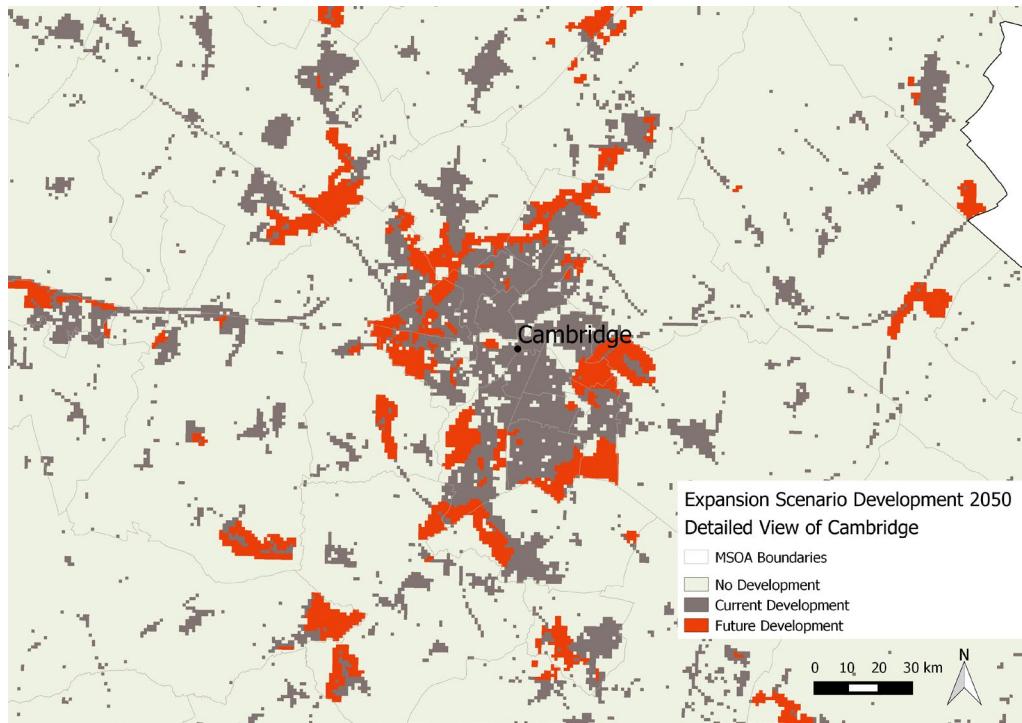


Figure 5. Urban development patterns arising around the city of Cambridge in the Expansion scenario.

order to fit the projected population in places like Oxford and Cambridge. The maximum increase is lower for the Baseline and Unplanned scenarios, as these do not attempt to concentrate population in desirable locations (e.g., around existing urban centers of public transport hubs). This demonstrates the trade-off between protecting greenspace, encouraging use of sustainable transport, and increasing development density.

Conclusions

This paper has made the case that demographic estimates and forecasts are an essential but often overlooked component of infrastructure planning projects. Indeed, they are a specialist component which often needs a dedicated team to estimate and present the data which feed directly into demand models. The paper sets out the detail of a workflow which deals with data download and consistency, the creation of synthetic microdata, the small area projection of people and households, and their allocation to 1-hectare grids based on land availability. We demonstrate how a range of scenarios can be created at high spatial resolution using these methods and that by considering land availability the impact of these scenarios can be checked (e.g. in terms of the amount of land that would need to be developed) and their feasibility assessed, in this case by looking at the change in population density that would result from the scenario becoming reality.

Returning to the four broad categories of small-area projection methods identified by Cameron and Cochrane (2017), the models presented in this paper takes advantages of the

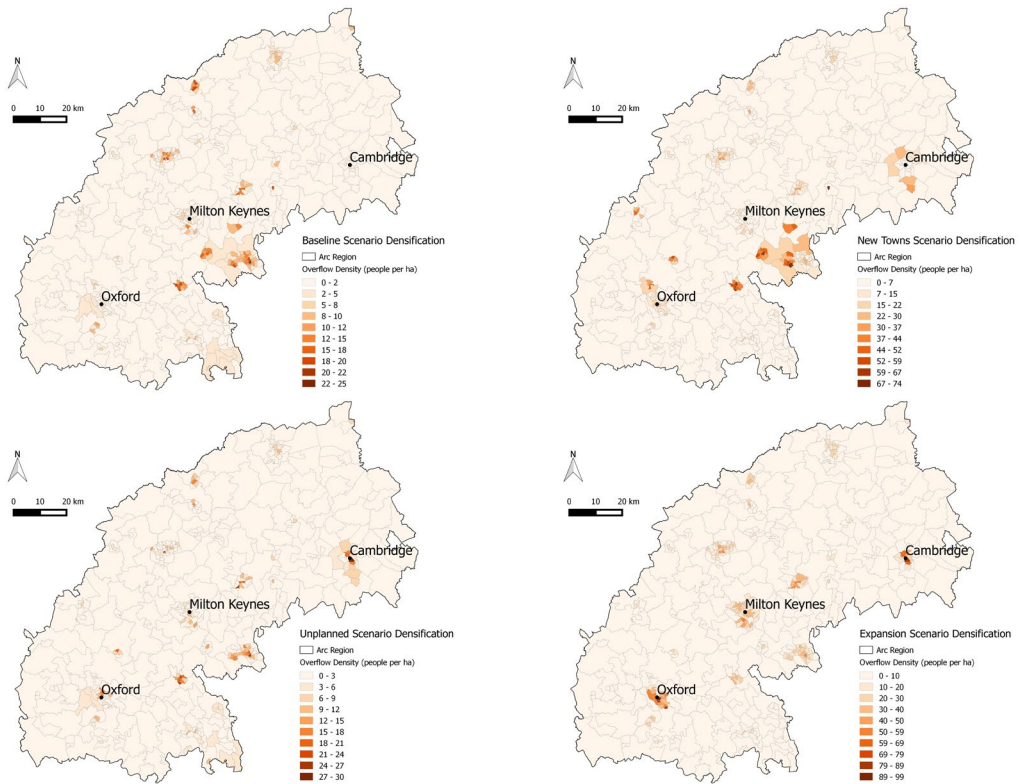


Figure 6. The required increase in population density above current observed density in each MSOA under the four scenarios for the Oxford-Cambridge Arc.

strength offered by each, which means that the system is greater than the sum of its parts. The naïve extrapolation and growth share methods that have been found to perform well in previous literature are used to support a robust baseline projection in our models by providing the headline constraints via the *ukpopulation* package. Because these headline (LA level) constraints are derived from data produced using a cohort-component model (produced by the United Kingdom NSAs), they are arguably more robust than simple extrapolations from past data. A statistical model is utilized in the scenario-generation module SIMIM, which incorporates important contextual variables and redistributes population accordingly. The strength of using SIMIM to produce counterfactual headline scenarios is that it allows the user to introduce variables which are usually missing from cohort component or extrapolative methods within the robust framework described. The CA-based urban development model used to allocate the small area population projections to spatial units is grounded in theory and is able to account for fine-scale local considerations in the way that other approaches are not. Underpinning all of this is a microsimulation approach, whereby individual level population and household data are produced. This provides the benefit of rich output detail as identified by Wilson et al. (2021) and the high resolution outputs are fairly novel in a small area projection context.

The models described in this paper are open source, meaning they could be applied to any infrastructure planning project and allow for a range of scenarios and options to be assessed.

This model structure could help planners to better assess the benefits and trade-offs which arise at a very detailed level of geographical disaggregation. This framing of the models as tools for assessing alternative scenarios is important because there is the potential for results to influence real-world events in so much that investment decisions based on simulated growth would serve to enable that particular outcome over other alternatives. The models can also be utilized in other noninfrastructure planning contexts where high-resolution demographic data are required, for example individual level data have been used as an input to a disease transition model (Spooner et al. 2021). This work is ongoing and as such various improvements and extensions are in progress. These include a stochastic projection model at Phase 3 of the workflow and the development of alternative scenario generation modules at Phase 4.

Funding information

The research described in this paper was supported by the UK EPSRC (Engineering and Physical Sciences Research Council) under grant EP/N017064/1 and Wave 1 of The UKRI Strategic Priorities Fund under the EPSRC Grant EP/T001569/1, particularly the “Digital Twins: Urban Analytics” theme within that grant & The Alan Turing Institute

References

- Ballas, D., G. Clarke, D. Dorling, H. Eyre, B. Thomas, and D. Rossiter. (2005). “SimBritain: A Spatial Microsimulation Approach to Population Dynamics.” *Population, Space and Place* 11(1), 13–34.
- Ballas, D., G. P. Clarke, and E. Wiemers. (2005). “Building a Dynamic Spatial Microsimulation Model for Ireland.” *Population, Space and Place* 11(3), 157–72. <https://doi.org/10.1002/psp.359>.
- Blainey, S. P., and J. M. Preston. (2019). “Predict or Prophecy? Issues and Trade-Offs in Modelling Long-Term Transport Infrastructure Demand and Capacity.” *Transport Policy* 74, 165–73. <https://doi.org/10.1016/j.tranpol.2018.12.001>.
- Briassoulis, H. (2000). Analysis of Land Use Change: Theoretical and Modelling Approaches. Regional Research Institute, West Virginia University. <http://www.rri.wvu.edu/WebBook/Briassoulis/contents.htm>.
- Cameron, M. P., and W. Cochrane. (2017). “Using Land-Use Modelling to Statistically Downscale Population Projections to Small Areas.” *The Australasian Journal of Regional Studies* 23(2), 195–216.
- Cameron, M. P., and J. Poot. (2011). “Lessons from Stochastic Small-Area Population Projections: The Case of Waikato Subregions in New Zealand.” *Journal of Population Research* 28(2), 245. <https://doi.org/10.1007/s12546-011-9056-3>.
- Carver, S. J. (1991). “Integrating Multi-Criteria Evaluation with Geographical Information Systems.” *International Journal of Geographical Information System* 5(3), 321–39.
- Cecchini, A. (1996). “Urban Modelling by Means of Cellular Automata: Generalised Urban Automata with the Help on-Line (AUGH) Model.” *Environment and Planning B: Planning and Design* 23(6), 721–32.
- Cheng, V., and K. Steemers. (2011). “Modelling Domestic Energy Consumption at District Scale: A Tool to Support National and Local Energy Policies.” *Environmental Modelling & Software* 26(10), 1186–98.
- Chi, G. (2009). “Can Knowledge Improve Population Forecasts at Subcounty Levels?” *Demography* 46(2), 405–27. <https://doi.org/10.1353/dem.0.0059>.
- Chi, G., and P. R. Voss. (2011). “Small-Area Population Forecasting: Borrowing Strength Across Space and Time: Spatial Population Forecasting.” *Population, Space and Place* 17(5), 505–20. <https://doi.org/10.1002/psp.617>.
- Clark, S. D., and N. Lomax. (2018). “A Mass-Market Appraisal of the English Housing Rental Market Using a Diverse Range of Modelling Techniques.” *Journal of Big Data* 5(1), 43.

- Clarke, K. C., S. Hoppen, and L. Gaydos. (1997). "A Self-Modifying Cellular Automaton Model of Historical Urbanization in the San Francisco Bay Area." *Environment and Planning B: Planning and Design* 24(2), 247–61.
- Couclelis, H. (1985). "Cellular Worlds: A Framework for Modeling Micro—Macro Dynamics." *Environment and Planning A* 17(5), 585–96.
- Debrezion, G., E. Pels, and P. Rietveld. (2011). "The Impact of Rail Transport on Real Estate Prices: An Empirical Analysis of the Dutch Housing Market." *Urban Studies* 48(5), 997–1015.
- Dennett, A. (2018). "Modelling Population Flows Using Spatial Interaction Models." *Australian Population Studies* 2(2), 33–58. <https://doi.org/10.37970/aps.v2i2.38>.
- Diamond, I., H. Tesfaghiorghis, and H. Joshi. (1990). "The Uses and Users of Population Projections in Australia." *Journal of the Australian Population Association* 7(2), 151–70. <https://doi.org/10.1007/BF03029362>.
- Druckman, A., and T. Jackson. (2008). "Household Energy Consumption in the UK: A Highly Geographically and Socio-Economically Disaggregated Model." *Energy Policy* 36(8), 3177–92. <https://doi.org/10.1016/j.enpol.2008.03.021>.
- Engelen, G., S. Geertman, P. Smits, and C. Wessels. (1999). "Dynamic GIS and Strategic Physical Planning Support: A Practical Application." In *Geographical Information and Planning*, 87–111, edited by J. Stillwell, S. Geertman and S. Openshaw. Berlin: Springer.
- Faust, K. M., D. M. Abraham, and S. P. McElmurry. (2016). "Water and Wastewater Infrastructure Management in Shrinking Cities." *Public Works Management & Policy* 21(2), 128–56.
- Faust, K. M., F. L. Mannering, and D. M. Abraham. (2016). "Statistical Analysis of Public Perceptions of Water Infrastructure Sustainability in Shrinking Cities." *Urban Water Journal* 13(6), 618–28.
- Ford, A., S. Barr, R. Dawson, J. Virgo, M. Batty, and J. Hall. (2019). "A Multi-Scale Urban Integrated Assessment Framework for Climate Change Studies: A Flooding Application." *Computers, Environment and Urban Systems* 75, 229–43.
- Fung, Y. H., and V. M. Rao Tummala. (1993). "Forecasting of Electricity Consumption: A Comparative Analysis of Regression and Artificial Neural Network Models." 1993 2nd International Conference on Advances in Power System Control, Operation and Management, APSCOM-93, Vol. 2, 782–7.
- Hall, J. W. (2019). "UK Reveals New Platform for Infrastructure Data Analysis and Simulation Modelling." *Proceedings of the Institution of Civil Engineers - Civil Engineering* 172(3), 102. <https://doi.org/10.1680/jcien.2019.172.3.102>.
- Hall, J. W., M. Tran, A. J. Hickford, and R. J. Nicholls. (2016). *The Future of National Infrastructure: A System-of-Systems Approach*. Cambridge: Cambridge University Press.
- Hamilton, C. H., and J. Perry. (1962). "A Short Method for Projecting Population by Age from One Decennial Census to Another." *Social Forces* 41(2), 163–70.
- Hardwicke, T. E., J. D. Wallach, M. C. Kidwell, T. Bendixen, S. Crüwell, and J. P. A. Ioannidis. (2020). "An Empirical Assessment of Transparency and Reproducibility-Related Research Practices in the Social Sciences (2014–2017)." *Royal Society Open Science* 7(2), 190806. <https://doi.org/10.1098/rsos.190806>.
- Hickford, A., T. Russel, J. Hall, and R. Nicholls. (2020). *A Sustainable Oxford-Cambridge Corridor? Spatial Analysis of Options and Futures for the Arc*. <https://www.itrc.org.uk/wp-content/uploads/2020/01/arc-main-report.pdf>.
- University Of Essex, I. F. S. (2020). *Understanding Society: Waves 1-10, 2009-2019 and Harmonised BHPS: Waves 1-18, 1991-2009, 13th ed. [Data set]*. UK Data Service. <https://doi.org/10.5255/UKDA-SN-6614-14>.
- Kanaroglou, P. S., H. F. Maoh, B. Newbold, D. M. Scott, and A. Paez (2009). "A Demographic Model for Small Area Population Projections: An Application to the Census Metropolitan Area of Hamilton in Ontario, Canada." *Environment and Planning A: Economy and Space* 41(4), 964–79. <https://doi.org/10.1068/a40172>.
- Lomax, N., and P. Norman. (2016). "Estimating Population Attribute Values in a Table: "Get Me Started In" Iterative Proportional Fitting." *The Professional Geographer* 68(3), 451–61.
- Lomax, N., P. Norman, P. Rees, and J. Stillwell. (2013). "Subnational Migration in the United Kingdom: Producing a Consistent Time Series Using a Combination of Available Data and Estimates." *Journal of Population Research* 30(3), 265–88.

- Lomax, N., and A. P. Smith. (2017). "Microsimulation for Demography." *Australian Population Studies* 1(1), 73–85.
- Lomax, N., and A. P. Smith. (2020). DAFNI Pilot 4: SPENSER—Synthetic Population Estimation and Scenario Projection Model. <https://dafni.ac.uk/wp-content/uploads/2020/05/dafni-pilot-4-dafni-hosts-population-forecast-model.pdf>.
- Lomax, N., P. Wohland, P. Rees, and P. Norman (2020). "The Impacts of International Migration on the UK's Ethnic Populations." *Journal of Ethnic and Migration Studies* 46(1), 177–99.
- Malczewski, J. (2004). "GIS-Based Land-Use Suitability Analysis: A Critical Overview." *Progress in Planning* 62(1), 3–65.
- Metz, D. (2012). "Demographic Determinants of Daily Travel Demand." *Transport Policy* 21, 20–5.
- Mohamed, Z., and P. Bodger. (2005). "Forecasting Electricity Consumption in New Zealand Using Economic and Demographic Variables." *Energy* 30(10), 1833–43. <https://doi.org/10.1016/j.energy.2004.08.012>.
- National Infrastructure Commission. (2017). Partnering for Prosperity: A New Deal for the Cambridge-Milton Keynes–Oxford Arc.
- ONS. (2020). Household Projections Across the UK: User Guide—Office for National Statistics. <https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationprojections/methodologies/householdprojectionsacrosstheukuserguide#comparability-summary>.
- Oshan, T. M. (2016). "A Primer for Working with the Spatial Interaction Modeling (SpInt) Module in the Python Spatial Analysis Library (PySAL)." *REGION* 3(2), 11. <https://doi.org/10.18335/region.v3i2.175>.
- Oughton, E. J., and T. Russell. (2020). "The Importance of Spatio-Temporal Infrastructure Assessment: Evidence for 5G from the Oxford-Cambridge Arc." *Computers, Environment and Urban Systems* 83, 101515. <https://doi.org/10.1016/j.compenvurbsys.2020.101515>.
- Rees, P., S. Clark, and R. Nawaz. (2020). "Household Forecasts for the Planning of Long-Term Domestic Water Demand: Application to London and the Thames Valley." *Population, Space and Place* 26(2), e2288. <https://doi.org/10.1002/psp.2288>.
- Rees, P., P. Wohland, P. Norman, N. Lomax, and S. D. Clark. (2017). "Population Projections by Ethnicity: Challenges and Solutions for the United Kingdom." In *The Frontiers of Applied Demography*, 383–408, edited by D. Cham: Springer.
- Rogers, A. (2008). "Demographic Modeling of the Geography of Migration and Population: A Multiregional Perspective." *Geographical Analysis* 40(3), 276–96. <https://doi.org/10.1111/j.1538-4632.2008.00726.x>
- Russell, T. (2019). Arc Scenarios v1.0.0. <https://github.com/nismod/arc-scenarios>. <https://doi.org/10.5281/zenodo.3529473>.
- Schlor, H., J.-F. Hake, and H. Kuckshinrichs. (2009). "Demographics as a New Challenge for Sustainable Development in the German Wastewater Sector." *International Journal of Environmental Technology and Management* 10(3–4), 327–52. <https://doi.org/10.1504/IJETM.2009.023738>.
- Shandas, V., and G. H. Parandvash. (2010). "Integrating Urban Form and Demographics in Water-Demand Management: An Empirical Case Study of Portland, Oregon." *Environment and Planning B: Planning and Design* 37(1), 112–28. <https://doi.org/10.1068/b35036>.
- Smith, A. P. (2018). "humanleague: A C++ Microsynthesis Package with R and Python Interfaces." *Journal of Open Source Software* 3(25), 629, <https://doi.org/10.21105/joss.00629>.
- Smith, A., R. Lovelace, and M. Birkin. (2017). "Population Synthesis with Quasirandom Integer Sampling." *Journal of Artificial Societies and Social Simulation* 20(4), 14. <https://doi.org/10.18564/jasss.3550>.
- Smith, A. P. (2017). "UKCensusAPI: Python and R Interfaces to the Nomisweb UK Census Data API." *The Journal of Open Source Software* 2(19), 408. <https://doi.org/10.21105/joss.00408>.
- Smith, A. P., and T. Russell. (2018). "ukpopulation: Unified National and Subnational Population Estimates and Projections, Including Variants." *Journal of Open Source Software* 3(28), 803. <https://doi.org/10.21105/joss.00803>.
- Smith, S. K., and J. Tayman. (2003). "An Evaluation of Population Projections by Age." *Demography* 40(4), 741–57. <https://doi.org/10.1353/dem.2003.0041>.

- Smith, S., J. Tayman, and D. A. Swanson. (2013). *A Practitioner's Guide to State and Local Population Projections*. Dordrecht: Springer.
- Spooner, F., J. F. Abrams, K. Morrissey, G. Shaddick, M. Batty, R. Milton, A. Dennett, N. Lomax, N. Malleon, N. Nelissen, A. Coleman, J. Nur, Y. Jin, R. Greig, C. Shenton, and M. Birkin. (2021). "A Dynamic Microsimulation Model for Epidemics." *Social Science & Medicine* 291, 114461. <https://doi.org/10.1016/j.socscimed.2021.114461>.
- Swanson, D. A., A. Schlottmann, and B. Schmidt. (2010). "Forecasting the Population of Census Tracts by Age and Sex: An Example of the Hamilton-Perry Method in Action." *Population Research and Policy Review* 29(1), 47–63.
- Tayman, J. (1996). "The Accuracy of Small-Area Population Forecasts Based on a Spatial Interaction Land-Use Modeling System." *Journal of the American Planning Association* 62(1), 85–98.
- Tayman, J., and D. A. Swanson. (1996). "On the Utility of Population Forecasts." *Demography* 33(4), 523–8.
- Tong, X., and Y. Feng. (2020). "A Review of Assessment Methods for Cellular Automata Models of Land-Use Change and Urban Growth." *International Journal of Geographical Information Science* 34(5), 866–98. <https://doi.org/10.1080/13658816.2019.1684499>.
- Valler, D., A. E. Jonas, and L. Robinson. (2020). "Evaluating Regional Spatial Imaginaries: The Oxford–Cambridge Arc." *Territory, Politics, Governance* 1–22. Epub ahead of print.
- Van Imhoff, E., and W. Post. (1998). "Microsimulation Methods for Population Projection." *Population: An English Selection* 10(1), 97–138.
- van Schroyen Lantman, J., P. H. Verburg, A. Bregt, and S. Geertman. (2011). "Core Principles and Concepts in Land-Use Modelling: A Literature Review." In *Land-Use Modelling in Planning Practice*, Vol. 101, 35–57, edited by E. Koomen and J. Borsboom-van Beurden. Springer Netherlands. https://doi.org/10.1007/978-94-007-1822-7_3.
- Wegener, M. (2021). "Land-Use Transport Interaction Models." In *Handbook of Regional Science*, 229–46, edited by M. M. Fischer and P. Nijkamp. Berlin Heidelberg: Springer. https://doi.org/10.1007/978-3-662-60723-7_41.
- White, R., I. Uljee, and G. Engelen. (2012). "Integrated Modelling of Population, Employment and Land-Use Change with a Multiple Activity-Based Variable Grid Cellular Automaton." *International Journal of Geographical Information Science* 26(7), 1251–80.
- Willis, R. M., R. A. Stewart, D. P. Giurco, M. R. Talebpour, and A. Mousavinejad. (2013). "End Use Water Consumption in Households: Impact of Socio-Demographic Factors and Efficient Devices." *Journal of Cleaner Production* 60, 107–15. <https://doi.org/10.1016/j.jclepro.2011.08.006>.
- Wilson, A. G. (1971). "A Family of Spatial Interaction Models, and Associated Developments." *Environment and Planning A* 3(1), 1–32.
- Wilson, T. (2015). "Short-Term Forecast Error of Australian Local Government area Population Projections." *The Australasian Journal of Regional Studies* 21(2), 253–75.
- Wilson, T. (2016). "Evaluation of Alternative Cohort-Component Models for Local Area Population Forecasts." *Population Research and Policy Review* 35(2), 241–61.
- Wilson, T., I. Grossman, M. Alexander, P. Rees, and J. Temple. (2021). "Methods for Small Area Population Forecasts: State-of-the-Art and Research Needs [Preprint]." *SocArXiv*. <https://doi.org/10.31235/osf.io/sp6me>.
- Wu, B. M., and M. H. Birkin. (2012). "Agent-Based Extensions to a Spatial Microsimulation Model of Demographic Change." In *Agent-Based Models of Geographical Systems*, 347–60, Springer.
- York, R. (2007). "Demographic Trends and Energy Consumption in European Union Nations, 1960–2025." *Social Science Research* 36(3), 855–72.
- Zipf, G. K. (1946). "P1P2/D Hypothesis: On the Intercity Movement of Persons." *American Sociological Review* 11(6), 677–86.