UNIVERSITY OF SOUTHAMPTON

FACULTY OF ENVIRONMENT AND LIFE SCIENCES

School of Geography and Environmental Science

Volume 1 of 1

An Ontology-Based Modelling Framework for Detailed Spatio-Temporal Population Estimation

by

Rebecca King

Thesis for the degree of Doctor of Philosophy

August 2018

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF SOCIAL, HUMAN AND MATHEMATICAL SCIENCES

Geography and Environment

Thesis for the degree of Doctor of Philosophy

AN ONTOLOGY-BASED MODELLING FRAMEWORK FOR DETAILED SPATIO-TEMPORAL POPULATION ESTIMATION

Rebecca King

The whereabouts of population changes over short time scales as people go about their daily lives. A requirement for very detailed population estimates that reflects this variation has been recognised for decades, with myriad application areas that could benefit from this in the public, research and commercial domains. Yet there remains a lack of suitable, extensible and transferrable methods for estimating population at the fine spatial and temporal scales of detail required for these applications. Such population estimation requires the integration of data from diverse sources including core geographic, statistical and the new and emerging sources from sensors and the internet. This integration includes creating appropriate linkages between the spatial, temporal and attribute data domains where these are related. Semantic web technologies provide a simple data model for the integration of such diverse data. Ontologies provide the ability to formalise the relationships between these data and make inferences through those defined relationships. This thesis presents a framework, or structure, into which new, evolving and alternative data can be worked with the goal of generating population estimates at very fine spatial (address level) and temporal (continuous) detail. The three-part modelling framework presented here integrates population in the spatial, temporal and attribute domains to estimate population counts at the level of addresses, on a continuous temporal scale. This thesis introduces, for the first time, the foundations of a semantic web-based modelling solution to this problem in the population estimation domain.

Contents

List of Tables

List of Figures

Declaration of Authorship

I, Rebecca King, declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research.

An Ontology-Based Modelling Framework for Detailed Spatio-Temporal Population Estimation

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. None of this work has been published before submission.

Signed: .....

Date: .....

Acknowledgements

Firstly, I would like to thank my academic supervisors Professor David Martin and Dr Nicholas Gibbins for giving me the opportunity to study with their constant and invaluable guidance and support over the last five years. Thanks also go to Jeremy Morley, Dr Jenny Harding and Glen Hart at Ordnance Survey who have, by turns, provided their valuable supervisory support.

I would also like to thank everyone, including those mentioned above, with whom I have worked within Geography and Environment, the Southampton Graduate School and Ordnance Survey for their intellectual and personal support.

Finally, I would like to thank Owen, Ellen and Bryn for their patience, understanding and support throughout this protracted process, and Sarah for her help in the final hours.

This research was funded by an Engineering and Physical Sciences Research Council (EPSRC) iCase Studentship at the University of Southampton and supported by Ordnance Survey.

Data acknowledgments: Ordnance Survey MasterMap Topography ™ and AddressBase Premium ™ under licence © Crown copyright and database rights 2015 Ordnance Survey. © Local Government Information House Limited copyright and database rights 2015. Census Output Area Boundaries, Workplace Zones and tables © Crown copyright 2011. UK Land cover map Edina supplied service © Crown Copyright. Time Use Survey 2014-2015 Edina supplied service © Crown Copyright. Google Street Map, Google Popular Times, Google Street View © Google. NHS data from https://www.england.nhs.uk/statistics/. Schools data from Gov.uk https://get-information-schools.service.gov.uk/.

Software: Ontotext GraphDB Free has been an invaluable tool for completing this work, and this work was partially conducted using the Protégé resource, which is supported by grant GM10331601 from the National Institute of General Medical Sciences of the United States National Institutes of Health.

Definitions and Abbreviations

A&E: Accident and Emergency

ABP: AddressBase Premium

AL2: Address Layer 2

AOI: Area of Interest

ATM: Automated Teller Machine

BLPU: Basic Land and Property Unit

CE: Communal Establishment

COWZ: Classification of Workplace Zones

DNF: Digital National Framework

DPA: Delivery Point Address

ESA: Eastleigh Study Area

FOI: Features of Interest

FS: Functional Site

GIS: Geographical Information Systems

GML: Geography Markup Language

GPS: Global Positioning System

GPT: Google Popular Times

GSV: Google Street View

HMO: Houses in Multiple Occupation

HSE: Health and Safety Executive

IoT: Internet of Things

LA: Local Authority

LPI: Local Property Identifier

MAF: Master Address File

MAUP: Modifiable Areal Unit Problem

MMS: MasterMap Sites Layer

MMT: MasterMap Topography Layer

MOC: Multiple Occupancy Count

MSP: MAF Structure Points

MTdb: MAF/TIGER database

MTUP: Modifiable Temporal Unit Problem

MSTUP: Modifiable Spatio-Temporal Unit Problem

NAG: National Address Gazetteer

NHS: National Health Service

NLPG: National Land and Property Gazetteer

NPD: National Population Database

OA: Output Area

OFCOM: Office of Communications

OGC: Open Geospatial Consortium

OGL: Open Government Licence

ONS: Office for National Statistics

OS: Ordnance Survey

OWL: Web Ontology Language

OWPA: Objects Without Postal Address

PAF: Postcode Address File

POI: Points of Interest

PSA: Portswood Study Area

RCC: Region Connection Calculus

RDF: Resource Description Framework

RDFS: RDF Schema

SIC: Standard Industrial Classification

SPARQL: SPARQL Protocol and RDF Query Language

SSA: Southampton Study Area

STA: Space Time Activity

STP: Space Time Points

TIN: Triangulated Irregular Network

TOID: Topographic Identifier

TS: Temporal Signature

Turtle: Terse RDF Triple Language

UPRN: Unique Property Reference Number

URI: Uniform Resource Identifiers

WZ: Workplace Zone

W3C: World Wide Web Consortium

XML: Extensible Markup Language

XREF: Cross Reference Table

Chapter 1 Introduction

The need for detailed spatio-temporal population data, that provide an estimate of population present in small geographical areas and at specific dates and times, has been recognised for several decades (Schmitt 1956), yet the availability of such data remains limited. In national and global datasets, spatial resolution is not high enough for many applications, and the temporal resolution tends to be limited to night-time only, or at best day-night or ambient (temporally averaged) populations (Sutton et al. 2001; McPherson & Brown 2004; Sutton, Elvidge & Obremski 2003). In order to meet the requirements of an increasingly wide range of applications, population models need to provide detailed spatial population distributions that go beyond the day-night models, and need to produce a full representation of population time (Martin, Cockings & Leung 2015). The growth in demand for high quality datasets is partly in response to new applications for the data, such as the need to plan networks capable for the Internet of Things (IoT) in emerging smart cities. Such technologies provision for, and monitor human mobility at very short time scales: minutes and hours, rather than months or years (Batty 2013). This requires very detailed spatio-temporal population intelligence to inform micro-geographic analysis. In this, and many other application areas, such highly detailed population data would therefore be of benefit. As Harper & Mayhew (2012, p.2) put it: “There is a demand for ever timelier and more detailed population data to satisfy a growing thirst for population intelligence”.

A number of new datasets that represent population are beginning to become available that have the potential to contribute to that intelligence. They come from a range of sources, including administrative, survey, social media and sensor data. Although these data are detailed in both spatial and temporal domains, and are complete in the sense that they are not sampled (Batty 2013), none of these sources can provide a complete picture of population distribution as no single source represents the entire population. Even seemingly ubiquitous datasets, such as Google Traffic or Popular Times, do not cover the whole population, and may have such great commercial value as to be available only to the organisation that collects these data, putting them out of reach for many potential applications. Each sensor based or social media dataset that may be available will represent a different, self-selected sample of the population (Jacques 2018) and as no single set of sensors will be able to provide a complete picture of population, an integrative model is required to provide a population estimate that brings disparate data sources together.

There is therefore a clear, yet currently unmet, requirement to integrate the intelligence gained from these sources with established geographic datasets in a single model that can provide population data that are detailed in both the spatial and temporal domains, for entire populations. This would represent a significant and valuable improvement on the currently available population models. One approach to modelling uses ontologies that lend themselves to

both data integration and the extraction of implicit knowledge from the data held within them, using inference. The development of such a model, that takes advantage of these capabilities, is the main focus of this thesis.

1.1 Population Data Application Areas

Such an integrating model could find application as a baseline for monitoring sensors (identification of anomalies by providing the order of magnitude of expected sensor data), for organisations who do not have access to sensor data, for machine learning training data or for informing land use classification.

Many organisations that have individuals as customers can benefit from knowing where they are and when they are there. Commercial, research and public service applications of time-specific population data are therefore very varied. Even businesses that operate solely on-line need be concerned with whether their customers have access to required internet services, which requires knowledge of customer locations.

Numerous commercial activities require access to detailed spatio-temporal population data. These include, but are not limited to location-allocation analysis for siting of new retail stores or service facilities (Deville et al. 2014), the provision of utilities and service delivery, such as mobile and fixed telecommunications, water, gas and distributed electricity generation and supply, that require dimensioning, planning and optimisation (Fry 1999; Garrity 2008). As well as these application examples concerned with the support of service provision, detailed population data are a resource in itself: “Knowing a customer’s location is an economic asset for a business” (Keßler & McKenzie 2018). While intelligence about the location of individual customers is not an aim for this research, the aggregated presence of customers is of value for activities such as market research, for understanding the customer base for products and services and then marketing to that mobile customer base where it congregates, or passes through.

Other application areas that fall more in the public services domain include those surrounding health, public safety, transportation, civil defence, security and education. Administrative functions requiring detailed intelligence about the location of people include resource allocation, construction of population ratios and urban and local planning and intelligence (Schmitt 1956; Mennis 2003; Bhaduri, Bright & Coleman 2005; Patterson et al. 2007; Ahola et al. 2007; Aubrecht et al. 2013; Harper & Mayhew 2012). All of these activities have customers and knowing where those customers are, at different times, can contribute to improving the services provided.

Planning smart cities, particularly their smart living aspects requires detailed knowledge about the population, what their requirements are and the various places that they spend their time, at the micro-geographic level. This is increasingly important as populations grow, and as they age and have specific requirements from the technologies that contribute to the smart city, such as the innovative controllers for the technological fabric of the smart city (and the smart home) that respond to the presence of people (Skouby et al. 2014). Planning adequate networks to carry IoT data that are essential to these technologies requires detailed knowledge of the spatio-temporal population patterns.

Research areas requiring detailed population data are many and varied, and include: spatial epidemiology (Linard & Tatem 2012); research and policy applications (Martin, Cockings & Leung 2015); research in criminology, exposure and vulnerability analysis (Martin 2016; Smith, Martin & Cockings 2012).

To demonstrate the importance of detailed population data, two of these application examples are discussed in more detail below. The demands on the population data remain the same despite the very different application areas.

1.1.1 Mobile Telecommunications

In the field of mobile telecommunications, population data are used for various different activities. Apart from sales and marketing, these activities also include network planning, optimisation and monitoring.

In the UK, regulatory bodies require certain minimum levels of population coverage for the mobile network (Ofcom 2012). Ensuring that these levels are met is a priority for network providers. This involves planning both improvement to the existing networks, and new technology layers, such as the fifth generation, or 5G, mobile network. Potential sites for mobile cell infrastructure are analysed to discover the network coverage they will provide, to prioritise areas that will serve the highest number of people (and raise the most revenue), and to ensure there are no areas where population will not be covered.

Network optimisation ensures that capacity is sufficient for the expected demand by improving performance of the existing network. Where optimisation will not meet demand, additional hardware must be utilised. Demand for network services changes over time, based on varying presence of population, so the need for detailed spatio-temporal population is clear for this activity, which ensures high quality service, while prioritising investment in the physical network for areas where population are present.

Network monitoring involves analysing coverage maps alongside population data to ensure that population coverage requirements are met. The UK regulatory body's minimum required population coverage level of "an area in which 98% of UK population lives" (Ofcom 2012) is based on residential populations, but it is in the interest of both the network provider, and the consumer to have adequate mobile network coverage everywhere that customers want to use their devices. This includes locations where users are engaged in residential, work and leisure activities, as well as in-transit and locations where people consume other public or commercial services. In order to provide the best service, the mobile network provider needs to understand the number of potential users in locations serving all of these functions, as many of these activities occur where there is no residential activity. This situation is common internationally, although industry knowledge indicates that the population thresholds may vary based on the settlement pattern of the country in question.

Such commercial requirements for population data are not often highlighted in the literature, but is apparent in recent moves by companies such as Facebook investing in developing population maps for network planning purposes. This mapping will be published as it is acknowledged that it has application areas including socio-economic research and risk assessment for natural disasters as well as the commercial application for which it was developed (Facebook 2016). Despite these wider applications, it is likely that commercial development and use of population data may more often be classified as commercially sensitive and therefore never enter the public domain.

1.1.2 Emergency Response and Planning

There is a bias towards public (rather than commercial) uses of population data within the literature, and within this body of work, emergency response and planning is frequently cited as a primary application area. This is not surprising given the variety and human impact of the events that this includes. Anthropogenic and natural hazards refer to events such as flooding, hurricanes, earthquakes, tsunamis, radiological releases, toxic chemical or biohazard releases, pandemics, fires, war situations or terrorist attacks (Zhang, Sunila & Virrantaus 2010; Tenerelli, Gallego & Ehrlich 2015; Ahola et al. 2007; Krisp 2010; Aubrecht et al. 2013; Smith, Martin & Cockings 2012; Renner et al. 2018). Activities for dealing with the consequences of these events are also varied, and can be broadly divided into two main areas: being prepared with emergency plans prior to an event occurring, and managing the immediate response to the event, both of which focus on protecting human safety as well as reducing damage to infrastructure and assets (Ahola et al. 2007). Being prepared for such events and having plans in place for how to respond can prevent a hazardous event becoming a disaster (Aubrecht et al. 2013) so these activities also consider the

means for minimising the potential severity of the event. Emergency planning involves assessing the severity of consequences and the likelihood of an event to assess risk.

In terms of protecting human safety, the number and location of people affected by an event is required in order to indicate the consequences. One of the core pieces of information for emergency planning is therefore a detailed and time-relevant population distribution (Aubrecht et al. 2013). This is also the case for emergency response activities, for identifying the numbers of people affected, which is time-dependent, in order to provide assistance.

There are two main activities highlighted here that require detailed knowledge of the spatio-temporal population distribution: the long-term planning emergency preparedness activity, and the short-term, time-critical decision making of emergency response (Ahola et al. 2007). Both can be planned for, and both require detailed spatio-temporal population data to determine numbers affected by an event. Emergency preparedness may involve activities such as planning the location of fire stations and medical facilities, or reducing the potential for bottlenecks in transportation networks in the event of an emergency. Emergency response is focused on activities such as delivering emergency services, providing medical assistance, evacuating at risk people or preventing a situation from deteriorating thereby putting more lives at risk. Highly detailed spatio-temporal population distribution data are key to these activities (Aubrecht et al. 2013) for identifying the numbers of people who are at risk, or who have already been exposed to a hazard.

1.2 Modelling Approaches

Two potential approaches for spatio-temporal population modelling are discussed below: Geographical Information Systems (GIS) modelling and Ontology modelling. The basic requirement of the modelling approach is that it must enable the integration of, and ability to make inferences across diverse and heterogeneous data from three domains:

1. Spatial domain: locations of human activity, in the form of addresses (throughout this thesis, addresses are defined as structured textual descriptions of spatial objects, including, but not limited to, those used in the management of postal delivery), cartographic objects (buildings) and Functional Sites (FSs, that are comprised of groups of cartographic objects), administrative data.
2. Attribute domain: functional class of the addresses, statistics, census and administrative data, properties of the cartographic objects.

3. Temporal domain: levels of occupation in the addresses at different times (based on values in the attribute domain)

This level of integration of data is a complicated process within GIS, as datasets need to be formatted to fit the specific GIS data model. The spatial and attribute domains are handled well, but the handling of the temporal domain is limited, tending to require time explicitly modelled on individual features using timestamps and as a temporal extension to the spatial database (Worboys & Duckham 2004, p369). The handling of spatio-temporal data continues to be recognised as an area for research in need of attention (Yuan 2015; Kwan & Neutens 2014). There are other disadvantages to using GIS for modelling population in this way. Firstly, GIS analyses require complex spatial analysis, which is computationally expensive and can produce very cumbersome databases. Secondly, once a GIS model has been created, it can be difficult to adapt or re-use the model, and, in particular, difficult to integrate new data as they become available, as this may require re-engineering of the existing model *and* the databases. Finally, GIS are not explicit about what is being represented, beyond the metadata associated with them that informs a human user. This means that inference within a GIS is not a straightforward process and requires connections between the data to be made by a human operator.

The ontology approach provides access to a simple data model capable of merging and modelling complex data and relationships (Allemang & Hendler 2011, p49; Powers 2003, p*ix*). As a result, data integration is a straightforward process. The spatial domain is handled differently to within a GIS, using qualitative reasoning for making inferences rather than complex spatial analysis that relies on quantitative representations of geography. It is possible to handle spatial relationships by explicitly stating these within the data model, and tools are available for limited spatial analysis. The data model used in the ontological approach is capable of handling time more flexibly, with the ability to use a hierarchical approach in modelling the temporal dimension of every feature. Where the best data are not available, it is straightforward to use alternative, more generalised data. The very simple data model that makes ontologies good for data integration also enables data and knowledge sharing. This data knowledge is held within the ontology as explicitly stated facts and relationships, which define what the data represent and how they relate to each other. This means that in comparison to GIS models, it is very straightforward to infer implicit information from the data. These advantages of the ontological approach seem to make this a natural choice for modelling the diverse datasets required for population estimation.

1.3 Research Aims

The purpose of this research is to explore the advantages of using an ontology for modelling highly detailed spatio-temporal population so that it can provide an improvement on the currently available data models.

> *The aim of this research is to develop a generalizable modelling framework for population estimation that enables high levels of spatial and temporal detail. The model must also have the ability to integrate data from diverse sources.*

The first of the stated model requirements, to enable high levels of spatial and temporal detail is to provide the full representation of population time (continuous temporal scale), at the highest possible spatial resolution (address level). This has not been achieved by previous efforts to estimate population at such high spatio-temporal resolution, tending towards either high spatial *or* high temporal resolution but not both (e.g. Martin, Cockings & Harfoot 2013; Greger 2014; Zhang, Sunila & Virrantaus 2010b).

The second requirement, the ability to integrate data from diverse sources, will utilise the simplest available data model so that any data source may be integrated into the model. These data sources may vary spatially, and be applicable at regional, local, or even individual address level.

The goal of the model is to be able to estimate population at a specific time and within a defined area using disaggregation from small area statistics, and allocation of data from address specific statistics. The modelling framework should therefore be capable of utilising detailed spatial and temporal data that can be aggregated for output, thereby allowing extraction of estimated population for any Area of Interest (AOI).

The intention is to develop the modelling framework in such a way as to enable the integration and use of commonly available datasets whilst allowing for variation in data quality, both within and between datasets. The model must therefore be generalizable in terms of the ability to translate it to other data environments, and to be able to make use of the best available data, whether or not its quality is less than that of the data used in this research.

It is an *a priori* assumption that the ontology modelling approach has much to offer the area of population estimation, due to the potential of the common framework for data integration, and the ability to make logical inferences on the data.

The second aim of this research is to investigate the potential role of using ontology and associated Semantic Web technologies for population estimation.

The focus is on proof of concept: that ontologies can be utilised to use the function of addresses, and therefore the activities that occur at them, to estimate population at those addresses, as this varies over time.

The innovation is in driving spatial resolution to a high level of detail, with the ability to add more detail for specific sites where data are available, so that individual sites do not have to use averaged population data if measured data are available. This is a novel approach, utilising ontologies within the population modelling domain. It responds to the need to disaggregate neighbourhood scale statistics to a finer spatial scale, in both the modelling and the data in order to produce a functioning model, rather than simply a collection of data in a GIS.

Potential new data sources are, and will continue to be, evolving in ways including adjustments to data offerings from suppliers, the opening up of national datasets and the availability of new sources of social media, survey and administrative data. This highlights the value of the development of such a flexible modelling framework.

1.4 Objectives

The above stated aims of this research require several objectives to be met, as described below. The first of these objectives arose from a suitability analysis of the core geographic datasets to be employed in the model.

1.4.1 Core Geographic Data Preparation

As the modelling framework is to be ontology-based, the spatial relationships must be represented as qualitative relationships so conventional GIS are required for data preparation. The first objective is therefore:

Objective 1: Prepare appropriate core datasets so that they can be utilised within the model.

This data preparation involves making spatial linkages between features in different datasets so that they can be represented qualitatively in the ontology. This process is outlined in Chapter 4: Data and Study Area, along with descriptions of the other, ancillary, datasets that are employed in the model, including data in the temporal domain.

In working towards this objective, an additional and significant objective was defined: to ensure that address data have classifications to an adequate level of detail. In the core geographic data sets employed, the addresses have a classification that is based on their function. This classification will be used to assign a *temporal signature* (pattern of occupation) to the address so that occupation levels can be estimated. Where inadequate classification is available, it must be possible to supplement the data to add additional information, leading to the following objective:

> *Objective 2: Supplement the classification of address data where this is required for population estimation purposes.*

The resulting process of generating residential address classifications that indicate dwelling type is outlined in Chapter 5.

1.4.2 Ontology Development

Ontologies and semantic web technologies provide a very simple data model, and the qualitative spatial reasoning that can be performed by these technologies offers a viable alternative to the current approaches that generally focus on quantitative spatial analysis. This research therefore includes the requirement to establish how these technologies can be utilised within the modelling framework, how well the temporal domain can be handled within an ontology, and what other benefit they can provide over current approaches (i.e. GIS and spatial analysis). This third objective is therefore:

> *Objective 3: Develop an ontological model that can utilise core data sources, and that is capable of integrating other data sources, including in the attribute and temporal domains.*

The ontology design, development and use, is described in Chapter 6: Ontology Development.

1.4.3 Population Estimation

The purpose of the modelling framework is to estimate population at the fine spatial and temporal resolutions already stated as a requirement for the various applications of population data. This leads to the final objective:

> *Objective 4: Ensure all arithmetic, aggregation and inference operations, required for population estimation, can be effectively applied using the ontology and semantic web technologies employed, and that it is possible to allocate population to the appropriate geographic features.*

This includes the ability to treat population sub-groups differently. The development of techniques for population estimation is described in Chapter 7: Population Estimation.

1.5 Out of Scope

The aims and objectives set out above focus on a proof of concept for a modelling framework for spatio-temporal population estimation. While the model must be capable of the appropriate calculations for population estimation, the estimation algorithm itself is not under development in this research. As such, the research focuses not on developing and demonstrating every aspect of population estimation, but on proving that all the required mechanisms are present in the modelling environment to enable disaggregation of small area populations to addresses, and the assignment of specific, populations to addresses, while accounting for temporal variations in occupation levels.

Although the modelling framework must be capable of handling time, in order that population can be estimated for specific times, the development of the temporal patterns of occupation is out of scope of this research. Several time profile datasets are already available from the Population 24/7 project at the University of Southampton (Martin, Cockings & Smith 2017), so their creation is not necessary. For proof of concept, the model needs to handle the temporal domain adequately, but the temporal patterns of occupation are not under development here.

Given that these two areas are out of scope, the production of a population data product is not one of the objectives.

The research will not involve any attempt to count people in individual addresses, as it is based on disaggregation of small area statistics to geographies at a higher spatial resolution. Spatial microsimulation, cellular automata or agent-based models will not be employed. These individual-based models that track or estimate the movements and activities of people remain impractical for large populations, they cannot achieve full population coverage and they can involve issues of geo-privacy (Richardson et al. 2015).

As this research is not concerned with measuring, modelling or simulating the movements of identifiable individuals, the modelling framework is focused instead on estimating where people are (within addresses) rather than the spatial behaviour of individuals. The subject of the model is the address spatial object, and not the individual person. It will therefore not be possible to compromise the geo-privacy of identifiable individuals from the results of this research.

There is no attempt to model Communal Establishments (CEs) or House in Multiple Occupation (HMOs) as there are limited data available to support this.

The research does not involve any attempt to handle changes in the data. The geographic and attribute data are considered as static for the purposes of framework development. The datasets employed are for specific reference years, and each data producer has its own mechanism for handling changes over time. For simplicity, spatial data will be considered as static, and data from a single reference date will be used. Likewise, the research is not concerned with harvesting dynamic data feeds to create a model capable of "now casting" (a modelling technique that, while not providing live situational results, allows short-term predictions of the current situation). The modelling framework however, should be capable of such data integration and it is clear from the start that this is a potential future area for research, rather than something to be incorporated in this initial modelling framework development.

No ground level transportation will be included in the model, as the focus is on ensuring population can be modelled at address and building objects. Air travel, which takes population out of the model, and underground travel for which detailed building data are not currently available, will also be excluded from this modelling process. The result is that some populations will be excluded from the model.

The research focuses on the use of semantic web technologies for qualitative spatial reasoning, and so spatial analysis is restricted to GIS, with no attempt to be made to incorporate spatial models in the modelling framework.

1.6 Report Outline

This thesis structure, set out in Figure 1, is as follows:

Chapter 2 contains a review of the literature on population estimation techniques and semantic web technologies. This includes a review of the key concepts of time geography, population redistribution and ontologies in spatial and temporal modelling. Semantic web technologies employed in the research are also presented here. This is followed by an introduction to the data landscape. Chapter 3 outlines the research design, providing detail of the approach to the modelling, which is in three parts.

Data assessment, preparation and supplementary classification of address data are described in Chapters 4 and 5 respectively. The approach to validation of the data load and classification are presented at the end of this chapter. The outputs of this part of the modelling framework feed

directly into the ontology, which described in Chapter 6. This gives an accounting of the central ontological model, and the data load, as well as an outline of the ontology validation using synthetic data. Chapter 7 describes how the model, once populated with data, can be used to estimate population through use of SPARQL Protocol and Resource Description Framework Query Language (SPARQL) scripts, and presents the model output alongside validation time-series that used mobile network data as a proxy for population.

Chapters 4, 5, 6 and 7 each present methods, results and validation of the different parts of the modelling framework.

Chapter 8 sets out a discussion of each of the three parts of the modelling framework, as well as an over-all assessment of its limitations and potential use in the population estimation domain. Finally, Chapter 9 presents conclusions and areas for further research.

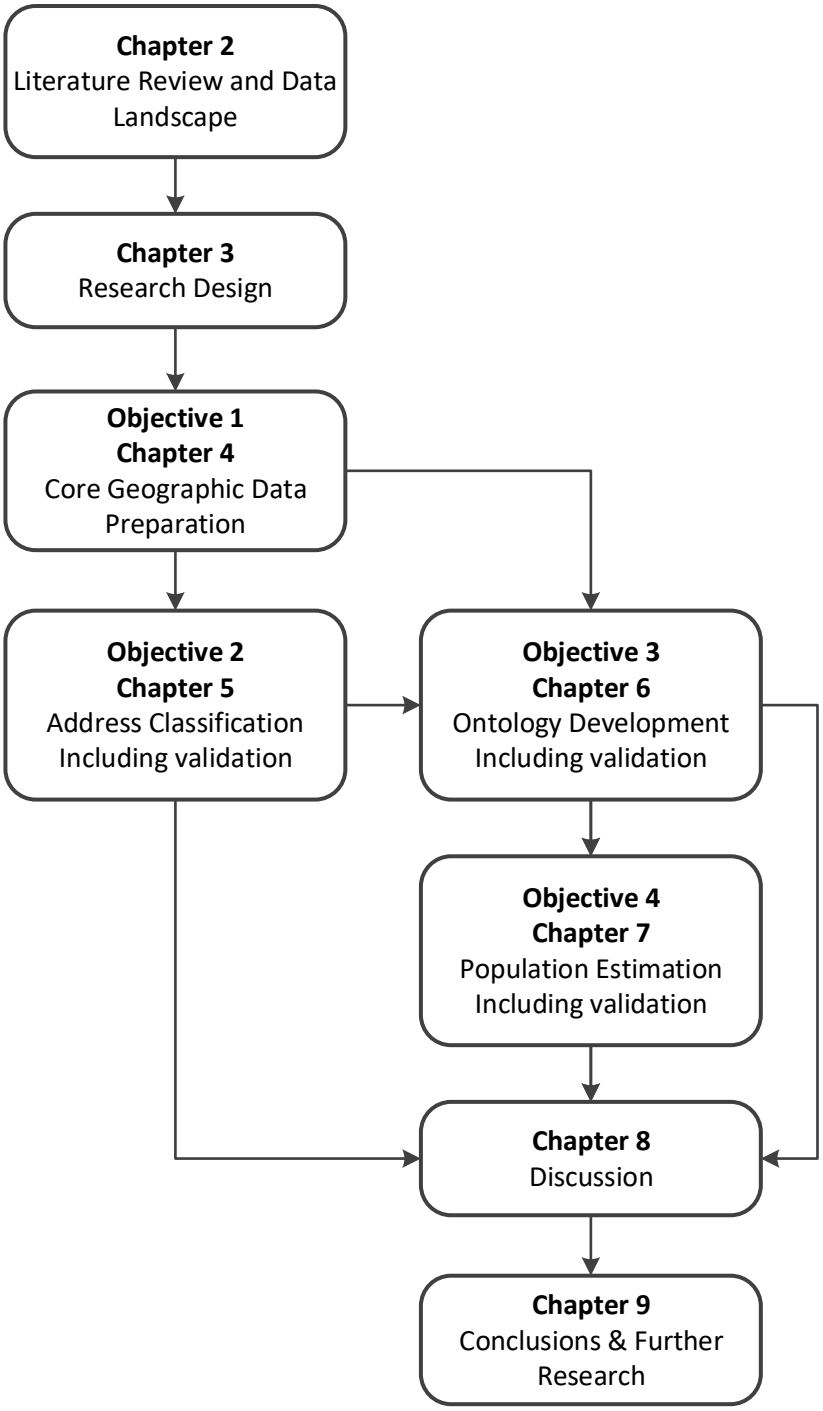


Figure 1: Thesis Structure

Chapter 2 Literature Review

Several aspects to the research aims and objectives laid out in the previous chapter need to be considered. They include the concepts associated with representing geographical features: consideration of the sources of population data; methods currently used for distributing these; and how the temporal aspect of data is handled with specific consideration to current approaches to redistributing population data that include the temporal dimension. The technologies associated with ontologies also need to be considered, as these are central to the functioning of a semantic web solution to spatio-temporal population estimation. Finally, the data that will be used within such a modelling framework need some attention. A review of the literature for each of these three areas is presented in this chapter.

2.1 Conceptual Review

2.1.1 Digital Representations of Geography

When people communicate geographic concepts to each other they can make themselves understood because they use natural language. When seeing and describing features in the landscape (such as a significant change in elevation), people parse it into discrete objects (such as a mountain), rather than trying to describe the continuous field that they see (Smith & Mark 2003). This is perhaps because in evolutionary terms objects have more meaning for us, as they have the potential to be food or predators (Peuquet, Smith & Brogaard 1998). It may also be the reason why natural language is better suited to describing objects than fields (Cova & Goodchild 2002).

The field and object based conceptualisations of the world are the two main models of reality used within GIS. Either model can be represented using vector or raster data structures (Goodchild 1992). Other models of space have also been suggested: the categorical coverage which contains bounded regions defined by values in the field, such as soil categories, and hard partitions which are defined by something other than the field values, such as administrative units (Peuquet, Smith & Brogaard 1998). Both of these can be represented as objects within GIS. Unless the key concepts of the field and the objects are well defined, it is difficult to model these concepts within a computer environment, as the natural language appropriate for human communication is too imprecise for use in computer models.

In the field view of the world, geographic features are represented in terms of their spatial distributions: the magnitude of the feature in question that is present (for example precipitation, temperature or elevation), in the spatial domain in question. The field view links the independent

locational reference frame with the dependent attribute domain and is a function, not a set of values (Worboys & Duckham 2004). As such, a field has several key characteristics: it assigns a value to every location because it is spatially continuous; the set of locations is theoretically infinite; the attribute can be discrete or continuous; any measurement scale can be used for attribution (binary, nominal, ordinal, interval or ratio); spatially, the field can be one, two or three dimensions. Importantly, representation of a field within a computer system must, by necessity, be approximate, and this is usually achieved using tessellations, point grids or irregular polygons representing average or dominant types (Goodchild 1992; Cova & Goodchild 2002). Fields can also be represented using Triangulated Irregular Networks (TINs) which allow for calculation of field values at any point by interpolation (O'Sullivan & Unwin 2003).

In the object view, space is described as a container, or a plane, populated by discrete objects with spatial reference (e.g. addresses, buildings, or field values sampled or averaged on a regular grid to create a raster (Fisher 1997)) that are commonly represented as points, lines and polygons (Goodchild 1992; Cova & Goodchild 2002). In the object view of the world, the features need not be connected and continuous (as they are in the field view), they may instead be scattered and separated (Smith & Mark 1998). As well as describing objects by their basic geometry, objects can also be described by their type. Three types of objects are suggested by Smith & Mark (1998) in their attempt to create an ontology of geographies: these are physical (e.g. river, building, bridge); those which are "imposed on the physical world only by human cognition" (e.g. bay, mountain); and those that "exist only as the hybrid spatial components of human cognition and action" (e.g. administrative boundaries). This object type influences the definition of the object boundary as either *fiat* (imposed by humans) or *bona-fide*. The boundary can be crisp and therefore suitable to be represented using points, lines or areas, or vague and best represented using fuzzy modelling (but often still inadequately represented using points, lines and areas).

A common example of a variable represented in the field view is the population density map. This is a continuous field that represents the ratio of population in a defined reference area (as hard partitions in the object view) surrounding that point (Nordbeck & Rystedt 1970). It is a value that is generated using a calculation on the reference area. A population density map is therefore a fiat field, as the values are imposed, and therefore not tangible in the same way as an individual unit of population is tangible (Peuquet, Smith & Brogaard 1998). The values of the population density map are determined by the size and shape of the reference area that is used in their calculation so for one population data set there can be more than one possible population density map (Nordbeck & Rystedt 1970; Peuquet, Smith & Brogaard 1998). Representation of the population density map as a raster, in common with all field representations, requires that the field is approximated into discrete values (Cova & Goodchild 2002) which could be number of

people in a raster pixel, or could be a statistical representation of the field values within a pixel, e.g. number of people per km^2.

2.1.2 Population Data

Human activities can be classified at a high level as either residential, non-residential or in-transit activities (Martin, Cockings & Leung 2015). Residential activities all occur in the place of residence. These could be either private residences, or Communal Establishments (CEs). The sum of residential population is the sum of the population (excluding visitors). Non-Residential activities include all activities that occur outside of places of residence. This includes work, education, leisure and consuming services (including retail, health services and public services), and is discussed below. In-transit activities are those involved with pedestrian, road, rail, water and air transport.

There is a range of sources of population data that can be used as a means for understanding where people are, including census, administrative and big data from social media, or streamed from sensors (Batty 2013).

2.1.2.1 Residential Population

In countries with accurate population registers (such as Finland and Norway), these can be the most successful means of geo-locating residential populations. In countries where there is no definitive national population register (such as the UK) it is necessary to integrate various administrative views of the population, such as those maintained by local government and the National Health Service (NHS). These can be successfully used as a source of residential population (Harper & Mayhew 2012). This approach has several issues, including delays in change of address details reaching the registers (Matheson 2014).

The most widely available comprehensive source of population data is currently census counts. Over 200 countries carry out a census and publish the population data (United Nations Statistics Division 2014), aggregated to small areas in order to remove the possibility of identifying individuals so that privacy is protected. These population data represent residential, or night-time population and when summed for a country will give a total population count, although these data do require adjustment to deal with undercounting, and that involves the use of administrative data (Shuttleworth & Martin 2016).

Unit postcodes are the base level of the postal geography in the UK, designed for the purpose of creating an efficient mail delivery service (Raper, Rhind & Shepherd 1992). These are not areas

per se but are collections of a small number of addresses (about 15) that can be used to construct areas. These areas tend to be socially homogeneous and they are used as building blocks for generating Census Output Areas (OAs) (Cockings et al. 2013) which are in turn designed to be as socially homogeneous as possible (ONS 2012b). In England and Wales, Census OAs are the smallest geographical areas at which a suite of census estimates is published by the Office for National Statistics (ONS), in Scotland by the National Records of Scotland (NRS) and in Northern Ireland by the Northern Ireland Statistics and Research Agency (NISRA). In common with several other countries, they include non-residential areas as they are also designed to be space filling, therefore covering the entire land surface (Cockings et al. 2013). The ONS also publish census headcounts for unit postcodes, the smallest geographical units for which population counts can be identified. These small area data therefore need to be disaggregated if population for smaller units (such as individual buildings or addresses) are to be estimated.

2.1.2.2 Non-Residential Population

Non-residential population includes those populations at work, education or engaged in other non-residential activities (except for transportation as this is dealt with separately). Data available for these populations can be classified according to whether they are aggregate data available for small areas, or site based data, available for individual sites, whether these are individual addresses, buildings or groups of buildings such as hospitals or schools.

In England and Wales, the ONS publish a workplace population count, which represents number of people working within an area, and is collected during the census. These figures exclude short term residents as they are based on individuals who were present in England and Wales, or intending to be present, for a total of 12 months or longer (ONS 2014). OAs were designed to have consistent population sizes, based on residential population. Alternative geographies are therefore required for generating consistent population sizes in terms of workplace areas, because residential and work activities occur at different locations (Martin, Cockings & Harfoot 2013). The smallest geographies at which these statistics are published are Workplace Zones (WZs), which have roughly consistent number of workers employed in each zone. These small areas are not the same as the residential OAs due to disclosure risks which arise in areas that are mostly residential and have small numbers of workers (ONS 2014) but they mostly do represent exact subdivisions, aggregations or matches to the OAs (Martin, Cockings & Harfoot 2013). In common with other census outputs, these small area data also therefore need to be disaggregated if population for smaller units are to be estimated.

Administrative estimates of population counts are available for hospitals, educational establishments and various commercial enterprises such as supermarkets. Sensors may be used for the data collection for some of these datasets, and may sometimes provide the counts of visitors to different types of site. These sensor data may be derived from, among others, mobile telephony, Wi-Fi networks, Global Positioning System (GPS) enabled devices, traffic sensors, or georeferenced social media posts, sensor networks recording environmental conditions, and volunteered internet based geospatial information (Kwan & Neutens 2014; Martin, Cockings & Leung 2015). Non-residential population may also be considered to include some leisure data sources such as visitors to indoor and outdoor leisure facilities and retail data sources such as footfalls in stores.

2.1.2.3 In-Transit Population

The detailed modelling of the number of people in transit and using various modes of transport is out of scope for this research and is dealt with in the substantial field of transportation planning. The model presented in this thesis does not include any transportation data, as the focus is on the occupation levels of addresses, but some sources of transportation data are discussed below.

Transportation data are available concerning population in transit from a variety of sources within the transportation sector. These include passenger numbers and footfall at stations for rail, numbers of vehicles on the road and numbers of passengers using ferries and visiting airports. For example, in the case of transportation hubs such as airports and cruise terminals, it is possible to model the detailed spatio-temporal patterns of population through openly available data such as flight schedules (Jochem, Sims, E. a. Bright, et al. 2013). These provide detailed, site-specific population counts. There are also definitive data sources for some in-transit populations, such as those from rail companies and government departments. In the case of road traffic in the UK, these data are available aggregated to average daily flows at survey sites (Department for Transport 2017) and it is this type of traffic data that is utilised in the Health and Safety Executive's (HSE) National Population Database (NPD) (Smith & Fairburn 2008), and by the Population 24/7 project (Martin, Cockings & Smith 2017), discussed later. These provide good background transportation layers for population estimation models.

The temporal granularity of this road survey data (annual update at best) and the inconsistent spatial granularity of cell based mobile phone data (areas ranging from less than $100m^2$ to tens of km^2) are too low and varied for detailed spatio-temporal modelling. There is, however, scope for improvements in travel data based on new, and existing, sensor technology (such as metro smart cards), and new approaches to processing the existing data (Batty 2013).

The sensor data may be real-time, near-real-time or aggregated and may represent different events, for example, measuring the number of people crossing the threshold of a shop, counts of the number of bus tickets sold at different times, traffic sensors, or mobile phone locational information etc.

It may be possible to estimate in-transit population counts using mobile network data for large areas (Deville et al. 2014) but these data represent a self-selected sub-population (Jacques 2018) so assumptions must be made regarding the travel habits of the remaining population. The existence of multiple network operators further reduces the size of the sample population if mobile network data are used. However, Google Popular Times (GPT) (Google 2018) and Traffic Layer are based on data from GPS in mobile phones, aggregated across devices and network operators, (thereby reaching a greater proportion of the population), including only those devices that are using location services (Barth 2009). These do not provide magnitudes of occupation, but do give an indication of patterns of occupation, and road use, through time, so they have great potential for measuring and estimating populations not at work or residences.

These additional sources of knowledge about the location of population are growing in number with the rise of smart technologies, so population estimation models need to be flexible enough to incorporate existing, and new data sources of this type.

2.1.2.4 Small Area Population Data

Despite their small spatial extents, the aggregated residential and workplace population data do not provide enough information for many applications (e.g. Freire et al. 2013; Greger 2014), for several reasons that can be categorised as spatial and temporal issues.

The spatial detail is lacking because census data tend to be aggregated for privacy reasons (Duke-Williams & Rees 1998) and so the spatial scale at which the data are published are larger than the scales at which they are collected. This decrease in spatial resolution of these data is dealt with using disaggregation techniques described in the next section. The small area nature of these data means that all locations fall inside a polygon (Martin et al. 2000). This allows them to be displayed quickly and easily as choropleth maps, but also means there will be no zero-value areas represented in the map. This model represents an object view of the world (Goodchild 1992), and assumes homogeneity in the population distribution within each small area, which clearly does not accurately represent the real-world.

There are other issues associated with aggregating population to small areas. These are largely associated with the Modifiable Areal Unit Problem (MAUP) in which the choice of boundaries for

spatial data aggregation may have a greater effect on the results of spatial analysis than the underlying distribution (Openshaw 1984; Fotheringham & Wong 1991). Related issues surrounding the use of choropleth maps as a visualisation tool for representing small area population values include the fact that larger areas lead to lower calculated population densities, both the size and shape of the area influence the values associated with it, and false spatial continuities and discontinuities are generated (Langford & Unwin 1994). These issues (aggregation, segmentation and zoning effects) related to the MAUP apply to the use of any discrete, irregular areas used for mapping population.

Two application areas for which small area statistics are inadequate were introduced in Chapter 1: mobile telecommunications applications and emergency planning and response applications. In mobile telecommunications in Great Britain (GB), the regulatory body, the Office of Communications (OFCOM) analyses coverage at residential delivery points, between which the small area populations are evenly distributed in the model (Ofcom 2012). This highlights the need for more highly detailed population maps: to overcome the inappropriate assumption of even population distribution within the small areas.

The spatial and temporal scales at which emergency events take place are varied. Flooding exemplifies this as it can occur at a variety of spatial and temporal scales – localised flash floods or floods covering entire counties or regions. Spatial scales can be large as in the 2002 and 2013 European floods in both of which flooding of the Elbe and Danube caused widespread damage (BBC 2013; Ulbrich et al. 2003), or small as in the Boscastle flash flood in 2004 (Burt 2005). Many of the emergencies highlighted earlier (such as chemical release, large-scale flood or earthquake) will result in an area requiring a response covering areas greater than the census tracts. However, for these, as well as for the finer scale hazards (such as terrorist attacks or smaller fires), the logistics of providing assistance to those affected requires population data at scales that are finer than the aggregated census data can provide.

Moreover, it is highly unlikely that the extent of a hazard's effects will exactly fit the boundaries for which the census data are published. As a result, sub-census geographies will be required for assessing population affected by a hazard, in order to focus response activities on those who are affected rather than those who are unaffected.

In order to mitigate these issues, and those associated with the MAUP, a better approach is to represent population at a sub-census level (Bhaduri, Bright & Coleman 2005), generating a field representation of the data at the highest possible spatial resolution, in which spatially homogeneous regions (e.g. a grid) are used (Goodchild 1992). This could be represented using either irregular zones or regular polygons, such as a raster. If the polygons or grid cells are

sufficiently small, this allows for data aggregation by any areal units. This approach will also overcome issues related to mismatches in spatial scales and extents of a hazard (or other analysis extent) and the published population data.

2.1.3 Population Re-Distribution Methods

The methods outlined in this section ignore the temporal aspect of the data and redistribute population recorded and published for a single snapshot in time. These methods can be applied to population data available as counts associated with small areas (fiat objects that have no explicit variability within them) and need to be re-distributed to smaller or different areas.

2.1.3.1 Spatial Interpolation and Disaggregation

There are well-established techniques for spatial disaggregation of census data published for small areas that have no continuous variability modelled within them, effectively increasing the spatial resolution of the published data.

Longley et al. (2011, p373) describe spatial interpolation as "a process of intelligent guesswork, in which the investigator (and the GIS) attempt to make a reasonable estimate of the value of a continuous field at places where the field has not actually been measured", making interpolation appropriate only for continuous field values such as population density.

In order to demonstrate the conceptual difference between spatial interpolation and spatial disaggregation, one can imagine two points, A and B which each have a measured population attribute. This value relates to the population at the point. If the two points have associated population *densities*, the density value at an intervening point can be estimated by interpolation, as this is suitable for continuous field values, using one of several possible models. The values at the measured points A and B remain the same, as they represent densities, not counts. The estimate can be based on a linear, piecewise or higher order function (Goodchild 1992), and, if the two points are known to be at the centre of a zone, it could be based on a distance decay function. The higher order functions that can be used for the interpolation of values between measured points vary in complexity and scope. Local interpolation models include the use of Inverse Distance Weighting (IDW), natural neighbours and TINs. Global interpolation models include the geostatistical Kriging technique, and the variational spline curves approach (Mitas & Mitasova 1999).

The difference with disaggregation is that the source values represent the *count* of population for the entire zone, and therefore the sum of counts of population in the entire zone will add up to

the original point value. The values for population at points A and B will not remain the same as the input counts but be spread across the zone. As such, interpolation techniques are appropriate for generating weightings to apply during the disaggregation of population counts, but not for disaggregating the data. The result of disaggregation is that the resolution of the data is effectively increased, and the area over which homogeneity is assumed can be decreased to encompass only the areas that are likely to be populated (Martin 1996). To increase the likelihood that population is redistributed to the right places, ancillary data can be used, such as, in the case of residential population, the extent of residential land use (these dasymetric mapping approaches are discussed in further detail in below).

Areal interpolation refers to the process of transferring values from one set of source areal units to a second, different set of target areal units (Goodchild & Lam 1980) which may be smaller than the source units, thereby increasing resolution of the data. It describes a different process to spatial interpolation introduced above (although spatial interpolation methods may be used in the re-distribution process). Several areal interpolation techniques exist that produce layers with irregular discrete regions, such as areal weighting and point interpolation methods described by Langford (2006) and Langford (2013) respectively. Both of these methods assume homogeneity in the source areas, and while they re-distribute population data to alternative geographies, including grids, they are simplistic.

More suitable methods exist for disaggregating population to uniformly sized grid cells which can be made small enough to remove many of the problems associated with the homogeneity assumption in the target polygons, i.e. if a grid cell is made small enough a cell is more likely to be homogeneous and the assumption is more likely to hold true.

The model being developed in this thesis requires techniques for disaggregation rather than interpolation, as the source data are counts of population rather than densities. The small area estimated population counts will be disaggregated and associated with individual locations (rather than grid cells), influenced by statistical ancillary data, akin to dasymetric mapping (described below). Some individual sites will also be assigned their own site-specific measured data.

There are two desirable requirements for a technique to redistribute population. First, the technique must allow use of ancillary data (such as land cover, land use or road networks) in order to guide the disaggregation and restrict the population in the target geographies, thereby allowing for zero population in target features and more accurately reflecting the real world. The use of ancillary data therefore reduces the assumptions of Tobler's first law of geography "everything is related to everything else, but near things are more related than distant things" (Tobler 1970), which do not necessarily apply to population (Martin 2016, p11).

Secondly, the method also needs to preserve the pycnophylactic (volume preserving) property (Tobler 1979), so that, where small area statistics are utilised, people are not removed from or added to the source zones. Several methods meet these criteria.

2.1.3.2 Methods of Spatial Disaggregation

A brief discussion of some of these spatial disaggregation techniques follows. All can utilise ancillary data and all preserve the pycnophylactic property. Firstly, in the network length algorithm, population is evenly distributed along a network such as a road network on the assumption that this represents the residential location of the population being distributed, thereby utilising road length as an ancillary variable. The population are then aggregated up to the target zones (Langford 2013; Xie 1995). Although there is a direct link between the ancillary layer and human activity, there are many issues here, not least the fact that not all roads are lined with houses for their entire length and so population will often be spread too thinly along the route and placed into target zones that they should not be in.

An alternative approach is to use “Surface Volume Integration using Ancillary Data” which uses a proxy for population density, such as schools or bus stops as described by (Zhang & Qiu 2011). Using these ancillary data, there are conflicting reports of whether this method results in better distribution of population, and this may well depend on the geographic characteristics of the study area in question (Langford 2013) so could not be assumed to give consistent results.

Dasymetric mapping involves utilising ancillary data to restrict the redistribution of the population to specific areas. The original technique used by Semenov-Tian-Shansky in 1911 (Petrov 2012) produced a population map showing discrete areas and their population densities and it was a similar technique subsequently used by (Wright 1936). This approach has since been refined using various different techniques resulting in the production of a population surface grid layer and applied at various scales, from local scale (Langford 2006; Langford 2013; Mennis 2003) to continental and global scales (Bhaduri, Bright & Coleman 2005; Bhaduri et al. 2007; Batista e Silva, Gallego & Lavalle 2013). Essentially, the number of people in each grid cell can be calculated by restricting the areas into which they can be distributed, and may use the information provided in the ancillary layer to proportionally distribute individuals, such as in multi-class dasymetric mapping (Mennis 2003). All methods of dasymetric mapping work on the assumption that the ancillary data indicate where people do and do not reside, and, in the case of multi-class dasymetric mapping, at what densities.

Binary dasymetric mapping uses an ancillary layer such as a classified remotely sensed image which indicates residential/non-residential areas to restrict the distribution of population

(Langford & Unwin 1994). In its simplest form (i.e. using land cover rather than land-use as described earlier), this method would exclude open spaces from the target zones, meaning that population could only be distributed amongst the built up areas. If a residential/non-residential classification (i.e. land use) is used, then the population should not be distributed into the public buildings either and this could produce a more representative residential population map.

Multiclass dasymetric mapping not only restricts the space that the population can be distributed within, but also introduces a weighting, based on the population densities believed to exist in different classes. These weightings can be derived in several ways. Through the use of empirical sampling of remotely sensed imagery, Mennis (2003) identified three population density classes: 'High', 'Medium' and 'No'. This was then used as the ancillary data for redistributing population, not just identifying the grid cells into which population is redistributed, but also guiding the proportion of the measured population to be assigned to each cell. Statistical regression can also be used to determine the proportion of population to be assigned to land cover classes. Here, the relationship between population density and land cover is modelled by either regional or global regression analysis (Langford 2006).

For both binary and multiclass dasymetric mapping, land cover derived from classified remotely sensed images is commonly used for the ancillary data, as this provides good geographic coverage, and there has also been a lack of availability other suitable ancillary datasets (Langford 2013).

Some of the dasymetric mapping techniques described above have been used for the generation of large-scale population maps with continental or global extents. Aubrecht et al. (2013) discuss some of these products in some detail, describing the Gridded Population of the World (GPW), which provides disaggregated national or sub-national population data for a 2.5 arc min grid. This is also combined with the Global Rural-Urban Mapping Project (GRUMP) data as an ancillary dataset, and produces a "Gridded Population of the World with Urban Reallocation" with a resolution of 30 arc seconds (approximately 1km).

Another large area population mapping programme is WorldPop (Stevens et al. 2015), which also maps age structures, births, pregnancies, poverty and urban growth. This is intended to amalgamate the smaller, continental based, AfriPop, AsiaPop and AmeriPop projects. The modelling approach uses census data, and settlement maps to generate a dasymetric map. The settlement maps are generated using a variety of local, open ancillary datasets for each country (including remotely sensed and geospatial datasets such as settlement locations, settlement extents, land cover, roads, building maps, health facility locations, satellite nightlights, vegetation, topography, and refugee camps) or using land cover and empirical determination of population

densities per land cover type. The result is a 100m dasymetric map with higher accuracy than other methods (GeoData Institute, University of Southampton n.d.). The data set is full open access and is able to be modified using new or better ancillary layers as they become available.

Two detailed settlement maps are currently in development, both of which use high-resolution satellite imagery and machine learning techniques to generate a high-resolution buildings layer. Firstly, International Earth Science Information Network at Columbia University and Facebook are developing a 5m resolution population dataset with limited extent (initially 20 countries), for the purpose of planning drone networks to provide internet connectivity, but with recognition of its usefulness in emergency planning and response (Facebook 2016). These application areas provide motivation for plans to evenly distribute population from administrative data (censuses) between the identified buildings (Facebook 2016). Secondly, the Global Human Settlement Layer (GHSL) will be up to 10m resolution, and global in extent. Population spatial modelling is cited as an application area for this layer, along with emergency planning and response (Pesaresi 2015), but distributed population will not be part of this product (Pesaresi et al. 2013).

The LandScan world population database is at a comparable resolution and utilises ancillary datasets for multi-dimensional dasymetric mapping, including elevation, slope, night-time lights, land cover, transportation networks and populated places (Aubrecht et al. 2013). Covering a smaller spatial extent (USA only), but at a higher spatial resolution of 3 arc seconds, or approximately 90m (Bhaduri et al. 2007), is LandScan USA which utilises additional ancillary datasets and specifically addresses the temporal dimension.

These building datasets will provide higher resolution ancillary data for dasymetric population models. While these products will mark a significant improvement in spatial resolution of ancillary data, they will not distinguish between residential and non-residential buildings, nor will they model population with any temporal information, other than that implicit in the census data.

These high spatial resolution ancillary data are not yet published, and existing global population products such as LandScan and GPW are all at a coarse spatial resolution unsuitable for many of the emergency planning and response activities described earlier. In most practical settings, it is not possible to consider evacuation plans when population data are available only for a 1km grid. In very dense urban areas, even the availability of a one hectare grid may well be inadequate for such an activity because human activity occurs at the building and sub-building scale. Average new-build residential densities in the United Kingdom (UK) of 42 dwelling units per hectare between 1999 and 2009, and densities of existing housing stock, in places, at over 80 dwelling units per hectare (Commission for Architecture and the Built Environment 2005) highlights the inadequacy of gridded population maps. Commercial addresses have similar density levels. It is

clear that while the knowledge of how many people might be in a hectare is useful, some applications require higher resolution. In order to attain higher spatial resolution that would serve the needs of all emergency planning and response activities, as well as many of the other applications already discussed, the spatial resolution of the data needs to be higher.

The highest spatial resolution for which population data can be conceptualised is that of the individual. Privacy and mobility issues make modelling large populations at this level both inappropriate and impractical, so aggregate data must be used, both within the model, and as the finest resolution for model output. The lowest possible aggregate level for both residential and commercial population is the address, and this is the preferred level for modelling. Addresses are sometimes at the sub-building level. Moving beyond the gridded approach to achieve the highest possible spatial resolution for population allocation therefore requires an address level model.

2.1.4 Temporal Considerations

The approaches to spatial disaggregation already discussed deal only with a snapshot value of population, that is a population measure recorded for a single reference date (e.g. the time of the census) and ignore all other temporal aspects of the data. Population distribution is subject to cyclical (Ahola et al. 2007) and diurnal patterns and ignoring these temporal patterns when utilising the data in many applications is inappropriate. Attempts to calculate ambient, or averaged, population (Sutton, Elvidge & Obremski 2003; Dobson et al. 2000; Bhaduri et al. 2007), or day-night populations (Sleeter & Wood 2006; Freire & Aubrecht 2011) acknowledge this requirement but do not model the detailed temporal variation in population counts.

In line with the common view that geographic phenomena are comprised of space, time and attributes (Langran 1992), it is generally recognised that time and space cannot be separated in the study of geography (Hägerstrand 1975; Pred 1977; Thrift 1977 and later Massey 2016; Peuquet & Duan 1995; Goodchild 2013) and it is possible to draw several general parallels between the two. Firstly, the field-object dichotomy holds true for both space and time (Peuquet, Smith & Brogaard 1998) with clock time representing a field of instants and events being equivalent to objects, such that an instant in time is equivalent to a point in space, and an event with duration is equivalent to a two dimensional object in space. Spatio-temporal phenomena can therefore be viewed from the field or object perspective and the dense field of time can be broken down into discrete objects (Galton 1998). These discrete objects could refer to either instances in time (events) or periods of time (e.g. 15-minute time slices).

Another parallel is that of temporal ordering. The relationships such as “betweenness” and contiguity apply in both space and time, and are absolute (Galton 1998). If an event occurs after a

previous event, and before a following event, it will fall between those two events. That relationship is absolute. This is the same for spatial relationships.

Finally, the language that we use around time and space are similar. We speak of long or short distances and long or short times. We speak of "nearly there" for events in time (e.g. "it's *nearly* my birthday") or space (e.g. "we are *nearly* at the motorway services"). Our expression of the relationships between times and spaces is also similar (e.g. "I'll take the cat to the vet *between* dropping the kids off at school and going to the supermarket" or "my house is *between* the two side roads"). Worboys & Duckham (2004) describe several more such metaphors in use in our language, highlighting our general understanding of the inseparability of space and time.

Space is a three dimensional phenomena that is often modelled in two dimensions representing the surface of the earth within GIS (with attributes, including height, sometimes referred to as an additional half, or third dimension). Time is a one-dimensional variant of space (Worboys & Duckham 2004, p372) , so time can be added to our understanding of space as an additional dimension. Cova & Goodchild (2002) refer to this addition in the field view of the world, but it is equally applicable to the object view, and in the object-field model, it may be possible to have objects separated in both space and time from the location in space and time on the field (*ibid*).

2.1.4.1 Time in Geography

In geography and GIS, time is dealt with in accordance with the way it is experienced by humans, that is as though it progresses in a linear fashion (Worboys & Duckham 2004, p372), always moving forwards. A working conceptualisation of how time works does not require a detailed understanding of how space-time flows; it only needs to work for modelling purposes, in the same way that we do not need to have a detailed understanding of space-time in order to model buildings as polygons in Euclidean space within a GIS.

Given that time and space cannot be separated in geographic study, they must be modelled in an integrated manner, studying the two together rather than studying them separately and then trying to integrate the results. This is expressed by Goodchild (2013) as a requirement for geography and GIScience to be the science of integration rather than the integration of science, on the basis that space and time together provide the context for integrating knowledge from different disciplines, and that representation of the variation in this context is required for that integration.

As suggested earlier, when thinking about time, the continuous field can be broken down into chunks or intervals, defined by their duration and with sharply defined start and finish times,

which are instants. In terms of handling temporal data, instants can be simply ordered, although it is more complex to order intervals as they may overlap.

In terms of temporal data, it is important to understand the scale and granularity of measuring, recording and modelling devices i.e. how short a time period can be measured, recorded or modelled, and how short is required for the job at hand (Galton 1998). The granularity of data recording may vary and this has the potential to influence how data can be utilised. For example, Charles-Edwards & Bell (2013), used a variety of data collection methods for estimating temporal patterns of university campus population. The recording methods used varied in that manual recording produced a tally for 15-minute intervals, and electronic sensors were used for continuous recording, the data from which was then aggregated to 15-minute intervals. In this study, the chosen interval revealed a change event (increases/decreases in population) in the 15 minutes before and after the hour, i.e. for half an hour, each hour. Data acquisition at the finest granularity allows for more flexible use of the data and the sensor data, which is continuously recorded and can therefore be aggregated after the recording with a temporal zoom-out as described by Pultar et al. (2010). This could potentially have revealed whether this change occurs for less than half of the time if it were aggregated to shorter time periods. In addition, the selection of interval length can have implications for which time interval events are assigned to. Greger (2014) used a temporal interval of an hour to model building occupancy, but it is clear that a person counted as entering a building at 8:59 will be counted as being in the building for the entire 8:00-9:00 period, and the author acknowledges the need for recording temporally finer-grained data.

This leads us to the Modifiable Temporal Unit Problem (MTUP, the temporal equivalent of the MAUP) which refers to issues associated with the temporal scale that can change the outcome of geographic analysis (Çöltekin et al. 2011). It is recognised that often, temporal data are selected based on their availability, but at the same time, selecting the appropriate temporal scale is essential for successful study of phenomena. It is possible to miss some events or cycles if inappropriate temporal scales are used (Jacquez 2011), or to incorrectly assume that two objects coexist at the same time (Langran 1992). Cheng & Adepeju (2014) detail factors influencing these effects as the scales of temporal aggregation (equivalent to spatial scale, or resolution), segmentation or, how the temporal dimension is divided (equivalent to spatial zoning effects) and boundary (of the temporal extent) effects and highlight the impact in the identification of space-time clusters, when different temporal scales, zones and boundaries are used.

This concept has been further developed by Jacquez (2011), to address both spatial and temporal scales, as the intersection of MAUP and MTUP as the Modifiable Spatio-Temporal Unit Problem

(MSTUP). This describes the situation where analysis results may be affected by the scale of either the spatial or the temporal data (Martin, Cockings & Leung 2015). Where the MAUP is at least partially mitigated by using data at the finest possible spatial resolution, and the MTUP is mitigated by using data at the finest possible temporal resolution, the MSTUP would require data at both finest spatial and finest temporal detail to mitigate its effects.

2.1.4.2 Time Geography

The influential Swedish geographer Torsten Hägerstrand suggested that insights into the main challenge of geography, to study "collateral processes within bounded regions", can only be understood by studying individuals and their daily spatial range (Hägerstrand 1976). This range influences the interactions (and processes) between individuals, whether these individuals are human or not, animate or inanimate. While focusing on local connections between individuals, this must be taken from a holistic perspective, acknowledging the transformative influence that modifying individual connections may have on other connections (*ibid*). The result is an activity perspective of geography, with activities constrained by space and time. In this view of the world, space and time are considered as resources. Human movement in space is constrained by available time, and the time available is limited by the size of spaces that must be covered in the course of completing activities. Space and time are therefore inextricably linked. This is described variously in terms of a space-time path which describes the spatio-temporal extent of these activities for an individual, and is enclosed by a space-time cube, or prism, and as an "aquarium", which describes the interactions between individuals (Hägerstrand 1970; Thrift 1977). In these, space is represented in two dimensions, and time is represented in the third dimension (Langran 1992). The space-time path provides a tool for the analysis of the interactions between individuals and their environment. This model is both spatially and temporally scale neutral as it works just as well at the scale of individuals on a minute by minute basis as it does for communities on an annual, or longer basis (Pred 1977; Thrift 1977). This time geography provides an alternative to the spatial cross-sectional view of the flow of events that is often used in geography (Hägerstrand 1973; Pred 1977) by dealing with time and space together. It reveals relations that can only beseen when individuals are studied in their context, that cannot be seen when they are studied in isolation (Lenntorp 1999). One of the principles of time geography is that every event has both temporal and spatial attributes. Events occur at stations (or domains) and where several individuals meet, this is termed a bundle (Pred 1977).

Time-geography has therefore been around for several decades; it was conceived in the mid-1960s and its application methods have been developed since then, with its spread influenced by the uptake of Hägerstrand's ideas by both Pred and Thrift (Pred 1977; Thrift 1977; Lenntorp

1999). Yet time-geography has not been much developed conceptually since the 1970s, despite the many disciplines to which the time-geographic approach could be applied. Although there have been some conceptual additions related to amalgamation (*ibid*), the core principles of time geography remain essentially the same and modelling remains restricted to only small numbers of individuals at a time. Yet there have been many practical applications of Hägerstrand's time-geography. Pred (1977) describes many of these, citing research in migration, urban growth, socio-technical ecology and traditional research themes of human geography, and describing how the approach has been used for accessibility studies in city and regional planning.

One of the reasons for the stalled development of the time geographic concepts could be the availability of detailed data. Kwan (2004) suggests that the combination of the high computational demand of time-geographic modelling, a lack of suitable data, analytical tools and visualisation issues (both cognitive and technical difficulty in visualising such data) have limited the development and implementation of time geography. Space Time Activity (STA) data that reflect the daily movements of large populations are now becoming available in the form of "big data" (Batty 2013; Miller 2005) sourced from GPS devices, mobile phones, Location Based Services (LBS) and systems such as Oyster cards used on public transportation in London. Some research has begun to utilise this, such as that by Çolak, Lima & González (2016) who used mobile phone traces to model transportation flows in urban areas and assess the potential of changing route choices to reduce congestion levels, an application area with the potential for short term urban planning suggested by Batty (2013). Another example is work by Huang & Wong (2015) who have used geo-tagged Twitter data to visualise space-time paths of individuals, revealing cyclical activity patterns that are not indicated in a one or two day diary approach to data collection, despite the relatively incomplete nature of the collected data. These time geography applications employ new data sources, better computing power and improved visualisation techniques.

The concept of the space-time prism has more recently been extended by Yu & Shaw (2008) to incorporate virtual spaces, and therefore to contract the space that human activity needs to cover in available times for certain interactions. Using virtual space frees up time for an individual to engage in other activities, such as travel to places previously not accessible within the available time resource, enabling more physical interaction between individuals. Miller (2005) highlights this shift in accessibility to stations and bundles through the development of telecommunications systems (and improved transportation networks) suggesting that there is increasing dissociation between places and activities and calling for increased efforts to develop individual-based models of individual activity within GIS. This would answer one criticism of time geography that it doesn't consider the individual as an acting subject (Hallin 1991). There is clearly potential to develop such individual-based activity models from new big data sources.

2.1.4.3 Time in GIS

The individual-based approaches described above deal with modelling the motion of active individuals. They constitute just one method of modelling time in GIS. Miller (2005) describes several of these people-based STA models including Shaw & Wang (2000), Shaw, Bombom & Yu (2008) and Wang & Cheng (2001). These are extensively covered by diary-based methods and time use surveys. Generic tools for analysing tracking data required for such approaches have, however, yet to be developed (Goodchild 2013).

Approaches to handling time in GIS take a place-based approach, modelling the places, rather than the individuals. These methods are classified in different ways by different authors, but they can be summarised into three main types, all of which are capable of incorporating state changes (i.e. changes in spatial or non-spatial attributes) caused by events, which lead to new versions of phenomena coming into existence (Langran 1992).

Firstly, a series of snapshots in sequence (Langran 1992; Worboys & Duckham 2004; Goodchild 2013), which is applicable to remotely sensed images, or for temporal sequences of polygon coverages. These snapshots are not integrated, so comparisons are required to evaluate for change. This approach can be adequately used in current GIS.

Secondly, modelling objects with changes, which, conceptually, can be achieved in several ways including base state with amendments superimposed (Langran 1992), where a set of polygons represents a start point and each change is stored as a new polygon, without changing the original polygons. Alternatively, the space-time composite can be used. This can be conceptualised as a three-dimensional space-time cube flattened to two dimensions in which new states become new objects (*ibid*). Each change generates a new update to existing polygons; new polygons have an attribute history describing the polygons that they used to be a part of. Object lifelines is another approach to modelling polygons with changes, in which objects can be created, changed or destroyed, and each of these changes is explicitly modelled as attributes (Worboys & Duckham 2004). The change types that can be recorded include creation and destruction, disappearance & reappearance, spatial change (shape, size, position), aspatial change (attributes), transmission, fission and fusion (clone, split, merge), mereological change (part-of relationships) and typological change (*ibid*). These two approaches add time as an additional dimension to the two spatial dimensions traditionally modelled in GIS. Time can, however, be added as two additional dimensions, one for database time, when the feature was encoded into the database, and one for event time when the feature was, or will be, in existence (Worboys 1994).

The final approach to handling time is a true spatio-temporal model in which space and time are integrated. In these models, events, actions and processes are explicitly modelled (Worboys & Duckham 2004; Goodchild 2013) as entities along with the objects that they act upon. Each event has a record along with its space-time location. In this model, events are linked spatially.

Spatio-temporal modelling should allow us to execute queries that answer questions related to where, when and what type a change is, as well as the rate and periodicity of the change. This should allow us to identify temporal patterns, trends and processes of change and potentially the causes of change (Langran 1992). The complexity of the spatio-temporal queries that can be answered increases with each of the modelling approaches, with the final, event based model allowing questions about events (i.e. change) to be answered.

Goodchild (2013) outlines several spatio-temporal data types. Firstly, tracking of objects, including humans, allowing questions about behaviour patterns to be answered. Temporal sequences of snapshots e.g. remote sensing imagery to detect change. Temporal sequences of polygon coverages, such as census tracts are non-overlapping polygons that change over time. Cellular-automata model space-time processes on a fixed raster (e.g. modelling land use transitions with urban growth) and agent-based models represent agents as discrete objects. Events and transactions modelled as individual records with their space-time locations and finally, multidimensional data such as environmental data intensively sampled through time at a set of fixed points. This variety in the types of space-time data make a single space-time GIS unlikely (*ibid*).

2.1.4.4 Example Spatio-Temporal Models

The above discussion has focused on conceptual work that uses a place-based approach rather than an individual-based approach. There are plenty of empirical works that provide examples of real implementations of these conceptualisations.

The Event Based Spatio-Temporal Data Model (ESTDM) stores time-stamped events for a specific thematic domain, recording only the changes in raster pixel values, using specific approaches to reduce the number of values that needs to be stored for each event (Peuquet & Duan 1995). This model allows spatio-temporal queries of the data, such as calculating change between two times, identifying locations that changed between two times or retrieving locations that have changed to a specific value at a specific time, but is developed for raster data.

The STA models developed by Shaw & Wang (2000), Shaw, Bombom & Yu (2008) and Wang & Cheng (2001) that were introduced in Section 2.1.4.3 use individual-centred (rather than location-

centred) modelling. These models use individual diary data, or automatically recorded data such as GPS tracks or credit card records. They generate space-time paths of individuals over time, as stepwise changes in two types of activity: staying at, or travelling between locations, based on the purpose of the activity. It is worth noting that an activity can only take place at the appropriate location for that activity, where the facilities, or the correct environment exist (Wang & Cheng, 2001). Practically, these model a set of event types in a relational database. These models also allow spatio-temporal analyses, such as spatio-temporal cluster analysis.

The three domain model described by Yuan (1994) uses a location-centred (rather than individual centred) and entity-centred snapshot model. The semantic, or attribute domain models atemporal and aspatial properties of the phenomena. The temporal domain contains temporal objects (points and lines), representing instance time and time intervals, as described in Section 2.1.4.1 and the spatial domain contains topological objects (points, lines, polygons, cells). In this model, all three domains can be maintained and managed separately and independently of each other. Using this approach, a place concept would be modelled in the semantic domain, linked to a spatial object in the spatial domain and the relevant temporal objects in the temporal domain. The general concepts can be applied to multiple temporal objects and to multiple spatial objects.

Pultar et al. (2010) propose (and implement) a spatio-temporal data model based on four dimensional Space Time Points (STPs) with their associated attributes (from the geo-atom theory of Goodchild, Yuan & Cova (2007), which can be used to generate either objects or fields (features) and which are grouped or collected into themes. In this model, space, time and attribute domains are treated as inseparable (as already discussed) which allows for query of one or two of these domains while holding the other(s) constant. Object interactions and topology can be queried to describe space-time relations between objects. For temporal query, this system (EDGIS) needs to step through all time-steps.

Each of these few examples of spatio-temporal models within GIS has their limitations, either in the types of data required for their application, or in the types of queries made possible by the modelling approach. Some of these models (STA, STP) are intended for modelling the interaction of geographic phenomena over time. Understanding the interactions are not necessarily the main purpose of the model.

These true spatio-temporal models such as the STA models, are only required if change is to be modelled (Worboys & Duckham 2004). Models can instead represent space and time without being true spatio-temporal models.

2.1.5 The Need for Highly Detailed Spatio-Temporal Population Models

Humans partake in many activities, including work, healthcare, education, transportation, residential and leisure activities. The start time and duration of these activities varies from person to person, depending on their circumstances, and therefore are best suited to modelling using time-geographic methods, but it is useful to group these activities in terms of the spatial and temporal scales of the activities, as described in Figure 2. For example, the activity “holiday” could be as short as a day or as long as a season, although typically this would only occur for a matter of weeks. It can be an activity that is carried out at home, but which frequently involves travel, sometimes locally to the nearest city, or rural area, and sometimes regionally, nationally, internationally or even globally. This describes the displacement of people when engaged in the “holiday” activity, i.e. their travel that is related to getting to the destination. During a holiday, individuals may stay in a single location, or indeed the holiday itself may be centred on the travel, hence the spatial and temporal extent of the “holiday” feature in Figure 2.

Taking an example of a family with children, weekday activities may occur at a variety of spatial scales. Travel to education or day care may be at the neighbourhood scale for young children, or, for older children, could extend to the nearest city. Once at an educational establishment, the individuals are effectively stationary when viewed at the neighbourhood level (as with a gridded population model), but not when considered at the building level. Even very young children move from one building to another during the course of a school day. For adults in a family, the “work” activity involves travel to locations, ranging from no travel if working at home, up to any scale (with long commutes involved) but once the travel has been completed, the “work” itself is often, but not always (as in the case of drivers, field engineers etc.), static in nature.

The temporal scale of human activity is also an important consideration. Many of the activities described in Figure 2, and other activities that people are engaged in are cyclical in nature (Ahola et al. 2007). The “work” and “education” activities tend to follow the same pattern, with individuals travelling at the same time of day, going to the same location and returning at roughly the same time (depending on their occupation and the industry in which they work). There are different patterns of activity at weekends and during holiday periods. Some human activities, however, may not fit into a cyclical pattern at all, such as a one-time visit to a tourist attraction. There may also be several activities occurring at a single location, within a building or a collection of buildings. A good example to demonstrate this is a university campus. These contain people engaged in a variety of occupations, including work, education, cultural and leisure activities (Charles-Edwards & Bell 2013). A campus may also include some residential population as halls of residence are sometimes present on-campus, and some population engaged in retail activity.

Hospitals are another good example, as they include populations who are workers, patients or visitors. It is important that all of these activities be considered in modelling population at the site.

The result is a constant shift in population throughout the day, highlighting not only the need for a highly detailed spatio-temporal population model, but one that can integrate different types of measured or estimated population data: those that relate to the day in-day out activities, and those that relate to the one-off activities. These will be used to address the limitations of existing residential based population models, or limited temporal scale models, which do not adequately reflect the variability in population distribution and may not allow cyclical activities to be identified.

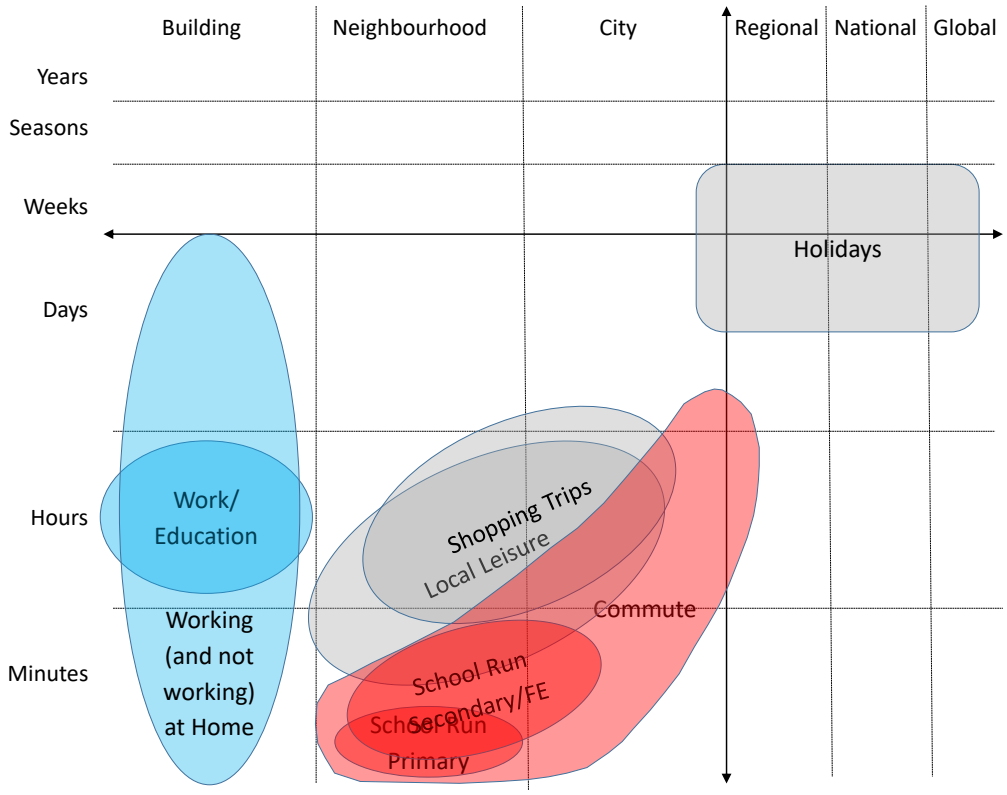


Figure 2: Spatial and Temporal scales of some common human activities. Transportation centred activities are displayed in red, with the activities that are at destinations displayed in blue or grey.

This may be achieved by modelling both where people are when engaged with the cyclical patterns of human movement (using residential and working populations, and transport data), as well as where they may be for non-cyclical activities (using specific site based data, such as expected visitor numbers at a tourist attraction). These data, when combined, could provide more spatially and temporally detailed population estimates.

This may begin to meet the need for a population model that can supply both high spatial *and* high temporal resolution detail. Population data are used as the starting point for emergency planning and response activities (Ahola et al. 2007), and also as the basis for calculating per population statistics such as crime rates (Malleson & Andresen 2016; Harper & Mayhew 2012) or health statistics (Harper & Mayhew 2012). Any improvement in the distribution of these data has the potential to lead to improvements for these application areas, as well as many of the application areas already discussed. It is possible to model incidents such as the extent of a toxic plume (McPherson & Brown 2004), or a flood (Smith, Martin & Cockings 2012), either in advance for planning purposes (Ahola et al. 2007; Aubrecht et al. 2013; Smith, Martin & Cockings 2016), or at the onset of an event, for response. Current methods of producing population data are lacking in flexibility in the spatial and temporal detail required to reflect the constant shift in populations, and for integration with the hazard model data related to these two aspects of planning and response. The examples of highly detailed spatio-temporal population models discussed below represent models in development or for very small areas, rather than models that can produce data for the relevant area in a timely manner. The result is that these data are often inadequate, if we are looking to evacuate people, or identify numbers of people likely to be seeking medical attention, or for rescue operations, or even for getting a fire engine to the right building and understanding the requirements that may be found there.

In this research, the purpose is to model where population are at different times. In the modelling framework presented in this thesis, this cannot be done while ignoring the reasons for the population being where they are. People are present at different types of places when engaged in different activities. They work in their workplace, and reside in their residences. They take leisure time at a variety of places and attend school at educational establishments. This model uses evidence from the data regarding how many people are engaged in those activities, and estimates how many people are present in the locations where those activities take place. The temporal aspects of the modelling framework are data driven, rather than interaction driven, but the data do not fit in to any of the temporal data categories already outlined. They are temporally variable attributes that can be applied to any number of spatial features, and are kept separate from the spatial data in all but a database linkage. The three-domain model is the basis for some population models (Ahola et al. 2007; Martin, Cockings & Leung 2015) and is the closest fit model to the modelling framework developed here. The model is centred on places rather than people, and population estimates are created for specific times, generating a series of snapshots.

2.1.6 Spatio-Temporal Population Models

Despite the previously discussed valuable techniques to re-distribute residential census data onto smaller spatial units, they remain inadequate for emergency planning and response activities because detail in the spatial scale is still insufficient, and the temporal detail that is required is absent. While information is gathered on place of work in the census, the outputs are related only to night-time (i.e. residential) and working daytime populations. Censuses are also usually carried out infrequently. In GB, the ONS publish mid-year estimates, calculated using a ratio change method (ONS 2017) and published annually, which ameliorates this issue.

In the emergency response application area, incidents arising from, or causing, hazards can occur at any time of day or night. This combined with the dynamic nature of population means we also need to model where people are at different times of day as they go about their different activities, thereby adding the temporal dimension to the spatial population model.

There has been some work to estimate ambient population that attempts to account for day and night-time population in one output (Sutton, Elvidge & Obremski 2003), providing a "temporally averaged" population density measure, but this is only suitable for larger scale applications that do not require estimates of populations at specific times. The previously discussed LandScan products represent ambient population (Bhaduri et al. 2007).

It has long been recognised that population is not static and that there are significant shifts of population towards the centres of metropolitan areas during the day (Foley 1952). As a result, there have been several studies looking at producing daytime population. An early investigation into potential methods by Schmitt (1956) examined the use of proxy datasets to estimate daytime populations. Much later, Sleeter & Wood (2006) utilised employee databases, workplace and educational data to indicate daytime populations, and distributed it using land use as ancillary data for dasymetric mapping.

Aubrecht et al. (2013) have expanded this to daytime and night-time population distributions, also recognising the need for very high spatial resolution down to the (sub-)building level. They used very high-resolution remotely sensed data (satellite and airborne laser scanning) to generate a land cover classification, and then combined this with postal addresses and census data to generate sub-building level population for daytime and night-time.

Another example, related to emergency planning and response comes from McPherson & Brown (2004) who utilised disaggregated residential data (using dasymetric mapping) and journey to work data alongside business directories to identify daytime and night-time populations exposed to a hypothetical toxic plume. This work has been further developed to assess the impact to

healthcare providers (McPherson et al. 2006) by utilising the journey to work data to demonstrate that the night-time or residential spatial extent of populations exposed during the day is much greater than if static population maps alone are used. While this day-night model clearly has benefit in modelling such scenarios, the output from these approaches supplies only two time slices. In reality, different sub-populations follow different daily routines, for example, a school pupil will be at school and home at different times to an employee, and again at different times to non-working populations. Associated leisure activities will also follow different temporal patterns. As a result, a finer temporal resolution is required that also accounts for the non-residential and non-working population.

All of the examples described above provide coarse, pre-defined temporal resolutions. In improving the spatial resolution of published population counts, the finest logical spatial resolution is the address level, and the most appropriate temporal resolution will meet the requirement that appropriately time-sliced data be available immediately as required. Addresses can be modelled in GIS as cartographic objects (points, lines and polygons). One approach to addressing the need for improved temporal and spatial resolution is therefore to model human activity at the address level.

Two very detailed site-specific population models have already been discussed, that is the university campus population model of Charles-Edwards & Bell (2013) and the airports and cruise terminal models of Jochem et al. (2013). Both of these models could inform a more general model, but scaling them up to cover a variety of different building functions within a single model would not currently be practical due to the large number of different types of sites. A more generic approach to achieving this level of granularity is required that can be applied to much larger numbers of places.

Greger (2014) has also modelled population at the building level in an area of Tokyo by estimating proportions of floor space within buildings that are used for different types of human activity (identified as residential or non-residential, from address data). Transient populations (i.e. customers or visitors in retail and leisure activities) are not included in this estimation. Only five non-residential usage categories are identified which limits the potential for accounting for different types of non-residential temporal variations in occupancy. Movement data was obtained through questionnaires. In this model, time is subdivided into one-hour slices for weekdays only.

Ahola et al. (2007), in modelling Helsinki's sub-populations (broken down by age) for the purpose of risk assessment and damage analysis produced a building level model with 14 time slices, applied to individual buildings, but this model is restricted to Helsinki. Zhang, Sunila & Virrantaus (2010) have extended this by developing a system for providing building level population data for

17 different time slices, for the purpose of crisis management in all of Finland. This object-oriented model takes population within buildings into account by classifying the buildings into 11 different spatio-temporal objects. It is not clear whether individual buildings within these objects have corresponding individual temporal characteristics assigned to them, but each of the different spatial objects does. This is a step further than the day-night models described earlier and is therefore of greater use to emergency planners and responders. Each of these two examples model populations only inside buildings.

In general, when the spatial resolution of the model is increased to individual building level, there tends to be a reliance on methods that are not generalizable (e.g. survey of individual buildings or detailed attribution using manual searches of the internet or directories). Consequently, the spatial extent that can be covered by these models is reduced as the data gathering exercise becomes unfeasible. This restriction can be overcome by removing the need for manual searches, and discovering the function of the building from the available data, in the case of the model presented here, that is address data with functional class.

2.1.6.1 National Population Database

There are several examples of population modelling at the building level, and at finer temporal scales than the day-night approach. Some of the different activities that people are likely to be engaged in are reflected in the NPD developed by Smith & Fairburn (2008). This geodatabase incorporates residential daytime and night-time population (using addresses and census small area estimates) as well as populations in the road transport system, retail areas, workplace and leisure facilities, with the aim of aiding the HSE to understand population at risk from incidents at hazardous facilities. The method for developing this involved applying generic population multipliers to the buildings, transport and land uses to estimate population density. The input data sources are at variety of scales, although the output population coverage is national, covering all of GB. Due to various constraints, the population estimates were based on potential occupation rather than counts (HSE 2017).

These data are at the building level, but the temporal scales of the different layers within the NPD do not reflect the complexity of human activities. For example, residential population is represented as daytime and night-time and the transport population is represented as an average daily, peak and maximum capacity value. The number of people is not, therefore, directly and explicitly linked to the time of day. Other limitations of this model include the limited number of population (activity) types. The locally sourced nature of the datasets deliberately highlights only major features. This database does not provide a default occupation level for buildings of

different types and relies on availability of detailed and specific source data. It does not model the population, but amalgamates and aggregates disparate data sources to give a well-informed picture, but is not capable of estimating population at different times.

2.1.6.2 Population 24/7

Adaptive Kernel Density Estimation is a disaggregation technique described by Martin (1996) and used in the SurfaceBuilder247 application (Martin, Cockings & Leung 2009). It requires a point dataset with associated population counts as the input. No associated source areas are required for this unconstrained method (although it is possible to use them as described below). It involves visiting each point in turn and estimating the most appropriate kernel size determined by the local inter-point distances. This kernel size is then used for calculating a distance decay function for the point. A weighting for all pixels surrounding the data points is then calculated. An assumption is made that the pixels closer to the centroid are more likely to contain the residential population than those further away and they are therefore given a greater weighting. Population is distributed according to these calculated weightings.

It is possible to constrain the population distribution so that the pycnophylactic (volume preserving) property is maintained. This constraint can be based on the areas associated with the source point data, but it is also possible to introduce ancillary data such as land use to further constrain the redistribution of population to residential areas, making this a dasymetric approach. It is worth noting that the area searched to calculate inter-point distances is user-specified and that this distance has a significant influence on the output population layer. As a result, and in common with many other methods of population disaggregation that have input parameters, this method cannot be considered to produce a definitive population layer (Goodchild, Anselin & Deichmann 1993). An exception would be where the parameters are derived from statistical analysis of the input layers alone (although this approach is subject to its own issues).

The Population 24/7 project at the University of Southampton (Martin, Cockings & Leung 2015) has developed this model into a three-domain modelling framework that enables fine spatial resolution modelling at any temporal scale by using continuous temporal data. This has not been applied at the building level, but has the potential to be. This model uses the SurfaceBuilder approach just described, and can be applied to large geographic areas. It re-distributes population in different subgroups between source and destination containers while preserving the volume of the population at all different levels in a hierarchy.

Population sub-group counts in source centroids are re-distributed amongst destination centroids based on time profiles of the destinations, and are disaggregated across the areas associated with

the destination centroids. The model utilises a background mask that represents the inverse of measured population data and transport networks (derived from the relative densities of road traffic from the national transport model). The background mask is a weighting layer and includes rasterised primary routes and calculated average annual daily flows of traffic, along with offsets in 17 time periods over a week.

This model has been used to produce population estimates at different times of day to assess vulnerable, at risk populations for flood (Smith, Martin & Cockings 2016) and in response to a hypothetical dam failure (Smith et al. 2015), and for assessing populations vulnerable to radiological hazards (Martin 2016). This model is transferrable outside of the UK, and has also been applied in Italy for integrated risk assessment (Renner et al. 2018). It is extensible in both time and space resolutions.

This approach, in common with many other spatio-temporal population models, involves best spatial scales approximating to the postcode level (averaging about 15 addresses). Sub-postcode geographies cannot be estimated using this model, although this is constraint of the available data, rather than the model. It does incorporate continuous time, so the time for the population estimation is specified at the point of analysis, and not restricted to specific times determined by data. The Population 24/7 model is very good at volume preserving, as known numbers of people taken from the census small area statistics are distributed to new destinations. The destination is determined by the time profile of all the destinations in the area.

The research presented in this thesis is a response to the need to drill down further into the spatial resolution, and apply these continuous temporal scales to a wide variety of building functions. None of the models described in this review can provide all of these capabilities.

2.1.7 Scale Considerations

Several scale issues need to be considered in building a population model. These include the scale of data collection and storage, the scale at which the model handles the data and the scale of the model output.

The first issue is the temporal granularity of data collection, that is whether data are recorded as events at an instant of time (the smallest possible discretization of the field), or for a pre-defined temporal discretization of the temporal field (as either a complete, or intermittent temporal coverage). Continuous recording may provide a value at the finest resolution of these time slices, or may record the times that events (such as a change in value) take place. Consider the data collection issues related to the data used by Charles-Edwards & Bell (2013) as already discussed.

These data represent individuals passing a point rather than counts of people at a site. In this case, the data were collected in either pre-defined 15-minute time blocks (manually) or as events (via sensor). One of the collection techniques (aggregating to 15-minute time blocks) prevents storage at the finest possible temporal resolution and limits the potential uses of the data. The aggregation and subsequent storage of the "continuously" recorded sensor data has the same limiting impact.

The second issue is the spatial granularity of data recording and storage. The finest possible granularity for recording of population data is that of the individual, but it may be recorded at the household, or address level. The issues surrounding the spatial scales therefore relate to whether the data are recorded at individual, household or address level and then aggregated up to small areas, as in the census, or whether spatially fine-grained data are available, such as number of visitors to individual retail stores, leisure facilities or other addresses.

Taking both of these aspects of data collection and storage together, it is well recognised that data should be recorded and stored at the finest possible resolution (Langran 1992), and this applies to both the spatial and temporal resolution of the data which cannot be separated in the same way that space and time cannot be separated (Hägerstrand 1976). Increasingly, spatio-temporal tracking data (especially from sensors) comes in the form of complete data with no need for sampling in the data collection (Batty 2013): increasing data storage capabilities and improved data analysis techniques mean that making use of high spatial and temporal resolution data is becoming possible. The importance of the temporal resolution is highlighted when attempting to identify cyclical phenomena, for which the temporal resolution is key (Jacquez 2011), as too coarse a resolution means that certain phenomena may be completely missed. This suggests that the higher the temporal resolution of data, the better. Higher temporal resolution also allows for temporal zooming, so that phenomena operating at different temporal scales can be identified. In terms of the data's spatial scales, the ability to categorise phenomena is scale dependent (Smith & Mark 1998). For example, in terms of the functions of places, at the correct scale, a high street contains a pedestrian (and therefore in-transit) population whereas the shops that line the high street have a retail function, but at larger scales, the high street and the shops may be categorised together as a single retail function.

The spatial and temporal scales that a model can handle are also relevant. Time-geography is recognised as scale independent (Pred 1977), meaning that there are no assumptions regarding scale inherent in the time-geographic models and they can be applied to different spatial and temporal scales equally well. According to Langran (1992), temporal objects can be treated irrespective of scale, so a model does not need to be adjusted to work at different temporal

scales, but many models may make assumptions regarding spatial scale meaning that they are most effective at a specific range of scales. The scales of model output are also important. For instance, in a spatio-temporal population model, it may be necessary to deliver data at the highest spatial resolution, such as buildings, or small pixels, or as aggregated data (the extent of which depending on application). Ideally a population model would be able to handle both.

The final question that needs to be considered here is what the data used in a model actually represents in terms of whether the population measure indicates the number of people inside a place or whether it relates to the number of people passing through a point. Census and administrative data such as patient registration data will represent amalgamated residential population counts, and therefore reflect the number of people in a small area or place during night-time. Sensor data however, may reflect the number of people passing through a point at different times of the day, but not the number of people within the place. This is especially true of transport data such as road traffic. For public transport usage data such as Oyster cards used in London, the data generated simply inform number of cards used to check in to and out of the system, but once in the transportation system these are not trackable and therefore do not tell us where the individuals using the cards actually are (Batty 2013). These sensor data could sometimes represent the number of people in a place; such would be the case with electronic entry cards to office spaces that record whether an individual is actually inside the place. This is an important distinction as it means that it is not always possible to count how many people are inside a place. In addition, counting numbers in and numbers out does not inform us as to how many were inside the place before counting began.

Whether the data, model and output capabilities are suitable will depend on the application.

2.2 Ontologies

In GIS, spatial relationships are represented quantitatively. They utilise coordinate systems and represent spatial objects numerically. Ontologies can provide an alternative, qualitative representation of spatial knowledge and information in computer systems that is more in line with human spatial cognition (Renz 2001) and is a deductive system, based on reasoning (Agarwal 2005).

In the philosophical domain, an ontology is defined as "the study and description of 'being', or that which can be said to exist in the world" (Gregory et al. 2011). In this domain, ontology is therefore concerned with describing *things*, and the relationships between *things* in such a way as to be concise and to reduce ambiguity associated with the often imprecise natural language used to describe things (as discussed in Section 2.1.1) . As such, clarity is one objective of developing an ontology. Ultimately, the aim is to represent a shared understanding of a domain.

In computer science another aim is to facilitate re-use of data and information gathered on the basis of different conceptualisations (Smith & Mark 2003; Agarwal 2005) as can happen in interdisciplinary research, for instance, where the same domain is approached from within different disciplines. To this end, an ontology is here defined as a "specification of a conceptualization**"** (Gruber 1993)**,** using "formalized vocabularies of terms… shared by a community of users" (W3C OWL Working Group 2012a)**.** This amounts to a formal description of the objects, concepts and other entities and the relationships between them so that there is no confusion about what is being modelled. By necessity, they are designed for a particular domain (Gruber 1993). We can therefore say that an ontology is an explicit specification of an abstract, simplified view of the world (Gruber 1995) that we wish to represent a specific domain of interest (Davies, Studer & Warren 2006).

That ontologies are formal specifications means that they permit reasoning by computer, which enables inferences to be made.

From the information science perspective developing an ontology is a pragmatic exercise and is achieved by formalising our conceptualisations of the world using logic (Smith & Mark 2003). For practical purposes, this development requires that we create "a statement of the necessary and sufficient conditions for something to be a particular kind of entity within a given domain" (Peuquet, Smith & Brogaard 1998). Not all attributes of the things being modelled need to be defined, but those attributes that make a class different from the other classes in the hierarchy must be modelled.

Qualities and relations can be added to the things that we are modelling (Peuquet, Smith & Brogaard 1998). As such, objects defined within an ontology will be defined by their properties, their class hierarchy and their relations, or interconnectivity (Raper 2005). The classes that describe *things* or objects that are described in an ontology can represent anything: they can be bona-fide or fiat (Smith & Mark 1998), as discussed in Section 2.1.1., or they could be activities or events. All of these *things* can be described as classes in a class hierarchy with defined relationships between them, and processes acting on them.

As such, an ontology is roughly equivalent to the schema in a relational database, including the formal definition of classes, which is comprised of a class hierarchy and clear statements that define which attributes and relationships can be present between instances of the different classes.

Ontologies have a close connection with knowledge representation and description logic and they are a long-standing approach used in the knowledge engineering community.

2.2.1 Ontologies for Spatio-Temporal Analysis

There are some particular issues for creating a spatio-temporal ontology. Geography suffers from a lack of commonly accepted terms because it is multidisciplinary (Agarwal 2005) and because there is no one methodology for data collection or naming of geo-phenomena (Raper 2005). This causes difficulties in agreement of terms including those in relation to definitions of classification of human activity. Development of ontologies can help with this issue by adding clarity to the terminology and definitions and providing a mechanism for alternative vocabularies to be used.

Geography also suffers from inherent vagueness in many of its definitions. Fisher, Wood and Cheng (2004), provide the definition of the extent of a mountain as an example. This vagueness is particularly present for features represented in the field view of the world. Peuquet, Smith and Brogaard (1998) provide a discussion on the considerations and possibilities for an ontology of fields, but without a solution. They point out that this is partly because of the gradation that is involved in the conceptual objects that are artificially created out of fields where “relatively stable spatio-temporal clusters”, or things (such as mountains), exist (*ibid*).

A related issue in spatio-temporal data is that of scale, which affects the categorisation and classification of phenomena (Agarwal 2005). This fuzziness is not particularly compatible with (but also highlights the need for) the explicit definitions required in an ontology and needs to be resolved.

One example of these specific issues is in the context of activity modelling: Miller (2005) describes two human activity classifications, the first including family, work, shopping, recreation and socialising, and the second including production, education, shopping, socializing, community activities, recreation, entertainment, worship and political behaviour. Miller points out that an unlimited number of classifications exist, but also that these activities occur at a few geographic locations and for limited temporal durations. Clarifying the classification is related to semantics and is part of the business of ontology: to clearly state the differences that make *things* distinct. The relationships between objects and activities would include, for instance, which activities tend to occur at a particular object type, such as a building, and at what times.

Ontologies provide several benefits in spatial and temporal analysis over using the relational database approach of most GIS. For instance, in the modelling of spatio-temporal population at the address level, based on functional classifications of all addresses within the buildings, the containment of addresses within buildings does not need to be spatially modelled, as long as it is represented qualitatively. This allows features to be linked via their relationships without the need for exact spatial modelling, which is especially useful where exact locations of features are not as important as the relationships between them. It also allows for a lattice of relationships (such as that required for modelling joint governance of administrative areas), rather than just hierarchical relationships (Frank 1998). The classification, resulting from combining addresses within individual buildings, may have any number of specific classes based on the different types of functions of the addresses (residential, commercial) and their proportions within the building (e.g. 10% commercial, 90% residential). This may be better modelled within an ontology than relational database. Spatio-temporal attributes may be modelled through inheritance rather than explicitly defined as (potentially many) attributes of each individual feature. The ontology provides the opportunity to describe the real world features that are being modelled, and the relationships between them, in a machine-readable manner with metadata incorporated in the data structure. It is also possible to account for data outside of the ontology's own domain by linking to external ontologies, opening up possibilities for future inclusion of data that are not yet available.

Ontologies have not yet been widely used with the geographic domain and there is no evidence of their use within the more specific domain of population estimation. This may be due to the interdisciplinary nature of population estimation ontologies, and the difficulties in defining geographic concepts to the extent that commonly accepted terms are lacking in the domain, as well as confusion about what ontologies are (Hart & Dolbear 2013). The lack of existing geographic ontologies means that for this research, a completely new ontology will be developed.

2.2.2 Ontology Types

There are several different types of ontology, with several different approaches to differentiating these types. One distinction is based on the formality of the ontology and another on its generality, yet another is based on whether the ontology is generic, normative or descriptive (Agarwal 2005). Van Heijst, Schreiber & Wielinga (1997) provide two typologies based on the amount and type of the conceptualisation and the subject of the conceptualisation. The former distinguishes between terminological, information and knowledge modelling ontologies, the latter between application, domain, generic and representation ontologies. In these typologies, the knowledge modelling ontology "specifies conceptualisations of the knowledge and are often tuned to a particular use of the knowledge that they describe" and domain ontologies "express conceptualisations that are specific for particular domains". These types of ontologies are re-usable in their domain. In GIScience, ontologies are used for three main reasons: for knowledge generation, domain specification and finally, information system development (Agarwal 2005).

The requirements of the ontology that will be developed in this thesis fit into these definitions as a formal, domain-specific ontology that is developed for a specific application: to estimate population at fine spatial and temporal scales.

2.2.3 Ontology Design Patterns for Modelling Spatial and Temporal Relationships

An ontology design pattern is "a reusable successful solution to a recurrent modelling problem" (Ontologydesignpatterns.org 2010). This section will provide a summary of some relevant design patterns and discuss which aspects of these will be adopted in the ontology design. The design patterns to be discussed begin with Allen's Interval Algebra, concerned with qualitative temporal relationships (Allen 1983). This is followed by a discussion of qualitative spatial relations: mereonomy which deals with part whole relationships (Casati & Varzi 1999) and mereotopology which is concerned with the boundaries and interiors of wholes, and the relations of contact and connectedness (Smith 1996). Region Connection Calculus (RCC) is the final discussion point, one example of which is RCC8 (Renz 2001).

Ontology can be used to model different types of spatial relationships, for example orientation, distance, size, topology, shape, and mereological relationships (Renz 2001). Each of these types of relationship require a set of binary relations in order to fully describe any of the possible relationships between two objects (Renz 2001), e.g. A is inside B and B contains A are defined as two separate relationships.

Qualitative temporal relationships can be modelled using the temporal logic of Allen's algebra in which there are 13 relations for intervals, two for points and 11 for interval-point combinations: x takes place before y; x meets y; x overlaps with y; x starts y; x during y; x finishes y; x is equal to y (Allen 1984; W3C 2014a). Reasoning about these relationships is to do with reasoning about ordered points (for both instants or start and finish instants that define intervals). The temporal is therefore one dimensional (Renz 2001), which makes it relatively simple to model mathematically, in comparison with two or three dimensional space.

Spatial relationships are more complex than the temporal relations described above, and fall into four groups: geometric, topological, mereological and mereotopological. Geometry is "concerned with the properties and relations of points, lines, surfaces, and solids" and "a particular mathematical system describing these properties and relations" (OED 2016). These relationships are affected by transformations of the underlying space. Topological relationships are those that do not change if the underlying space is subjected to topological transformations. Topological properties include inside/outside, intersects and connected (Worboys & Duckham 2004).

Mereological relationships are those concerned with part-whole relations (Casati & Varzi 1999). Varzi (2016) indicates that these include the relation of part to whole as well as the relations of parts within a whole. Varzi (2016) also points out various features of mereological relationships, based on some examples (below) that highlight that the topological connectedness relationship is not necessarily present when one thing is part of another thing and the mereological relationship can be temporal. These relations may be related to real or abstract entities, and mereological relationship without a topological relationship are common, some examples are presented in Table 1.

Mereotopological relationships are those that refer to both topological and mereological relations and are an extension to the mereological relationships, describing the topological relationships between wholes, parts, parts of parts and the boundaries between parts. These relationships are the key to making spatial inferences within ontology. For mereotopological relationships, the two dimensional qualitative spatial relationships can be modelled using the spatial logic of Region Connection Calculus (RCC8) which are a fundamental part of human conceptualisation of space and topological relationships (Renz 2001). Eight spatial relationships, shown in **Error! Reference source not found.**, that correspond to Allen's interval algebra are defined : disconnected (DC); externally connected (EC); equal (EQ); partially overlapping (PO); tangential proper part (TPP); tangential proper part inverse (TPP^{-1}); non-tangential proper part (NTPP); non-tangential proper part inverse ($NTPP^{-1}$) (Renz 2001). In this logic, all spatial objects are treated as regions.

Table 1: Some examples of relationship types

Statement	**Relationship**
The handle is part of the mug.	Mereotopological
That area is part of the living room.	
The outermost points are part of the perimeter.	
The remote control is part of the stereo system.	Mereological
The left half is your part of the cake.	
The cutlery is part of the tableware.	
The contents of this bag are only part of what I bought.	
The first act was the best part of the play.	Mereological, Temporal

Table 2: Examples of inference made from different relationship types through qualitative spatial reasoning, using several facts to infer a new fact

Relationship	**Example**	**Inference**
Mereological	Building A is inside Functional Site 1 Functional Site 1 is inside AOI 1	Building A is inside AOI 1
Mereotopological	Building A is adjacent to Building B Building A is adjacent to Building C	Building A is a Terrace
Topological	Address 1 is inside Building A Address 2 is inside Building A Address 3 is inside Building A	Residential Addresses in Building A are Flats

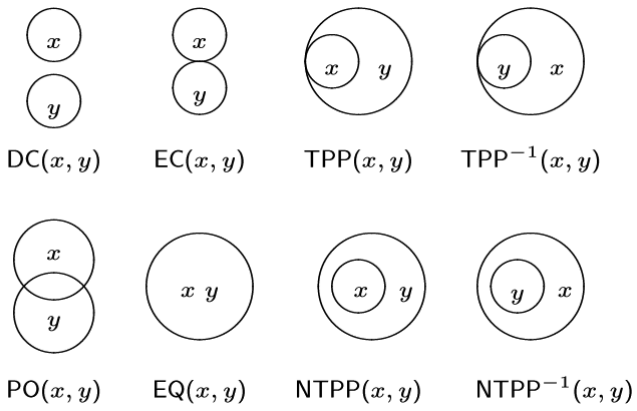


Figure 3: Two-dimensional examples for the eight base relations of RCC8. From Renz (2001)

GIS typically use only geometry and topology for spatial analysis. However, modelling those features from GIS in an ontology, while not allowing the relationships to be identified (as that requires spatial analysis), would allow these mereotopological relations to be qualitatively represented, including, for example, addresses contained by buildings, or groups of features that comprise another feature. If data representing these features and relations have database links, the spatial analysis of GIS is no longer required for expressing these relationships in an ontology, and geometry need not be stored for the types of analysis that can be carried out within the ontology. A fully functioning spatio-temporal system would include queries and methods using both the qualitative spatial and temporal logic. This would allow queries and methods including cluster analysis, involving spatial and temporal proximity analysis that is not included in these two algebras, as the qualitative spatial relationships are based on connectivity.

2.2.4 Semantic Web Technologies

This section's purpose is to introduce the Semantic Web technologies that are used in this research.

On the web, database schemas and Extensible Markup Language (XML) provide structure, but neither consider inference and have no way to express meaning in the data. The Semantic Web is an extension to the web that gives this well-defined meaning to information so that it can be processed automatically (Berners-Lee, Hendler & Lassila 2001). It uses ontologies to facilitate the interchange of data on the web, and combines ontologies with hypertext (texts containing links to other texts (W3C 2003)) and Linked Data (collections of interrelated data sets on the web (W3C 2015)). The ontology approach is more expressive, allowing meaning to be clearly expressed. This leads to several advantages in using ontology over, for example, relational databases: a focus on formal semantics to represent meaning, sharing of formally defined information and the ability to make inferences that are otherwise concealed in the data.

While a vocabulary specifies which classes are expected in the domain, and what properties are required, an ontology defines how these terms can be used with each other. An ontology has four components, including the vocabulary: explicit definitions of concepts (or classes) and their hierarchy, relations between these classes (properties), axioms and instances (individuals) (Davies, Studer & Warren 2006). The first three of these components contribute to an approximate equivalent of a schema definition in a relational database, and the instances are the equivalent of data tables in the relational database. Table 3 sets out the different components along with their equivalent relational database components.

Ontologies are built using a variety of Semantic Web technologies, at the core of which is the Resource Description Framework (RDF). Below is a brief overview of the different technologies employed in the development and implementation of the ontology that is central to the modelling framework for spatio-temporally detailed population estimation.

2.2.4.1 Uniform Resource Identifiers (URIs)

URIs are a means to identify things, or resources, that are modelled within the web and the Semantic Web. URIs provide a simple universal syntax that identifies the name and location of a file or resource (Berners-Lee et al. 1998). The URIs may be used to identify any of the resources that are referenced from within the ontology. These may include classes, datatype properties (attributes), object properties (how classes and instances relate to each other) and the objects themselves (individuals) that are described by the data.

The use of such a global set of identifiers facilitates data sharing. In addition, using URIs to identify the resources means that not all data need be stored in the triple store. Some URIs may be stored externally as Linked Data.

Table 3: Components of ontology, and their approximate equivalents in a relational database.

<table>
<tr><th>Component</th><th>Explanation</th><th>Relational Database Equivalent</th></tr>
<tr><td>Classes</td><td>Explicit definition of concepts</td><td rowspan="3">Database schema</td></tr>
<tr><td>Class hierarchy</td><td>Sub-classes (including inheritance)</td></tr>
<tr><td rowspan="2">Axioms</td><td>Definitions of allowed statements about the properties of, and relations between, classes</td></tr>
<tr><td>Statements about the properties of, and relations between, instances of classes</td><td rowspan="2">Data tables</td></tr>
<tr><td>Instances</td><td>Data representing real-world instances of things</td></tr>
</table>

2.2.4.2 Ontology Languages

The Semantic Web ontology languages that are utilised in this research are presented here so that each new language described builds on the previous one, adding expressivity and complexity. All the languages are used together to produce an expressive ontology that meets the needs of the modelling framework and examples are presented to demonstrate how the languages are utilised.

Each of these examples can be used independently to create just a few RDF triples (described below) to be added to a triple store.

2.2.4.2.1 Resource Description Framework (RDF)

Many Semantic Web applications are based on the accessibility of, and integration of Linked Data, all in a common format. Both data and relationships among the data need to be made available for Linked Data to be accessible. The common format allows data conversion, or on-the-fly access to databases (W3C 2015).

RDF is such a common language for representing data on the Semantic Web. It is a knowledge representation language and a standard model for data interchange on the Web. The World Wide Web Consortium (W3C) states that "RDF has features that facilitate data merging even if the underlying schemas differ, and it specifically supports the evolution of schemas over time without requiring all the data consumers to be changed" (W3C 2014b). As a data model, RDF is independent of the syntax used to represent the data, and of the subject area in which it is applied. RDF is therefore a flexible model that will allow data to be brought together from disparate sources, which may change over time. This makes it an ideal data model for representing the variety of continuously evolving and growing data sources that may be utilised in the modelling framework for population estimation.

The *triple* is the fundamental unit of RDF: all data are represented in terms of triples. DuCharme (2013, p2) clearly describes triples: "Each triple is like a little sentence that states a fact. We call the three parts of the triple the subject, predicate, and object, but you can think of them as the identifier of the thing being described". The things that are described may be resources, property names or property values. Triples can be represented as RDF Graphs. An example of how a triple might look written both as a graph and in English is set out in Figure 4.

RDF allows a URI to be defined as a resource. URIs can be used for any part of the subject-object-predicate portion of the triple. RDF also provides the property class, and the subject, predicate and object properties. In this way, URIs can be used to represent any of the components of the RDF triple. RDF also includes the ability to state that a resource is an instance of a class. This allows triples to be constructed with the subject, predicate and objects defined according to the RDF standard, and the ability to state class membership of the resource.

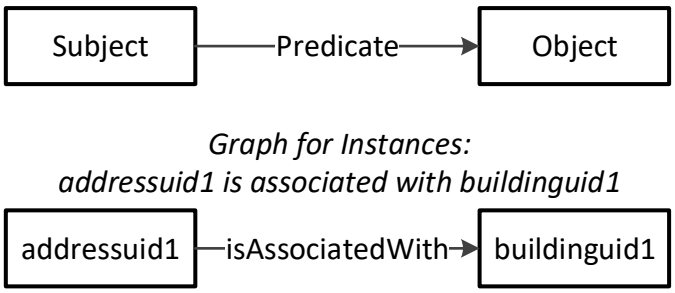


Figure 4: Graphs of RDF Triples with examples of how the triple looks written as a sentence. A set of RDF triples is an RDF Graph. The node-arc diagrams are illustrations of graphs: the two nodes represent the subject and object, and the arc represents the predicate. The middle graph represents a relationship between two classes, and the bottom graph represents the relationship between two instances of these classes.

The use of RDF triples and URIs makes for an easily extensible ontology, as it can be a trivial exercise to bring additional data in to the ontology. There is also no need to generate a table column for every row of data, regardless of whether data are available for that row, but only the instances with additional attributes.

2.2.4.2.2 Serialisation

RDF is serialized (for storage) using one of a number of syntaxes including RDF/XML, Terse RDF Triple Language (Turtle), N-Triples and RDFa. The syntax used in examples presented here use the Turtle syntax, which is widely used for reading and writing triples by hand, as it is in plain text (Hart and Dolbear, 2013, p75).

RDF uses prefixes to abbreviate URIs. In the examples below, `pop:` is the prefix referring to the domain with the full namespace `http://www.w3.org/example`. The RDF, RDF Schema (RDFS) and Web Ontology Language (OWL) resources, also described below, have standard namespace prefixes that are also present in the following two examples. In Code Example One, block 1 defines prefixes for the namespaces that are required; block 2 defines the address and building classes using the RDF construct type and the OWL specification class. Block 3 creates an instance of an address and a building using the RDF construct “type”, and then specifies that an *isAssociatedWith* relationship is present between the address instance and the building instance. The definition of this relationship requires RDF Schema.

2.2.4.2.3 RDF Schema

RDFS builds on RDF, adding functionality and expressivity: specifically, classes and datatypes can now be defined. Sub-classes and sub-properties can also be specified, which is essential for defining a class hierarchy. Utilising RDF Schema (RDFS) enables, for example, a building polygon class to be defined as a sub-class of a topographic area polygon class. Instances defined within sub-classes will have allowed attributes and relationships that distinguish them from other sub-classes.

RDFS also includes the ability to define the domain and range of a property so that class membership may be inferred from the use of a property. In Code Example Two, block 2 shows how the property *isAssociatedWith* is defined with respect to the classes that it relates to one another. In this code sample, the relationship is first defined as an instance of type *ObjectProperty*. So that the address class has an *isAssociatedWith* relationship with the building class, the domain and range of the *isAssociatedWith* property are defined as address and building, respectively. The inverse relationship is also defined in a similar manner, in block 3, which also specifies that this is the *inverseOf isAssociatedWith*.

2.2.4.2.4 Web Ontology Language (OWL)

OWL is an ontology language, and a vocabulary extension to RDF. It adds expressivity so that more detailed knowledge about the domain of interest can be stated, or inferred, thereby overcoming some of the limitations of RDF (Hart and Dolbear, 2013, p161). OWL is description-logic based and provides constructs for defining property characteristics, and complex classes, property restrictions and equivalence relations (W3C OWL Working Group 2012a). It also provides the ability to set classes as disjoint from each other. This means that if an instance is given membership of a class, it cannot also be given membership of the disjoint class. For example, if a resource is defined as a topographic Area, the same resource cannot also be defined as a Functional Site (FS) if these two classes are defined as disjoint.

OWL is utilised in the ontology developed as part of this modelling framework because of these additional constructs, which are required for some of the representation. Some of these can be seen in the examples in the code examples, specifically in the definition of the classes, the object property (relationship between classes) and the inverse relationship.

2.2.4.2.5 Code Example One

This code example demonstrates how classes and instances are defined in RDF, but also needs to utilise OWL functionality.

```
# block 1: define the namespaces:
@prefix pop: <http://www.w3.org/example> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .

# block 2: define the classes:
pop:address rdf:type owl:Class .
pop:building rdf:type owl:Class .

# block 3: create the instances:
pop:addressuid1 rdf:type pop:address .
pop:buildinguid1 rdf:type pop:building .
pop:addressuid1 pop:isAssoicatedWith pop:buildinguid1 .
```

2.2.4.2.6 Code Example Two

This code example demonstrates how a property is defined using RDFS range and domain.

```
# block 1: define the namespaces:
@prefix pop: <http://www.w3.org/example#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .

# block 2: define the isAssociatedWith property:
pop:isAssociatedWith rdf:type owl:ObjectProperty .
pop:isAssociatedWith rdfs:domain pop:address .
pop:isAssociatedWith rdfs:range pop:building .

# block 3: define the inverse of the isAssociatedWith
# property:
pop:hasAssociatedPlace rdf:type owl:ObjectProperty .
pop:hasAssociatedPlace owl:inverseOf pop:isAssociatedWith .
pop:hasAssociatedPlace rdfs:domain pop:building .
pop:hasAssociatedPlace rdfs:range pop:address .
```

2.2.4.3 RDF Triple Stores

Triple stores are specialised databases tuned for storing and retrieving RDF triples. They may also provide parser and serialisation facilities for populating the store and publishing information from the store, as well as a query engine. RDF stores also include the ability to merge two datasets together: any resources with the same URI are considered to be equivalent (Allemang and Hendler, 2011, p54) so querying data from different data stores is straightforward.

In practical terms, triple stores contain both the ontology and the data, and can perform some reasoning based on RDFS or OWL. This allows some inferences to be made as well as checking for inconsistencies in the ontology or the data (although the extent of this capability varies between triple stores).

Unlike a relational database, in which rows of existing data that have no link to newly added data will need blank cells in their data tables, a triple store only stores available data, with no need to specify absence of attribute data where they are not available.

2.2.4.4 Reasoners

Semantic reasoners are tools that check ontologies for inconsistencies by inferring logical consequences from the set of axioms in the ontology, and are also used to infer new knowledge from the asserted facts (Abburu 2012). The strategies used by reasoners include forward chaining and backward chaining.

Forward chaining involves applying inference rules to the explicitly stated facts to generate new facts. This process is repeated until no further facts can be inferred (OntoText 2016) and is appropriate when the knowledge base is large and requires many queries. The alternative approach is backward chaining, which involves starting with a fact or query and attempting to prove it through examination of the ontology. This is an appropriate approach where there are fewer queries that do not require full entailment of the ontology.

The greater the expressiveness of the ontology language, the greater the difficulty of reasoning over the representations built using that language so using the greater expressivity of OWL makes reasoning more difficult (Brachman & Levesque 1984). Ontology design therefore requires an awareness of the trade-offs between the modelling requirements of the domain, expressivity and complexity of the language, and the effect that the size of the datasets has on the cost of reasoning.

2.2.4.5 SPARQL

SPARQL Protocol and RDF Query Language (SPARQL) is a W3C standard for querying RDF data and is to RDF databases what Structured Query Language (SQL) is to relational databases. It is a query language that can be used across diverse data sources, including local and remote RDF data (DuCharme 2013; W3C 2008) as it includes the facility to query named graphs identified by URIs. It can be used for selecting, constructing, describing, inserting and deleting data from an RDF database. As with SQL, keywords that refine the queries can be applied, such as filters and joins.

SPARQL queries utilise graph pattern matching. If the RDF graph dataset contains all the ontology and data associated with that ontology, when a query is executed, instances in the RDF graph dataset are retained in the query results where they match the graph pattern specified in the query. Where there is more than one graph pattern to match, SPARQL will return instances where all of these patterns are matched. In the SPARQL code example below, the first pattern selects all instances of the type address and stores these in the variable *a* (denoted by the ? prefix). The second pattern match selects all instances of the type building and stores these in the variable *b*. Finally, the instances in variable a are restricted to all addresses that have an *isAssociatedWith* relationship with a building.

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX pop: <http://www.example.org/pop#>

select * where {
     ?a rdf:type pop:address .
     ?b rdf:type pop:building .
     ?a pop:isAssociatedWith ?b .
}
```

2.2.5 Research Specific Benefits of Ontology

In summary, there are several benefits of using an ontology for this research. The first is in the formal specification of the features being modelled and the relationships between them. This removes any confusion about what is modelled thereby facilitating data re-use or re-purposing both within the population estimation domain, and outside of it. Beyond the scope of this research, such data re-use can provide a two-way benefit: it is possible, and a simple exercise, to re-use data from other sources and the data within this model can be easily utilised by other ontologies. The first of these two situations is of interest within this domain: as more and more data come available over time, the ability to incorporate the knowledge contained within them is already present in the technology.

Another key benefit is that there is no need for quantitative spatial analysis in the ontological model and subsequent query of this, reducing the computational burden. This approach can take full advantage of the database linkages that are present within and between data sets. Design patterns are available for modelling the relationships between features: the mereological (part-whole) relationships of RCC5 (Cohn et al. 1997) will be used within the proof of concept, and although the additional mereotopological (adjacency) relationships of RCC8 (Cohn & Renz 2008) will be in the model, they will not be used in this research.

Time can be modelled within an ontology, by making connections between the data in a way that GIS cannot yet manage because of their “snapshot” nature.

The main benefit is the functionality that the ontology and semantic web technologies add, beyond data sharing, linkages and re-use. The ability to make inferences from the data present and to reveal knowledge that is very difficult to garner using a relational database is of most significance to this research.

One final benefit is related to the open world assumption of the web. That is, we must always assume that more information may come to light (Allemang & Hendler 2011, p10). Database technologies generally use a closed world assumption: that something is assumed to be false if it is not held within the database (Hart & Dolbear 2013, p52). This can lead to very large and normalised relational databases. This open world assumption means that new knowledge can become available, and although we cannot know how that new information will fit into our model, the simple RDF data model means that should the information be relevant, it can be incorporated. This is very relevant to a modelling framework that must be extensible, with the ability to integrate new data as and when it becomes available, as the data available for population estimation are currently incomplete.

2.3 Data Landscape

Data products are constantly evolving to meet demand, causing changes to the data that are available within any one country: the data landscape. It is therefore necessary to assess the currently available data while recognising that many changes are likely to occur in the future. The modelling framework that is being developed in this research is intended to be relatively generic, utilising broad data categories as input, since the key aim is to prove the concept that the detailed spatio-temporal patterns of population can be modelled at the address level, using ontology, rather than to focus on the data. The modelling framework will draw on the fundamental concepts already reviewed and should not be influenced by the specifics of data availability, although its implementation may need to be adapted to fully integrate future data. This section therefore presents the relevant data available at this current time.

Given the need to increase the spatial resolution of population data, there is a clear case for developing an address level model. This is the highest resolution that can be modelled without tracking individuals. The relationships of addresses with buildings are also important as intelligence about the addresses can be inferred through these relationships.

A review that examines availability of addresses and buildings data is therefore necessary. This must consider the modelling framework's requirement with regard to these data. Available population data sources also require consideration. Some basic requirements must be met by these three datasets. First, it must be possible to identify a classification from the address dataset, whether this is supplied from an external source or delivered with the data. It must also be possible to re-distribute the population data where it is supplied as small area statistics and the address classification must provide a means to influence the population allocation process.

The data environment for the implementation of the modelling framework is an area in the South of England. Data sets that cover this region may have extents specific to the region, to England and Wales, to GB or to the UK.

2.3.1 Address Data

2.3.1.1 Address Requirements

For using addresses to identify the functional classification of buildings, there are three main requirements for the address dataset. The first of these is that a classification attribute is present or can be derived analytically, and that this can be used to identify the function of the address. This must be available for residential and commercial addresses as these two functions serve the

majority of human activity. Any shortcomings of these classifications may be partially dealt with using geo-processing techniques used to supplement classifications.

The classification itself needs to be detailed enough to allow the association with different types of activity occurring at the address. For example, a classification of “Commercial” will be inadequate as there are many different types of commercial activity, each with different temporal patterns of occupation. In this example, the sub-classes may include “Office”, “Leisure” or “Agricultural”. Each of these sub-classes will also need a further breakdown. For “Commercial, Leisure” classes, this could include cinemas, golf courses or leisure centres. Each will have distinct patterns of occupation, or Temporal Signatures (TSs), and the classification must allow for this to be identified. The classification available in the data should be considered as input to the functional classification used in this methodology rather than as a definitive classification in its own right, as different datasets have different classifications. In addition, the implementation of this classification must be complete with all addresses giving the required level of detail. The classification requirements are discussed further in Section 4.1.1

The spatial relationships between the addresses and the buildings will be used in the model, so the second requirement is that spatially accurate locations of the addresses are present through locational attributes or a cross reference to the relevant building’s unique identifier. Here spatial accuracy means to be located within the appropriate polygon representing the building, if such an association exists. If a spatial object contains more than one address, the number of addresses that fall within the object must be identifiable in order to identify blocks of flats or flats above retail or commercial properties. This can only be done if the addresses are located within the correct building polygons.

An ideal dataset would also indicate the presence of CEs and HMOs. CEs provide managed residential establishments and include prisons, university halls, care homes, hotels or bed and breakfast accommodation with 10 or more guest beds and army barracks (ONS 2012a). HMOs are properties that accommodate more than one unit of living accommodation, they share one or more of the basic amenities, such as toilet, washing and cooking facilities, and they can be buildings or parts of buildings, individual self-contained flats, or converted buildings (The National Archives 2004). By this definition, an HMO may include shared accommodation such as student houses. While CEs are often identifiable by their name, HMOs can be considered to contain hidden households as there is no external sign of multiple occupancy (Calder 2009), and as such they are very difficult to identify.

2.3.1.2 Definitions, Standards and Availability Outside Great Britain

The European INSPIRE Directive describes an address as "an identification of the fixed location of a property", and it is defined as the "location of properties based on address identifiers, usually by road name, house number, postal code" (European Commission 2014).

The definitions of addresses used by Australia and the US include this requirement, plus a means of referencing both addressable and non-addressable objects. Australia's definition states that an address relates to a property or parcel, so a building need not be present, in addition, the address may not necessarily receive mail, and the standard (AS/NZS 4819:2011), implemented by both Australia and New Zealand, allows for alias addresses, locality address or street addresses (PSMA Australia Limited 2014).

The US standard for addresses definition states that an address "specifies a location by reference to a thoroughfare or a landmark; or it specifies a point of postal delivery" (FGDC 2010). Again, this standard allows for the use of aliases and both postally and non-postally addressable objects (FGDC 2010).

In Australia, the Geocoded National Address File (G-NAF) is the definitive address product. It is produced by *PSMA Australia Limited*, which is a joint venture between Australia's government mapping agencies to provide a national street-level digital map database. Its purpose is to provide an authoritative address database and thereby improve efficiency and accuracy of existing spatial analyses as well as to enable novel products and services (Paull 2003). The product contains residential, commercial and other addresses (PSMA Australia Limited 2014). The data model is very similar to that of the AddressBase product in GB (see Section 2.3.2.3 below) including a core address record and additional tables containing related attributes (*ibid.*).

In New Zealand, the address product is included in the Landonline product which includes land parcels, addresses and roads, with the addresses fully integrated with the land parcel information (not buildings). Landonline includes all the major elements of addressing, it is based on parcels and also includes legal parcel descriptions (Land Information New Zealand 2010). Like the Australian product, there is no classification of the addresses as the data are used to provide a land parcel framework for use in property related GIS (*ibid*).

Similarly, in the US, the MAF/TIGER database (MTdb) incorporates the key address data, the Master Address File (MAF). The addresses were integrated with the TIGER/Line database containing streets and administrative boundaries down to block level and was developed for the purpose of census administration (Penn State College of Earth and Mineral Sciences 2014). This is maintained by the US Census Bureau and contains both all known *living quarters*, and employer

addresses (U.S. Census Bureau Fact Finder n.d.). Although the MTdb does include precise address locations, these are confidential and so they are not published (Penn State College of Earth and Mineral Sciences 2014). The street and address range data can, however, be used for the purpose of geocoding individual addresses (U.S. Census Bureau 2013). The MAF part of the database includes MAF Structure Points (MSP) which are the records of the structure on the ground (including coordinates), with associated MAF Units which are the individual properties. Associated with each MAF unit can be one or more address records thereby allowing for alias addresses. The MSPs can also be grouped together into *Facility Records* containing more than one MSP. The MAF includes attributes that represent unit type, which is a high-level functional classification, with more detail available, particularly for residential MAF units.

2.3.1.3 Address Data in Great Britain

These European, Australian and US standards contain all of the same elements as BS7666:2006, the British Standard, that defines an address as "a means of referencing an object for the purposes of unique identification and location" (Katalysis Limited 2014) and provides a standardised format for storing details of every property. It was developed for the creation of the National Land and Property Gazetteer (NLPG), discussed below. The standard is based on the Basic Land and Property Unit (BLPU) which represents an area of land that is uniform in terms of property rights or physical features, occupation or use (National Land and Property Gazetteer 2007). Each BLPU has a Unique Property Reference Number (UPRN), a spatial reference, and at least one address associated with it, which in turn links to a street record (*ibid*). BS7666:2006 includes the location as a spatial reference in its definition. This is more than is encompassed by the INSPIRE directive definition. It also explicitly defines the address as relating to an object, rather than a property (land parcel, building or portion of building such as a flat).

In this context, an address can relate to a property, but it could also relate to many other object types that cannot be occupied and do not receive mail. These are non-postally addressable entities, such as communications masts, parks or agricultural buildings. Some of these do not need an estimated population count, as they are not occupiable. Addresses are therefore not only places of business or residence, but can also include many other features. This differs somewhat from the common perception of *address*, which most people consider a way of finding a building, for residential, commercial or public use.

BS7666 includes no differentiation between the classes of the properties or whether they are occupied, developed or vacant or any distinction between postally addressable properties and non-postally addressable entities (National Land and Property Gazetteer 2007). A postally

addressable object can be identified from a number or name, such as a house number and street name. The example of a communications mast demonstrates than an address can relate to a non-postally addressable real world object. Such an address may be on a street with an associated name, but does not necessarily need to be located within a property that has an associated number: it may be located in the corner of a field that has no building or road name associated with it.

Three source address lists have contributed to the creation of the definitive address dataset in GB: The National Address Gazetteer (NAG), which is available as Ordnance Survey (OS) AddressBase range of products. These are the NLPG, the Postcode Address File (PAF) and OS's MasterMap Address Layer 2 (AL2). None of these three datasets provides a comprehensive address list, and only the PAF is still available for use. Descriptions of all three of these datasets are provided below for context.

2.3.1.3.1 National Land and Property Gazetteer

The NLPG was a national address list that provided unique identification of land and property (Intelligent Addressing 2010). It was initiated in 1999 and was completed in 2008. It was intended as the master address dataset for England and Wales to provide unique identification of land and property. This was a dataset that was continually updated (daily) by the Local Land and Property Gazetteers (LLPGs) custodians, and was maintained to BS7666:2006 (see below) by 348 Local Authorities (LAs) (National Land and Property Gazetteer 2014). The NLPG provided addresses for sub-divisions of addressable objects (e.g. flats and industrial premises), non-postally addressable objects and it also allowed multiple address references for each property. There was also a classification for each address (Intelligent Addressing 2010). This data set does not include Scottish address data so includes only England and Wales. The lack of differentiation between the property types (postally addressable and non-postally addressable) meant that this dataset alone could allow for the identification of occupiable addresses from the address list, although the information that was contained within it is incorporated in the NAG as described below.

2.3.1.3.2 Postcode Address File

The PAF is owned and managed by Royal Mail. It contains 29 million postal addresses and was created to aid in the efficient delivery of mail (Royal Mail 2014). As such, the PAF contains addresses that represent *delivery points* for mail and so contains information that is geared towards this function. It contains all of the attributes required to deliver mail, including building names and numbers, the street names and locality, the post town and the postcode and also

organisation name, alias data (where a single property unit can be addressed in more than one way) and unique identifiers (Katalysis Limited 2014) and also includes large and small business codes. The PAF is continuously updated by Royal Mail.

The PAF file lacks a functional classification, so it would not allow the association with appropriate TS for each address. There is also no locational information contained within this file. As attaching addresses to cartographic objects requires the ability to locate an address along with its classification, the PAF file alone will not suffice.

2.3.1.3.3 Ordnance Survey MasterMap Address Layer 2

OS MasterMap AL2, which is now discontinued, contained approximately 29 million addresses grouped into three themes. The postal theme contained 27 million residential and commercial premises, which were sourced from the PAF. OS added classification and spatial references (derived from OS surveys) to these addresses. The PAF delivery points were not necessarily related to objects in the MasterMap Topology layer (e.g. temporary buildings and houseboats). The non-postal theme contained 1.5 million “miscellaneous” premises such as churches and car parks, which are sourced from OS MasterMap Topography Layer (MMT). These were Objects Without Postal Addresses (OWPA). The Multi-Occupancies without a Postal Address (MOWPA) theme contained residences in multi-occupied buildings that were not listed in the PAF but were contained in Royal Mail’s Multi-Residence (MR) file (Ordnance Survey 2011). These were not HMOs but would include buildings containing self-contained flats in converted or purpose built buildings that share a front door.

These addresses were attributed with cross references to other datasets (such as MMT layer or PAF), and one or more classifications. The classifications used were related to the source data, so each address could contain a base classification (from OS) as well as one or more classes from the National Land Use Database (NLUD, which includes a classification of land use and land cover) and the Valuation Office Agency (VOA, which deals with valuations of property in support of taxation and benefits calculations), along with a classification confidence measure (Ordnance Survey 2011).

The information that was contained within the AL2 is now incorporated into the NAG. It is not currently clear whether the additional information in the NAG is necessary for a better understanding of human activity at a location, although the AddressBase Premium (ABP) product (described below), which is based on the NAG, is considered the most comprehensive address product at this time.

2.3.1.3.4 ONS Census Address List

For the 2011 Census, the ONS generated a residential address list for the purpose of managing distribution and collection of the census forms. The source of the data was the PAF along with locational and high-level classification (e.g. residential, commercial) information supplied by AL2 dataset. This was textually and spatially matched to the PAF and NLPG and mismatches were processed through various means including field checks (ONS 2012a).

This residential address list was complemented by the creation of a Communal Establishments Address Register (CEAR) generated from over 200 sources including LAs (ONS 2012a). This identified the class and size of different CEs such as prisons, hospitals, nursing homes and university halls of residence. As some CEs were not covered at all in national products, other data sources such as 'Edubase' were used, as well as information from LAs and county councils (*ibid*).

The result was a spatially referenced and classified (albeit at a high level) address register that was very close to complete which would have been useful for identifying residential addresses classifications. This dataset is however unavailable for use, as it was maintained only until census day (Calder 2009), and was subject to data protection issues. As there will be a census in 2021, it is likely that this work will be replicated in some form, with the production of a residential address register and a CEAR (or equivalents). It would be useful for the future if this resulted in a comprehensive dataset that continues to be maintained beyond census day.

2.3.1.3.5 National Address Gazetteer and AddressBase

Since 2011 there has been a single definitive national address register, the NAG which is managed by GeoPlace, a limited liability partnership (Ordnance Survey 2013b), and which brings together the available national datasets described above (NLPG and PAF, as well as relevant data from AL2). Prior to this date there was no relationship defined between the NLPG and the PAF (Katalysis Limited 2014). GeoPlace's role is central governance of the NAG and this involves taking data feeds from several organisations.

Royal Mail supplies the PAF containing 29 million postal addresses without locational information. Local Governments supply standardised LLPGs, which are merged to form the NLPG. Scottish Local Governments (of which there are 32) supply the One Scotland Gazetteer (OSG) via Forth Valley GIS. These land and property gazetteers contribute the most up-to-date street names and numbers and these are updated when new properties are generated through new build or subdivision of existing properties. There are also data feeds from other agencies including the

Highways Authorities in England and Wales, fire, police, national park, conservation boards and passenger transport authorities (Ordnance Survey 2013b).

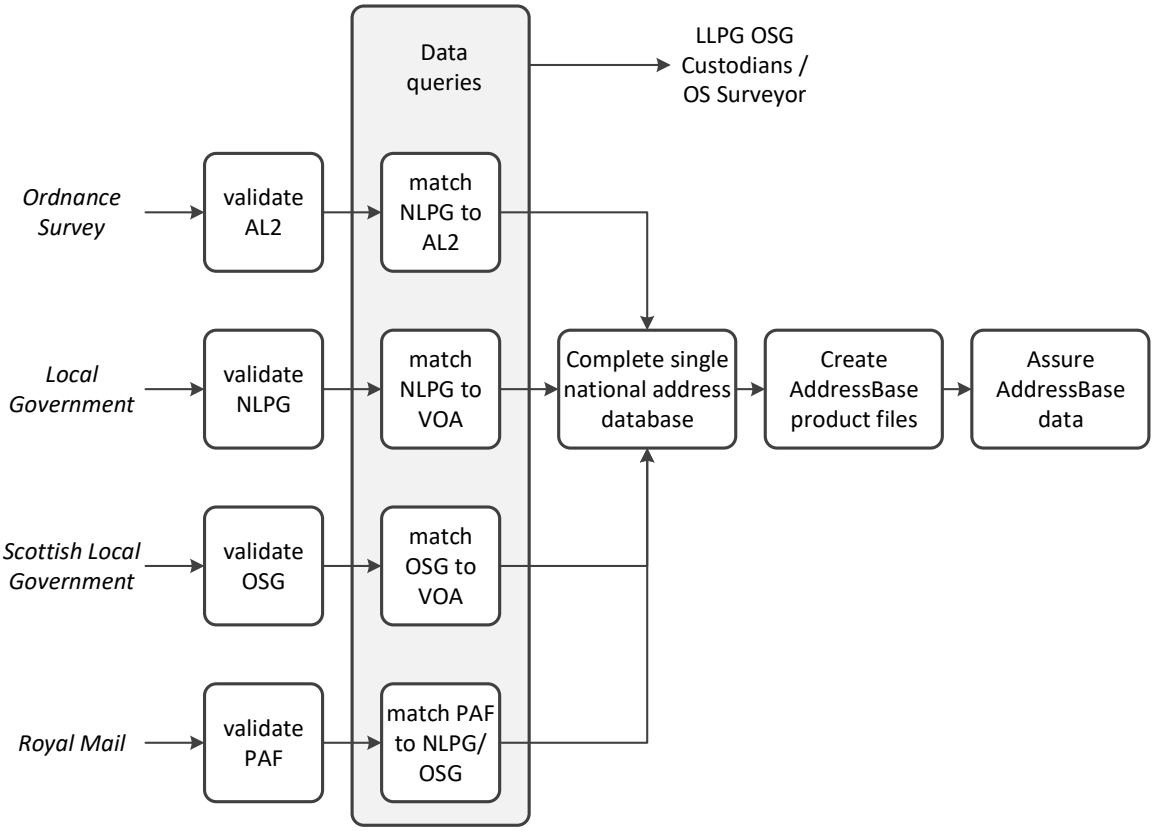


Figure 5: AddressBase Production. This figure, recreated from the AddressBase Products User Guide, indicates the processes and agencies involved in the production of AddressBase products. Source: (Ordnance Survey 2013b)

Figure 5 gives an overview of GeoPlace's governance activities. These include validating each incoming data feed, matching datasets together, compiling a single national address database and creating the AddressBase product files which are then supplied to OS for distribution. A key part of this process is that persistent key identifiers, the UPRNs, are assigned to each address record. As such, records in the AddressBase layer will never have their unique identifier modified. It is therefore possible for several address records to be held for a single property when amendments have been made, or for alternative addressing for example, such as where a BLPU has an alias address (e.g. a house on a corner can be addressed using more than one street name). This UPRN is also used to join related address records across tables supplied with AddressBase, or across

datasets (Ordnance Survey 2013b), this includes MMT Layer and this is key to linking these datasets.

When LA addresses are received by GeoPlace, they are textually matched to the PAF addresses. If a match is found, the MMT Topographic Identifier (TOID, as a Cross Reference Table (XREF) record) is added to the NAG address associated with the LA address. MMT TOIDs are only present in the NAG where there is a match made. If no match is made, there will be no MMT TOID associated with the address in the NAG. Where the PAF address is textually matched to an LA address, the spatial reference is sourced from the AL2 data.

The current output from this process is the OS AddressBase range of products, of which ABP is the most detailed. Because of the many organisations contributing intelligence to the product, there are complex licencing issues. Ownership of the data is part Royal Mail (which is privately owned) and part OS (a public corporation), who both generate revenues from the sale of the constituent parts.

The NAG, and therefore AddressBase, provides the most comprehensive address list available in GB at this time, and is therefore the preferred address data source for identifying classification for the addresses and providing textual address information.

2.3.1.3.6 Information Not Available in AddressBase Premium

The ABP product includes a multi-occupancy field that gives details of the buildings that contain more than one address. As already mentioned, an ideal dataset for identifying classification of addresses would also include an indication of the presence of an HMO. HMOs and CEs are difficult to identify, as demonstrated in the generation of CEAR created by the ONS for the 2011 census (Calder 2009) which did not include HMOs. The ABP product would be more satisfactory for the purposes of locating addresses if it contained this additional information. It may be possible to source this information from LAs, as HMOs must be registered. It is possible that the future use of administrative datasets and the potential creation of an up-to-date and maintained CEs register for future censuses could meet this need.

2.3.2 Topographic Data

2.3.2.1 Topographic Data Requirements

For modelling population at the address level, there are several key features required of a large scale mapping dataset. The first is that it must be possible to identify the objects of interest and

their topology, so the data must be vector (or it must be possible to convert to vector from raster), with individual polygons, lines or points representing different features. This will enable each building to be evaluated individually, as an object onto which additional attributes can be attached. All buildings and open spaces should therefore be uniquely identifiable from their attributes and geometry, as discrete objects. This will prevent ambiguity in associating addresses with building objects. If the building objects are polygons, this will also allow the geometry of the objects to be used as a means to estimate building footprints. It would be helpful if a building height attribute were present so that building capacity could also be estimated, and proportional numbers of occupants estimated.

Secondly, the data should be a topologically complete dataset, i.e. with correct topological structures in place to enable analysis of adjacency, number of neighbours and containment relationships prior to loading data into an ontology.

Finally, it would be helpful if the layer also provides a means to group features into a FS (several buildings with the same function, such as a hospital) thereby allowing these sites to potentially be treated as single entities. The layer must also include (or be compatible with other layers containing) places that may contain people that are not addressed, such as parks and other open spaces.

This data may be a cartographic dataset, or included in a digital cadastral database, which represents digital boundaries of legal property boundaries and, sometimes, building outlines. A digital cadastral map that does not include building outlines is of less use for identifying addresses, as a property boundary will extend far beyond the building where human activity occurs (especially in the case of properties such as farms). This would skew redistribution of population according to area or volume of the property boundary rather than the occupied buildings. The non-geographic cadastral information is also of less importance for this research as land rights are not relevant.

2.3.2.2 Topographic Data in Great Britain

2.3.2.2.1 Ordnance Survey's MasterMap Product

OS MasterMap has the primary purpose of providing "the most detailed topographic data available of the physical environment of Great Britain" (Ordnance Survey 2010, p16). It was developed in response to the need for such a dataset that realistically represents the real world, for use in GIS, and is designed to facilitate the Digital National Framework (DNF). The DNF is intended to enable the integration and sharing of multiple sources of geographic information. The

MasterMap product suite facilitates this through linking different datasets and reducing ambiguity. This is largely achieved by the fact that every feature within the dataset has an associated TOID attribute, which stays with the feature throughout its life cycle and will never be re-used. These TOIDs offer explicit links to features in other layers, including ABP which is attributed with a cross reference to the MasterMap TOID (Ordnance Survey 2014).

MasterMap is available under an evolving licence from OS, but key parts of this layer (property extents and TOIDS, both of value to this modelling framework) are to be made available under the Open Government Licence (OGL), as part of a project with the Geospatial Commission (Ordnance Survey 2018).

2.3.2.2.2 MasterMap Topography Layer

The topography layer is one of four layers that comprise the MasterMap suite of products (along with the Integrated Transport Network and Imagery Layers and the Water Network), and it represents real-world features that appear in the landscape such as buildings, roads and land, as well as those fiat features such as administrative boundaries or mean high water. Of the six feature layers supplied with the MMT, the topography area feature layer is of most interest as it is comprised of nine different themes: Buildings, Land, Water, Administrative Boundaries, Heritage and Antiquities, Rail, Roads, Tracks and Paths, Structures and Terrain and Height (Ordnance Survey 2014). The first three of these themes are of particular interest as places that can be occupied by people, and are described in more detail in Section 4.2.2.

2.3.2.2.3 MasterMap Sites Layer

The OS MasterMap Sites Layer (MMS) is another part of the MMT product and is generated through a combination of automatic (to extract site names) and manual (to construct the site) methods (Ordnance Survey 2013d). It is a polygon layer with associated access points and routing points which represent collections of real world objects that have meaning as individual sites (Ordnance Survey 2013d). Therefore, a school in the MMS layer is a polygon encompassing all of the relevant topography features that together make up that school such as buildings, car parks and playing fields. This layer separates the form from the function of the features and represents the place, rather than the constituent features of the site.

2.3.3 Address and Topography Data Combined Requirements

It is necessary that the topography and address datasets can be combined in order to aggregate addresses according to the polygon features that they fall inside. This must be achievable using either cross-referencing of the objects through database links, or through spatial analysis.

The location of the addresses will indicate which building polygon they fall inside, and therefore indicate whether there are several addresses in a single building.

Data availability varies between UK countries, with geographic datasets supplied by OS covering GB, and Office of National Statistics datasets covering England and Wales. Separate and comparable census products are available for Scotland so a complete GB data census dataset could be compiled.

2.3.4 Population Data

The different sources of residential and workplace population were discussed in Section 2.1.2. The geographical layers containing WZs and OAs are available alongside statistical tables that relate to these geographies.

The data to be utilised in this research includes, for residential population (for OAs), the count of individuals by their age by single year (table QA103) and the count of individuals living in different accommodation types (or dwelling type, in table QS401). For working population (for WZs) the count of people working in different industries (self-reported by Standard Industrial Classification (SIC) Section in table QS605) and count of people by their employment status (table QS601). There are many more statistical measures available for each of these geographies, many of which could be usefully employed in this modelling framework.

There are many options for ancillary data that can provide visitor population intelligence for specific sites; many have been outlined in the previous sections. They include administrative, survey, social media, mobile telephony, Wi-Fi networks, GPS-enabled devices, traffic sensors, or georeferenced social media posts, sensor networks recording environmental conditions, and volunteered internet based geospatial information. Only a small subset of these ancillary data sets will be utilised as the focus of this research is on the modelling framework rather than the data that can be integrated within it.

2.4 Definitions of Key Concepts

Bona-fide and Fiat Boundary

A *bona-fide* boundary is one that represents a real world physical feature. It relates to an object in the real world and reflects physical discontinuities. A *fiat* boundary is one that is based on a human construct such as an administrative boundary.

Cartographic Object

The aim of this research is to model the sophisticated spatio-temporal patterns of population distributions at the level of the individual cartographic object. Cartographic objects are representations of discretised (i.e. object) features modelled within a GIS. These may be polygons representing areal features such as buildings or open spaces, lines representing railways or roads, or points representing the location of addresses. For simplicity, these will be concerned with buildings unless stated otherwise, although they could also include other cartographic objects such as open spaces (parks, car parks, beaches or plazas). In the ontological context, objects may represent regions, boundaries, parcels of land, water-bodies, roads, buildings, bridges, and so on, as well as the parts and aggregates of all of these (Smith & Mark 1998). In both GIS and the ontological context, the cartographic objects may represent fiat or bona-fide objects.

Building

Physically, a building is generally considered as a stand-alone structure in the built environment. However, in terms of the data layers used in this research, they are polygons that represent buildings as objects. The boundaries of the building polygons may sometimes represent individual properties, or may represent a change in the physical structure of the building. A block of buildings with two front doors will be divided into two separate polygons. If a section of the block of buildings has a different level (e.g. is elevated off the ground), this is also represented as a separate polygon. Only the physical structures that can be determined from the street are recorded in the building polygons.

Functional Site

A Functional Site is a collection of topographic features that together serve a single function, such as the buildings, car park and playing fields at a school, the runway, car parking and terminal buildings at an airport, or the various buildings that make up a hospital. The features that comprise these sites may contiguous or not. For example, a university could be made up of several individual campus sites.

Address

The definition of address is discussed in some detail in Section 2.3.1.1. In GB the address is "a means of referencing an object for the purposes of unique identification and location" (Katalysis Limited 2014) and is based on BLPUs (and therefore properties). These are points with spatial references and attributes that fully describe the address.

Functional Classification

The functional classification will apply to addresses (for which there are several sources: including the address data, Points of Interest (POI) data, SIC etc.) and buildings that contain those addresses. The functional classification of a building can be one class, or a combination of different classes in various proportions. Such a classification will not be definitive because classifications tend to be developed according to application (and therefore varies according to their source), and because classification is, in part, dependent on scale.

Occupiable Address

In this research, an address (and therefore the building that it is associated with) is considered as occupiable if people can use the address for any activities. As discussed, some addresses, such as advertising hoardings, cannot be used for any activities, and are therefore not considered as occupiable.

Household

A household is considered as a person, or group of people who either share a meal at least once a day, or share the living accommodation as their main place of residence (Department for Communities and Local Government 2012). The household includes Houses in Multiple Occupation (HMOs).

Dwelling

A dwelling is defined as "a self-contained unit of accommodation" (Department for Communities and Local Government 2012), in which all the rooms including kitchen and bathroom are behind a single door only for use by members of the household that live in the dwelling.

Dwelling Type

Dwelling type represents the physical structure of a building that is used for residential purposes. Dwellings in the UK are typically detached, semi-detached, terraced or flats. They may also be characterised by the number of storeys, with bungalows being single storey. For the purposes of

this research, the definitions reflect the physical form of the building. Detached properties are a single self-contained dwelling unit in a single building. Semi-Detached properties have two dwellings in a single building joined by a party wall, but otherwise not connected internally, and accessed by separate entrances. Terraces have more than two dwellings in a single building, with separate entrances from the street, and divided vertically. Flats are recognised as being difficult to define (Department for Communities and Local Government 2012). For the purposes of this research, the Department of Communities and Local Government definition is not suitable. This states that flats are defined as part of a building that is divided horizontally (suggesting that at least two storeys are present). However, given that for residential buildings, the building polygons represent individual "houses" (with their own front door), if there is more than one address inside the building, and at least one of them is residential, this is considered a flat. The dwelling type "houses" include single storey bungalows. Dwelling types in this research include detached, semi-detached, mid-terrace and end-terrace. These dwelling types may be appended with either 'flat' or 'bungalow'.

Communal Establishment

Communal Establishments are defined by the ONS as "an establishment providing managed residential accommodation. 'Managed' in this context means full-time or part-time supervision of the accommodation. Communal establishments include sheltered accommodation units (including homeless temporary shelter), hotels, guest houses, B&Bs and inns and pubs, and all accommodation provided solely for students (during term-time)" (ONS 2015). The key here is that the establishment is managed. Different definitions exist, CEs, however, are not dealt with in this proof of concept research.

House in Multiple Occupation

HMOs are houses with at least three people who form more than one household, and who share a kitchen or bathroom facilities (MHCLG 2016). These are very difficult to identify, as there is no outward sign on the building to indicate that it is a HMO.

Temporal Signatures

The modelling framework will involve applying estimates of daytime and night-time populations, in the form of TSs, to buildings that represent SHAs. TSs describe the occupation rates of SHAs at specific times during the day and may be linked to a class of place, an individual building or a single address. The TS can be considered as a graph, with time on the *x* axis and per cent of maximum capacity occupied on the *y* axis. The TSs used for estimating population will need to represent different changes at a variety of temporal scales: minutes, hours, weekdays and

weekends, and will need to encompass seasonal and term-time changes in building use. The TSs will be based on the type of human activity that occurs within the different buildings.

Visitors

Visitors are people present at an address who are not there for residential or work purposes. For example, at a school, visitors include pupils attending school during school hours, and parents escorting children to and from school at the start and end of the day, or visiting the school for other reasons. At a hospital, visitors include patients (in-patients, out-patients and accident and emergency), along with anybody who is accompanying those patients. It does not include workers at the site, or those who are resident there and therefore counted in the census as such. In retail or commercial addresses, visitors are those people consuming the services at the address. At residential addresses, visitors are those people present who do not reside there, and are not working there as a matter of course.

Ontology

An ontology is an "explicit and formal specification of a conceptualisation of a domain of interest" (Davies, Studer & Warren 2006). The formal specification means that it is useable by a computer. An ontology is an explicit specification of an abstract, simplified view of the world that we wish to use to represent a specific domain of interest.

Classes

Classes are sets containing individuals that share certain characteristics. A class is a concrete representation of the concept. In ontology, class definitions need to ensure that the set is given a clear and unambiguous definition.

Class Hierarchy

Classes can have super and sub-classes which generates a class hierarchy. Further down the hierarchy, each member of a class will inherit the relationships defined on all of its super classes.

Properties

The relationships defined on classes are known as properties and may be object properties or value properties.

Object properties are binary relations meaning that an object property links two individuals within the ontology. These individuals may be in the same class, or different classes depending on the

relationship that is defined. The object properties are defined on the classes meaning that any individuals from the two classes can have that relationship.

Value properties are binary relations that link an individual with a value such as a text string or an integer.

The classes that can have the relationships are defined using domain and range settings for each of the properties. In the subject-predicate-object representation used for the ontology, individuals in the domain classes can be the subject, the relationship is the predicate and individuals in the range classes can be the object.

Class Restrictions

Class restrictions can be used to infer classes from the properties (relationships) already defined within the schema. These may be based on object properties (so if an individual has a certain type of relationship with another individual then it belongs to the inferred class), or they may be value properties (if the individual has an attribute value within a certain range then it is a member of the inferred class).

Resource Description Framework

The Resource Description Framework (RDF) is a knowledge representation language, used for expressing the ontology design and the data.

Triples

The subject-predicate-object triple is the fundamental unit of RDF (Davies, Studer & Warren 2006) and is the means of representing classes, class hierarchy, individuals and relationships within the ontology, as described in Figure 6.

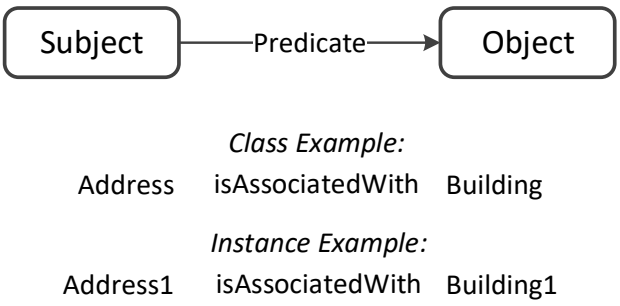


Figure 6: The subject-predicate-object triple.

The standard representation for ontologies.

Triple Store

Triple stores are specialised databases for storing RDF triples. In practical terms, the triple store stores the ontology as well as the data and can perform some reasoning.

2.5 Chapter Review

This chapter has reviewed current methods of population re-distribution and the concepts that must be considered when approaching the problem. Ontologies have been introduced, along with the semantic web technologies required for implementing an ontological model, and finally relevant data sets have been reviewed, with particular focus on the data environment in England and Wales. This is followed by a brief definition of the key concepts that need consideration, or will be utilised in the research design.

The next chapter will introduce the research design and lay out the methods that have been followed in order to meet the aims and objectives laid out in Chapter 1.

Chapter 3 Research Design

This chapter provides an outline of the research design, comprising a brief description of the modelling framework and a conceptual comparison between this and the Population 24/7 model (Martin, Cockings & Leung 2015).

The aim of this research: *to develop a generalizable modelling framework for population estimation that enables high levels of spatial and temporal detail,* is a response to the limitations to spatial and temporal scales at which existing models operate. In particular, there is a tendency for an increase in scale to cause limitations to the spatial extent at which modelling can be performed. For example, the scales at which the Population 24/7 model (discussed in Chapter 1) can effectively work are limited because the model functions only within a region, as there is no adequate means to incorporate visitors to the region, or to account for people leaving the region. Additional visitors can be modelled, but they must be sourced from an origin (outside of the region) and distributed to a destination (inside the region). The model output is a grid so sub-grid population estimations cannot be made. In other models (e.g. Ahola et al. 2007; Zhang, Sunila & Virrantaus 2010; Greger 2014), the trade-off for very high spatial and temporal resolution tends to be a reduction in the spatial extent that can be modelled or the level of detail available in attribution. Where the spatial extent is increased, the spatio-temporal detail, by necessity, must be reduced.

The modelling framework presented here focuses on laying the foundations for an alternative approach to address these limitations of scale, using an ontological approach. The benefits of the ontological approach have already been discussed in Section 2.2, and, in using ontology instead of GIS in this framework, there is one other key benefit. The use of qualitative spatial reasoning means that the representation used for spatial objects is of less importance than the relationships between them: it does not matter whether a building is represented as a point or a polygon, so long as addresses that are associated with the building have that relationship explicitly defined. This means that the data and modelling scales may be increased (or decreased) as required so long as the relationships are defined explicitly.

This ontological model is centred on addresses, which reduces the necessity to model the complexity of the built environment. An alternative approach may be focused on building polygons. However, this is a more complex modelling exercise, partly due to the existence of multi-functional buildings. For example, in the real world, multi-functional buildings such as tower blocks containing retail, office, leisure and residential spaces can be modelled just as easily using the address focused approach, as the topologically less complex detached housing. This address-focused approach allows the model to utilise site-specific temporal activity data at the very local

level, and provides the ability to apply known population patterns for an address to other similar addresses.

This proof of concept modelling framework uses fine-grained spatial data to provide a platform for fine-grained temporal data that is starting to come on-line from various sources, including big data such as that produced by social media, and data sourced from the web.

3.1 Approach Overview

The following sets out some of the concepts and considerations that have influenced the research design.

Addresses represent the bundles and stations of time geography (Hägerstrand 1976; Pred 1977), and as such are the main features of interest in this modelling framework, being the places that people spend their time. In the available data described in Section 2.3, all addresses are associated with a cartographic object, such as a representation of a building, which has a clearly defined spatial extent. The main objects to be modelled are therefore addresses which are *fiat* objects and topographic objects which are the crisp boundaries of physical *bona-fide* objects (Smith & Mark 1998). The model presented here is a place-based object model. Estimated population is an attribute of these features but is modelled using the third type of object: the Temporal Signature (TS).

The TSs used in Population 24/7 (Martin, Cockings & Leung 2015) represents the estimated proportion of maximum capacity that is present at any one time, with population estimates for different addresses represented on a graph with occupation rate on the *y* axis and time on the *x* axis. These TSs will be used to represent a variety of time scales such as minute, hours, days, weeks and seasons. In practice, the available data are likely to have been aggregated into discrete, non-overlapping, interval objects. TSs differ depending on the function of the address that they model. For instance, a library has specific opening times, and busy hours within those opening times. These opening times and busy hours differ considerably from those of other leisure addresses, such as cinemas.

In this model, the focus is on usual cyclical patterns of human activity, for example work and home (Pultar et al. 2010), and those activities that many people are engaged in at once (e.g. leisure, retail). The less frequent, and non-cyclical activities that individuals engage in alone will not be modelled in detail as this research is focused on proof of concept. For each functional class, there may therefore be up to three activity types modelled: residential, work and visitor. Any non-work and non-residential population measure at an address is considered to be a visiting

population. For example, a school may have a working population for the staff and the students may be considered visitors. A single address may have several visitor populations, for example a hospital may have visitors who are patients (with sub-categories in-patients, out-patients and emergency care patients), and also visitors who accompany or actually visit those patients in hospital. Each of these activity types has its own TS (or perhaps more than one), depending on the patterns of activity at the address, for example people accompanying an out-patient during office hours will have different visitor patterns to those who accompany emergency patients at any time of day or night. Each activity has its own population data sources. Residential and working population are aggregate data and as such will be *disaggregated and allocated* to individual addresses based on their functional classification. The values that are allocated to the addresses, in this case, represent the maximum capacity of the addresses. Maximum visitor capacity of specific addresses is either taken from measured data, or estimated based on the height and footprint of the building with which the address is associated. Site-specific visitor populations, such as those at hospitals, education and retail sites, are based wherever possible on measured data.

If the addresses are associated with a single building object, the population estimation will be applicable to the individual building. If however, the address is associated with a group of buildings that comprise a FS, the population can be distributed between all the buildings that are in the FS.

The model operates only in two spatial dimensions, with temporal information representing a third dimension (as an attribute on the two dimensional spatial objects). Addresses are well suited to a two dimensional model as they can be represented as points, and the human activities occur within the buildings, or other topographic objects, that are represented in two dimensions. Although buildings have a height attribute in the topographic data and could be spatially represented in three dimensions, for allocating population to individual buildings in an ontology, the height attribute is used only for estimating the capacity of the addresses within buildings, so it does not need to be modelled spatially.

At this time, temporal occupancy data are not available for many addresses or functional classes. Therefore, a means to add an assumed TS to each individual address is required. For Commercial addresses, this will be achieved by using a default occupancy pattern for different groups of functional classes, and assigning this to addresses that do not have an address-specific TS. In time, not all addresses will require the default TS as real data may be available, but for each class (or group of classes), a default TS will be made available.

3.2 Model Requirements

In order to facilitate population estimation within a predefined area and at a predefined instance in time, the ontological model itself has several requirements:

Firstly, in view of the ability to increase or decrease the spatial and temporal scales at which the model may operate, it must be possible to utilise data from any environment, with a variety of scales in both spatial and temporal domains.

Secondly, it must be possible for the framework to take advantage of the ability to link features for which population data are available, with the population data itself. For example, linking statistical data with the small areas for which they are published and enabling aggregation to the AOI.

Thirdly, the ability to disaggregate population from small areas to addresses must be present. This involves explicitly linking the modelled features via spatial and non-spatial relationships. For example, addresses must be explicitly stated to be associated with a particular building, or a group of buildings (a FS). It must then be possible to redistribute population between relevant features using the explicit relations between them. This will be achieved by associating addresses with the statistical regions that they are associated with, and then distributing a proportion of population to each of the addresses within that statistical region. A means of aggregating the population counts for groups of addresses must also be present.

Finally, the ability to assign several TSs to addresses is essential, allowing data from different sources, and for different activities, to be utilised. In practice, this means that it is possible to model workers separately from visitors and residents, or to model different demographic groups separately.

The modelling framework set out below does not develop all of these features, but does ensure that the mechanisms are present for all of these features to be developed.

3.3 The Framework

The modelling framework is in three distinct stages, outlined in Figure 7. Firstly, data preparation, a stage that will be adapted to the data environment. Secondly, ontology development and data load which will be broadly the same in any data environment, and finally population estimation, which also remains the same wherever the model is applied.

Data are present in the spatial, attribute and temporal domains. Data in the spatial domain include addresses, large scale mapping (topography), and census geographies. The first two of these are used to identify linkages between addresses and cartographic objects in a GIS. These spatial objects and their relationships are then modelled in the ontology. The attribute domain includes residential and workplace population counts from the census data. These are combined with TSs to generate a TS for each group of functional classes. Residential occupancy may vary according to census OAs, according to average occupancy by dwelling type. Combining the relationships between spatial objects and the TSs allows a time sliced population to be extracted from the modelled data for a specified area.

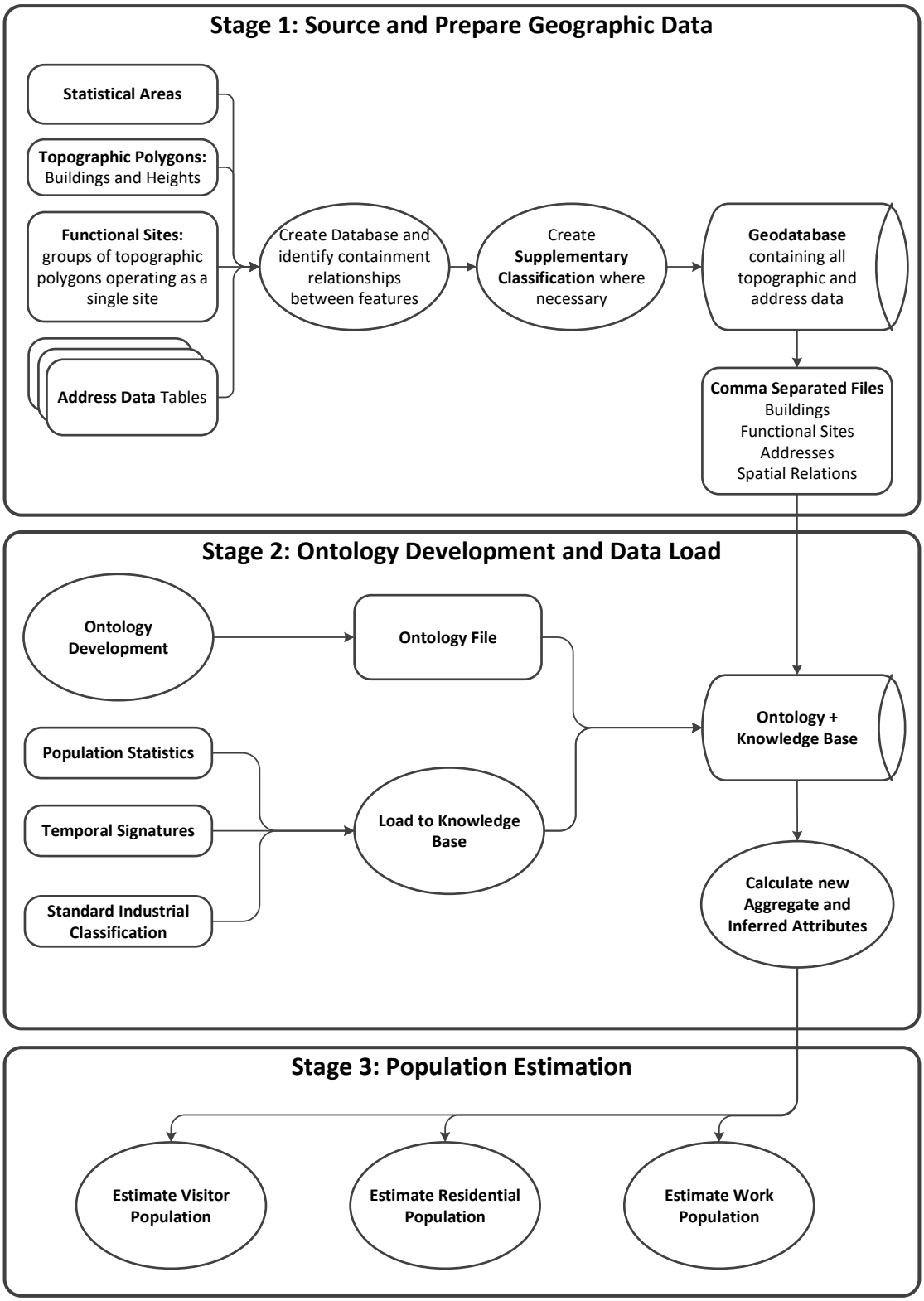


Figure 7: Research Methodology. There are several different ways that the activities could be split between GIS and ontology technologies. The approach taken here is to retain all traditional GIS activity within the GIS technologies and to keep the ontology activities restricted to those that demonstrate clearly the advantages of using the ontology.

3.3.1 Stage One: Data Preparation

This stage utilises ArcGIS and Python Scripts to prepare, load and link locational data (addresses, large-scale mapping and POIs) for specified study areas. The data are first assessed to ensure their suitability for use within the modelling framework. Geographic data sources are then integrated, where appropriate. This stage produces a set of comma-separated (.csv) files that represent all of the addresses, buildings, FSs and the spatial and non-spatial relationships between these features.

For consistency in the data quality, and to ensure all of the relevant information is available for Stage Three, this stage also involves calculation of a supplementary classification that identifies dwelling type for residential addresses.

The data and study area are discussed further in Chapter 4. The supplementary classification of addresses is discussed in Chapter 5.

3.3.2 Stage Two: Design and Populate Ontology

This stage begins with the logical design of the ontology, identifying the *things* that need to be modelled, and the relationships between these things.

This stage utilises OWL in the Protégé ontology editor (Musen 2015) to create the ontology, and uses OntoText's GraphDB Free to convert comma-separated files output from the first stage into an RDF triple store. Once the ontology has been populated, and some preliminary calculations made on the data, the population estimation stage can begin. The concepts that need to be modelled, and the ontological requirement laid out below is presented in Chapter 5.

3.3.3 Stage Three: Estimate Population within Area of Interest

Census OAs are modelled within the ontology with attributes that can be used to indicate the population capacity that should be assigned to each residential address based on the OA aggregate population value and average occupancy by dwelling type. These will be used to estimate residential maximum capacity.

Similarly, census WZ are modelled within the ontology with attributes that allow the estimation of working population capacity of each commercial address. This is based on measured counts of workers in each Standard Industrial Classification (SIC) Section, as reported in the census. The model provides the capability to take the relative size of the commercial addresses into account, although this is not implemented here.

These disaggregations are sufficient to demonstrate the ability of the modelling framework to distribute population between addresses, in different ways depending on address function.

The population is then estimated for the specified time by querying the ontology. The query needs to identify the functional class of each address, which will indicate the TSs for the address and the *per cent* of maximum capacity that the address is expected to contain at the time of day, for each of the three main activities: worker, residential and visitor. Finally, the population at the requested time is calculated and aggregated within the AOI. In this research, the AOI is a WZ because each AOI can be characterised by its Classification of Workplace Zones (COWZ) type.

All of these calculations are executed using SPARQL queries within GraphDB, on the triple store generated in Stage Two. The detail of the population estimation techniques is presented in Chapter 7.

3.3.4 Validation

Validation is carried out separately on each of the three Stages of the modelling framework. Stage One supplementary classification is validated using Google Street View (GSV) to inspect a sample of the classified addresses. Visual interpretation of GSV has previously been successfully applied to manual image interpretation for finer details than required for this validation (Diakakis et al. 2017). Stage Two is validated by comparing synthetically generated simple data analysed in ArcGIS with the data in GraphDB. The output from the third and final stage is the population estimation, which is validated against temporally detailed data provided by a mobile network operator. These data represent the number of devices on the network, and are aggregated to WZ geographies.

3.4 Conceptual Comparison with Population 24/7

In common with Population 24/7, this modelling framework aims to facilitate extraction of estimated population data for any space-time slice using disaggregation of small area population data to finer scales. Both methods use the allocation of known population counts for estimating the number of people at different locations. There are, however, some conceptual differences between Population 24/7 and the framework presented here.

The first and most significant difference concerns the general approach. Population 24/7 moves counts of people from origin to destination containers. Counts of population in origin containers reflect the residential population. Numbers of people at destinations are estimated by re-distributing these origin counts, based on intelligence from several sources of measured

population. The total population in the origin containers can be reallocated between the containers but cannot be lost or gained, so the method is volume preserving. Allocation is calculated based on the temporal occupation patterns of the destinations, described in a TS. This *top-down* approach uses as much data as is available to disaggregate the population that is published as small areas.

In contrast, the modelling framework presented here, while addressing the same requirements, takes a *bottom-up* approach. It estimates counts of population that would be expected to be at individual addresses based on the proportion of the maximum capacity that is likely to be present, which in turn is based on the TSs associated with the address. TSs are sourced from the Population 24/7 NRT project (Cockings et al. 2017). Maximum capacity may be based on the size of the building with which the address is associated (accounting for any other addresses also associated with the same building), or taken from measured maximum counts. Measured population at addresses or FSs can be used in the place of estimated values, or for the generation of TSs that may be applied to similar sites. In common with Population 24/7, small area residential and work statistics are used for estimating populations engaged in those activities. This modelling framework is not concerned with where the people have come from, only with where they are likely to be at the specified time. It is using information from the built environment combined with TSs and known population counts from register data to estimate the population.

The Population 24/7 model is very good at volume preserving, as known numbers of people taken from the census small area statistics are distributed to new destinations. The population at destinations is determined by the time profile of all the destinations in the area. By contrast, this modelling framework does not attempt a full accounting of population and does not attempt such a reconciliation of the population data, as the bottom up approach does not lend itself to that. Conceptually, TSs in this modelling framework are a notional capacity of the address and the model is driven by the size of the containers rather than the available population from the statistics. This approach means that visiting populations from outside or inside the AOI are dealt with intrinsically, based on the population container size and the occupation levels.

Population 24/7 utilises a background mask that represents the inverse of measured population data and transport networks (derived from the relative densities of road traffic). The background mask is a weighting layer and includes rasterised primary routes and calculated average annual daily flows of traffic, and offsets in 17 time periods over a week. By contrast, this modelling framework does not attempt to account for transportation and an extension to the framework to account for linear transport features will be necessary for this.

This approach is in the same research space as Population 24/7, and is closely related to it, but sits alongside that research as an alternative approach to population estimation.

3.5 Chapter Review

This chapter has outlined the general approach to meeting the model aims to create a generalizable modelling framework for population estimation, at high levels of spatial and temporal detail. The focus is on the investigation into the use of semantic web technologies in this research domain, and on proof of concept rather than attempting to produce a fully functioning system, or a detailed data product. The central activity is the ontology development, which provides a structured model into which to load and integrate the data (prepared in Stage One), and the simple data model to facilitate the population estimation (tested in Stage Three).

The following chapters present the three stages that comprise the modelling framework. Data preparation and the supplementary address classification are presented next, in Chapters 4 and 5, followed by a detailed ontology development outline in Chapter 6, and the approach taken to test for appropriate functionality in Chapter 7.

Chapter 4 Data and Study Area

This chapter introduces the data used in the modelling framework, and provides an explanation of how the study area was selected. The datasets, set out in Table 4, include core geographic data sets, attributed statistical regions and source data for TSs and validation.

The geographic data have been tested for their suitability for the modelling framework and for their combined functioning, for the purpose of spatial analysis. The methods used for that suitability testing are presented in this chapter.

The sourcing, pre-processing and quality checking of data for the population estimation brought to light the need for the supplementary classification of the address data. This process is necessary despite the very high quality data available in the GB data environment. It is therefore likely to be required in other data environments where the quality may not be so high. The process is described in Chapter 5. The methods used for sourcing and preparing the geographic data, as well as generating the supplementary classification, are laid out in Figure 8.

The types of data required, and their possible sources in GB have already been discussed in Section 2.3. The potential core geographic datasets that are assessed include OS MasterMap Topography (MMT) with building height and Functional Sites (FS) as well as AddressBase Premium (ABP).

Residential and working population are sourced from the Office of National Statistics, from the decennial census: the small area geographies and a small selection of their associated statistics are utilised. Finally, the TSs are largely sourced from the Time Use Survey (2014-2015) via the Population 24/7 NRT project as discussed in Section 4.5.4.1, with some additional TSs generated from GPT, Gov.co.uk data on school capacities and NHS statistics for hospital data. These hospital population datasets are not tested for their suitability, as they are official statistics. Their use within the modelling framework is described in this chapter.

Definition of the study area is also discussed in this chapter.

Table 4: Overview of Model Data Sources

Type	Data Set	Product Name	Reference Date	Extent Used	Source	Licencing	Use
Core Geographic Data	Topographic Data	MasterMap Topography	2014	Southampton Study Area	Ordnance Survey	Under Licence	Topographic and Address Data for explicit definitions of spatial relations
	Building Heights						
	Functional Sites						
	Addresses	AddressBase Premium					
Statistical Regions	ONS WZs	Workplace Zones	2011	England & Wales	Office for National Statistics: NOMIS	Open	Small Areas with population statistics to be disaggregated
	ONS OAs	Output Areas					
Tabular	ONS WZ Statistics	WP601EW Employment Status		South West			
		WP605EW Industry					
	ONS OA Statistics	QS103EW Age By Single Year					
		QS401EW Accommodation Type People					
Temporal Signatures	Hospitals	A&E	2018	Regional A&E and Trust Level Data	england.nhs.uk		Number of patients at hospital site
		In Patients					

Type	Data Set	Product Name	Reference Date	Extent Used	Source	Licencing	Use
		Out Patients		Trust Level Data			
	Schools	Gov.uk schools data		Southampton and Hampshire	Gov.co.uk		Number of Pupil Visitors at each school
	Selected Commercial Addresses	Google Popular Times	2018	5 addresses only	Google Popular Times	Public Domain	Estimate TS
	Time Use Survey	Time Use Survey	2015	Great Britain	Population 24/7 (UK Data Service)	Under Licence	Identification of SIC Groups and their TS
Validation	Mobile Network Data	Device Location	1st February to 30th April 2017	Southampton Study Area	Mobile Network Operator	Not Able to Publish	Validation of

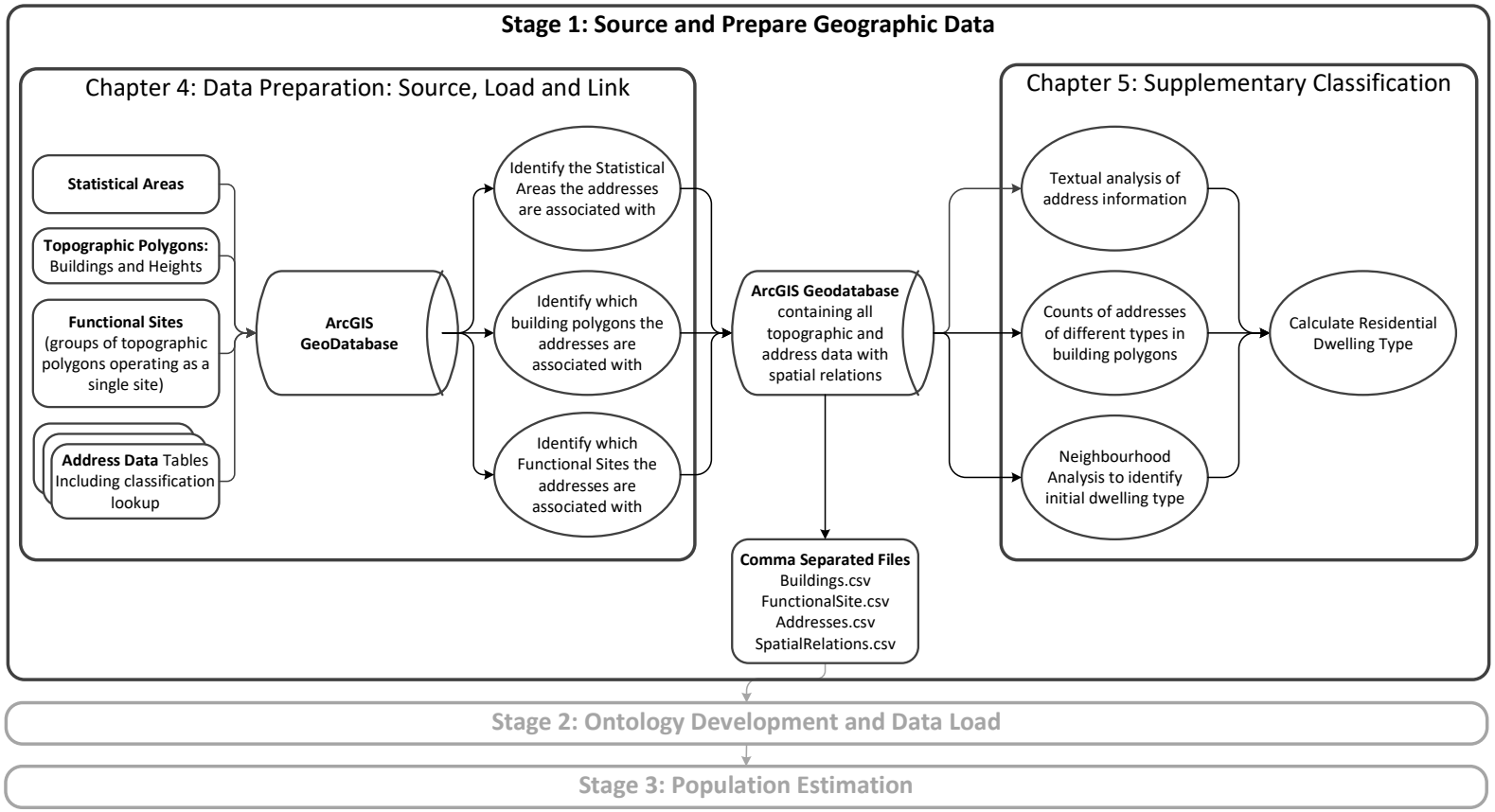


Figure 8: Data Preparation in the context of the modelling framework: Source and Prepare Geographic Data. Comprising activities in Chapters 4 and 5

4.1 Addresses

The address dataset selected for suitability assessment is ABP. The model requirements for this dataset are set out below.

4.1.1 Address Data Requirements

The primary feature for estimating population is the address. Summing the population estimation for addresses within buildings is a more straightforward exercise than attempting to construct a TS for every type of multi-functional building. It also requires fewer assumptions to be made with respect to both multi-functional classification, and the occupation patterns themselves.

The use of address functional classification is explicitly attributed within the ABP data. Addresses also contribute spatial and attribute information related to the size and type of the building with which they are associated. For example, the number of addresses within a building and whether the building is characterised as business, private or a combination of the two.

In cases where the residential classification is inadequate, the address data should also provide the ability to supplement the functional classification by inspecting the textual content of address fields. The address attributes may hold clues about the function of the address within the textual information, e.g. the word “apartment” or “flat” in the first line of an address indicates that more than one address can be found within a single building object.

4.1.2 Structure of the AddressBase Premium Product

AddressBase Premium, the most detailed product in the AddressBase range, is supplied under licence as a relational database. It consists of one main Building Land and Property Unit (BLPU) table, with several other associated tables (Ordnance Survey 2013a). Figure 9 provides details of the different tables present in the ABP dataset. The core of the ABP product is the BLPU which is a spatially referenced, uniquely identifiable record representing a single address (Ordnance Survey 2013a) The BLPU is linked to the other attributes contained within the ABP via a key (usually the UPRN) and relationships which are mandatory or optional depending on the linked table, as indicated in Figure 9 (*ibid).*

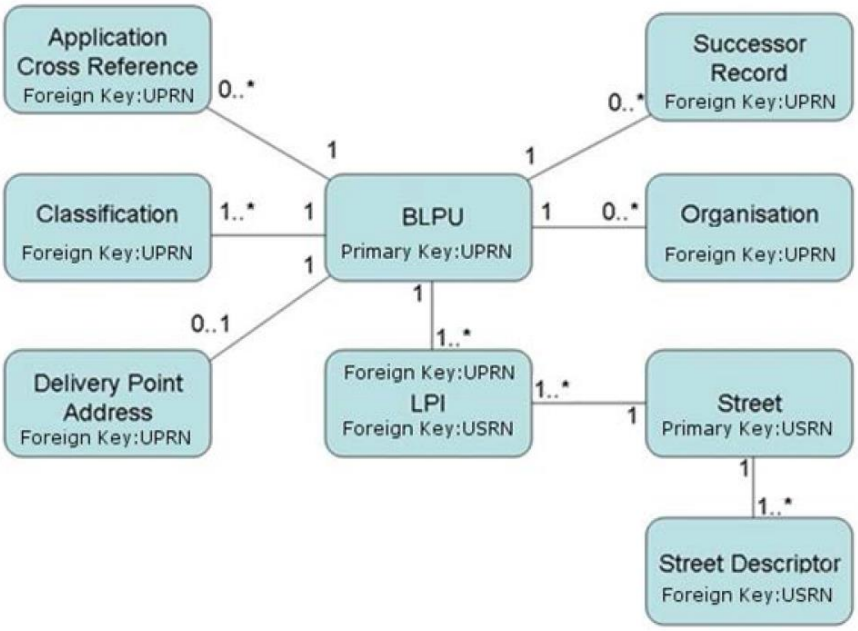


Figure 9: AddressBase premium Model Overview The key relationships between tables in ABP. Tables are linked to the BLPU table via the UPRN with various different relationships (Ordnance Survey 2013a).

The BLPU includes attributes such as a Multiple Occupancy Count (MOC). This is a count of all child UPRNs and may aid in identifying multi-occupancy buildings i.e. those buildings that have more than one *address record* associated with them, such as self-contained flats within a single building, (this is not the same as HMOs, which are shared households with only *one address,* such as student houses). There is also a Parent UPRN field, which indicates whether a BLPU is part of a building with more than one address, i.e. a multi-occupied building. This table also includes a BLPU state field that indicates whether a property is under construction, in use, unoccupied, no longer existing or whether planning permission has been granted (Ordnance Survey 2013a).

An Application Cross Reference (XREF) table allows linkage via the UPRN between a single BLPU and related records in other databases. For example, the BLPU may have an entry in the XREF table that supplies the TOID of the related polygon in the MMT layer and another entry that supplies the TOID of the related feature in the AL2 (Ordnance Survey 2013a). The BLPU includes spatial coordinate attributes, and can therefore be used to generate point features.

A classification table supplies the classification of the buildings via the UPRN key. The classification is a key feature of this dataset that needs to be assessed for suitability and relevance to the methodology. The classification is supplied at one of four hierarchical levels: primary,

secondary, tertiary or quaternary. Table 5 provides a complete list of the primary classification codes (for a more detailed classification description see Appendix A).

Table 5: Primary Classification and Source of addresses in ABP.

Primary Code	Description
C	Commercial (attracts non domestic rates and/or use is of a business nature)
L	Land
M	Military, Military Defence Site
O	Other
P	Parent Shell (property shell, where the BLPU does not attract its own cross references, or street shell)
R	Residential
U	Unclassified
X	Dual Use (a temporary classification for BLPUs attracting both Council Tax and NNDR pending creation of separated BLPUs, and also for Living/Work units that attract both residential and commercial rates)
Z	Object of Interest

For each BLPU, the class is defined by the local government contributing authority, except for those in class “Other” that are defined by OS. The relationship between BLPU and classification allows more than one classification per BLPU, and there must be at least one classification per BLPU (Ordnance Survey 2013a). The classification will be considered as adequate if it is available at the tertiary level. This is the level at which dwelling type is indicated for residential addresses, and has enough detail to distinguish different activity patterns for commercial addresses. The primary and secondary classification levels are inadequate for determining the activities occurring at the different SHAs.

To demonstrate the requirement for tertiary level classification, the Commercial primary class has 14 secondary classes associated with it: Agricultural, Community Services, Education, Hotel/Motel/Guest House, Industrial (manufacturing), Leisure, Medical, Animal Centre, Office, Retail, Transport, Utility, Emergency Services, and Information. In itself, the secondary classification will also not have enough inherent information to determine the functional class at those locations. Considering the Leisure secondary class as an example, a cinema (CL07CI) will have a very different pattern of human activity to, for example, a Cricket Facility (CL06CK). The detail required is therefore at least to the tertiary classification.

Buildings classified as Commercial at the primary level could include a very large range of activities. Three such types are offices (secondary classification) with high densities of people occupying them, agricultural classes (secondary classification) such as fisheries (tertiary classification) which are likely to have few people working at them, and other agricultural classes (tertiary classification) which may have many or few people engaged in the commercial activity, depending on the time of year. This again highlights the need for tertiary classifications to be present for addressing questions relating to human activity.

For residential classes, 'Dwelling' is the secondary class (other secondary residential classes include 'Ancillary Building', 'Car Park Space', 'Garage', 'House with Multiple Occupation' and 'Residential Institution') a tertiary classification is also required in order to identify the dwelling type of the residential property with secondary classification RD.

Entries in the BLPU table have an optional relationship with entries in the Delivery Point Address (DPA) table. This means that a BLPU does not need to have an associated DPA (which is the case for non-addressable objects), and can have a maximum of one DPA associated with it. The DPA table contains the address, including organisation name (Ordnance Survey 2013a). Non-addressable objects cannot be occupied; they include addresses in classes such as Automated Teller Machines (ATMs) and Advertising Hoardings. These do not need an estimated population.

An organisation table includes the organisation name, linked via the UPRN. This is the name on the building fascia (Ordnance Survey 2013a). It is not mandatory for a BLPU to have an organisation record associated with it, but there may be more than one record for each BLPU.

The BLPU is also linked to a table containing Local Property Identifiers (LPI). The BLPU must have a record in the LPI table, and can have more than one LPI record. This table includes a field indicating the building name or description (Ordnance Survey 2013a). The LPI table links to street records (including street classifications and start/end coordinates), which in turn link to street descriptor tables.

4.1.3 Alternative Address Classifications

The classification of addresses could be provided by an alternative source, such as POI data supplied by OS in the PointX dataset. These have their own classification, which differs from the ABP classification and links to SIC codes via a lookup table supplied with the data. This would however, require suitable database or spatial linkage between the BLPU and the POI so that an unambiguous connection could be established.

Alternatively, the classification of establishments monitored by the Food Standards Agency (FSA) may add further information. Again, these different classifications can only be used where there is a precise and unambiguous method of linking the datasets.

Finally, the classification could be supplemented from within the address data itself, using the supplied attributes. For example, the first line of an address may include an organisation name, which can supply clues to the function of the address. The most obvious examples are organisation names that include words such as “nursing home” or “guest house”.

Where the selected address database does not have adequate tertiary level classifications, alternative classifications could therefore be used.

The data (addresses, buildings and census) selected for the modelling has a clear heritage that should not be downgraded by adding unnecessary uncertainties from data of unknown provenance. As this research is concerned with proof of concept, rather than introducing potential error through the assumptions that would need to be made with data that are not explicitly linked, and in light of the need for the modelling framework to be applicable in alternative data environments, a decision has been taken not to supplement the commercial classification of addresses. The result is that some of the commercial addresses will be modelled at inadequate classification detail, which provides an opportunity for demonstrating the application of the modelling framework for data with variable quality.

4.2 Large Scale Mapping

The large-scale mapping dataset selected for suitability assessment is the OS MMT layer with OS MMS layer, which is supplied under licence. The modelling requirements for these two datasets are set out below.

4.2.1 Large Scale Mapping Data Requirement

Large-scale mapping is required to provide detailed cartographic objects that can be used for identifying the explicit spatial relationships with addresses. The large-scale mapping is combined with the address data using spatial joins and database linkages. The functional classification of the addresses within buildings is used to indicate counts of addresses used for different purposes within buildings. These counts, in turn, are used for calculation of the relative size of addresses, as well as whether buildings are multi-functional. Because this research is a proof of concept, the cartographic features used are limited to building areas, and the calculation of relative address size is deliberately simple.

The areal component is important as it can be used, in conjunction with height attributes where these are available, to estimate a floor area for the cartographic object.

The building objects can therefore be used for the following purposes:

1. Estimating the floor-space in a building from its building height and its area.
2. Identifying where flats are present but are not identified from the address information alone, by using topological analysis to find multiple residential addresses present within individual buildings.
3. Identifying, through topological analysis, where the presence of multiple addresses in a single building indicates a multi-functional building.
4. Contributing, by means of topological analysis, to the tertiary level classification of the residential addresses that do not already have this (see Chapter 5).

This final use of the building objects will utilise similar methods to those of Orford & Radcliffe (2007) which involved using topology (adjacency and number of neighbours) to identify dwelling type (detached/semi-detached/terrace/flats) which is equivalent to the tertiary level of classification on addresses. This will not be necessary for all residential buildings: approximately 30% of residential addresses within the Southampton Study Area (SSA) have only secondary classification.

The purpose of the Large Scale Mapping FS layer will be to identify those groups of area polygons that constitute a single site of human activity for the purposes of identifying a TS. It is possible, for example, that population data are available for individual sites but not for the constituent buildings, such as in a school. In these cases, the population could be distributed in a number of ways.

4.2.2 Features of the OS Master Map Topography Layer

Each theme within the MMT area layer does not constitute a classification; instead, it is a set of features grouped together into *Descriptive Groups* in order to make selections within the data easier. Figure 10, shows the area features within a small section of the topography layer, displayed in different colours according to their themes and their descriptive groups. There are 21 different descriptive groups. The feature code attribute identifies unique combinations of the feature type, descriptive group and descriptive term attributes (Ordnance Survey 2014). The descriptive term adds further classification information.

The Buildings Theme (displayed in light orange in Figure 10) represents “roofed construction, usually with walls and being permanent” (Ordnance Survey 2010, p29), and these include

buildings with a calculated area of greater than eight square metres. There is no distinction made between different functions of the buildings (i.e. residential, commercial or industrial) and the class is not limited to brick buildings. Permanent mobile and park homes are also included, as are glasshouses (greater than 50 square metres), archways and covered passageways (Ordnance Survey 2014).

Two additional themes are of interest for the extensibility of this research, but are not implemented here. The Land Theme (displayed in green in Figure 10) represents "man-made and natural features that delimit and describe the surface cover, other than routes of communication and buildings" (Ordnance Survey 2010, p30). This layer includes parks, playing fields, football pitches, gardens, car parks, woodlands. All of these listed features are places where human activity takes place, and are therefore of interest for modelling population. Open spaces are not currently modelled in this framework, although as previously mentioned, the model is capable of this. It is worth noting that it is possible for the land features to overlap in some circumstances and this has implications for processing of the geographic data when analysis is carried out.

The Water Theme (displayed in blue in Figure 10) is also of interest because it includes swimming pools, lakes and lochs, and rivers and ponds, which are also places that people visit.

4.2.3 Features of the OS MasterMap Sites Layer

There are seven themes within the MMT Sites layer: Air Transport, Education, Medical Care, Rail Transport, Road Transport, Water Transport and Utility or Industrial. There are 30 attributed *functions* within these themes.

A single site may be attributed with more than one function. It may also be contiguous or disparate (as in the case of a University with several campuses). All sites have their own TOID and the same level of versioning information as the other MasterMap features (including those in the MMT layer). Each site also has an attribute UPRN that links the site to the address of the primary building within the site (Ordnance Survey 2013c).

Figure 11, shows examples of two MMS sites within the Exeter sample dataset. The label is the Distinctive Name attributed to the feature. This figure demonstrates the number of topography features that can be incorporated into a single FS. Topography features from more than one theme can be incorporated into a single FS.

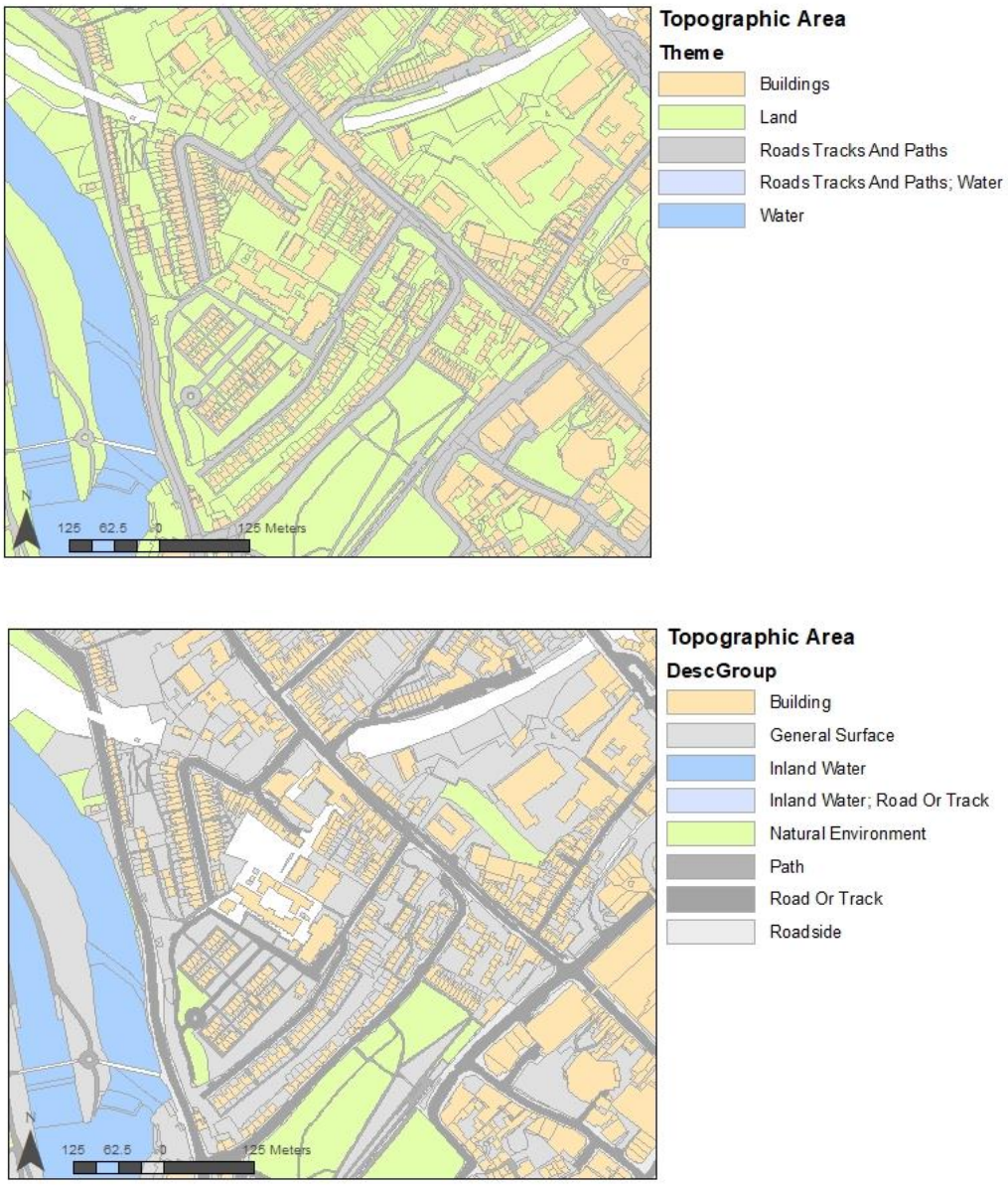


Figure 10: MasterMap Topography Layer Topography Area features by Theme (top) and by Descriptive Group (bottom). The polygons in the Topography Layer do not overlap and are each assigned one or more of nine themes. In the top map, six themes are represented, and there are features in the sample data that are assigned both the *Land* and the *Roads Tracks and Paths* themes (these are displayed as members of the Land theme). In the bottom map, 11 descriptive groups are represented, and there are features classified with more than one descriptive group (*Path* and *General Surface*), displayed here as members of the Path group. The area covered is part of the standard Exeter sample data area provided by OS. *© Crown copyright and database rights 2015 Ordnance Survey.*

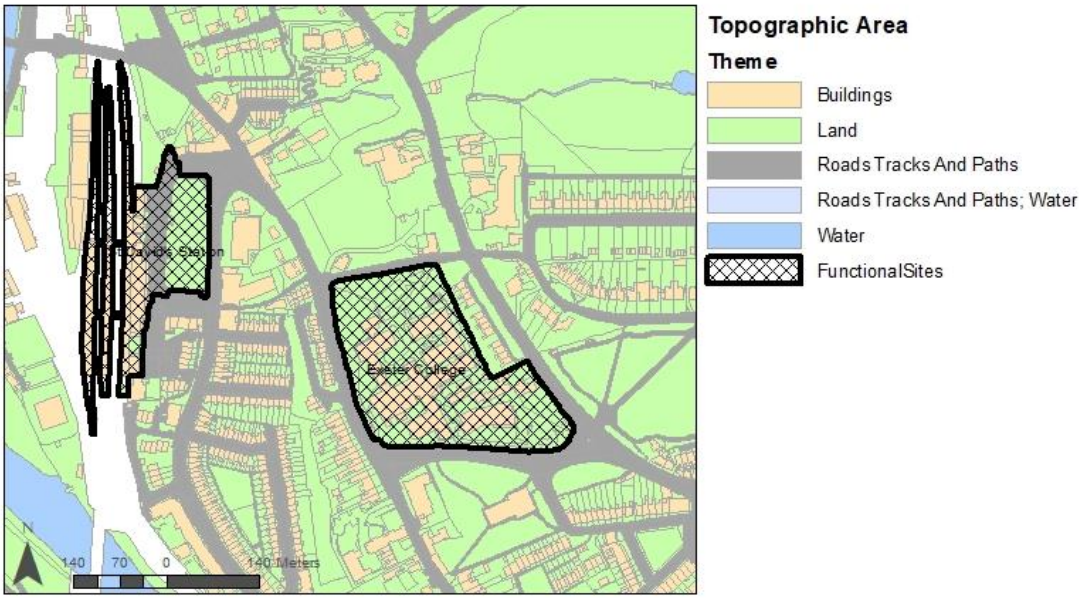


Figure 11: MasterMap Sites Layer Examples All features that are contained within the site are completely contained within it; there are no overlapping Topography Area features for any of the MMS sites in the sample dataset. *© Crown copyright and database rights 2015 Ordnance Survey.*

4.3 Linking Topological and Address Data

Address data can be joined to the topological features in one of two ways. The first is by database join utilising the XREF attribute within the ABP layer (using table join in ArcGIS), and the second method is via spatial join in ArcGIS. The spatial join method should only be used where there is no database join available.

4.4 OS Change Only Updates

OS Products are delivered as complete datasets on first delivery and updates are delivered as Change Only Updates (COU). In a COU, only records that have changed are supplied. This includes insert, deletion or change of features. Set procedures can be followed to update the core dataset with these changes. All of the data delivered by OS for this research were delivered as a one-time only delivery. This means that change update procedures are not required for any of the datasets.

4.5 Population Data

Population data are divided into residential, working and visitor activities in the modelling framework. The residential and workplace population are to be disaggregated from their small

areas to the addresses within those small areas to represent the maximum capacity of those addresses. Both residential and workplace population data, are supplied by ONS via the NOMIS website (ONS 2018a) under the OGL. Many statistical tables are made available through this website. Two workplace population datasets and two residential datasets have been selected for use in the modelling framework. These are the minimum required for population estimation in this data environment, although with further development the population estimation algorithms could utilise many more of these datasets.

4.5.1 Residential Population

Two residential population tables are used:

1. Age by Single Year QS103: These data, derived from the date of birth question in the census, are supplied for 2011 OAs in the following categories: All categories: Age; Age under 1; Age 1; …; Age 99; Age 100 and over (Nomis 2017a). They will be aggregated within the model to reflect different broad activity patterns as indicated in Table 6.
2. Accommodation Type People QS401: These data, supplied for each OA, provide an estimate of usual residents by the dwelling type of the accommodation that they live in (Nomis 2017b). The source data are aggregated within the model according to the rules in Table 7. Dwelling type has been identified for the topographic data that addresses are associated with, so this can be used to distribute population accordingly.

Table 6: QS103: Age by Single Year Aggregation into different activity groups

Aggregated Variables	**Activity Group**
Age 0 to 4	Pre-school
Age 5 to 9	Primary School Age
Age 10 to 17	Secondary School Age
Age 18 to 24	Student or Employed
Age 25 to 64	Working Age
Age 65 and over	Retirees

Table 7: QS401: Accommodation Type aggregations: provides an estimate of usual residents by dwelling type

QS605 Category	Aggregation
Unshared Detached	Detached
Unshared Semi-detached	Semi-Detached
Unshared Terraced (including end-terrace)	Terrace
Unshared Flat, maisonette or apartment: Total	Flat
Unshared Purpose-built block of flats or tenement	
Unshared Part of a converted or shared house (including bed-sits)	
Unshared Flat In commercial building	
Unshared Caravan or other mobile or temporary structure	Temporary
Shared dwelling	

There are several additional residential population measures, which could be utilised in a more sophisticated algorithm to estimate the maximum capacity of addresses than that applied in this modelling framework. Examples include:

1. Communal Establishments KS405UK. These data could contribute to a more realistic estimate of maximum capacity for those addresses that are recognised as CEs. These include those in the specific commercial residential classes that can be used to identify those CEs.
2. Age By Residence Type LC1104: usual residents by residence type (household or communal resident) and by age. These data could contribute to a demographic sub-group breakdown.
3. Rooms Bedrooms and Central Heating KS403EW: average rooms by household, average bedrooms by household. These data could potentially supply a more detailed estimated maximum capacity for addresses.
4. Adults not in employment KS106EW: includes long-term health disability and dependent children. These data can indicate a change to the assumed TS for the addresses inside the OA.
5. People aged 16-64 Living in One Person Household QS117EW. This table could be used to adjust the maximum capacity of addresses.

6. Household Type QS116EW, indicating the living arrangements of the inhabitants, e.g. cohabiting couple, lone parents, married couple families. This could be used to adjust the maximum capacity of addresses.
7. Household Size QS406EW. These data indicate not the address, but the household, as defined by the ONS (see Section 2.4).
8. Dwellings QS418EW, which indicates shared or unshared dwellings. Assuming they are correctly identified, this could be used to indicate the population counts to distribute between HMOs.

4.5.2 Workplace Population Data

Workplace populations are published for WZs: small area polygons. One workplace population dataset is utilised in this implementation of the modelling framework: QS605 Industry: which contains estimates that classify usual residents aged 16 to 74 in employment the week before the census in England and Wales by industry (Nomis 2016), classified by SIC Section. These industries are self-reported by individuals completing their census form. These data can be aggregated within the model according to the "SIC Groups" described in Section 4.5.4.1. They can then be disaggregated to the addresses whose classification belong to the SIC group.

Table 8: QS605: Usual Residents by Industry (SIC Sections)

KS605 Variable: SIC Section	**Industry**
All categories	All
A	Agriculture, forestry and fishing
B	Mining and quarrying
C	Manufacturing
D	Electricity, gas, steam and air conditioning supply
E	Water supply; sewerage, waste management and remediation activities
F	Construction
G	Wholesale and retail trade; repair of motor vehicles and motor cycles
H	Transport and storage
I	Accommodation and food service activities
J	Information and communication
K	Financial and insurance activities
L	Real estate activities
M	Professional, scientific and technical activities

KS605 Variable: SIC Section	Industry
N	Administrative and support service activities
O	Public administration and defence; compulsory social security
P	Education
Q	Human health and social work activities
RSTU	Other

There are several additional workplace population measures, which could be utilised in a more sophisticated algorithm to estimate maximum worker capacity of commercial addresses than that applied in this modelling framework. These include:

1. Employment Status WP601: this indicates the number of individuals in full time and part time employment, those employed and self-employed, and the number of full time students who work in each WZ.
2. Hours Worked WP604: estimates that classify the workplace population in England and Wales by hours worked, which could prove useful for building TSs.
3. Distance Travelled to Work WP702: estimates that classify the workplace population in England and Wales by distance travelled to work, which could prove useful for generating a background layer.
4. Sex by single year of age WP1101: estimates that classify the workplace population in England and Wales by sex and by single year of age, which would be useful for demographic breakdowns or workplace populations in relation to the age by single year breakdown of residential population.

The COWZ categorises WZs for which workplace population are published, based on a range of 48 census workplace, residential and socio-economic variables including workplace populations, ethnic groups, level of qualification, socio-economic class, travel to work and employment characteristics (Cockings, Martin & Harfoot 2015). The result is a breakdown into 29 types of WZ in seven super-groups (Table 9). These are useful in the modelling framework as they allow the characterisation of the WZs used for validation.

Table 9: Classification of Workplace Zones

COWZEW_SGN	COWZEW_GN
Manufacturing and distribution	Business parks
	Industrial units
	Manufacturing, energy and utilities
	Mining and quarrying facilities
Metro suburbs	Cosmopolitan metro suburban mix
	Independent professional metro services
	Metro suburban distribution
	Suburban metro infrastructure
Retail	Eat, drink, shop and be merry
	Low density retail and wholesale
	Market squares
	Multicultural urban high streets
	Shop until you drop
	Traditional high streets
	Rural with core services
	Rural with mining or quarrying
	Rural with non-local workers
	Traditional countryside
Servants of society	Highly qualified workforces and professional services
	Large scale education
	Major hospitals
	Public administration
	Non-metropolitan suburban areas
	Primarily residential suburbs
Top jobs	Administrative centres
	Big city life
	Global business
	Regional business centres
	Science and business parks

4.5.3 Visitor Populations

Visitor populations differ from the residential and workplace populations as they are not all used to represent the maximum capacity of the address.

One example of available visitor population is schools. These require staff and pupil counts, as well as visitors at specific times of day. In this model, for simplicity, these are restricted to staff and pupils. Staff populations are provided by the workplace population and numbers of pupils are sourced from gov.uk schools data. These indicate the number of pupils, capacity of the school, and the age ranges that attend the school (which could be used for demographic sub-group estimation). These data are individual site maximum capacities.

Other addresses may have data available on the web. In the modelling framework's current form, these need to be converted to a maximum capacity and a TS. This is an area for future development: to incorporate population counts directly from visitor statistics published as linked date on the web.

4.5.4 Temporal Signatures

TSs are essentially a graph of percent occupation of an address, a site or a functional class of addresses. In this implementation, the TS can use non-overlapping, variable length intervals, which could last for hours or minutes, but are sub-day lengths. Each interval therefore requires an indication of the day type that it represents. The calendar day is not sufficient as activity patterns are influenced by national holidays, school and university vacations and the timing of Christmas and Easter. The day types used by Highways England (Highways England 2015) are therefore used in this modelling framework and are presented in Table 10. While these are not implemented in this model (in favour of day types aggregated from the Time Use Survey, see Section 7.3), some of the data have been aggregated using these day types. A lookup table is therefore used to match them to the implemented day types. These Highways England day types can be used in the generation of an almanac that indicates the day type for each day of the year so that this can be matched to the available data.

TSs can be subdivided into those that represent a default to use where an address or class specific TS is not available, and those that represent a specific address.

Table 10: Traffic England Day Types (Highways England 2015)

0	First working day of normal week;
1	Normal working Tuesday;
2	Normal working Wednesday;
3	Normal working Thursday;
4	Last working day of normal week;
5	Saturday, but excluding days falling within type 14;
6	Sunday, but excluding days falling within type 14;
7	First day of school holidays;
9	Middle of week: school holidays, but excluding days falling within type 12, 13 or 14;
11	Last day of week: school holidays, but excluding days falling within type 12, 13 or 14;
12	Bank Holidays, including Good Friday, but excluding days falling within type 14;
13	Christmas period holidays between Christmas day and New Year's Day;
14	Christmas Day/New Year's Day.

4.5.4.1 Commercial Temporal Signatures from SIC Groups

Where an address does not have its own TS, or one for its functional class, a SIC Group's TS is applied.

Nomenclature Statistique Desactivités Économiques dans la Communauté Européenne (NACE) is the European Union industrial classification system (ONS 2007). There is European legislation in place that requires the UK to keep up to date in line with NACE. The SIC is therefore based on NACE. It is a hierarchical five-digit system in 21 Sections (indicated by letter A-U). There are 88 two digit divisions and 272 groups (the third digit) followed by 615 classes and 191 subclasses.

The Time Use Survey 2014-2015 is a large-scale household survey, based on time diaries from 10,000 households, in which daily activities are recorded at ten minute intervals, and a work schedule is completed by those respondents who are in work (Gershuny & Sullivan 2017). There are 16000 diary entries which, as part of the Population247NRT project (Cockings et al. 2017) have been used to calculate an estimated temporal signature for each of the 21 high level SIC Sections. This is achieved by diary aggregation to represent the typical daily working patterns for each Highways England day type, for all respondents working in each SIC Section. The SIC Sections

were then clustered based only on their activity profiles and the clusters aggregated to create relative activity levels for each of the different clusters. The result is eight SIC Groups, each with its own distinct temporal activity patterns for workplace attendance. These are: ABDE, C, F, GI, HJ, KLMN, OPQ, RSTU.

The SIC Group of each tertiary ABP class is determined using a pragmatic, common-sense approach. In the absence of existing data, the most appropriate SIC Section for the tertiary class is assigned to the ABP class. The SIC Group is then applied based on the SIC Section assigned. The lookup table can be found in Appendix A.3.

4.5.4.2 Commercial Temporal Signatures for Individual Addresses or Sites

4.5.4.2.1 Hospitals

Hospital data are the most complex of those modelled in this framework. The data have been generated for Southampton General hospital only. The hospital has workers, residents, Accident and Emergency (A&E) patients, in-patients who stay overnight, outpatients who have appointments during the day, individuals visiting or accompanying those patients, and workers and visitors who attend addresses within the hospital site, but are not the hospital address (e.g. a café or gift shop within the hospital grounds).

Data about the temporal patterns of visitors to Southampton General Hospital have been gathered from the NHS website. Most data available here are available at regional level, trust level or Clinical Commissioning Group level. These aggregate statistics are published each month, with a core dataset that is consistently published, and a slightly different set of aggregated statistics published.

The data utilised in this implementation are sourced from the NHS as described in Table 11. The assumptions made in estimating accompanying visitors are likely to lead to over-estimates of the population at the hospital site. The crude assignment of the trust level data to in-patients and outpatients will also lead to some over-estimate of the population on site. This has been accepted as a limitation in this data as any alternative would involve the introduction of more assumptions and potentially a less constant error. However, the general pattern of population estimates should be reflected in the estimates, despite any differences in the magnitude.

Table 11: Hospital Visitor Data Sources: including the assumptions made for accompanying visitors.

Data Type	Source	Description
In-Patients	Trust level data	The number of day-only beds, overnight beds and occupation rates are used to estimate the number of in-patients.
Outpatients	Trust level data	Total referrals over the published time period are scaled to an hourly rate, assuming a constant rate, and length of appointments.
A&E	Regional and trust level statistics	Regional statistics have a daily/hourly breakdown of A&E attendance. Trust level data has hourly attendance across all days. Each hour has been allocated a population count based on the percent of attendances for all days scaled to the trust level data for a daily/hourly breakdown. Values take account of average waiting times at this site.
Visitors: In-Patients	In the absence of a data source, these have been estimated using a pragmatic approach based on the assumption that patients will have an accompanying visitor during visiting hours.	Estimated one visitor per inpatient during the day at visiting times, one per 10 in-patients in the later evening
Visitors: Outpatients		Estimated one visitor per outpatient
Visitors: A&E		Estimate one visitor per A&E patient

The total visitors to the hospital is the sum of each of the separately estimated patients and visitors for A&E and non-A&E. These do not have an impact on the working population at the hospital, which are handled separately in the model. The published statistics can then be used to identify the maximum capacity of the hospital, and the TS represents the percent of this maximum capacity that is present at any time. These data are constructed TSs plus maximum capacity for an individual site.

4.5.4.2.2 Individual Addresses in Other Commercial Classifications

These data include the production of TSs, counts of population or capacity for individuals belonging to classes that are not covered already. The examples in this implementation include a large supermarket, library, public house, bingo hall and medical centre. These data have been manually sourced from GPT, with estimates based on the graph displayed on the page when the specific address is found using search. GPT data are based on Android device presence, which is

subject to similar bias as mobile phone data described by (Jacques 2018). These approximations give hourly estimated TSs over the course of a week. Examples that demonstrate the variety in these TSs are presented in Figure 12. There is no indication of the magnitude of these data, so estimates of maximum capacity for these types of sites are based on size of the address within the building (assuming equal proportion of all addresses within each building).

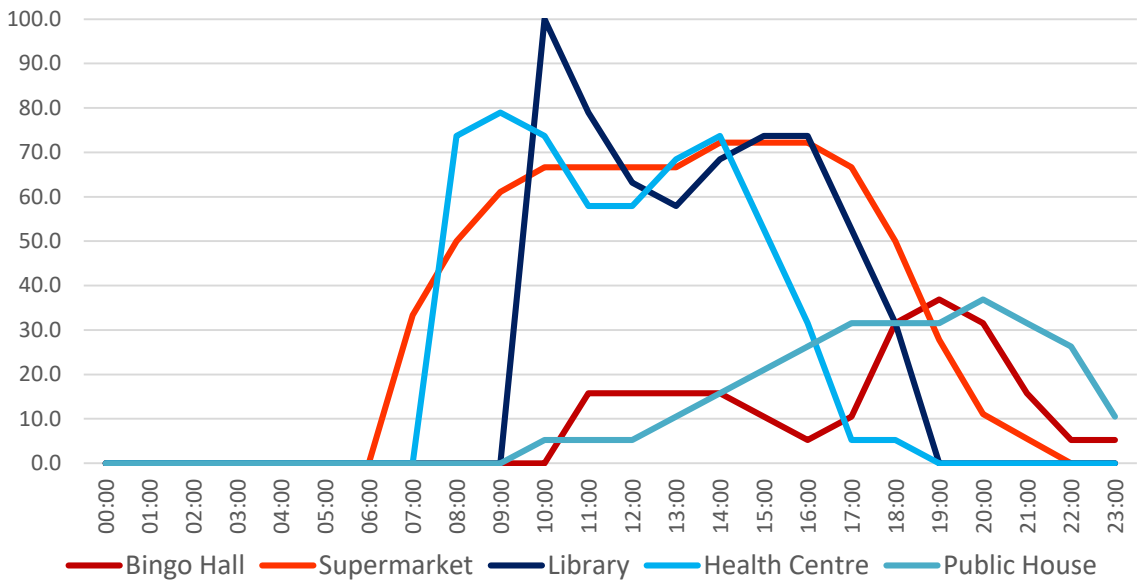


Figure 12: Example Temporal Signatures, sourced from Google Popular Times (Tuesday only)

Once again, there are too many missing values to get these TSs and magnitudes of populations exactly right, so the pragmatic approach is to use the available data to prove the modelling framework is appropriate for the population estimation and data integration task, by modelling the patterns of occupation only.

4.5.4.3 Residential Temporal Signatures

Residential TSs are not used in Population 24/7 because it is a source-destination model. They are therefore calculated as the inverse of the SIC group TSs for this implementation of the modelling framework. This may lead to a degree of double counting, particularly for leisure activities. Ideally, the residential population would be calculated as the inverse of workers *and* visitors to the addresses. However, this is not currently dealt with in this manner, so visitors at addresses may also be counted at their residence.

The variety of the sources of population data and TSs exemplifies the need for integration of data from different types of sources. These are all pre-processed in one way or another, but the model is also capable of integrating linked data (i.e. those with a URI for an identifier) from the semantic web.

4.6 Study Area

4.6.1 Study Area Selection Process

Selection of the study area is based on several requirements. These are:

1. Data quality, particularly for address classifications, which vary geographically depending on the LA supplying the classification, attributes.
2. Building height data availability, which varies geographically as the building heights data were still being captured at the time of data request.
3. Variety of land use and facilities, as a greater variety of land use and facilities included in the study area gives greater opportunity for demonstrating the abilities and shortcomings of the modelling framework being developed.
4. The region in which the study area falls, as a local study area is more accessible for ground-truth data collection for the purpose of model validation.

The study area was carefully chosen to include a variety of land use types to ensure that the model is able to demonstrate its ability to work on any address class. The first stage of selecting a study area involved identifying different land use types that need to be included for effectively testing the modelling framework. To ensure broad applicability of the methods developed it is desirable that a study area should include as many as possible of the following features:

a. Urban and dense urban including town centre
b. Suburban including mixed residential
c. Rural to include rural residential and rural commercial buildings if possible
d. Industrial: both light and heavy if possible
e. Educational establishments: schools, university
f. Hospital and other medical buildings
g. Leisure facilities: cinema, bowling, restaurants, sporting facilities
h. Business parks, Leisure Parks, Retail parks
i. Other retail: high street, suburban retail
j. Places of worship
k. Transportation: motorway, railway station, bus station, airport, seaport

Five prospective areas were selected, all of which included both a substantial urban area and surrounding rural land use. These were restricted to Southern English towns so that multiple visits for ground truth exercises would be possible, should this become necessary. The five candidate towns were Swindon, Oxford, Portsmouth, Southampton and Bristol. Prospective study areas were defined around each of these to incorporate as much variation in land use types in the smallest possible area to ensure that data quantities do not become prohibitively large, in consideration of available computing resources.

As the address classification data are key, the quality of this dataset was assessed for all of the prospective study areas. An initial data availability and quality survey was carried out by OS, revealing that several of the areas had very good commercial address data. They have high proportions of addresses that have detailed classifications to the tertiary level, the requirement for which is described in Section 2.3.1.3. The Southampton area also had a higher proportion of residential addresses with the more detailed, tertiary level classification.

Data availability was also a consideration in study area selection. As the address and large-scale mapping datasets are available for all of GB, the availability of building height attributes on the large-scale mapping that had some influence over the choice of study area. At the time that source geographic data was ordered (June 2015), the Southampton Study Area (SSA) also included a greater area that has height values available for the buildings data within MMT layer. All of the other study areas had height data available only within the town boundary lines whereas for Southampton, this data extends into the rural areas as well as covering Hedge End to the east of the city (Ordnance Survey 2015a). OS offices are located in Southampton, which may influence the level of detail in the data for this area.

4.6.2 Description of Study Area

Southampton is a city with approximately 245,000 inhabitants (Southampton City Council 2015), located on the south coast of England. The SSA illustrated in Figure 13, provides a great variety of different features that are interesting from a functional (and modelling) perspective. These include a local ferry terminal, cruise terminal and docks to the south of the study area, the Eastleigh railway works and marshalling yards, a significant stretch of the M27 motorway and Southampton Airport. The city centre includes a shopping centre and the large West Quay shopping mall, and there is an out-of-town shopping centre at Hedge End. The study area also includes Southampton General Hospital to the west, and two universities (the University of Southampton and Southampton Solent University) with associated halls of residence (CEs) as well as many HMOs. The study area includes large urban and suburban residential areas, with a

sizeable area of rural land use deliberately included. According to the Land Cover Map 2007 (Centre for Ecology and Hydrology 2008), approximately 8% of the SSA is urban, and about 28% is suburban land cover. There are many primary and several secondary schools within the study area. There is also industrial activity occurring within the city, including those activities in the docks area. A variety of leisure facilities is included: both isolated individual indoor facilities and a leisure complex in the city centre as well as outdoor facilities.

The entire study area is 17km (East to West) by 11km (North to South) in size (LL 437000, 109000 , UR 454000,120000) and includes all of the required elements of an ideal study area as described above and confirmed using spatial analysis of the addresses and visual assessment of topography maps of the area.

A further four, smaller study areas have been defined within the larger SSA for the purpose of model development and testing. The first two areas cover approximately 1km square around the Portswood and Eastleigh areas of Southampton (Portswood Study Area, or PSA and Eastleigh Study Area, or ESA). PSA and ESA each contain some MMS sites as well as a mix of residential and commercial properties, and each falls within a different LA. Eastleigh is 26% urban and 33% suburban, and Portswood is 17% urban and 76% suburban. As previously indicated, the data quality varies geographically according to the LA that is supplying the data and this is apparent in these two smaller study areas as described later. PSA and ESA are used for development of the data preparation as well as the ontology development. Two slightly larger areas that include several complete WZs have been defined around Southampton General Hospital (Shirley Study Area) and several educational establishments serving different age groups (Itchen Study Area) and these are used for generating model output as they contain features that are interesting for population estimation.

4.7 Data Preparation

Data that cover the SSA are prepared in two distinct stages. Firstly, acquisition, preparation, loading and linking of the data, and secondly assessment of the data once the first stage is complete. These two stages are outlined below.

4.7.1 Acquire, Prepare, Load and Link Locational Data

4.7.1.1 Data Delivery

OS delivered the core geographic datasets described in Table 12 in June 2015. The data cover the full extent of the study area, and in some cases extend beyond the study area because data are extracted by tile, or because some included features fall partly outside of the defined study area. A full description of these datasets, the data load process and the subsequent data assessment follows.

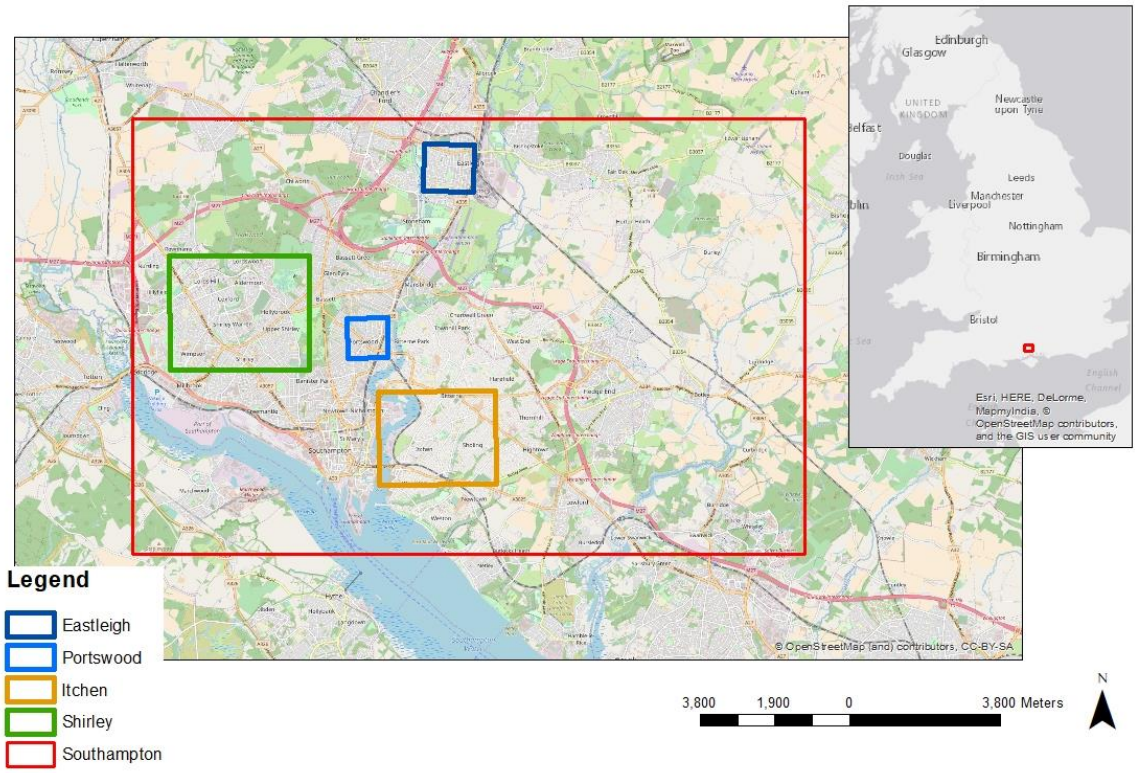


Figure 13: Southampton, Itchen, Shirley, Portswood and Eastleigh Study Areas

Table 12: Data Supplied by Ordnance Survey for the SSA

Data Layer	Features of Interest	Date	Usual Update Frequency	Delivery Format
Master Map Topography	Areas, particularly buildings	Extraction Date 11/06/15	6 week refresh	GML
Master Map Topography Sites	All functional site areas	Release Date April 2015	Bi-annually	GML
AddressBase Premium (Epoch 33)	BLPUs with address detail and classification	Supply Date 16/07/15 Release Note from June 2015	6 weekly	CSV relational database

4.7.1.2 MasterMap Topography

The process illustrated in Figure 14 was designed for the load of all required MMT data, including the topography layer, the MMS layer and the building height attribute. The process incorporates the OS instruction for data load of MMT layers, which is achieved manually, and processes developed within a python script to link the three data sources, appending building height and

MMS site attributes to the areas in the topography layer, as indicated by those processes with bold outlines.

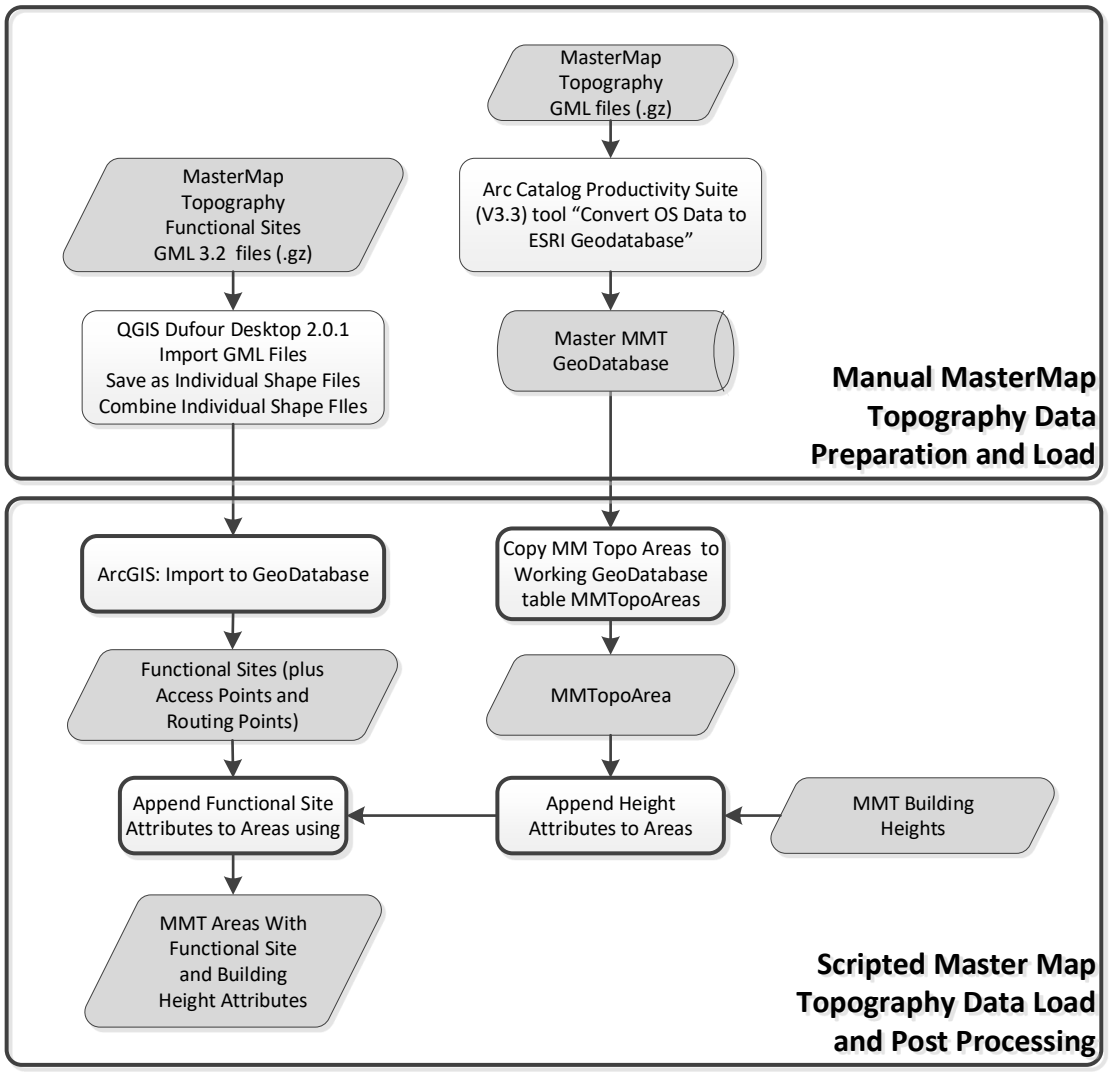


Figure 14: MasterMap Topography layers data load and post processing method Bold outlines are processes that are contained within the main *Data Load and Preparation* python script.

4.7.1.2.1 Initial Data Load

The MMT data layers were delivered in Geography Markup Language (GML) format, which were then manually imported into a master ArcGIS File Geodatabase using the ArcCatalog Productivity Suite (V3.3). For estimating population within at the address level, only the area layer and the FS layer (which is delivered separately) are of interest.

A python script that utilises the *ArcPy* python module (and that is written to incorporate all *Data Load and Processing* operations) was then run. ArcPy is a site package that allows the same geographic analysis used by ArcGIS to be used in python scripts. This copied the area data layer

into another, working geodatabase so that the initial data conversion and load does not need to be repeated, as this process is time consuming.

Figure 14 also describes how MMS sites were loaded into the working geodatabase. These were delivered as tiles of complete polygons (i.e. no single site polygon will be split across more than one tile). As a result, the tiles required merging. This was done in QGIS, as ArcGIS 10.2 is unable to access GML 3.2 format in which the Sites layer was delivered. The resulting shape file was then loaded directly into the working geodatabase for analysis alongside the other imported layers.

4.7.1.2.2 Post Processing

The *Data Load and Preparation* script also includes load of the height attribute comma separated table to the geodatabase for the buildings within the area layer. The attributes were appended to building features within the area layer using a database join.

Additional post processing involved appending the attributes of MMS sites to the areas, as this enables immediate identification of area features that are within a FS, using attribute selection rather than spatial query.

4.7.1.2.3 Extraction of Features of Interest

The Features of Interest (FOI) for classifying buildings are the buildings within the area layer. However, there is a case for also including open areas within the FOIs. This would include roads, pavements and other features that individuals may occupy in transit between buildings, but also features that are destinations in their own right, such as parks, beaches, shopping plazas etc. In a fully developed modelling framework, these will each need to be treated separately.

The FOIs within the MMT Areas layer are therefore buildings. Initially the feature extraction extends to selecting only the buildings from the area layer, and saving this to a separate layer within the geodatabase (using query Theme = 'Buildings' OR Theme = 'Buildings; Roads Tracks And Paths' OR Theme = 'Buildings; Structures'). This occurs within a python script for extracting AOI data.

All features within the MMS layer are of interest so no process is required to extract FOIs from this layer.

4.7.1.3 Result of Data Load

Table 13 describes the result of the data load process for all delivered data, including the different types of MMS sites present. The majority of the MMS sites are educational (207 out of 279), and within this educational function, the majority are primary education (134). 44% of all of the area features are buildings. Figure 15 shows these three topography layers for a small extract of the study area (approximately 400m x 350m in size).

Table 13: Features delivered in the MasterMap Areas layers

Data Source	Number of Records	Note
MMT Areas	486847	14280 of these areas are a part of a FS
MMT Buildings	214733	2723 of these buildings are a part of a FS
MMS	279	Based on the Functional Theme attribute, there are 207 Education sites; 1 Air Transport site; 27 Medical sites; 20 Rail Transport sites; 4 Road Transport sites; 13 Utility or Industrial sites and 7 Water Transport sites.

4.7.1.4 AddressBase Premium

The process illustrated in Figure 16 was designed for the data load and post processing of ABP data. The process combines OS instruction (and supplied scripts) for data pre-processing with processes developed within a python script to load the address data into the working geodatabase, generate points from the BLPU table, and append required attributes to the main BLPU layer.

4.7.1.4.1 Initial Data Load

The load process for ABP involves manually merging the delivered geographically chunked files using a supplied *gawk* script, adding headers to the individual csv files using another script supplied by OS, and finally splitting the delivered .csv files by record type, using a final *gawk* script, also supplied by OS.

Using the *Data Load and Processing* python script, these tables were then imported into an ArcGIS geodatabase. The BLPU table was used to generate a point feature layer and the additional tables were linked to this (or each other) using relates (or joins) within ArcGIS. A classification scheme file was supplied by OS. This was also loaded directly into the geodatabase and linked to the BLPU layer using the classification code attribute.

(a)

(b)

(c)

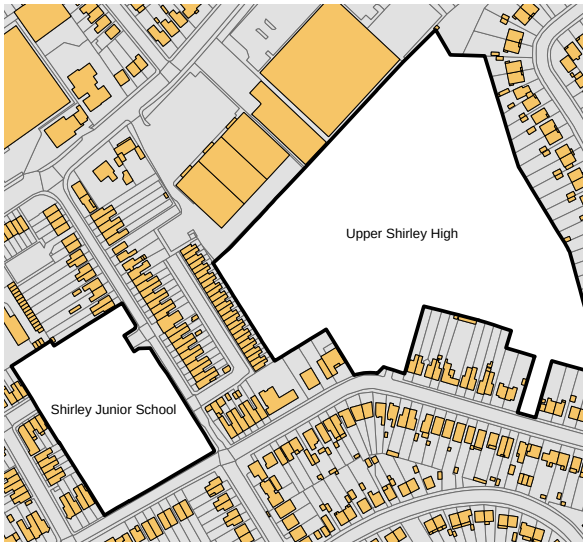


Figure 15: MasterMap Topography layers data load and post processing result: for a small extract of the study area (a) shows all polygons within the areas layer; (b) overlays all polygons coded as buildings in the Theme attribute, in orange; (c) adds the MMS polygons to this as a thick black outline *© Crown copyright and database rights 2015 Ordnance Survey.*

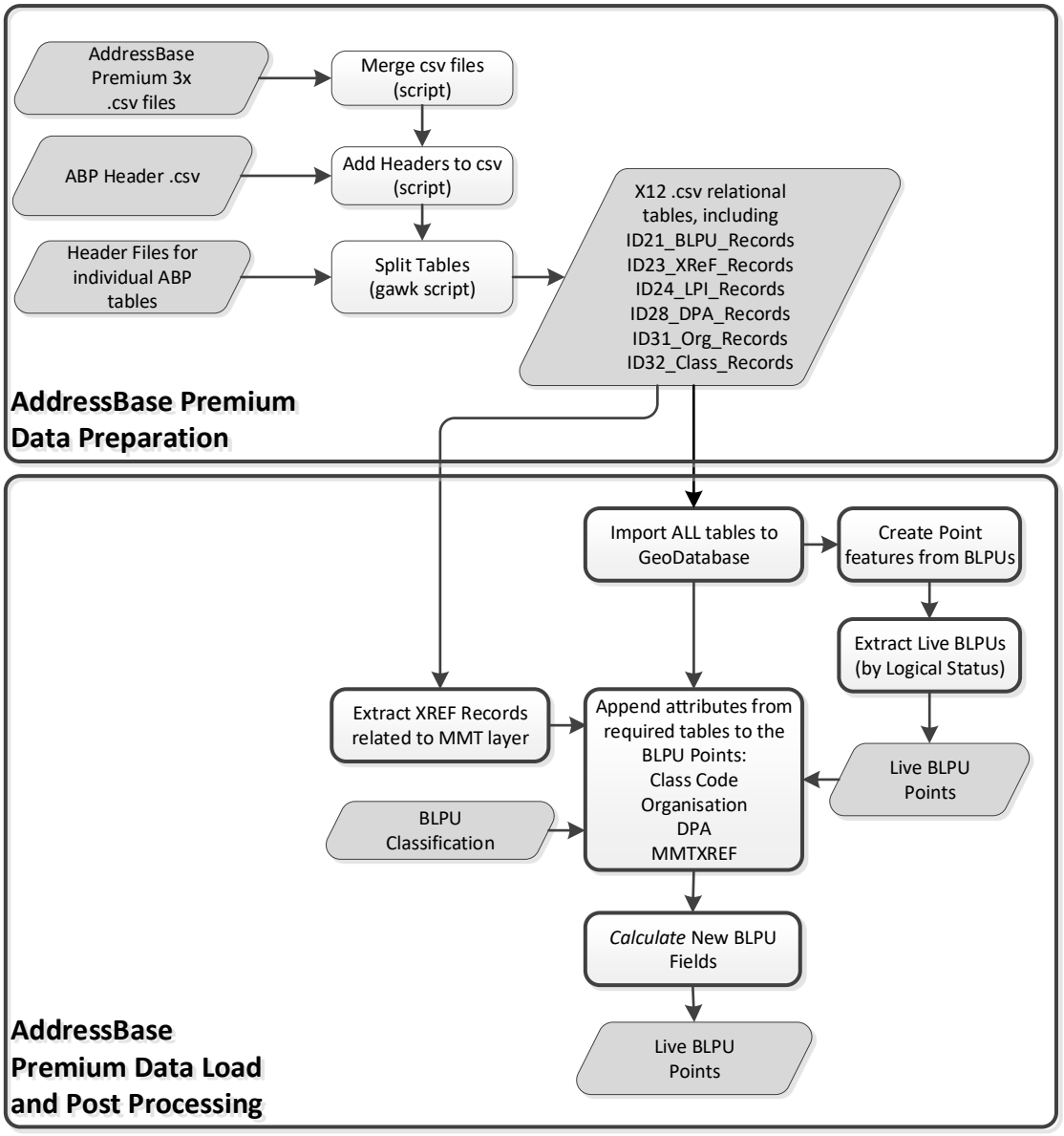


Figure 16 Address Base Premium data load and pre-processing method. The data preparation uses scripts supplied by Ordnance Survey. The bold outlined processes are part of the main Data Load and Preparation python script.

Once this data load process was completed, the relationships were manually tested to demonstrate numbers of BLPUs with associated rows in the different tables. The number of rows and the relationships between the different tables are detailed in Table 14. There are over 200,000 BLPUs, and not every BLPU has a link to each of the relational tables. While the stated relationship between BLPUs and classification is that every BLPU has one or more classifications, in this study area, the relationship is one-to-one.

Table 14: Records delivered in AddressBase Premium tables within the SSA.

Table	Number of Records	Note
BLPU	201501	
XREF	671666	This relationship is one BLPU to zero or more XREFs 16572 XREFs refer to MasterMap Topo features
LPI	206894	This relationship is one BLPU to one or more LPIs 5393 more LPIs than BLPUs
DPA	165523	This relationship is one BLPU to zero or one DPA. There must be 35978 BLPUs with no DPA
Organisation	7160	This relationship is one BLPU to zero or many Organisations
Classification	201501	This relationship is one BLPU to one or many Classifications
Street	5354	This relationship is one to many LPIs have one Street record
Street Descriptor	5354	This relationship is one Street has one to many Street Descriptors
Successor Record	0	

4.7.1.4.2 Post Processing

Post-processing of the ABP data was also carried out within the data load python script. The cross references in the XREF table that refer to the MMT layer were extracted and saved out to a new table.

Selected fields from the relational tables were copied into the BLPU layer (such as address fields from the DPA layer, cross reference from the XREF table, class code from the classification table, class descriptions from the classification lookup table and, finally, organisation from the organisation table). Within the same python script, a spatial join was executed to identify the MMT feature that the BLPU falls inside and selected fields (TOID, Theme, Descriptive Group, Descriptive Term, Make, Physical Level, Physical Presence and Broken) were copied from the MMT layer into the ABP layer. Two fields were then calculated: best level of classification (whether this is primary, secondary, tertiary or quaternary), and a TOID comparison field that specifies whether:

a) the XREF TOID is absent
b) the spatial join TOID differs from the database linkage TOID
c) the spatial join TOID is the same as the database linkage TOID taken from the XREF relationship.

This processing means that it is possible to quickly analyse and summarise the BLPUs by their attributes.

4.7.1.4.3 Extraction of Features of Interest

The Features of Interest (FOIs) within the BLPU layer are primarily live BLPUs. These were identified using the "logical status" attribute. This selection is displayed in Figure 17 (b), and it is apparent that there are few non-live BLPUs delivered in the ABP product.

For estimating how many people may be occupying addresses, the FOIs are those addresses that people might be at for residential or non-residential purposes as these are the addresses to which population will be distributed later on. A BLPU does not have to be addressable to be occupiable: there are addresses within seven different primary classes (including Commercial, Land, Object of Interest and Residential) within the subset of addresses that are not addressable. Human activity may occur at these locations and so all addresses are considered as potential FOIs. Some addresses are not occupiable and are therefore of no interest in terms of distributing population. For example, addresses that are advertising hoardings etc. are of little interest, as these will have neither residential, workplace nor visitor populations distributed amongst them.

All of the data assessment described in Section 4.7.2 is based on *all* BLPUs as this is preferable to excluding specific addresses based on their attribution at this early stage of the modelling.

There is, however, a need to identify buildings that contain residential addresses for the topological analysis that will supply a tertiary classification where this is missing in the supplied data. This is necessary because nearly 30% of all residential addresses have only a secondary level classification, i.e. to the level of "Residential, Dwelling" with no indication of the dwelling type. As there are official statistics available that indicate average occupancy by dwelling type, these residential BLPUs were selected (Primary Code = 'R') and saved out to a new feature layer within the geodatabase. This FOIs extraction process is detailed in Section 4.7.2.2.1, as this occurs in the AOI/FOI extraction python script.

4.7.1.4.4 Results of Data Load

The results of the ABP data load and feature extraction process are illustrated in Figure 17, with those addresses included in the residential extract shown as the red point features in (c).

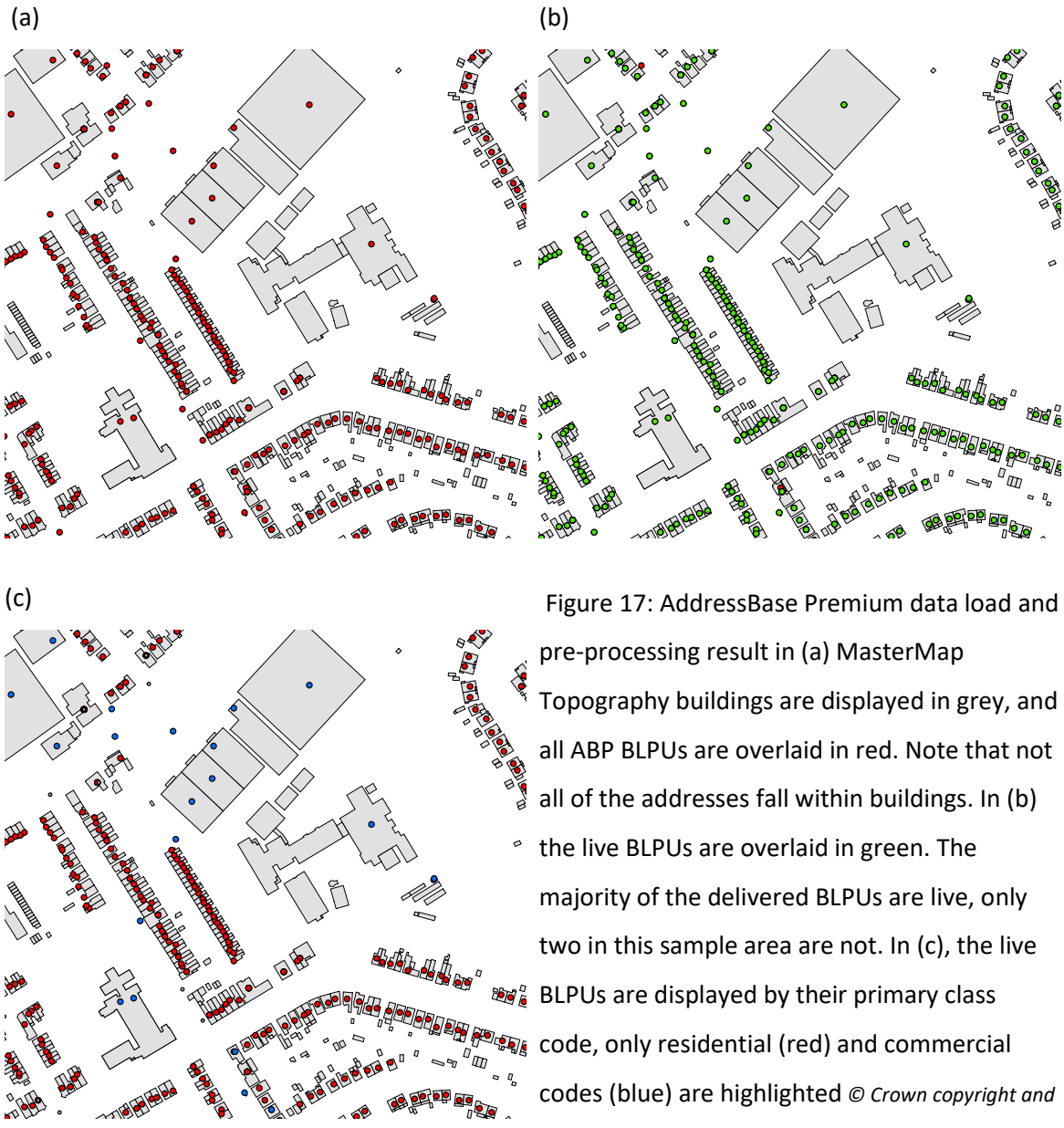


Figure 17: AddressBase Premium data load and pre-processing result in (a) MasterMap Topography buildings are displayed in grey, and all ABP BLPUs are overlaid in red. Note that not all of the addresses fall within buildings. In (b) the live BLPUs are overlaid in green. The majority of the delivered BLPUs are live, only two in this sample area are not. In (c), the live BLPUs are displayed by their primary class code, only residential (red) and commercial codes (blue) are highlighted *© Crown copyright and database rights 2015 Ordnance Survey. © Local Government Information House Limited copyright and database rights 2015*

4.7.2 Assess Locational Data

This section outlines data assessment and analysis that has been undertaken, with the intention to check that *these* data, i.e. the SSA data, are fit for the purpose of developing a modelling framework for spatially and temporally detailed population distribution. Based on the data requirements already outlined, there follows a series of questions that need to be asked of the data. Considering the aim of this research, which involves using address function to enable temporally detailed population estimates to be made, it is essential that the function of individual

addresses can be identified and that associations can be made between the addresses and the buildings to assess for multi-functionality. Address functions therefore need to be summarised for each building to derive this information. The questions that need to be asked of the data relate to whether the addresses have suitable classification attributes, whether these can be supplemented in some way if they are not suitable, and whether they can be used for identifying multi-functionality or more information about the individual addresses.

4.7.2.1 Suitability of Locational Data

4.7.2.1.1 Addresses

The addresses need to be supplied with a sufficiently detailed classification for identifying the function of the address. The required class detail needs to be at a level that will enable a TS to be derived so that population within the address at any time can be estimated. For residential properties, this involves specifying not only the residential function, but also the dwelling type so that census average occupancy counts per dwelling type can be used for the estimation. These requirements suggest the following questions need to be asked of the address data:

1. Is the classification supplied with the address data consistently at a high enough level of detail for the analysis required to apply functional classes to the buildings?
2. If the classification scheme is inadequate, is it possible to derive class detail from attributes contained supplied with the address data?

4.7.2.1.2 Topography Layers

There are two topography layers to consider. They will need to provide building features as individual polygons. There are several questions that need to be asked of these layers, in combination with the address layer:

1. Is it possible to link the address layer and the topography layer so that addresses may be associated with the correct building?
2. If residential address classifications are inadequate, is it possible to extract relevant building features and derive information from their topology that could be used for classification of these addresses?
3. Can the address data, combined with the topography layers, also provide an indication of where many addresses are inside single buildings as well as indicating the presence of multi-function buildings?
4. Is it possible to identify multiple addresses at individual locations?
5. For MMS layer, what is the relationship between the topography layer and the sites?

4.7.2.2 Assessment Methods

4.7.2.2.1 Area of Interest Data Extraction

The OS extraction process of data for delivery is such that some features may fall outside of the SSA. However, it is important that where links are added to the data, they refer only to data that is within the SSA so that there are no loose ends in the later processing. A process has therefore been developed to extract data that will form a complete set of inputs for the model, and ensure that for each feature in the topography layer, all related addresses are extracted. This process is described in Figure 18 that also describes the FOI extraction processes already discussed. The following description refers to the SSA, but the process is developed such that repeating it for other study areas (such as the four smaller study areas highlighted in Figure 13) is straightforward.

As the MMS sites may be multi-part, it is possible that a FS that overlaps the SSA boundary could have a part that is completely outside of the boundary. Initially, visual checks were undertaken to confirm that this is not the case. The MMS sites that intersected the SSA were then selected and saved out to an AOI FS layer. The union of the AOI FS layer and the SSA was calculated and used for selecting areas from the MMT Area layer, which were in turn used for selecting BLPUs. The Southampton AOI is therefore defined as the area covered by MMT Areas that intersect or are within the intersection of the SSA and MMS sites.

The resulting AOI layer is not a simple rectangular shape, but follows the boundaries of all of the intersecting areas of that rectangle as Figure 19 clearly shows.

The process also includes the FOI selection of addresses and topographic areas from the AOI and therefore enables faster process development on smaller study areas with varied features.

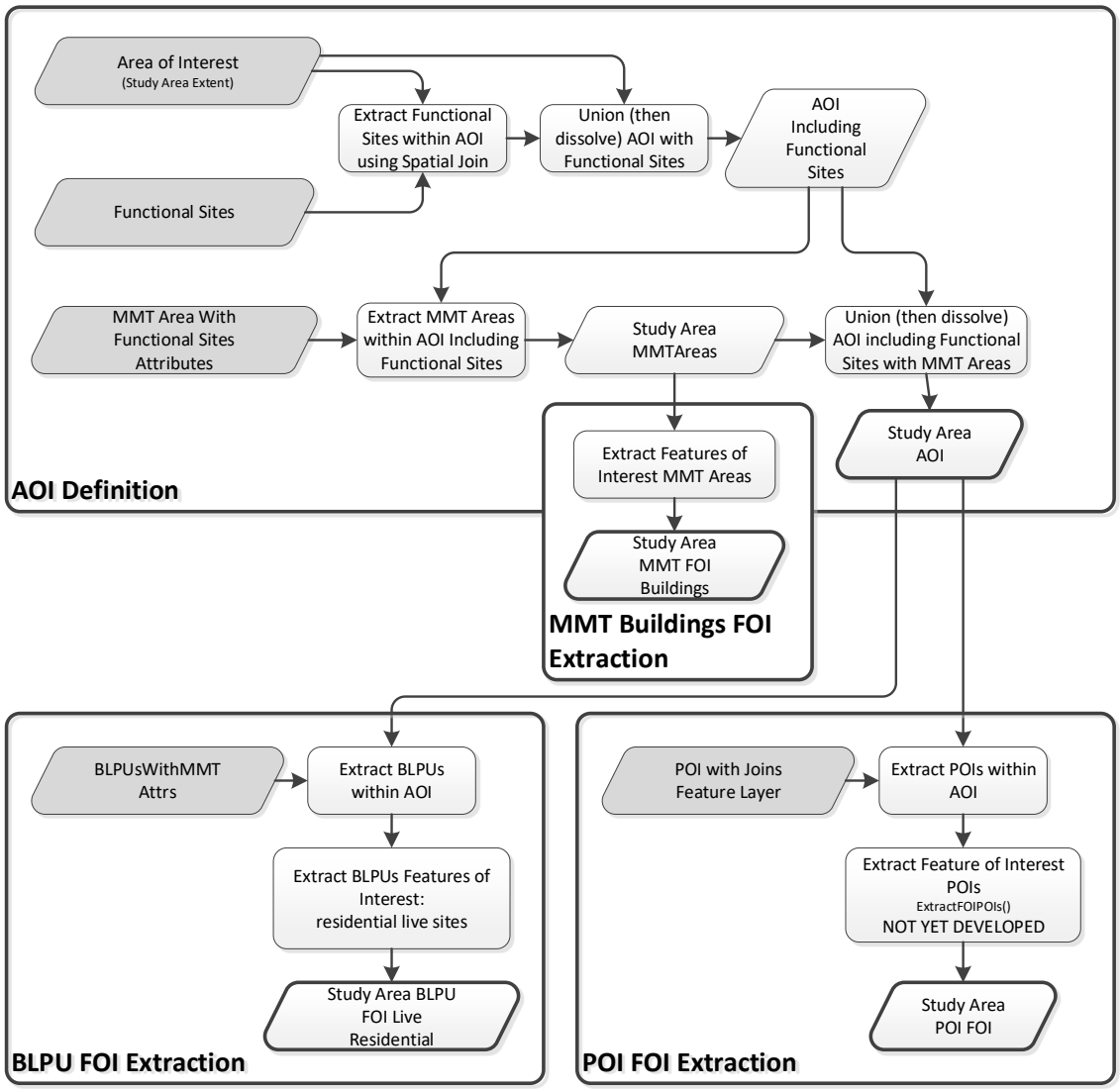


Figure 18: AOI and FOI data extraction method, SSA example

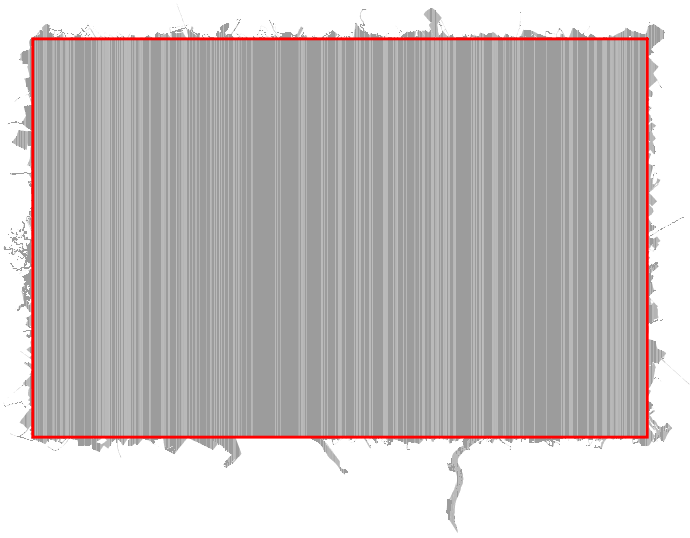


Figure 19: AOI data extraction result: The SSA (in red) and the Southampton AOI (grey)

4.7.2.2.2 Addresses

Analysis of the address data and associated attributes can be used for testing whether the classification supplied with address data are at a consistently high enough level of detail and whether it is possible to derive additional class detail from the attributes within the address data. This analysis was carried out in Microsoft Excel, by producing pivot tables of the best classification levels, and by counting the numbers of textual address attributes that are populated for each address.

4.7.2.2.3 Topography Layers

To answer the question of whether it is possible to link addresses and buildings within the topography layer, pivot tables were generated from the addresses indicating whether the address spatial and database joins both related to the same building feature.

The question of deriving topological information was tested by running the ArcGIS functions that would be required for such an analysis (spatial join, multipart to single part and polygon neighbour) and then processing the output to make sure that the neighbours can be identified from it.

An indication of whether the address and topography data combined can provide information that can identify the location of multi-functional buildings can be gained from further Excel analysis. This is assessed by using the building polygon unique identifier (in this example this is the TOID attribute) from the database or spatial join that was carried out as part of the data loading process.

Multiple addresses at individual locations (i.e. within single buildings) were identified from the address table output from the data load process. Once again, a pivot table was used to count the number of addresses inside each building, and then a histogram constructed indicating the number of buildings with different counts of addresses within them.

4.7.2.3 Assessment Results and Discussion

All of the analyses below relate to features selected within the entire SSA only, excluding addresses and topographic features that were supplied outside of this area due to the OS data extraction process.

4.7.2.3.1 Addresses

It has already been established that a tertiary level classification is required for commercial and residential addresses. A pivot table showing the number of BLPUs in each ABP primary class was created (see Table 15). This reveals that in general the Commercial classification is good (with nearly 90% of BLPUs having a tertiary classification), and the majority of the Residential addresses have a tertiary classification. However, there are nearly 27% of residential addresses that have only a secondary classification. With primary classification of "Residential", a secondary level classification may be "Dwelling". This is inadequate for redistribution of population according to dwelling type. The implication is that *it is necessary to carry out topological analysis on the residential buildings in order to classify the residential buildings by dwelling type*. This allows the ABP classification to be supplemented with this additional classification.

Table 15: BLPU classification levels . Number of BLPUs with each level of classification: Pivot table grouped by MasterMap Theme for the SSA.

	Primary		**Secondary**		**Tertiary**		**Quaternary**		**Total Count**
Row Labels	**Count**	**%**	**Count**	**%**	**Count**	**%**	**Count**	**%**	
Commercial	375	2.5%	364	2.4%	13887	92.6%	364	2.4%	14990
Dual Use	31	100.0%		0.0%		0.0%		0.0%	31
Land	315	15.0%	427	20.3%	1033	49.0%	332	15.8%	2107
Military		0.0%	14	100.0%		0.0%		0.0%	14
Object of Interest	1	0.4%	123	49.8%	56	22.7%	67	27.1%	247
Other (OS Only)		0.0%	320	17.3%	1529	82.7%		0.0%	1849
Parent Shell	79	0.7%	10533	99.3%		0.0%		0.0%	10612
Residential	4	0.0%	42723	26.8%	116641	73.2%		0.0%	159368
Grand Total	**805**	**0.4%**	**54504**	**28.8%**	**133146**	**70.4%**	**763**	**0.4%**	**189218**

Displaying the BLPUs in the PSA and ESA according to their most detailed classification level reveals that the classifications of residential addresses within the PSA are nearly always at an adequate, tertiary level of detail, while those in the ESA are not (see Figure 20). This is a consequence of different LAs supplying the AddressBase products with different quality of data and reveals that the requirement for using supplementary classification is spatially variable.

For an initial assessment of address textual attributes, a pivot table was created to reveal the number of BLPUs that have each of up to 9 DPA fields populated (see Table 16). The DPA fields

contain the textual address information (the fields are Organisation Name, Department Name, Sub Building Name, Building Name, Building Number, Dependent Thoroughfare Name, Thoroughfare Name, Double Dependent Locality, Dependent Locality, Post Town, Postcode, Postcode Type. The table below shows how many fields are populated in each of the MMT Primary Classes). There are 12 DPA fields that will never all be populated for a single record.

29325 BLPUs that have no DPA fields at all. Of the remaining addresses, the vast majority have five or more fields of textual address information that may be used for generating alternative, or supplementary classifications.

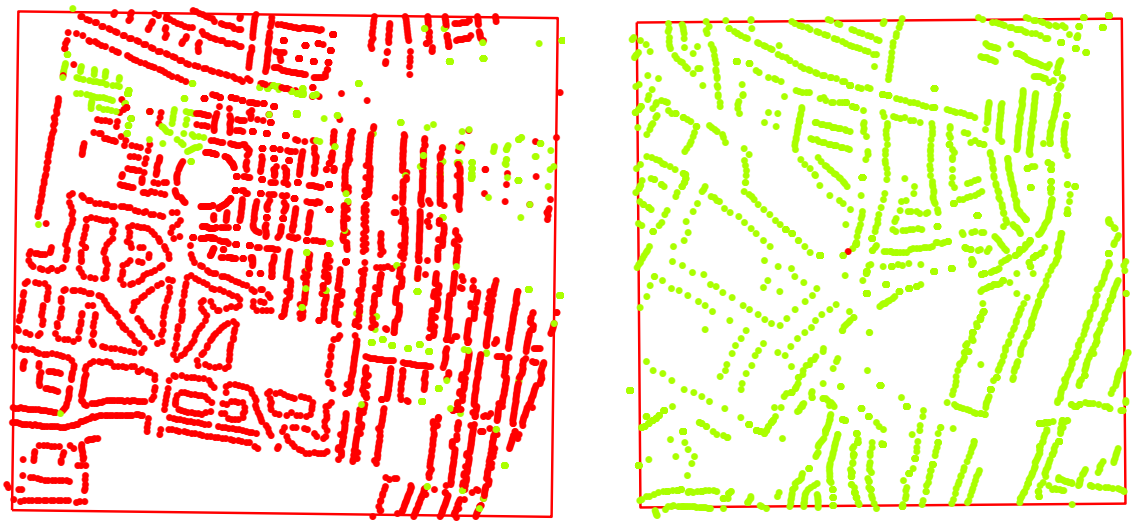


Figure 20: Geographic variation in BLPU classification levels for residential addresses Eastleigh (left) and Portswood (right). Addresses are shown in green if they have a tertiary classification and red if the best level of classification is secondary.

Of the 8619 Commercial BLPUs in the SSA that have no populated DPA fields, 320 have a best classification level of primary and 223 have a best classification of secondary. These will require further classification. These are geographically distributed throughout the study area more or less evenly. 42 of these 543 sites fall inside a FS polygon and a further 92 have an Organisation attribute that may contribute to the classification process, leaving 409 that will require some other means of classification that does not include textual analysis on the DPA fields. This is approximately 2.7% of Commercial BLPUs.

Table 16: BLPUs and populated DPA fields (column headings). Grouped by ABP Primary Class (row headings), and broken down by the best classification that is present on the BLPU feature (primary, secondary, tertiary or quaternary) for Commercial and Residential properties in the SSA

BLPU Primary Class and Best Class	Count of DPA Fields Populated												
	0		5		6		7		8		9		Total Count
	Count	%	Count	%	Count	%	Count	%	Count	%	Count	%	
Commercial	8619	0.6	1153	0.1	3325	0.2	1371	0.1	516	0.0	6	0.0	14990
1	320	0.9	20	0.1	27	0.1	5	0.0	3	0.0		0.0	375
2	223	0.6	12	0.0	46	0.1	49	0.1	33	0.1	1	0.0	364
3	8016	0.6	1006	0.1	3121	0.2	1274	0.1	465	0.0	5	0.0	13887
4	60	0.2	115	0.3	131	0.4	43	0.1	15	0.0		0.0	364
Residential	7743	0.0	84455	0.5	58183	0.4	8704	0.1	282	0.0	1	0.0	159368
1	4	1.0		0.0		0.0		0.0		0.0		0.0	4
2	557	0.0	7201	0.2	33998	0.8	925	0.0	42	0.0		0.0	42723
3	7182	0.1	77254	0.7	24185	0.2	7779	0.1	240	0.0	1	0.0	116641
Grand Total	**16362**	**0.1**	**85608**	**0.5**	**61508**	**0.4**	**10075**	**0.1**	**798**	**0.0**	**7**	**0.0**	**174358**

Of the 6371 Commercial BLPUs in the SSA that have some populated DPA fields, only 196 (approximately 1.3% of all Commercial BLPUs) require further classification as their best level is only primary or secondary (six fall inside FS polygons which will supply a separate classification). These BLPUs seem to be clustered in the North and the East of the study area. The implication here is that *the benefit of developing a textual analysis tool for classifying Commercial BLPUs is limited to improving classification for less than 200 BLPUs within this limited dataset.* It is unknown whether this is the case for other areas of GB within the same data products.

4.7.2.3.2 Topography Layers

Where the functional classification supplied by the address data, for residential addresses (see below) is not available, can the cartographic objects supply further information?

A pivot table (Table 17) was generated from the BLPUs showing the "TOIDCompare" field that was calculated based on the TOID identified through spatial analysis and the TOID identified through the database linkage (from the XREF table). Nearly a fifth of all BLPUs have either no XREF TOID or different TOIDs in the analysis. There are several reasons why this could occur. Firstly, some addresses are present in the LA lists but are not present in the PAF. Both of these data sources are

used in the generation of the address data and when this circumstance occurs, no XREF TOID is associated with the BLPU. Another reason for mismatched TOIDs is a disagreement between the spatial reference supplied by the LA, and the spatial reference used the OS. Finally, a BLPU can be placed so that it is in the centroid of the BLPU extent, and on occasion this falls outside of the building itself.

The implication of this analysis is *that relying solely on the XREF TOID in the BLPU to link the addresses with area polygons in the MMT will not be adequate*. For the residential BLPUs where 90% of BLPUs have a matching TOID, this is less of a problem than it is for the Commercial BLPUs where only 42% have a matching TOID. The absence of an XREF TOID can be addressed using spatial joins, but the accuracy of this will be difficult to measure. Where TOIDs disagree, on the advice of OS, the database join should be used.

Table 17: Database linkage versus spatial analysis for building associations between BLPUs and building polygons for the BLPUs in the SSA

	Different TOIDs		**Matching TOIDs**		**No XREF TOID**		**Total Count**
BLPU Primary Class	**Count**	**%**	**Count**	**%**	**Count**	**%**	
Commercial	882	7%	5625	42%	6980	52%	13487
Dual Use	3	10%	27	87%	1	3%	31
Land	4	0%	5	0%	1740	99%	1749
Military	1	7%	1	7%	12	86%	14
Object of Interest	8	3%	86	35%	153	62%	247
Parent Shell	59	1%	900	8%	9651	91%	10610
Residential	5475	3%	143583	90%	10213	6%	159271
Grand Total	**6432**	**3%**	**150227**	**81%**	**28750**	**16%**	**185409**

The question of deriving information from the topology that can be used for classification of addresses where the supplied classification level is not detailed enough is relevant only to residential addresses, as the dwelling type is linked to number of residents in available census data. A script has been written that will do this topological analysis and add a dwelling type to the addresses, which is then inherited by the building containing the address. This script has not been run on the entire SSA due to the large quantity of data, but only on the two smaller PSA and ESA study areas, which have classifications supplied by different LAs. Portswood residential addresses mostly have a tertiary classification, but the Eastleigh address classifications tend to be only at

secondary level. For a full description of the topological analysis, see Chapter 5. *The conclusion from this initial topological analysis is that it is possible, and feasible to derive useful information from the topology that can supplement the residential address data where only a secondary classification is available.*

Investigations into the identification of multiple addresses inside single buildings can involve a number of different measures used to identify buildings containing flats and multi-functional buildings. These include the count of BLPUs inside buildings, the presence of parent shells within buildings, the MOC attribute on the address, the presence of textual identifiers that may indicate the presence of a flat, and the classification supplied with the building.

For identifying multi-functional buildings, a pivot table has been generated that shows the number of addresses in multifunction buildings. In Table 18, the first column details the number of functions (primary classes) associated with the different BLPUs that fall inside the individual buildings. The other two columns show the count of buildings with that number of functions, and the count of BLPUs that fall inside the buildings with that number of functions. The maximum number of functions of the buildings is 5 (3 buildings containing 148 BLPUs), the vast majority of addresses fall inside single function buildings that are residential or commercial (123381 buildings containing 136332 BLPUs). Of the remainder, as expected, the next highest combination of functions is Commercial/Residential.

Considering these investigations, it will be possible to identify flats and multi-functional buildings from these data.

Table 18: Multi-functional buildings: number of functions within individual buildings, and the number of BLPUs within those buildings in the SSA.

Number of BLPU Primary Classes inside Building	**Number of Buildings**	**Number of BLPUs**
1	123381	136332
2	6877	44810
3	384	7319
4	17	609
5	3	148
Grand Total	**130662**	**189218**

For another quick test of whether the data can be used to identify multiple addresses within individual buildings, an analysis of the tabular data, which involved counting the number of BLPUs per MMT TOID (based on database linkage), was carried out. Not all of the TOIDs will necessarily be buildings. The results of this analysis can be seen in Table 19.

The results of these investigations suggest that these data can be used to identify multiple addresses at individual locations.

The MMS layer maps the extent of FSs such as airports, schools, hospitals, ports, utility and infrastructure sites (Ordnance Survey 2015b). This sites layer is used to identify sites comprised of several buildings (or non-building) polygons, for modelling.

The relationship between the topography layer and the sites was assessed by running a python script that runs a spatial join adding the FS attributes to any topography areas that fall inside (not completely inside) that site. The output feature layer was then visually inspected to check for incomplete polygons within each FS.

There are 307 sites, which contain 15907 Topography Area features. Some of these fall entirely outside of the SSA. There are 142 Education sites within the SSA. Of these, 91 are concerned with primary education. There are ten higher or University Education sites. Figure 21 is a map of MMS sites around the main campus of the University of Southampton (which is outlined in cyan). A single feature can be comprised of multiple polygons. The University of Southampton (indicated in the “stakeholder” attribute which links regionally-disparate sites) has seven separate polygons which include three features with distinctive name of “The University of Southampton” (two of which have multiple polygons), as well as one each for “Avenue Campus”, “Bassett House”, “Glen Eyre Hall” and “St Margaret’s House”.

Table 19: BLPU counts per TOID in the Southampton Study Area

Number of BLPUs	*Frequency*
1	106696
2	5574
3	1119
4	749
5	211
6	570
7	104
8	220
9	168
10	109
11-20	726
21-50	235
51-100	40
101-200	24

Table 20: Summary of MasterMap Sites by Functional Theme in the Southampton Study Area

Functional Theme	**Count**
Air Transport	1
Airport	1
Education	142
Further Education	7
Further Education, Secondary Education	1
Higher or University Education	10
Non State Primary Education	4
Non State Primary Education, Non State Secondary Education	1
Non State Secondary Education	3
Primary Education	90
Secondary Education	17
Special Needs Education	9
Medical Care	18
Hospital	7
Medical Care Accommodation	11
Rail Transport	13
Railway Station	13
Road Transport	3
Bus Station	1
Coach Station	1
Road User Services	1
Utility or Industrial	5
Electricity Distribution	2
Electricity Production	1
Gas Distribution or Storage	2
Water Transport	6
Passenger Ferry Terminal	2
Passenger Ferry Terminal, Vehicular Ferry Terminal	1
Port Consisting of Docks and Nautical Berthing	2
Vehicular Ferry Terminal	1
Grand Total	**188**

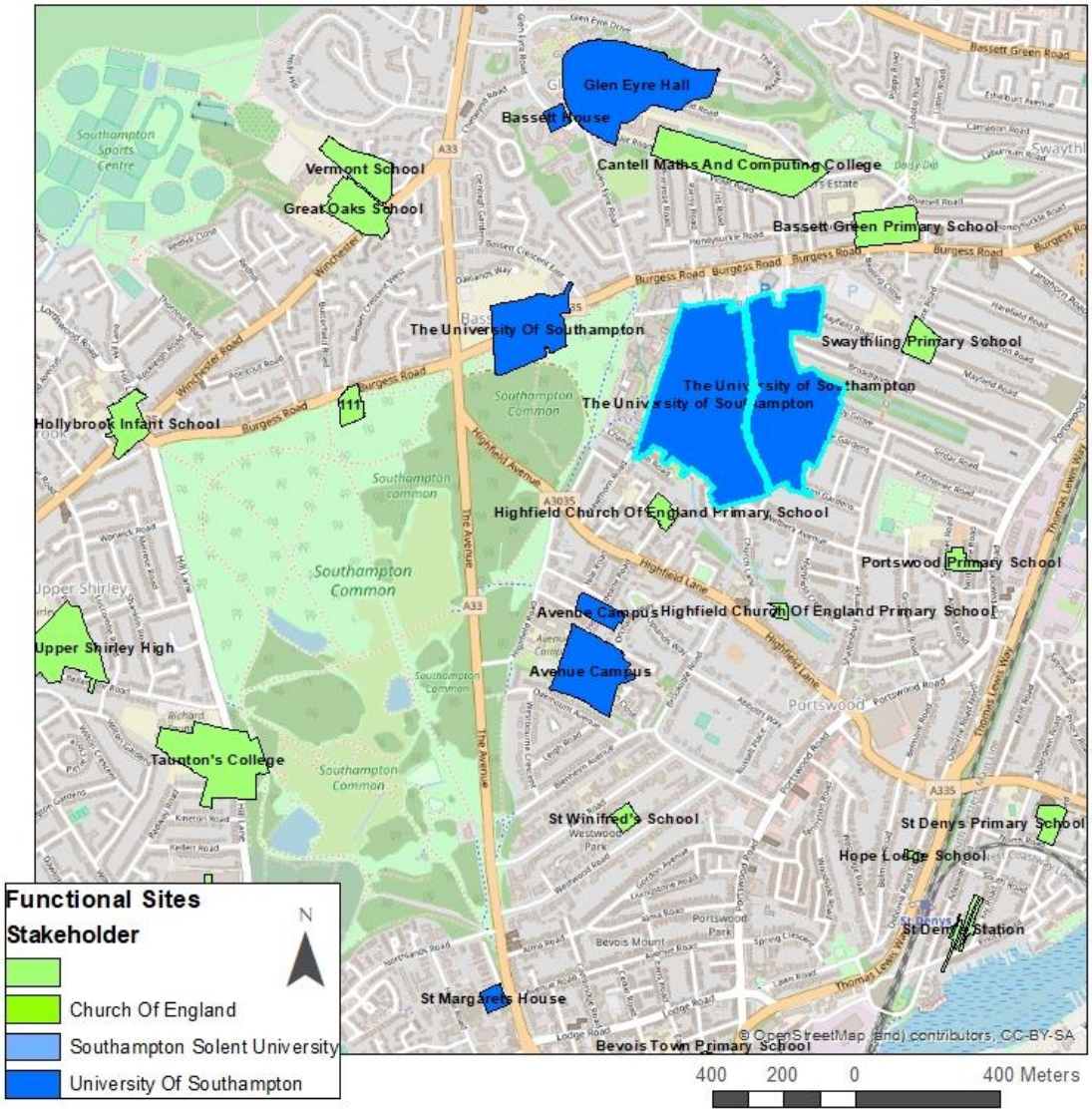


Figure 21: MasterMap Sites around the University of Southampton, coloured individually by TOID to show how one FS can be multi-polygon. *© Crown copyright and database rights 2015 Ordnance Survey*

4.8 Chapter Review

This chapter has described the data selected for use within the model, and how those data were assessed for their suitability and loaded to geodatabases for spatial analyses. The core geographic data sets (ABP and MMT, with MMS) are supplied by OS and have been processed ready to load into the ontology. Another stage of data preparation is now required due to the geographically varied data quality of the most essential of these datasets, the addresses. The next chapter presents a method for generating this supplementary classification for residential addresses, appropriate to these current datasets.

Chapter 5 Address Classification

In the previous chapter, the variability in the quality of address classification delivered in the ABP product was highlighted. This leads to the need to supplement these classifications. This chapter presents the methodology used for supplementary classification of the residential addresses to a tertiary level so that dwelling type is identified for all residential addresses. In order to demonstrate the ability of the modelling framework to handle variable data quality, commercial address classifications will not be supplemented.

5.1 Address Classification Method

It is clear that address occupation patterns differ between these two main classes, even at the most basic, primary level of classification (i.e. residential or commercial). If a classification is present that indicates whether an address is residential or not this enables the different treatment of different addresses according to their function. The greater the detail of the classification within the class hierarchy, the more inherent information is available about the address function, and therefore the temporal signature, of the addresses. The classification of addresses is therefore key to this modelling framework.

The tertiary classification of residential addresses reflects the dwelling type (building form) rather than the building function (which is reflected in the secondary classification). For commercial addresses, the form of the building is not indicated by any of the classification levels, only the address function.

A tertiary level classification is therefore required for distribution of population amongst residential addresses because residential population will be redistributed from small area statistics taking account of the dwelling type from the census data, which includes average residents per dwelling type in different OAs.

The residential address classification methodology employs a three-stage process, as described in Figure 22. This process has been developed within a python script, which utilises the ArcPy module but avoids using ArcPy where spatial operations are not required.

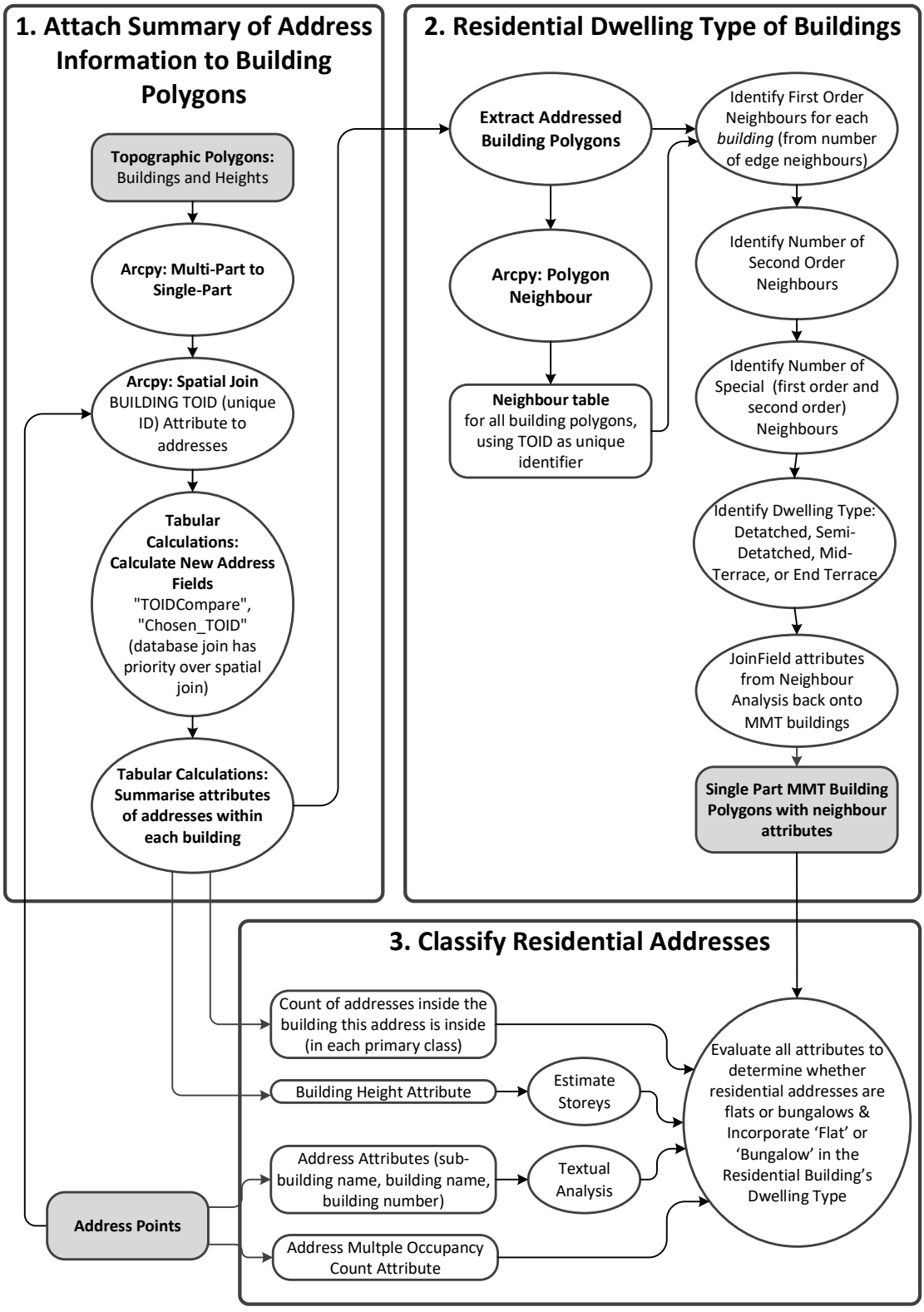


Figure 22: The residential address classification process, summarising the three stages of the process. Apart from a few ArcPy functions, only core python and open source library functions are used in the analysis. The grey shaded rectangles represent geographic datasets.

In the SSA, approximately 27% of the residential addresses have a classification that is only to secondary level. For commercial addresses, 2.4% have only secondary level classification and 2.5% have only primary classification. The completeness of the classification varies geographically, so there are some areas (notably the ESA illustrated in Figure 13 and Figure 20) that have a much larger proportion of addresses classified only at primary or secondary level. While dwelling type only applies to residential addresses, the topological analysis requires all buildings to be assessed, and all addresses to be included in the analysis.

5.1.1 Attach Summary of Address Information to Building Polygons

The aim of this stage of building classification is to make sure that the addresses are summarised for each building, so that it is possible to identify buildings that contain residential addresses and combination of address types. This stage involves linking BLPUs and buildings using a spatial join. As previously discussed, the ABP data includes a cross reference table, which links individual BLPUs to specific MasterMap topology area features. However, as has already been indicated, some of the BLPUs do not have a polygon associated with them in the database. The spatial join enables the identification of the building polygon with which the BLPU is associated. A preferred building TOID for the BLPU is calculated based on the database linkage taking precedence over the spatial join result. A process that summarises all of the addresses within each building is also implemented during this stage of the analysis. This yields a count of addresses in each primary classification within each building and adds attributes containing lists of BLPU TOIDs, their class codes and TOID comparison measures for all BLPUs within the building. These attributes are appended to the single part building polygons.

The address class codes that are contained in these lists are at their finest detail level. In some cases, this may be to quaternary level. There are 563 possible classifications. The 304 possible tertiary classifications are listed in Appendix A.1.

5.1.2 Residential Dwelling Type of Buildings

The aim of this second part of the building classification method is to identify the dwelling type of buildings containing residential addresses in order to supplement the classification information available. A process utilising topological and textual analysis has been developed to specify whether a building is detached, semi-detached, mid-terrace, end-terrace, bungalow or a flat, which is the equivalent of a tertiary classification for residential dwellings.

As discussed in the Conceptual Review, this dwelling type has previously been identified by (Orford & Radcliffe 2007), using a four stage process for identifying dwelling types using topological and textual analysis of AL2 and MMT layers, the output of which was assessed using census data. AL2 did not include a classification, and has now been superseded by the AddressBase product suite. The four stages were to identify first and second order neighbours, identify trivial divisions, identify flats using one of three different measures and to identify commercial properties. This method has been adapted here to enable use of the newer ABP data layer and the Building Heights data that are now available for the MMT Layer. Only two of these stages are now required as improvements made between AL2 and AddressBase products include the identification of address classification at least at the primary level, so residential and commercial addresses are already identified in the data.

The first part of this analysis identifies first order and second order neighbours for each building using the output from the ArcGIS Polygon Neighbour tool. The output of this tool is a neighbour table, which specifies, for each polygon, all other buildings that are neighbours, along with the length of the boundary that is shared. Building polygons are identified by their TOID. From this table it is possible to identify first order neighbours (those that directly adjoin a building) and second order neighbours (neighbours of neighbours). For the purposes of this part of the building classification, function is not relevant because only the building form is being determined. In addition, a building's neighbours are relevant and must be accounted for whether they are themselves residential or commercial. Buildings are therefore included in this analysis regardless of the primary function of their associated addresses.

This process differs from that of Orford & Radcliffe (2007), as neighbours with trivial boundaries (those that share less than 15% of the total length of a building's perimeter, or that are less than 4m long) are not treated differently to all other neighbours. This is because the absence of a BLPU within a building outline is assumed to indicate that the building is not the main building with which an address is associated, but most likely an outbuilding such as a garden building or a garage. In some cases, this may also indicate a more complex structure such as a section of a building that is only present at first floor or higher, such as that described in Figure 23 (a) and (b), or that is an extension. All of the buildings that do not contain address points are therefore excluded from the neighbour analysis. Consequently, there are occasions where a neighbouring building polygon is ignored in error. Examples include those in Figure 23. The methodology was also tested using trivial boundaries, but there were some obvious misclassifications of dwelling type present using this approach that were absent when buildings that do not contain addresses were excluded, as described in Section 5.2.

(a)

(b)

Figure 23: Example of complex residential topographic structures : (a) shows the MMT Buildings, with emboldened labels being the last four digits of the building TOID and the non-bold labels being the last four digits of the address chosen TOID. From (b) (Google 2016), it is evident that these building polygons do not all have the same configuration as depicted in (a). The ground floor for two of these polygons has a "garage" function, and in the centre, it is a driveway. One of the garages is represented as a part of the building polygon with TOID ending 9040. This is also an example of the BLPU being associated with a building that spatially, it is not inside, but that the database linkage indicates the correct building polygon.

Table 21: Rules for identifying dwelling type from number of first and second order neighbours

Dwelling Type	**Number of First Order Neighbours**	**Number of Second Order Neighbours**
Detached	0	
Semi-Detached*	1	1
End Terrace	1	>1
Mid Terrace	>1	
*** first order neighbours = 1 and second order neighbours = 0 is an additional possibility if the trivial boundaries method is used, and indicates a semi-detached building.**		

5.1.3 Flats

The aim of this part of the analysis is to classify the residential addresses with a final dwelling type, so that average population by dwelling type can be used in the disaggregation algorithm for population estimation.

For addresses with residential primary class, this identifies whether the address is a flat. As previously discussed in the Conceptual Review, a flat is defined in such a way that it must be contained within a dwelling containing more than one storey and must be a self-contained part of a horizontally divided building (Department for Communities and Local Government 2012). This means that there must be more than one address present within a building in order for one of those addresses to be a flat, and at least one of those addresses must be residential. The measure used to identify flats is therefore the presence of more than one residential address inside a building, or at least one residential address where there are more than one addresses present.

If either of these conditions are true, then the address gains a value in the *isFlat* attribute, otherwise the value remains null. This process also involves the consideration of several building and address attributes, each contributing a different value to the *isFlat* attribute so that the reason why a flat is classified as such is recorded. These contributions to the identification of flats are recorded in the *isFlat* attribute in such a way that they can all be isolated to evaluate the reason why any individual address is recorded as a flat and can contribute towards confidence in that classification process.

Textual analysis of the attribute values to identify whether the sub-building attribute on the address (which in itself will identify whether a building has been sub-divided) contains certain text strings, such as ‘Flat’, ‘Apartment’ or ‘Towers’. This text string would usually indicate that an address relates to a flat. The count of the indicators found within each *sub-building name* attribute is recorded in the *isFlat* attribute. Other address fields (*building name*, *building number* and *department name*) are also assessed for indicators, using the same approach to textual analysis and the count of indicators found is recorded in the *isFlat* attribute.

The building height, where available, can be used to estimate the number of storeys that a building has, using average building heights as identified by Orford (2010), rounded to next highest 50cm. There are alternative approaches to estimating building heights such as that used by Alahmadi, Atkinson & Martin (2013). Wu, Blunden & Bahaj (2018) automatically estimate the geometry of buildings from LiDAR data and then estimate number of storeys by using a non-linear relationship between height and number of storeys. This calculated number of storeys is most useful if there is a need to flag whether a property is a bungalow. It is not used for identification

of flats in this process because the other methods of identifying flats are more reliable (see Section 5.4).

The MOC attribute delivered with the ABP product is also accounted for in this process. The presence of a MOC greater than one is recorded in the *isFlat* attribute.

A final, probabilistic classification (at tertiary level) was appended to each of the residential addresses, taking into account the calculated dwelling type (including calculated *IsFlat* attribute), and the delivered ABP classification.

Table 22: Rules for identifying addresses that are flats by calculating the IsAFlat value. Each contributing condition can be isolated from the isFlat value, allowing assessment and validation of the method, and the ability to identify why any individual building has been classified as a flat. For example, if this is only because of their estimated storeys value.

Attribute	**Value**	**IsItAFlat**
Number of residential addresses inside building containing this address	>1	+1000000
Number of non-parent shell BLPUs inside building containing this address	>1	+2000000
Address Sub-Building Name Textual Analysis Count of Indicator Words	>1	+ count * 100000
Other Address Textual Analysis Indicator Words	>1	+ count * 10000
Estimated Storeys (from building height)	>3	+ estimated storeys * 100
Multiple Occupancy Count	>0	+ Multiple Occupancy Count

5.2 Residential Address Classification Results and Evaluation

Inspection of Figure 24 suggests that the dwelling types are generally assigned appropriately to buildings, with mid and end terrace, semi-detached and detached appropriately assigned most of the time. It is not possible to know, simply by viewing this map whether flats and bungalows are appropriately assigned, and this requires validation (described in Section 5.4).

Figure 24: Example of classified area within the Eastleigh Study Area. *© Crown copyright and database rights 2015 Ordnance Survey. © Local Government Information House Limited copyright and database rights 2015*

The outputs from the classification of residential addresses in the two smaller study areas (PSA & ESA) are described in Table 23 and Table 24. These compare the tertiary classification delivered with the ABP product (rows) with the supplementary classification calculated using the method described (columns). Classification delivered with the AddressBase data is used only at the primary level for identifying residential properties. In these tables, the residential tertiary classifications that are not identified in the topological assessment (Care / Nursing Home, Communal Residence, HMO classes, Residential Education and not classified) are removed from the table. Flat dwelling types have been amalgamated, as have terrace dwelling types.

In the PSA (Table 23) there is a high proportion of residential properties that are delivered with a tertiary class. This area is therefore ideal for testing the ability of the model to identify a building class accurately at this tertiary level. More than 95% of these residential addresses have been assigned the correct dwelling type as compared to the ABP class, indicating that this may be a suitable process to use when there is no other dwelling type information available, but that there is still scope for improvement of the process. In the ESA the majority of residential classes do not have a tertiary class in the ABP data, meaning that this small study area requires validation. This has been completed using internet searches on a sample of addresses, and is described in Section 5.4.

Table 23: Portswood Study Area topological dwelling types, where the majority of the residential addresses are supplied with a tertiary classification. Dwelling types not identified in the topological analysis have been removed from the table. Terrace and Flat dwelling types have been amalgamated for the purpose of this comparison.

ABP Tertiary Classification	**Calculated Supplementary Classification**				
	Detached	**Semi-Detached**	**Terrace**	**Flat**	**Not in Building**
None	1				
Detached	**305**	2	2	12	
Semi-Detached	6	**572**	10	16	
Terraced	3	13	**416**	27	
Self-Contained Flat	1	8	7	**1353**	11

Table 24: Eastleigh Study Area topological dwelling types , where the majority of the residential addresses have no associated tertiary classification. Dwelling types not identified in the topological analysis have been removed from the table. Terrace and Flat dwelling types have been amalgamated for the purpose of this comparison.

ABP Tertiary Classification	**Calculated Supplementary Classification**				
	Detached	Semi-Detached	Terrace	Flat	Not in Building
None	96	1055	1652	795	12
Detached	**7**	3	10		1
Semi-Detached		**9**	26		4
Terraced			**26**	2	
Self-Contained Flat		1	4	**651**	12

Figure 25: Results of topological analysis excluding smaller polygons. By (a) excluding buildings not containing addresses, and (b) excluding neighbours that have trivial boundaries, but analysing all buildings including those that do not contain addresses. The main differences within this small portion of the PSA are highlighted with red circles.

Visual inspection of Figure 25 highlights some exceptions. Circle 1 shows where excluding buildings not containing addresses has misclassified two blocks of flats as detached, and using the trivial boundaries method classified them as semi-detached. In both cases the flats are correctly identified but the building form is incorrect. Circle 2 shows where using the trivial boundaries method misclassified a mid-terrace building as semi-detached because the universally applied shared boundary length threshold is inappropriate for this building.

Circle 3 shows an example of the trivial boundaries method inappropriately classifying buildings as mid-terrace because of the specific configuration. This occurs where two neighbours are on the same side of the building, and are themselves adjoining. Here the first order neighbour is the same as the second order neighbour. It may be possible to mitigate this issue by checking each first order neighbour TOID against each second order neighbour TOID. Alternatively a different algorithm could be used, such as the shortest path Floyd-Warshall algorithm (Cormen, Leiserson &

Rivest 1990, p558), which could be used on entire blocks of adjoining buildings to identify how many buildings away each in the block is from each individual in the block.

Another example of common misclassification that is present in the results are where archways are present, and linking two other buildings that have associated addresses. In the first example, the buildings are joined, but not at ground level, so the MMT represents the archway as a separate building polygon. The classification algorithm could be amended to merge polygons without addresses that adjoin polygons with addresses. This may introduce some uncertainty, as there is no attribution within the data to indicate with which neighbour an un-addressed building polygon should be merged. The impact of incorrect merging may be minimal because the topological relationships may remain the same. These possible solutions are an area for further research.

The extent of these misclassifications is presented in Section 5.4.

5.3 Potential for Address Misclassification

Although it is not common, there are some instances where MMT areas overlap. In these instances, the approach to processing the data will assign addresses to the first polygon in the database list. In the PSA, there is one example where a BLPU falls inside overlapping MMT areas. The output from the spatial join (joining MMT Polygons to ABP BLPU points) contains two rows for this BLPU, one for each of the MMT areas. Subsequent processing of the addresses uses only the first of these polygons in the list. Joining self-intersected MMT areas to the ABP BLPUs within the SSA, reveals that there are 172 addresses that fall inside more than one polygon in the entire SSA (80 of these are commercial addresses).

This method does not identify HMOs and CEs, and while some of these are identified in the ABP tertiary classification, further development will be required to identify these, as this is not the focus of the model development.

There are some instances where, for example, a block of flats will have its addresses located in a lobby and therefore need to use the TOID identified in the database join rather than that identified in the spatial join information, such as that in Figure 23. In these instances, the use of database joins (that link the address to the TOID of the building that the flats are actually inside) rather than spatial joins (that link the addresses to the lobby polygon) properly assigns the address to the correct building.

It is worth noting that the misclassifications produced using this method are not gross misclassifications. For example, terraces being misclassified as semi-detached will have a minor impact on the population estimation for those properties. There are no instances where a single address building has been misclassified as an HMO or CE (such as student halls) which could have a significant impact on the population estimation for the addresses in that building. This should be a consideration should this method be applied in a different domain.

5.4 Residential Address Classification Validation

This chapter has described the process used for creating a supplementary classification for residential addresses where the supplied classification is not to the required standard for population estimation, i.e. to tertiary level, which indicates the dwelling type of the address. This section outlines how the results of this supplementary classification have been validated to ensure the process is sound.

5.4.1 Overview of Method

The two small study areas (about 1km^2, see Figure 13) were processed to identify the supplementary classification for every address within the study area. The PSA had almost complete coverage at tertiary classification, while the ESA was supplied with addresses that had mostly only secondary classification. The results of this preliminary assessment are clearly highlighted in Figure 20.

The PSA was used as an indicator of the success of classification from the data itself. In over 95% of cases, the supplementary classification that had used only primary class to identify dwelling type agreed with the supplied tertiary classification.

In the ESA the approach to validation was to generate a stratified sample of buildings using the ArcGIS add-in SamplingTool_10.esriAddIn (developed by National Oceanic and Atmospheric Administration (NOAA) Biogeography Branch (Buja & Menza 2013)), within the small study areas and to visit these sites using GSV (Google 2016) to evaluate whether or not the supplementary classification was accurate.

The sample population was buildings within the study area rather than addresses, as the address classification was derived from building topology with contributions from spatial relationships with addresses and their attribution. The buildings can be seen from outside, whereas the addresses cannot.

5.4.2 Sampling Process

There were 3476 buildings in the sampling frame. Of these 3230 were linked to residential addresses. A 5% sampling rate therefore means that 162 buildings are included in the sample. There are 15 possible attribute values for the supplementary classification. A stratified sample was generated, but with over-sampling for the building classifications that are least represented in the population. For example, the mid-terrace special classification appears only 26 times in the sampling frame which would lead to a stratified sample size of only one building (between three classes) if 5% of the population are sampled. A minimum sample size of five was therefore required for all classes. The stratified and adjusted sample sizes are indicated in Table 25 and Figure 26 shows a map of the sampled buildings with their supplementary classification.

Table 25: Stratified sample sizes, adjusted to over-sample under-represented classes

Dwelling Type With Flats Attribute Value	**Number of Buildings in ESA**	**Stratified Sample Size**	**Adjusted Sample Size**
Detached	55	3	5
Detached – Bungalow	53	3	5
Detached – Flat	67	3	5
End-Terrace	299	15	15
End-Terrace – Bungalow	96	5	5
End-Terrace – Flat	98	5	5
Mid-Terrace	1238	62	62
Mid-Terrace – Bungalow	96	5	5
Mid-Terrace – Flat	97	5	5
Mid-Terrace-Special	3	0	3
Mid-Terrace-Special – Flat	6	0	5
Semi-Detached	1012	51	51
Semi-Detached - Bungalow	61	3	5
Semi-Detached – Flat	49	2	5
Total	**3230**	**162**	**181**

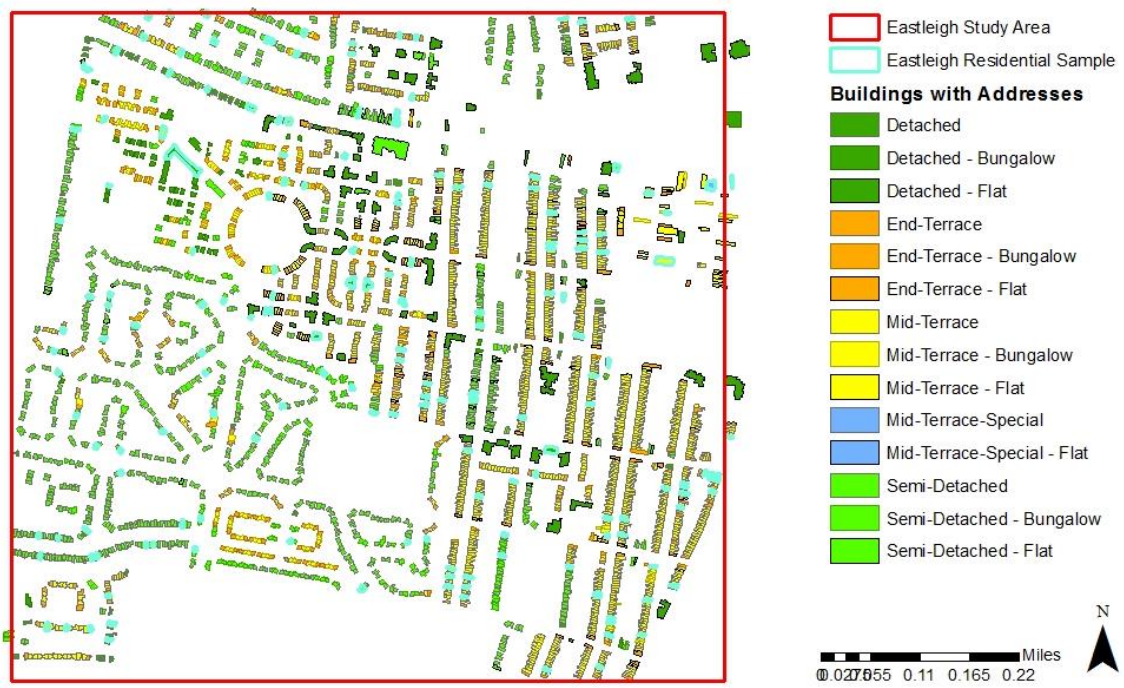


Figure 26: Sampled buildings in each calculated dwelling type in the Eastleigh Study Area.

5.4.3 Summary of Results of Validation

As previously mentioned, GSV was used to assess whether the dwelling type for addresses in buildings containing at least one residential address was accurate. The results recorded in a comprehensive document containing the evaluation output for each of the buildings (presented in part, in Appendix B). A summary of these results, classified into eight possible outcomes are presented in Table 26.

GSV image capture dates vary between October 2008 and May 2016. 163 out of the 181 are from May 2014, April 2015 or May 2016. All but 48 are from dates before the MasterMap data was delivered (July 2015).

Notable results include:

1. 132 out of the 181 sampled buildings (72%) are correctly classified to tertiary level.
    a. 49 out of the 181 sampled buildings were not correctly classified to tertiary level:
    b. 24 (13%) are incorrectly classified due to building height issues (of these 18 have been addressed using a code tweak which assumes all buildings to be non-bungalow unless clearly stated by the building height).
    c. 16 (9%) of errors are caused by linked, archway or missing address data faults (see below for explanation).

2. 9 (5%) are unknown as they are flats that can't be identified from GSV.

Table 26: Summary Information from supplementary classification validation

Classification Validation Result	
Row Labels	**Count of OBJ ID**
Archway	7
Data Error (height missing)	16
Data Error (height)	7
Data Error (missing address next door)	1
Linked	8
Processing Error (height). Caused by Data Error (height)	1
Unknown	9
Valid	132
Grand Total	**181**

There are four main issues highlighted here: Firstly, bungalows: the recorded heights are to the eaves of the buildings, so they are being classified as semi-detached or terrace but not bungalow as the height values are greater than the threshold for single storey buildings. Secondly, two storey buildings: some buildings have no heights or heights that are clearly in error (e.g. 0.7m) meaning that they are being incorrectly classified as bungalows. This is common with new buildings that were building sites when the heights were calculated. Thirdly, linked buildings: where detached or semi-detached buildings are linked with neighbours through a trivial boundary such as that created by garages, or entrance porches. Finally, archways: where an archway generates a separate polygon between two terraced properties that does not contain an address and therefore causes terraced properties to be classified as semi-detached or detached.

The issues with linked buildings could be minimised by identifying trivial boundaries and accounting for this in the processing. The issue with archways could be dealt with by identifying the building polygons that are archways (attributed within the data) and merging them with one of their neighbours. This would solve the classification problem, but this would affect the building capacity calculations that are needed later in the modelling process, and uncertainty about which neighbour to assign the building polygon to would be introduced.

5.5 Chapter Review

This chapter has described the essential process of calculating supplementary classifications for residential addresses. This facilitates a demonstration of how the modelling framework can disaggregate small area residential population data from the decennial census to addresses. The level of classification provided by this process is adequate for the disaggregation according to the available census data, and it produces a high success rate of up to 95%.

The output from this process is an additional attribute on the point based ABP address data, which, after this stage, have been exported from the geodatabase as .csv files ready to be loaded into the ontology. The development and implementation of the ontology is laid out in the next chapter.

Chapter 6 Ontology Development

This chapter focuses on Stage Two of the modelling framework adopted for time-specific population estimation. As set out in Figure 7 (page 88), the ontology follows preparation of input data and supplementary classification of residential addresses within a GIS (the process for which is dependent on the data environment, and is described in Chapter 4 and Chapter 5).

There follows an introduction to how the concepts outlined in the literature review translate to usable technologies, and how they are applied to this application domain. This is followed by a description of the overall design of the ontology, and then explanations of specific design decisions that have been made. Finally, details of the implementation used to prove the concept of using ontology for estimation of population are presented. The implementation is based on data for the SSA, but the ontology, as the central novel development in the modelling framework, is designed in such a way as to be transferrable to other data environments.

6.1 Use Case for Ontology

To prove the concept that semantic web technologies and ontologies can be utilised as part of the modelling framework for population estimation the model must demonstrate that it is possible to re-distribute population from statistical regions to addresses based on attributes, and to be able to extract occupation levels for individual addresses. It therefore requires the following capabilities:

1. The ability to distribute population statistics for a statistical area across appropriate groups of addresses within that statistical area.
2. The ability to treat addresses differently depending on whether they are associated with a FS that has a multiple building topology.
3. The ability to treat addresses differently depending on whether they have a site-specific TS or will need to use a more general, functional class based TS.
4. The ability to use more than one TS if there is more than one required, i.e. worker, resident and visitor TSs.

The ontology needs to be designed in such a way that it can be queried using a list of addresses within an AOI, and a specific time as input. It must then be possible to:

1. Select each of these addresses based on the address type (as described in Table 27) and then identify:
    a. classification of the address
    b. attributes including size of the building that the address is associated with

c. number of other addresses within that building
d. statistical region with which the address is associated
e. number of addresses of the same type within that statistical region that the population values need to be distributed between.

2. Identify the appropriate TS for each address, whether that is a site-specific or generic TS (based on functional class).
3. Extract the occupation level (as a per cent of full capacity occupation) of the address from the TS.
4. Calculate the estimated occupancy (as number of people) of the address.
5. Sum the population for all of the selected addresses.

Table 27: Specific address cases that require population estimation

Primary Class	In FS	Specific TS	Activity	Potential Data Source
Commercial	No	No	Workers	From census data WZ statistics
			Visitors	Where the address is retail, this is estimated based on other studies, i.e. WZ retail types
		Yes	Workers	From census data WZ statistics
			Visitors	From building estimated floor area and Google Popular Times
Residential	No		Residents	From OA stats
				In Institutions classes
Commercial	Yes	No	Workers	From census data WZ statistics
			Visitors	May be multiple TSs (e.g. at hospitals where there are in-patients, outpatients, A&E, visitors) Note some places will have visitors depending on their type – look at LEISURE classes for defaults for this
			Residents	
		Yes	Workers	From census data WZ statistics
			Visitors	May be multiple TSs (e.g. at hospitals where there are in-patients, outpatients, A&E, visitors)
			Residents	
Residential	Yes	No		

The data sources being used for the population estimation are detailed in Chapter 4, and these are summarised in Table 4 on page 96. While the FS data in the GB data are specific to this data environment, the membership of a building to a group of buildings with an over-arching function

(such as a hospital, school or university) is not unique to the GB data environment. The presence of several addresses with different functions within a single building is also not unique to the GB data environment. Likewise, addresses will fall inside statistical regions that contain many addresses. These relationships are key to estimating population at the address, building or wider AOI and their common presence in different data environments makes this modelling framework transferrable, despite variation in the types of available data in other environments. This may also make the modelling framework suitable for use with other data sources within the GB data environment.

In order to estimate population within an AOI and to ensure that all different address types are accounted for, queries acting on the sub-classes of addresses described in Table 27 are required. These sub-classes represent the complete set of residential and commercial addresses.

6.1.1 Why Semantics are Important to This Use Case

The aim of the modelling framework is to enable estimation of population at very fine spatial and temporal levels of detail (i.e. at the address level at any time of day or night). The modelling framework is intended to enable the allocation of available population, from a variety of sources, to the addresses within an AOI, based on the address function, which has a direct impact on the occupation levels of the addresses.

There are several reasons why this modelling framework is centred on an ontology. To recap, the first reason is that the use of Semantic Web standards, including URIs that have been adopted for this purpose, and RDF, RDFS and OWL means that it is very easy to integrate data in these formats from different sources. In this domain, the geographical features (addresses) serve as the link between other data sources. It is a trivial exercise to add more data to the ontology, using these instances of addresses. The ontology is therefore easily extensible (i.e. easy to add more data, and handle change) so will enable integration of new and evolving datasets (which is very important given the speed of data evolution).

Secondly, the formally defined semantics can be used to provide meaning to the data, allowing hidden information to be revealed through inference. Such inferences can be made using these semantic web tools in several ways. The techniques that are employed in this ontology are discussed in the next sections of this chapter.

A third reason to use ontology is because inferences can be made using this approach. Using GIS for such an application would require a significant investment in linking databases, identifying relations, developing GIS models, writing pieces of code and associated activities. In an ontology

that accurately reflects the relationships between “things”, the declarative definition of these relationships is already present, meaning that further spatial and table analyses on individuals is of less importance. For instance, in the GIS processes for supplementary classification of residential addresses, a count of addresses with each primary class inside each building, is generated by using spatial analysis or database linkage. This is stored in a table that is then appended to the original buildings layer, and each address then uses its association with the building polygon to use these address counts for the identification of flats in the supplementary address classification. This process is greatly simplified when transferred to the ontology. The count of addresses of relevant classes within each building can be calculated as part of a SPARQL query and the count attributes do not necessarily need to be stored. The RDF data model is simple, and does not require additional “work-arounds”, such as joining the output of one analysis back to the source dataset, to achieve full integration of the datasets, unlike the specific workings of complex GIS data models and software. This example is even more relevant where address and building datasets are not as well integrated as in the datasets used here, and where the database linkages are not present.

The ontology design described below is a domain specific, lightweight representation ontology: it is designed to answer specific questions within a specific domain (Hart & Dolbear 2013; Martinez-Cruz, Blanco & Vila 2012). It is a novel case of population modelling in the context of the Semantic Web. There follows an explanation of the ontology design and implementation.

6.2 Ontology Design Process

The process of designing the ontology schema is to first identify the key concepts that need to be modelled to ensure that the ontology can answer all questions that we want it to answer. Secondly, these key concepts need to be arranged into a taxonomy, indicating super and sub classes. Thirdly, the key relationships between these classes are identified and modelled, and finally restrictions on classes are identified and modelled. These stages are detailed below. Each one is guided by the required use within the population estimation domain.

6.2.1 Key Concepts (Objects) and Class Hierarchy

For the purposes of this model, the concepts have been abstracted as far as possible and grouped into three high level classes: Regions, which includes both statistical regions and cartographic objects, Places and TSs. Each has its own subclasses. For instance, subclasses of the Region super-class include cartographic objects, areas of interest and statistical regions. Instances of the Region

classes relate to polygons and instances of the Places class relate to address points as modelled in the GIS. Instances of the TS class relate to tabular data rather than geographic features.

These concepts have been organised into a taxonomy of classes, which is described in Section 6.3.1. Everything within the taxonomy is a sub-class of *All Things*. This is the case in all ontologies (Smith & Munn 2008).

6.2.2 Key Relationships

The process of creating an ontology requires that relationships between the defined classes are identified. These are discussed in detail in Section 6.3.1. Briefly, they are divided into spatial relationships, which are based on RCC8 (Cohn & Renz 2008), and are used for qualitative spatial reasoning, and non-spatial relationships, which mostly relate to the TSs. These clearly defined relationships provide the ability to make inferences from the defined spatial properties. The qualitative representation of the relationships between the classes is therefore of great importance for population estimation.

The spatial relationships between instances of the classes are provided by GIS as the output of spatial analysis, or through database linkage, which, in this data environment, is available in the form of the lookup table between addresses and topographic polygons. Using the ontological approach, the focus is on the relationships that are already provided, so there is no need to represent the geometries within the ontology.

Some of these relationships between instances are not identified by spatial analysis, but by database connections during Stage One of this modelling framework. The spatial relationships remain the same whichever approach was used to define them. For instance, spatial analysis was used to link a topographic area to a FS via a containment relationship, as this was not attributed in the topographic areas data layer. It would be possible instead to attribute the topographic areas layer in the source dataset, which would allow use of a database linkage instead of spatial analysis. The result is the same: the topographic area is considered *part of* the FS. The relationships have been defined on the highest-level classes within the taxonomy. Making the properties as abstract as possible in this way reduces the size and complexity of the ontology and increases its re-useability.

This ontology assumes that places are points, and regions are areas. There may be some circumstances where this is not strictly true. In an alternative model, a place may be represented by a region. There are circumstances where a region could be considered as a collection of other geometries, as sometimes happens with UK postcode geographies, where a unit postcode is

actually a collection of points, rather than a polygon, but is sometimes treated as a polygon enclosing all of those points in the collection. One of the benefits of using the ontological approach is that the representation of these geometries becomes less important than the relationships between them.

In the hierarchy developed here, the spatial relationships between points (places) and polygons (topographic areas) are restricted to the containment relation (*isAssociatedWithRegion*) described in Section 6.3.1, and this will hold true as long as addresses are represented as places continue to be represented as points. In future developments, certain types of place may be represented by polygons, in which case additional spatial relations may be required in the ontology, including those that represent adjacency (neighbourhood), and partial overlaps. Should further research involve extending the ontology to include linear features representing transportation links, the ontology may also need to be extended to incorporate qualitative linear reasoning.

Some of the sub-classes within the taxonomy could be defined not by explicitly stating them in the triple store, but by adding a class restriction. For example, the Residential sub-class of BLPU is defined as a BLPU that has a 'class' attribute value of 'Residential'. Such a class restriction means that the class membership is calculated by properties other than *rdf:type* class membership triple. These class restrictions are not used in this ontology, as the cost of logical reasoning required is too high in this instance, due to the complex ABP hierarchical classification. This is a clear example of the reasoning capabilities and computational expense of the ontology languages and software influencing the design of the ontology. Instead of using class restrictions, the addresses are explicitly assigned a tertiary class membership on import to the ontology.

6.2.2.1 Ontological Limitations

One of the means to make inferences within an OWL ontology is by using a role inclusion axiom. This steps through a chain of properties or relationships. For example, if an address is associated with a topographic area, and the topographic area is part of a FS, then logically the address is also part of the FS.

It would be useful to be able to represent building neighbourhood relationships via a role inclusion axiom as this would allow more of the analysis to be executed in the ontology rather than in the GIS part of the modelling framework. This has not been possible due to limitations of OWL. The role inclusion axiom is capable of identifying first order (immediate) neighbours from a neighbourhood list output from GIS, as well as second order neighbours (neighbours of neighbours). Using the logic in the available reasoner however, this also allows a building to be related to itself as a second order neighbour, unless second order neighbourhood is defined as

disjoint with the building. This produces a logical conflict that cannot be resolved, and is therefore not allowable in OWL and the reasoner fails.

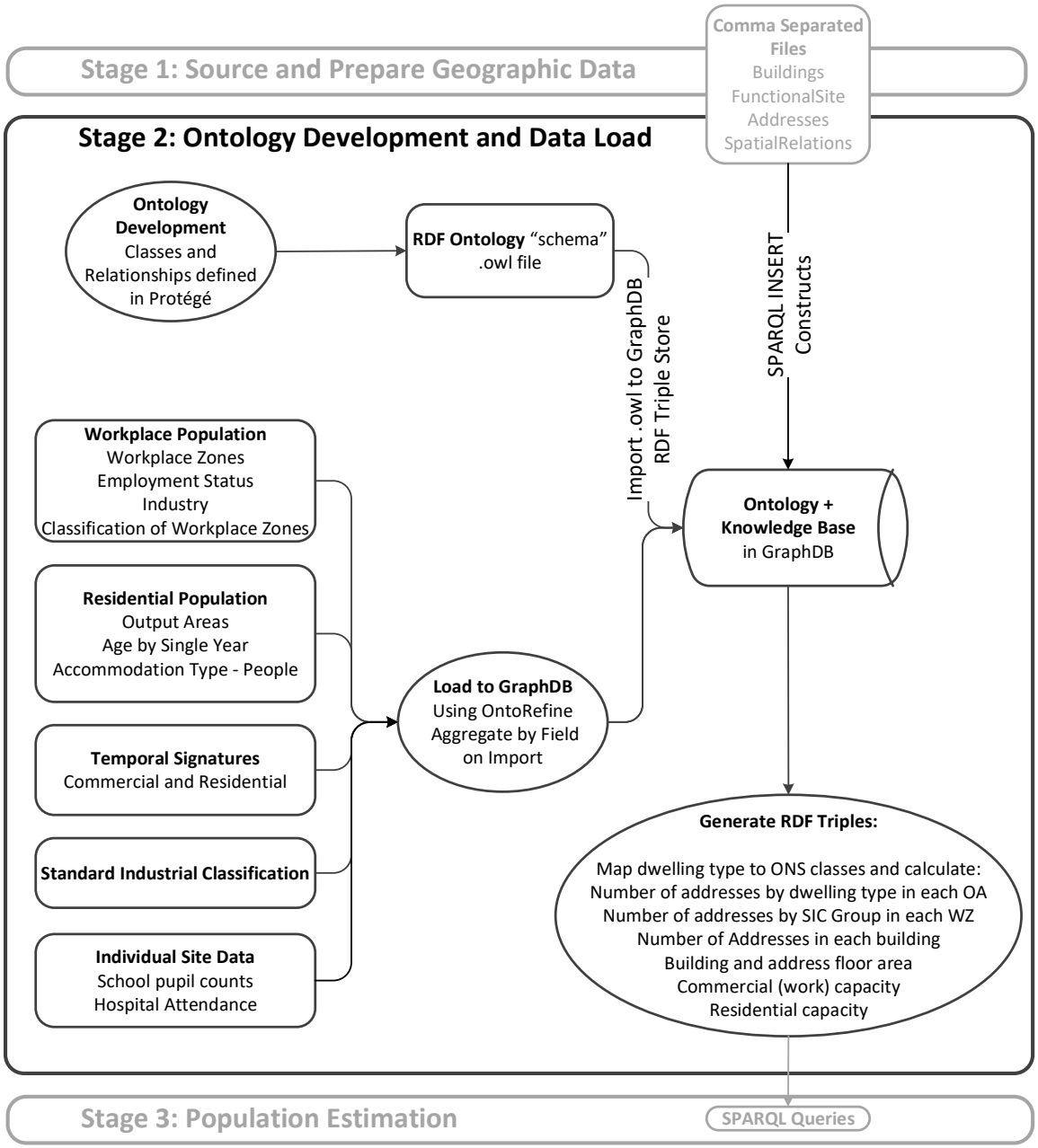


Figure 27: Ontology Build in the context of the modelling framework. The technologies employed are indicated for the different stages of the ontology build.

6.3 Ontology Development Process

The practical need for both straightforward ontology development and inference mean that both Protégé and GraphDB ontology software (see below) have been utilised. The ontology schema was developed using the visual tools available in Protégé, and checked for inconsistencies using the HermiT reasoner. This ontology was then exported as an OWL RDF file (.owl), and imported to

GraphDB using the RDF Import tool. The ontology and data were tested for inconsistencies using the TRREE Engine by importing instances of the classes and relationships from (synthetic and then real world) data files exported from ArcGIS. This is discussed further in the Ontology Validation section at the end of this chapter.

Figure 27 illustrates the flow of activities involved in building the ontology. The input data from Stage One of the modelling framework are in .csv format and contain the qualitative relations between instances of addresses, buildings and FSs, as well as between addresses and statistical regions. The ontology is modelled using Protégé to define the concepts and relationships which are stored in a .owl file which is imported into GraphDB. The GIS data are then imported into GraphDB to create the triple store containing both the schema and knowledge base.

6.3.1 Software Used for Ontology Development and Implementation

The development of the ontology for detailed population estimation utilises several pieces of software. Below is a brief introduction to this software.

6.3.1.1 Protégé

Protégé is a desktop (or web-based) ontology editor that can be used to define the classes, relationships and instances in an OWL ontology. It was developed by Stanford University School of Medicine, as a free open-source tool (Musen 2015). The ontology development described in this thesis uses Protégé 5.2.0 with the HermiT 1.3.8 reasoner.

Protégé provides a good platform in which to develop an ontology and ensure that there are no logical inconsistencies in the ontology itself. However, Protégé's SPARQL plug-in does not allow inference, which is essential to population estimation in this environment. Protégé does allow use of SNAP-SPARQL (V4.2) as an alternative plug-in to using SPARQL. SNAP-SPARQL provides inference capabilities in Protégé, but has expressive limitations in terms of the SPARQL queries that may be asked (some queries that cannot be asked in SNAP-SPARQL are presented in Section 6.6.3).

6.3.1.2 GraphDB

GraphDB is an RDF triple store, developed by OntoText with a free to use version that has been used in this modelling framework. It allows for streamlined use of Linked Data datasets with local datasets. GraphDB can perform RDFS and OWL inferencing, which is a requirement of the model being developed. The inference is performed by the reasoner "Triple Reasoning and Rule

Entailment" (TRREE) Engine. This engine uses forward chaining, which is appropriate to the large knowledge base that needs to be uploaded to the GraphDB triple store, and the number of queries required.

6.4 PopOnt

The aim of the ontology development and build is to facilitate the estimation of the population within an AOI. To this end, when an AOI is identified, all addresses within that area can be used to generate a .csv file that can then be converted to RDF in GraphDB and used for pattern matching to restrict population estimation to the specific list of addresses.

This section sets out the ontology design that has been developed in order to answer this specific question. The top-level taxonomy is presented in Figure 28. The key concepts: Regions, Places and TSs are represented here as a top-level class, along with the relationships between these classes. Two additional top-level classes are also presented: the Interval and SIC Group classes. The remainder of this section provides a discussion of these classes and their relationships in the taxonomy.

6.4.1 Classes

6.4.1.1 Place

Place is the central concept in the ontology. These are places where people engage in activities, such as residential, working and leisure activities. They may be addresses (as in this research), POIs, potentially outside areas such as beaches or parks, or even events, present for a limited time only.

There are no data properties (attributes) set on the top-level Place class because attributes will be dependent on the data environment and any dataset used as a sub-class of place will have its own set of attributes. Places, however, will always have the *isAssociatedWithRegion* relationship with Regions, and the *isModelledBy* relationship with TS as presented in Figure 28 (a) and (b) respectively.

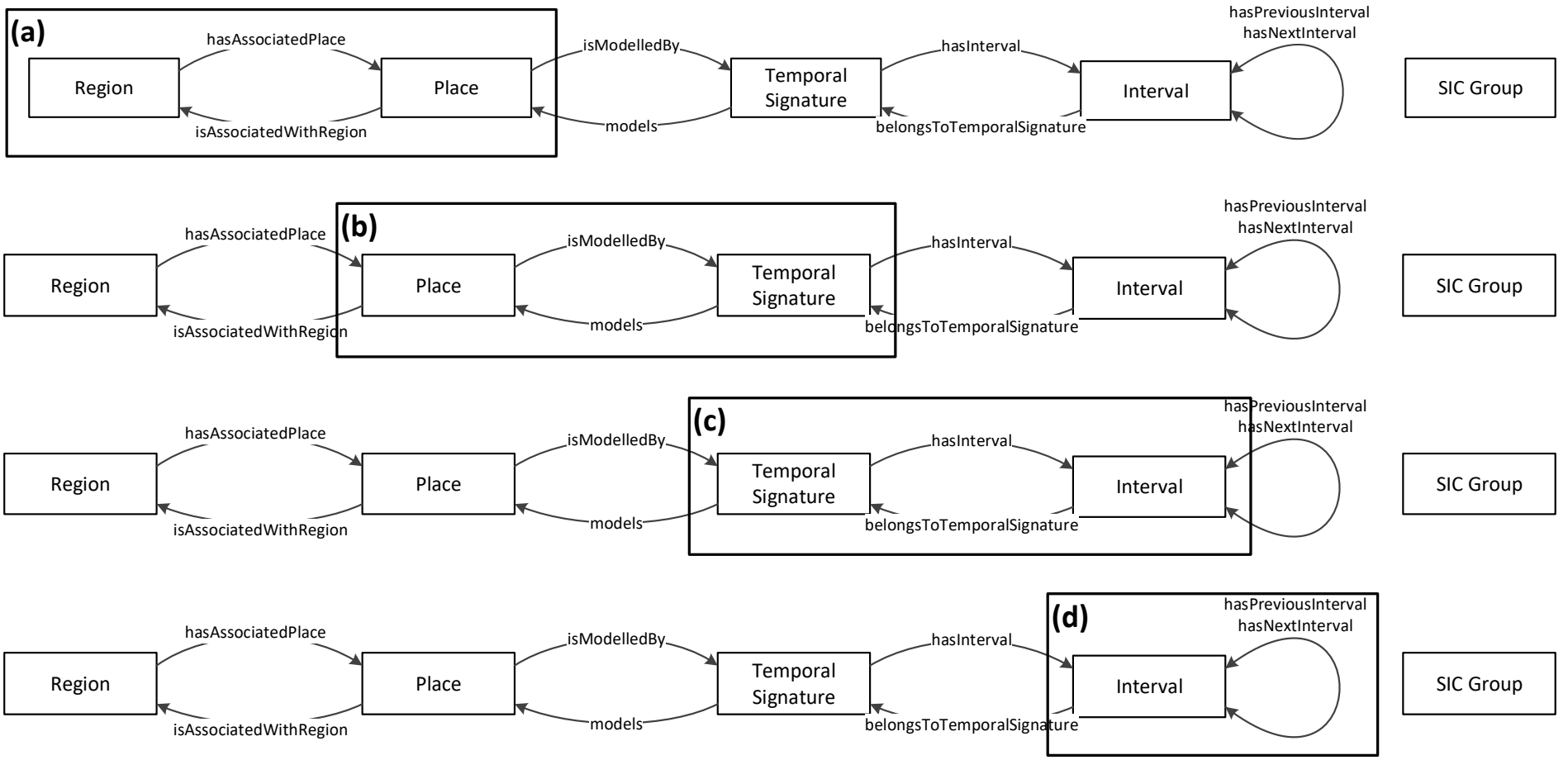


Figure 28: The top level classes representing the key concepts in the ontology, and their relationships : (a) hasAssociatedPlace, (b) isModelledBy, (c) hasInterval and (d) hasPreviousInterval.

6.4.1.2 Place Sub-Classes

Figure 29 shows the structure of the Place sub-classes. Place has a sub-class Address (in this case BLPU), which in turn has its own sub-classes in a hierarchy that mirrors the hierarchical classification of the address data. The two most relevant classes within this hierarchy are Commercial and Residential BLPUs. These are primary classes. Each has sub-classes that mirror the secondary classes in the ABP classification scheme, and each of these has sub-classes that mirror the tertiary ABP classification scheme.

The BLPU class has many data properties including UPRN (the unique identifier for the BLPUs) and the address details, which are taken from the source data. Capacity for residential, work and visitor populations are also present, to be calculated using SPARQL queries described in Section 6.5.

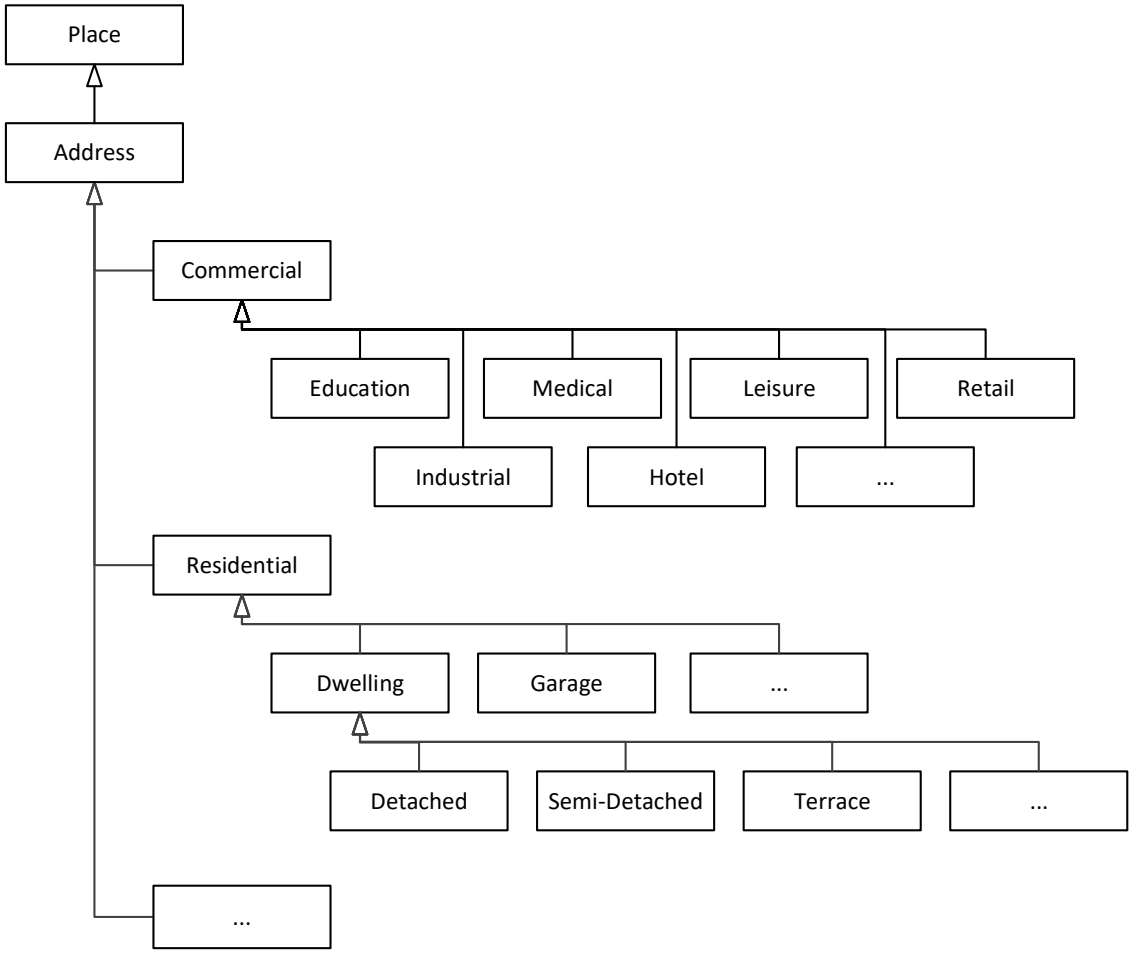


Figure 29: Sub-classes of the Address super-class

6.4.1.3 Region Sub-Classes

Regions can be cartographic objects, statistical regions or any other relevant area type such as an AOI. The hierarchy of Region sub-classes is presented in Figure 30. The model contains two sub-classes of statistical region corresponding to the most detailed level at which the GB data are available (OA and WZ). The hierarchy of statistical regions could easily be extended to include the wider statistical areas if those data were also required, or alternative statistical regions used for producing alternative population statistics, such as Nomenclature of Territorial Units for Statistics (NUTS) areas.

At the top-level, Places are related to Regions via an *isAssociatedWithRegion* relationship and its inverse *hasAssociatedPlace* as presented in Figure 28 (a). The sub-classes also have this same relationship with the Place class (and its sub-classes), inherited from their super-class.

AOI is included in the model to add flexibility: if an AOI is comprised of a set of polygons, then the addresses that have an *isAssociatedWithRegion* relationship with those polygons can be selected, rather than having to load an additional RDF file to the triple store, containing all of the address triples.

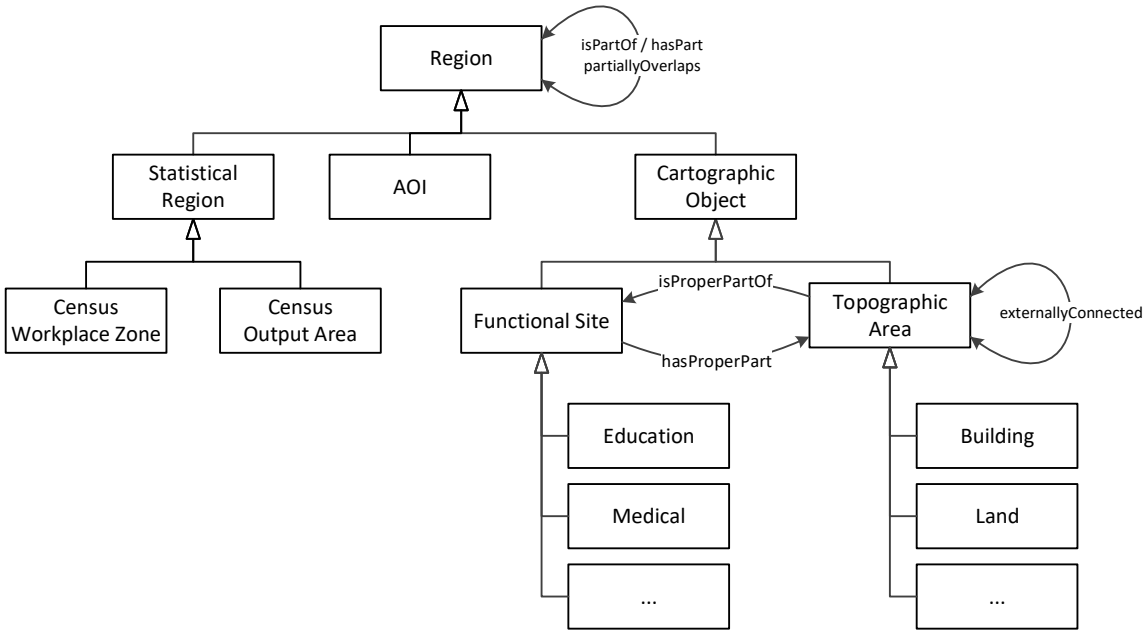


Figure 30: The Region Hierarchy: sub-classes and relationships between regions within the ontology.

6.4.1.4 Temporal Signature

TSs are related to Places via an *isModelledBy* relationship, and its inverse *models*. One instance of each TS class is created for each of the activities that the TSs model: work, residential and visitor. An extract of a TS input file is presented in Table 28, demonstrating the structure of the input file. It is comprised of a list of time periods, with the per cent of a notional capacity that is expected to be present within the address at that time.

6.4.1.5 Temporal Signature Sub-Classes

The sub-classes of TS mirror the sub-classes of the Address class. Where TSs are available for specific Address classes, these can be used via the sub-property of *isModelledBy*, which relates the sub-class of address to the relevant sub-class of TS, as shown in Figure 31.

An example of how this works might be for a bowling alley (tertiary class), will have a bowling alley TS, but it must also have the Commercial Leisure (secondary class) TS and the Commercial (primary class) TS, inherited from their super-classes, because it's further down the hierarchy. The sub-properties do not override the super properties. They are present in addition to them, so the act of applying a TS to a place must involve the selection of the most specific TS. In this way, the addresses with the less detailed classifications will gain relationships to the less detailed TSs. The use of sub-properties allows more flexibility in querying the data.

TSs have only one attribute: *hasActivityType*, which explicitly states the activity type, and therefore the data sources that are to be used for population estimation, and to which each TS should be applied.

Table 28: Extract from Temporal Signature: SIC Group ABDE, Day Type 1, 2-hour Interval duration. The ABDE column provides a per cent occupation level, to be applied to the notional capacity of the address.

Time	Duration	Next Interval	Day Type	ABDE
00:00	360	06:00	1	0.0
06:00	120	08:00	1	2.0
08:00	120	10:00	1	6.1
10:00	120	12:00	1	10.2
12:00	120	14:00	1	9.2
14:00	120	16:00	1	6.1
16:00	120	18:00	1	10.2
18:00	120	20:00	1	4.1
20:00	120	22:00	1	1.0
22:00	120	00:00	1	0.0

6.4.1.5.1 Places and Temporal Signatures

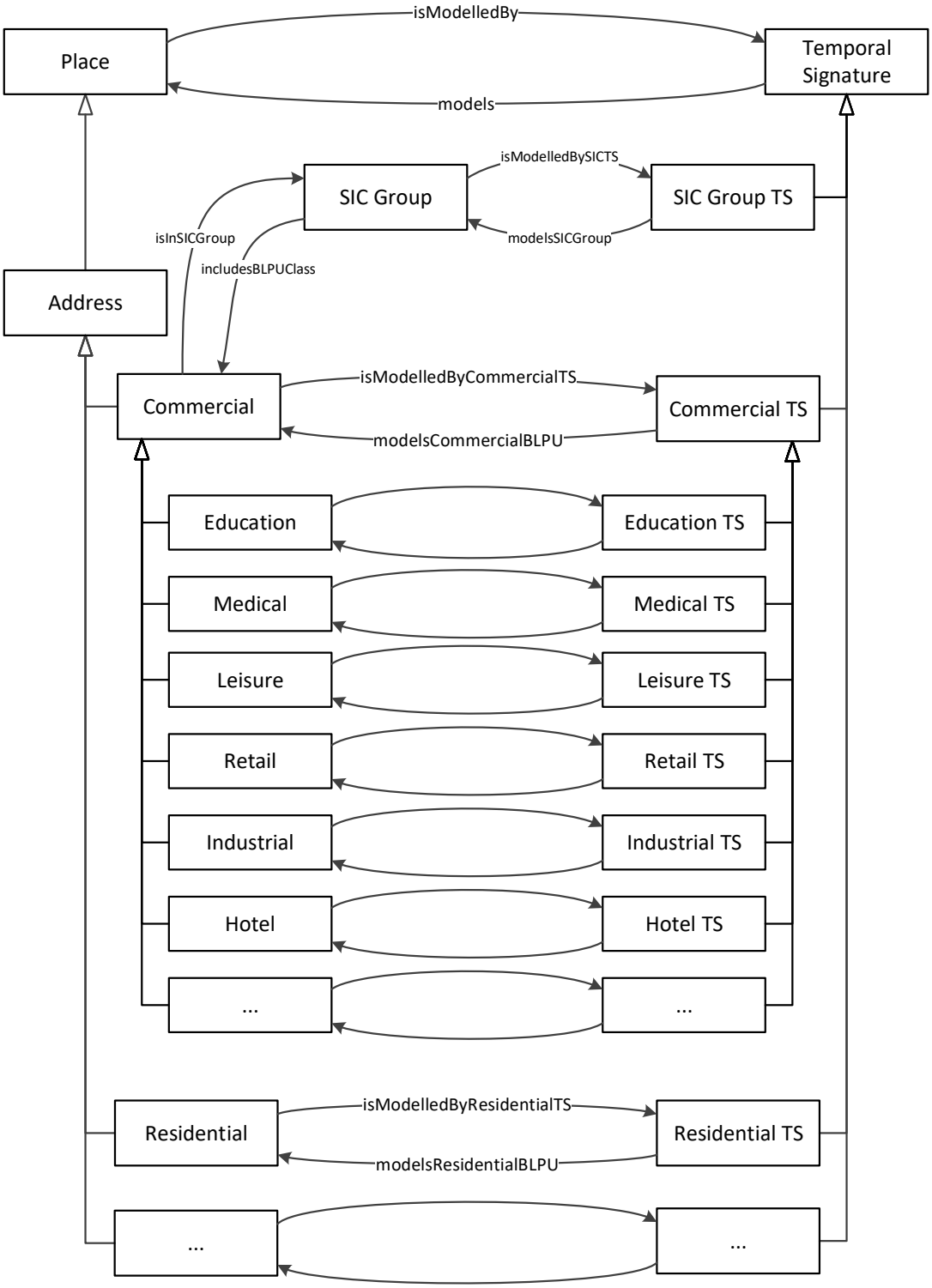


Figure 31: Sub-Classes of Places and Temporal Signatures, demonstrating only selected Residential and Commercial sub-classes of BLPUs and the *models/isModelledBy* properties and sub-properties. The TS sub-class structure mirrors that of the AddressBase class hierarchy.

6.4.1.6 SIC Group

SIC Group is a top-level class included so that an alternative approach to linking Address and TS can be taken. The SIC Group data source is discussed in Section 4.5.4. These are used in the absence of class-specific or site-specific TSs. The SIC Group class has no relationship with any of the other top-level classes, but only with the Commercial BLPU subclass.

The practical use of SIC Groups is discussed in further detail in Section 6.5. It is expected that class specific TSs will not be available for all, or even any, of the functional classes that are indicated in the address data. The model has therefore been designed to allow the use of lookup tables to determine alternative TSs for addresses.

6.4.1.7 Interval

The Interval has a *belongsToTemporalSignature* relationship with the TS class (as shown in Figure 28(b)). A TS may be comprised of many intervals. These have previously been created as a sequential list of times and an occupation level associated with each time, such as that in Table 28. In the ontology, the interval between these times can be of variable length, as the model does not rely on a sequential list of intervals to step through until the correct time period is identified, but instead uses pattern matching to identify the nearest previous time interval to the specified time. In this way, if a building has the same level of occupation over an extended time period, only one interval is required. More details of the time handling are presented in Section 6.6.3.

The attributes of the Interval class are of special interest, as they enable time handling. These attributes are calculated on import of the TSs to the RDF triple store, should sequential searches be required, and are:

1. *hasPreviousInterval* and its inverse *hasNextInterval*
2. *hasPrecedingInterval* and its inverse *hasFollowingInterval*

These have been designed in such a way as to ensure mechanisms can be put in place to step through a time series, forwards or backwards. Given a specific time, it is also possible, using SPARQL, to select all of the intervals with an earlier time, and then find the “maximum” of these times. This provides the Interval in which the specified time falls and is essential for selecting the occupancy rate from a TS.

6.4.2 Relationships

6.4.2.1 isAssociatedWithRegion and hasAssociatedPlace

A place can have an *isAssociatedWithRegion* relationship with one or more Regions. This relationship is so named because in real datasets, address points may not actually be located within a building, but may have a database linkage to the relevant building. This relationship while spatial in nature does not need to be modelled spatially within the ontology, and neither do the regions and places. In this implementation, a place is a BLPU sourced from the ABP dataset.

The primary reason for modelling this relationship is so that BLPUs can be associated with Building polygons, a relationship that can be identified through database linkage or through spatial analysis in Stage One of the modelling framework. This *isAssociatedWithRegion* relationship is defined as a functional relationship. In this model, a place can therefore be associated with only one topographic area in each class, although its inverse *hasAssociatedPlace* relationship is not functional, because a region can have this relationship with many instances of place. This means that the model as implemented assumes that each sub-class of the Region super-class has no overlapping polygons.

6.4.2.2 isModelledBy and models

The object property *isModelledBy* and its inverse *models* describe the relationship between places and TSs**.** Within the population estimation domain, a place is modelled by a TS, and a TS models a place. Within the ontology, this is only defined on those BLPU sub-classes that may be occupied in the real world, and no TS may be created for non-occupiable classes. Places may have an *isModelledBy* relationship with one or more TSs. These represent the different activities that may occur at the place: residential, work and visitor, and there may be more than one TS for each activity, e.g. it is possible for a residential addresses to have a different TS for school aged children than for retired individuals. The inverse of these two relationships will be inferred if not explicitly stated. More on this in the next section.

6.4.2.3 hasInterval and belongsToTemporalSignature

Each TS can be related to many Intervals via the *hasInterval* relationship. The taxonomy with regard to TSs is designed in such a way as to allow a single Interval instance to be used in more than one TS.

6.4.2.4 isAssociatedWithFS

A FS, as a collection of topographic areas, may have one or several addresses falling inside its topographic areas. However, the relationship between the addresses and the FSs is not explicitly stored. This *isAssociatedWithFS* relationship is inferred via a role inclusion axiom. The role inclusion axiom is an explicitly stated relationship that involves stepping along a chain of properties to make an inference that is not explicitly stated. For example, if a point *x* is associated with region *r*, and region *r* is inside region *s*, we can infer that point *x* is associated with region *s* by stepping along the chain of different properties, and do not need to store this information explicitly. This means that new information related to an individual site (such as a university campus or a hospital) can be accommodated as it becomes available, allowing it to be propagated through to all other objects with relationships to the site.

Bearing in mind that *isAssociatedWithRegion* is intended to indicate an "is inside" where data are not spatially aligned, if a topographic area is a part of a FS and an address is associated with the topographic area, then it follows that the address is also associated with the FS.

This means that if population data are available for a single address in a FS that contains several buildings, it is possible to infer which buildings those data refer to using the role inclusion axiom, so that population can be re-distributed across them all. This is a functional property. By definition, a FS contains several buildings and an address instance can only have one *isAssociatedWithRegion* relationship with one building instance. A place (address) can therefore only be associated with a single FS.

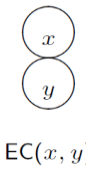
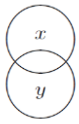
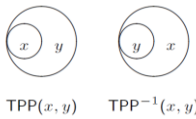
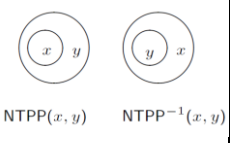
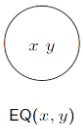
6.4.2.5 hasPrimaryAddress (inverse of isPrimaryAddressFor)

This relationship relates a FS to an Address specified as its primary address. Where data are available for a single FS, the primary address will be used to allocate the population across the various buildings that the FS is comprised of.

6.4.3 Spatial Relations

Figure 30 shows the spatial relations between the different types of region that are modelled in the ontology. These relations are based on RCC8, and the explanations for the limited use of these is laid out in Table 29.

Table 29: Relations from RCC8 relevant to this modelling domain. Only the Proper Part relation is relevant, so this is the only relation modelled in the ontology. These relations are consistent with the Open Geospatial Consortium (OGC) mapping of properties to RCC8 (Hart & Dolbear 2013, p96). Diagrams after Renz (2001, p.44).

Relation (RCC8)	**Example**	**Explanation**
DC disjoint		It is not necessary to model this relation within the ontology
EC externally connected (i.e. touches)		Adding this relation enables neighbourhood to be modelled therefore enabling the ability to identify e.g. a parade of shops.
PO partially overlaps		This relationship could be present for buildings that partially overlap the AOI, and is explicitly modelled here, as the AOI could be represented as a region, set of regions, or a set of addresses.
TPP (TPPi) tangential proper part (and its inverse)		In the ontology, it does not matter whether the part is tangential or not, so this is represented as **Proper Part (TPP & NTPP)**. *isProperPartOf* is equivalent to *within*, and *hasProperPart* is the equivalent to *contains*. Within and contains are familiar spatial relationships represented in common GIS software. These relations are transitive.
NTPP (NTPPi) non tangential proper part (and its inverse)		
EQ equals		A FS by definition contains more than one MMT object so this relationship is not relevant in the topographic areas. It is not relevant to this domain if instances of the other classes are equal.

The difference between the relationships *isProperPartOf* and *isPartOf* must be noted. A FS, by definition, cannot be a single topographic area such as a single building, or a single field, but must be comprised of more than one topographic area. In mereology, proper part is a subset of part, which cannot also be the whole. A topographic area cannot, therefore, have an *isPartOf* relationship with a FS as this requires the FS to be allowed to be constructed from a single part.

An AOI, however, could be constructed from a single topographic area, for example a building that is a block of flats or a shopping mall. It could also be constructed of several areas within any defined polygon. The relationship *isPartOf* allows for both circumstances.

Spatial relations will be stored explicitly in the triple store where they apply. This will allow, for example, addresses associated with buildings to be identified using the property “hasValue” within OWL.

The spatial relations that have been modelled are *externallyConnected*, *partiallyOverlaps* and *isPartOf* and *isProperPartOf*. The *partiallyOverlaps* relation is included for the same reasons outlined for the inclusion of the AOI class (to add flexibility and test the possibilities of the ontology). The externallyConnected relation is included so that building polygons can be grouped together in examples such as a parade of shops, where the building polygons could be selected based on adjacency and the addresses within the parade treated as a single unit, although this is not implemented in this research as it is out of scope.

In logical terms, *isProperPartOf* is a sub-property of *isPartOf*. It is possible that if e.g. a *isPartOf* b, then a could also be equal to b. However, the *isProperPartOf* relation cannot also be equal to. Because a FS must be comprised of more than one Topographic Area, the *isProperPartOf* relationship applies here, as it is not possible for a single topographic area to be equal to a FS. However, it would be possible for an AOI to be equal to a WZ or an OA, or a FS. The *isPartOf* relationship is therefore applied at the Region super-class level.

Cartographic objects will have either an isPartOf or a partiallyOverlaps relationship with the AOI, but never both relationships. This is in line with the RCC8 relations PO and TPP/TPPi as described in Table 29. The data used in this model are known to have correct relationships as they were generated using high quality source data and GIS processing, and the partiallyOverlaps property is not required in any of the testing, as the AOIs are equal to WZ, so only the isPartOf relation is used in this model.

6.4.4 Design Overview

Figure 32 brings all of the classes and relationships together to form a representation of the entire ontology. The sub-classes included in this diagram are kept to a minimum for simplicity, and class attributes are not displayed (For a full diagram of the ontology design, including attributes, see Appendix C). All of the relationships that are defined in the ontology are represented here.

6.5 Using PopOnt

6.5.1 Study Areas

Within the extended SSA, four smaller study areas have been selected in order to build and test the ontology in different circumstances. The first two are the Portswood (characterised in COWZ as a mix of residential suburbs, cosmopolitan metro suburban mix and multicultural high streets, as well as large-scale education) and Eastleigh (largely residential or traditional high streets and large scale education) study areas that were used for the residential address supplementary classification. The second two are in Shirley (containing Southampton General Hospital) and Itchen (which contains several schools for different age groups). All study areas have been loaded into their own data repository so that testing and validation can be performed separately under the different conditions.

6.5.2 Schema Load

The schema refers to the ontology: the assertions about which classes are allowed, which attributes can be present on the different classes, and what the relationships are between them. Loading the schema into GraphDB requires only the .owl file developed and saved in Protégé.

6.5.3 Knowledge Base Assertions

For each of the four study areas, the knowledge base assertions include the RDF triples that generate instances of the data, their attributes and the relationships between the instances.

The process of loading data into GraphDB involves first loading the data file (using OntoRefine). The data load tool allows for selective load of columns and rows of data from the .csv files. Once the data are loaded to GraphDB, SPARQL scripts are used to create the RDF triples from the loaded data file and insert these to the triple store. Using SPARQL to generate of the RDF triples means that filters can be applied and fields can be aggregated before any triple is actually written to the data repository.

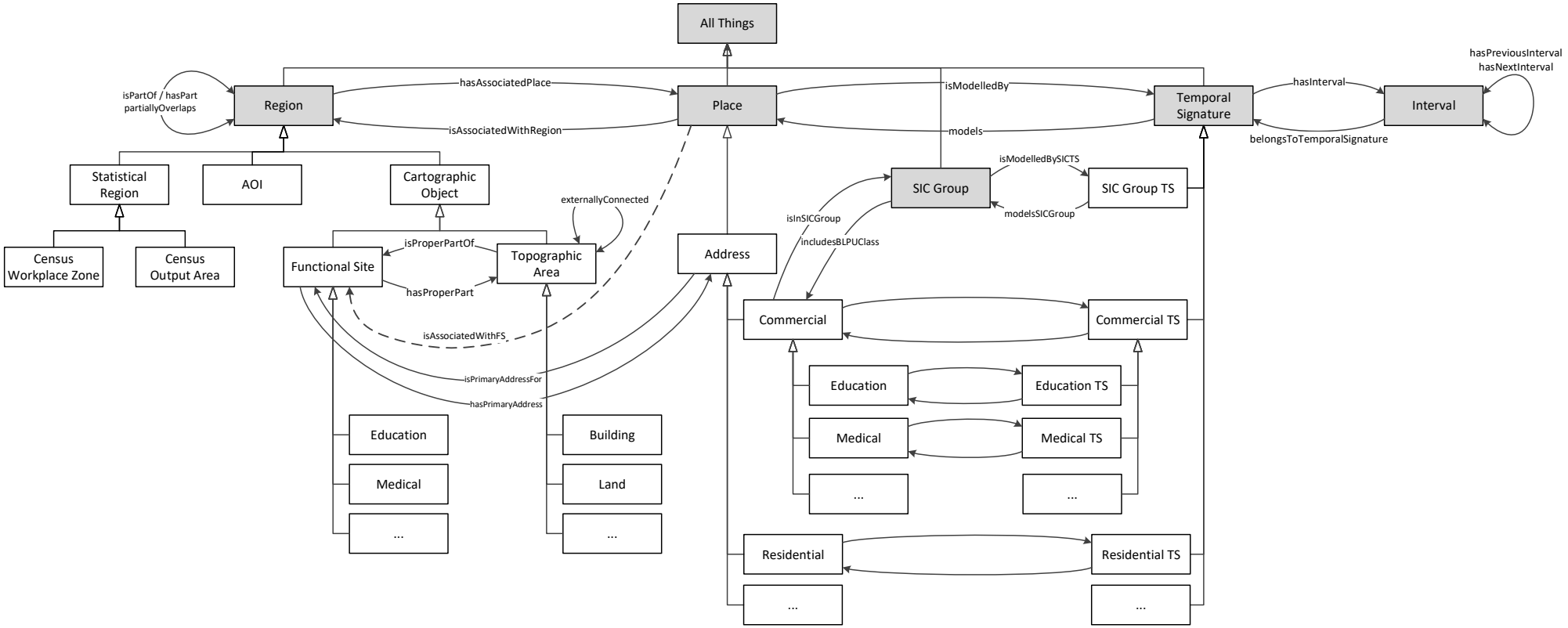


Figure 32: Ontology Design: the complete taxonomy, for the ideal ontology in this domain. The full diagram as implemented (with adjustment for the limitations of OWL) and including attributes, can be found in Appendix C.1.

6.5.4 URI Naming Conventions

URIs are generated for each instance of a class within the same SPARQL scripts that also load and calculate data. URI naming conventions as used in this modelling framework are presented in Table 30. These conventions have been chosen so that there is a clear ability to deconstruct the URI to identify both the class, and the instance to which the URI refers. While there is a clear argument for using random URIs (Hart & Dolbear 2013, p88), in this instance, the data are not linked and open, and clarity is important.

Table 30: URI Naming Conventions used in this model

Class	URI Naming Convention	Justification
Temporal Signature	**TS_Residential**	Only one residential TS
	TS_Class_*activityType*	Each class of Place can have its own TS for workers, visitors and residents.
	TS_SICGroup_*groupname _activitytype*	Each SIC Group has a default work TS. SIC groups are relevant only for the work activity.
	TS_*URI_activitytype*	For individual instances, where specific TSs are available.
Interval	**Int_TS_Residential.***daytype. hour.min*	All Intervals for residential belong to the only residential TS, and all refer to the "residential" activity. Using day type, hour and minute in the URI means that the URI can be deconstructed (or constructed) to discover which time interval this interval instance refers to.
	Int_TS_Class_*ClassCode. ActivityType.daytype.hour.min*	All intervals for Class TSs.
	Int_TS_SICGroup_*groupname. activitytype.daytype.hour.min*	As with the residential intervals, the naming convention is used so that the URI can be deconstructed or constructed to identify either which time interval this instance refers to, or to select the interval instance from specified time.
	Int_TS_UPRN.*uprn. activitytype.daytype.hour.min*	For TSs related to instances rather than entire classes
Topographic Areas	**TOID_***toidvalue*	The TOID is a unique identifier so is used for the URI. The TOID prefix identified the type of the instance from its URI
Functional Site		

Class	URI Naming Convention	Justification
Address	**UPRN_***uprnvalue*	As with the TOIDs, UPRN is a unique identifier and can be used to identify the instance from the URI if a UPRN is known.
Output Area	**OA_***oaid*	OAID is a unique identifier. OA indicates the class of the instance that the URI refers to.
Workplace Zone	**WZ_***wzid*	WZID is unique identifier. WZ indicates the class of the instance that the URI refers to.

6.5.5 Prerequisite Calculations

Once the source data have been loaded, there are a set of calculations that can be made on the knowledge base. These could easily be scripted into the population estimation scripts described in Chapter 7, but for the purpose of a proof of concept, they have been separated into their own scripts so that each can be independently checked. These calculations are described in Figure 33.

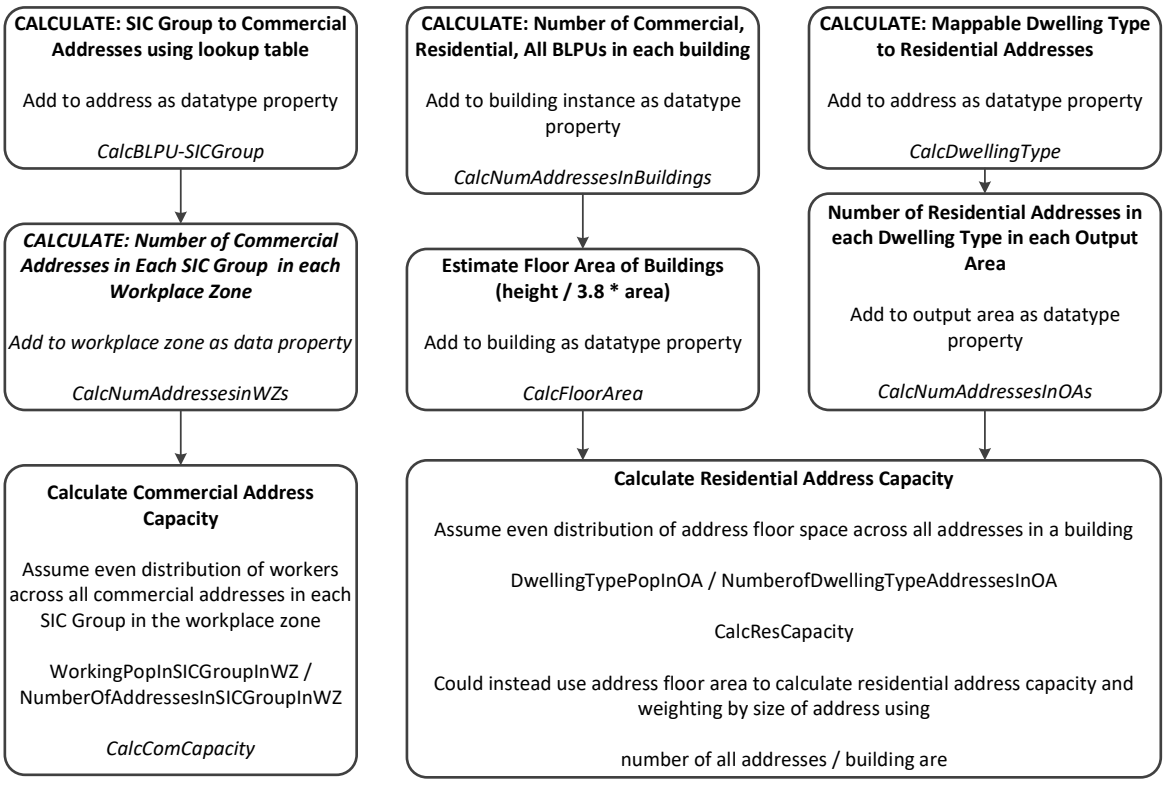


Figure 33: The Prerequisite Calculations on the Knowledge Base: to simplify the population estimation scripts.

Table 31: Datasets loaded to the data repositories within GraphDB

Dataset	**Extent**	**Source**	**RDF Instances and Properties Added**
Temporal Signatures: SIC Groups and Residential	Universal (the same anywhere in this data environment)	Spreadsheet containing TS based on data from Time-Use Survey	Creates residential, work and visitor instances of TS for each SIC Group. Instance URI has TS_*SIC_Group_activityType* prefix. Separate script creates a single instance of residential TS.
Interval Instances	Universal		Loads contents of a single file containing all SIC Group TS data. Generates instances of Interval, with properties indicating duration, occupancy, day type, next interval, and relationship *belongsToTemporalSignature*.
Output Areas	SSA	Attribute Table of OAs	Creates OA instance for all the OAs in the input csv file, with unique identifier attribute. URI for OA is based on OA ID, and has the prefix OA_.
Workplace Zones	SSA	Attribute Table of WZs	Creates WZ instance for all the OAs in the input csv file, with unique identifier attribute. URI for WZ is based on WZ ID, and has the prefix WZ_
COWZ Classes	SSA	Table of COWZ classes	Creates COWZ class data property on existing WZ instances only.
Functional Sites	SSA	Attribute Table of FSs	Create instances for all FSs with data properties for unique identifier (TOID), function and theme, as well as primary address (linking to URI based on UPRN), stakeholder and distinctive name where these are available. Instances are created within their sub-classes, whose names are based on the Theme of the FS. The URI is based on the TOID, with TOID_ as the prefix.
Address Instances	X4 study areas	Attribute Table of BLPUs	Create instances of BLPU sub-classes, by constructing sub-class name based on class code. Adds *isAssociatedWithRegion* relationship, and tertiary classification as well as dwelling type to the instance. URI is based on UPRN and has prefix UPRN_.
Building Instances	X4 study areas	Attribute Table of MMT Areas	Creates instances of buildings. The URI is based on the TOID and the instance has the prefix TOID_.

Dataset	Extent	Source	RDF Instances and Properties Added
Building Relationship with Functional Sites	X4 study areas	Buildings output from GIS process	Adds *isProperPartOf* FS relationship for Building instances where it exists.
Building Area Attributes	X4 study areas		Adds building area properties for Building instances from the attribute table.
Building Height Attributes	X4 study areas		Adds building height properties for Building instances, from the attribute table where these exist.
Building Dwelling Type Attribute	X4 study areas		Adds dwelling type property for Building instances where this exists.
Address associated with OA	SSA	Table of containment relationships for OAs and WZs	Adds the containment relationship Address *isAssociatedWithRegion* OA where the Address instance exists thereby restricting RDF triple creation to those instances within the study areas.
Address associated with WZ			Adds the containment relationship Address *isAssociatedWithRegion* WZ where the Address instance exists thereby restricting RDF triple creation to those instances within the study areas.
QS401 OA Accommodation Type	SSA	Table of Statistics attributes for OAs	Loads ONS statistics to RDF triples where the WZs or OAs exist (i.e. the SSA). The published statistics are aggregated on import so that • data properties for number of people in each accommodation in the OA are created rather than the published accommodation types • data properties for demographic groups rather than individual ages • data properties for SIC Groups are created rather than for individual industries • data properties for broad employment status groups are created rather than for the published ONS employment status'
QS103 OA Age by Single Year			
QS605 WZ Industry		Table of Statistics attributes for WZs	
QS601 WZ Employment Status			
Schools	Itchen Study Area	Attribute Table for individual schools	Loads pupil number properties to schools using URI constructed from the UPRN field in the Schools data to identify the Address instance.

Dataset	Extent	Source	RDF Instances and Properties Added
Southampton General Hospital	Hospital Site Only	Attribute Table for single site	Creates Hospital Visitor TS specific to Southampton General Hospital primary address. Creates interval instances for hospital visitor TS.

6.5.6 Results of Data Load

The result of the data load and calculations described above is a fully functioning RDF triple store, checked for inconsistencies, and with some data properties pre-calculated to improve the ease of population estimation.

6.5.7 Handling of Data Inconsistencies

Logical inconsistencies in the model are identified by the Protégé reasoner prior to load into GraphDB. Logical inconsistencies in the data are handled by GraphDB. Where there is a logical inconsistency in the data (e.g. a URI is declared as two disjoint classes), they are revealed when SPARQL scripts to insert the RDF Triples are executed.

6.6 Ontology Validation

To ensure that the ontology is functioning as expected, synthetic test data have been generated specifically for this purpose and then imported to a data repository within GraphDB. Queries were first constructed to test whether the data are in the correct format, and then to test whether inferences, aggregations and arithmetic operations can be effectively carried out. This process was then repeated on a very small sample of real-world data.

There follows a description of the synthetic data creation, and how this was tested against real data.

6.6.1 Rationale for Use of Synthetic Data

The purpose of the ontology is to enable use of the pre-prepared data to facilitate the estimation of population at fine spatial and temporal levels of detail. To this end, there are certain questions that need to be answered, and these have already been outlined in Section 6.1. Specifically, it must be possible to re-distribute population from statistical regions to addresses, based on attributes of those addresses, and to be able to extract the occupation levels at a specific time.

In order to answer the questions that need to be asked of the ontology, it must be possible to query the knowledge base to filter query results, and to perform aggregation, arithmetic and inference operations and to be able to write the results of these back into the RDF triple store. These capabilities are all exploited in the SPARQL scripts.

In order to test the capabilities of the ontology, very simple data need to be available. Real world address and building data are highly complex, with many configurations of building morphology present in the buildings data, and many different combinations of address types present in the address data. At times addresses are linked to buildings only through database linkage: a relationship not always present when spatial analysis is used to find connections between features. FSs have database linkage with buildings, but buildings do not have a link to FSs. FSs may be non-continuous. An AOI may overlap with buildings, whose addresses may fall inside or outside of the AOI. Consequently, even when a very small sample area is used, the data complexity can make it very difficult to unpick the relationships that are present between the real world objects. For testing the ontology itself, it is best to use the simplest data that have the desired relationships for modelling for population estimation.

For this reason, synthetic data have been generated specifically to test the queries and inferences that need to be made within the ontology. These data mimic the real-world data environment, but are simplified so that they are manageable. The relationships are exactly those required to test various scenarios. The volume of data is therefore greatly reduced. The attribution is reduced to the minimum required for inferences and the complexity is removed. The relationships that are specifically modelled for this purpose are:

1. there are buildings that overlap the AOI
    a. a building overlaps the AOI but its address falls outside of the AOI
    b. a building overlaps the AOI and its address falls inside of the AOI
2. there are buildings that fall completely within the AOI
3. there are buildings that fall completely outside the AOI
4. an address point is associated with a building polygon that it does not geographically fall inside
5. there are buildings containing several addresses with a variety of functional classes
6. a FS contains more than one building polygon but only one address
7. a building has several addresses that are associated with it

The data layers presented in Figure 34 comprise the entirety of the synthetic data. They mirror the real-world data layers and attributes, and include simulated addresses, with address classification attributes; building polygons with theme attribute; FSs with theme attribute and an AOI covering a variety of data scenarios. The scenarios, along with validation results, and pseudo-code are presented in Table 32. These required testing within the ontology to ensure that the relevant inferences can be made, and that queries return the appropriate information. The process of validation went through the following stages:

1. The data instances and relationships were initially created manually in the Protégé ontology editor, to test for inconsistencies within the data and the model.
2. The SPARQL queries were generated in GraphDB and the results recorded.
3. The data layers were re-created as spatial entities in ArcGIS and their attribute tables and relationship tables exported to .csv files. These files were used to create instances of the various classes and sub-classes within the ontology using the OntoText import tool.
4. The SPARQL queries developed in Step 2 were applied to the GIS generated data to test for consistency with data from Stage One of the modelling framework.

6.6.2 Synthetic Data Validation Results

The results of the SPARQL queries that act on the manually created synthetic knowledge base are identical to the results of the queries as applied to the knowledge base generated using ArcGIS and imported using GraphDB tools. This proves that the export process from Stage One of the modelling framework is appropriate.

The results of the SPARQL queries prove, when checked against the synthetic data, that the SPARQL queries are giving appropriate responses.

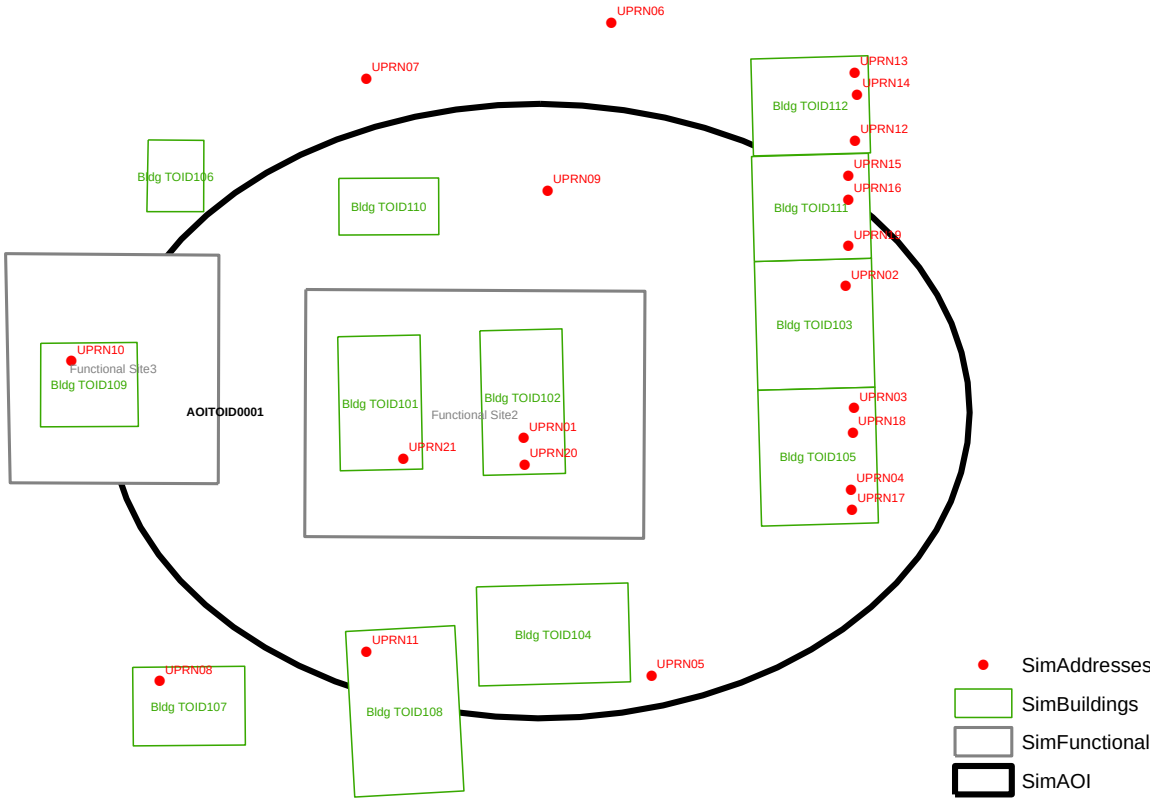


Figure 34: Synthetic Data for testing the ontology design

Table 32: Scenarios for ontology testing using synthetic data. There are 23 BLPUs in the synthetic data.

Scenario to Test	**SPARQL Pseudo Code for Query**	**Result**
There are addresses both inside and outside an AOI – is it possible to select only addresses within the AOI	address is a type of BLPU and address isAssociatedWithRegion AOITOID0001	Returns 13 results – all correct: BLPU 1, 2, 3, 4, 5, 9, 11, 16, 17, 18, 19, 20, 21
There are addresses both inside and outside an AOI – is it possible to select only addresses within the AOI, by primary class	address is a type of R_BLPU. and address isAssociatedWithRegion AOITOID0001	Residential: Returns 7 results – all correct: BLPU 3, 5, 9, 11, 16, 17, 19 Commercial: Returns 6 results – all correct: BLPU 1, 2, 4, 18, 20, 21
There are buildings that are inside and outside the AOI – is it possible to identify the buildings with which the AOI residential addresses are associated?	aoiAddress is a type of R_BLPU and aoiAddress isAssociatedWithRegion AOITOID0001 and building is a type of BuildingsMMTArea and aoiAddress isAssociatedWithRegion building	Returns 6: 3 (bldg. 5), 5 (bldg.4), 11 (bldg. 8), 16 (bldg. 11), 17 (bldg. 5), 19 (bldg. 11) Removes address 9 as this is not inside a building

Scenario to Test	SPARQL Pseudo Code for Query	Result
Some buildings overlap the AOI – is it possible to select all addresses in the buildings the AOI addresses are associated with, whether or not they are in the AOI themselves?	aoiAddress is a type of BLPU and aoiAddress isAssociatedWithRegion AOITOID0001 and aoiAddress hasBLPUClassCode "Residential" building is a type of BuildingsMMTArea and aoiAddress isAssociatedWithRegion building address isAssociatedWithRegion building and address bpop:hasBLPUClassCode "Residential"	Returns 7: 3 (bldg. 5), 5 (bldg.4), 11 (bldg. 8), 15 (bldg. 11), 16 (bldg. 11), 17 (bldg. 5), 19 (bldg. 11) This is correct – address 15 in building 11 is actually outside of the AOI and would not necessarily have been listed in the AOI addresses.

Some addresses fall inside an FS –is it possible to select only addresses in the buildings the AOI addresses are associated with INCLUDING ONLY/NOT INCLUDING FS Addresses	aoiAddress is a type of BLPU
	and aoiAddress isAssociatedWithRegion AOITOID0001
	and aoiAddress hasBLPUClassCode "Commercial"
	
	building is a type of BuildingsMMTArea
	and aoiAddress isAssociatedWithRegion building
	
	address isAssociatedWithRegion building
	and address hasBLPUClassCode "Commercial"
	MINUS{
	building is a type of BuildingsMMTArea
	and building isPartOf fs
	}
		Correct result returned

Scenario to Test	SPARQL Pseudo Code for Query	Result
The BLPU count for different primary classes in regions is required to identify e.g. estimated floor area of an address – get BLPU Counts Inside Buildings	Count (grouped by building): address isAssociatedWithRegion building and address is a type of BLPU and building is a type of BuildingsMMTArea UNION Count (grouped by building): address isAssociatedWithRegion building and address is a type of R_BLPU and building is a type of BuildingsMMTArea UNION Count (grouped by building) address isAssociatedWithRegion building and address is a type of C_BLPU and building is a type of BuildingsMMTArea	24 results – correct

6.6.3 Real-World Data Validation Results

The final stage of ontology development is to validate the model to ensure similar results are being returned as compared with the results of ArcGIS queries. This level of validation is executed on some real-world data on a subset of the Itchen Study Area data, which has its own data repository within GraphDB. This repository contains 300 addresses, 221 buildings with addresses and one FS.

The SPARQL scripts developed for the final data load have been tested in the small sample data load. The results of the data load validation are summarised in Table 33.

Table 33: Results of validation of real-world data load

Script	Adds # instances/triples	Validation technique and Result
Load Addresses	1570 statements added – are 267 BLPUs in input csv. All have UPRN, class code, class description and chosen TOID. 20 have no DTWFlats attribute.	Checked results in .csv files output from ArcGIS: 1570 rows of data in .csv –All have UPRN, class code, class description and chosen TOID. 20 have no DTWFlats attribute
Load Buildings	413 statements added –and 413 buildings in class hierarchy	Checked results in .csv files output from ArcGIS: 413 rows of data in the input .csv
Load Buildings FS Attribute	23 statements added	Checked results in .csv files output from ArcGIS: 23 buildings in input csv that have a fs_toid attribute (i.e. that are inside FSs)
Load Buildings Height Attributes	For Itchen Mini: 2035 statements added	Checked results in .csv files output from ArcGIS: there are 6 buildings in input csv that have no height attributes – there are 5 height attributes and 413 buildings = 407 * 5 rows to add = 2035
Load Buildings Dwelling Type Attribute	For Itchen Mini: 198 statements added	Checked results in .csv files output from ArcGIS: 198 rows in input csv, all have a DwellingTypeWithFlats attribute (does not include buildings that don't have addresses)
Load OAs	Added 1212 statements – oaIRI only – no extra fields	Confirmed count in source data: 1212 OAs in SSA
Load WZs	Added 358 statements – wzIRI only – no extra fields	Confirmed count in source data: 358 WZs in SSA

Script	Adds # instances/triples	Validation technique and Result
Load FS	Added 10 statements – there are 2 FSs created when the LoadBuildingsSPFS is run (because *isProperPartOf* relationships that are added can only have FSs as their predicate), although that these are FS is inferred. This code makes it explicit and adds all the relevant attributes.	Count of FS in study area in ArcGIS: 10 FS in the validation study area
Load COWZ	Loads 1432 statements – 4 attributes x358 WZs. = all of the WZs in the study area –	There are 1212 OAs in input data, the COWZ attributes are loaded for all WZs not just those in the small study area
Load QS401	Adds 7272 statements – 6 per OA (1212*6)	1212 OAs in input data
Load QS103	Adds 8484 statements – 7 per OA (1212*7)	1212 OAs in input data
Load QS605	Adds number of WZs (358) * number of new attributes (8) = 2864 statements	358 WZs in input data
Load QS601	Adds number of WZs (358) * number of new attributes (4) = 1432 statements	358 WZs in input data
Load SIC Group TS	Creates 30240 statements: 3360 for each of the 8 SIC Group + 3360 for residential	
Load BLPU-OA Attribute	Generates 267 new statements	267 BLPUs in repository
Load BLPU-WZ Attribute	Generates 267 new statements	267 BLPUs in repository
Add SIC Group to Individual Addresses	Creates 31 new statements	There are 35 commercial BLPUs in ItchenMini Queried to explain difference selects the 4 Commercial BLPUs that are missing: 2 have tertiary class CZ01 - signage, one has CR11 - ATM, and one has no tertiary class, but its class code is CS – storage land. There is no SIC Code in the lookup table for these classes as they are not occupiable.

Script	Adds # instances/triples	Validation technique and Result
Add Mappable Dwelling Type to Addresses	Adds 247 statements	This matches the input data which has 267 BLPUs – the missing 20 have no dwelling type as they are not occupiable (telecommunication, street record, advertising hoarding, post box etc.)

6.7 Chapter Review

This chapter has provided an explanation of the design decisions made in the development of the ontology and highlighted some of the practical issues with, and reiterated the advantages of semantic web technologies within this application domain. Despite some limitations of the technologies, the techniques employed in the ontology have been successfully demonstrated as appropriate as all data load and query validation shows.

After this stage, the data required for proof of concept that these technologies incorporate all of the necessary components for population estimation have been loaded into the ontology, and have appropriate relationships defined. The final part of the modelling framework involves generating SPARQL queries to estimate the population in a specific set of addresses at specific times and this is covered in the next chapter.

.

Chapter 7 Population Estimation

The previous chapters have outlined the processes used to prepare the input data and supplement classifications where necessary, detailed the design and build of the ontology and the population of the triple store that will be the target for SPARQL queries designed to estimate population within an area at a specific time of day.

On completion of Stage One and Two, population data have been sourced from the census working and residential populations. Site-specific data for schools (capacity only), one hospital (capacity and TS) and five other commercial properties (TS only), have been obtained from alternative sources. Small area statistics, site-specific population data and TSs have been loaded into the triple store along with the geographic data and relations, and prerequisite calculations have been carried out adding additional data to the repository (see Figure 35).

The final part of the modelling framework involves estimating the population. This is done in one of two ways, depending on the available data. The first is by estimating maximum population for individual addresses using disaggregation of small area statistics. TSs are then applied to these maximum capacity values. The second approach, to be used when single site commercial data are available is to use a specific TS for the address. In this case, the capacity will be sourced either from the TS (if these are sourced with counts rather than percent occupation) or based on the size of the address, as a proportion of the building that it is related to.

This chapter describes the methods used for estimating population at specific times for specific data situations. The methods set out below have been implemented to demonstrate that the ontology approach, and the technologies employed are suitable for estimating spatio-temporal population. A fully-developed estimation algorithm is not an aim or objective of this part of the modelling framework, so it is not expected that the estimated population counts are accurate, but that the broad changes in the counts over time can be demonstrated. To this end, there are a series of requirements for the design of the ontology and query of the data repository. These are set out in Section 6.1 on page 165. They are the ability to disaggregate small area statistics to addresses, the ability to treat addresses differently depending on whether they are part of a FS or whether they have site-specific TSs, and the ability to use more than one TS for a single address. How the ontology is designed in order to meet these requirements has already been discussed in Chapter 6.

This chapter focuses on how the data repository can be queried to estimate population for an AOI. It begins with a review of the requirements of the algorithms in the context of the proof of concept, and continues with a description of the method of querying the data repository with a

breakdown of the algorithms used for population estimation. Finally, the validation of the model output is discussed, with results presented graphically.

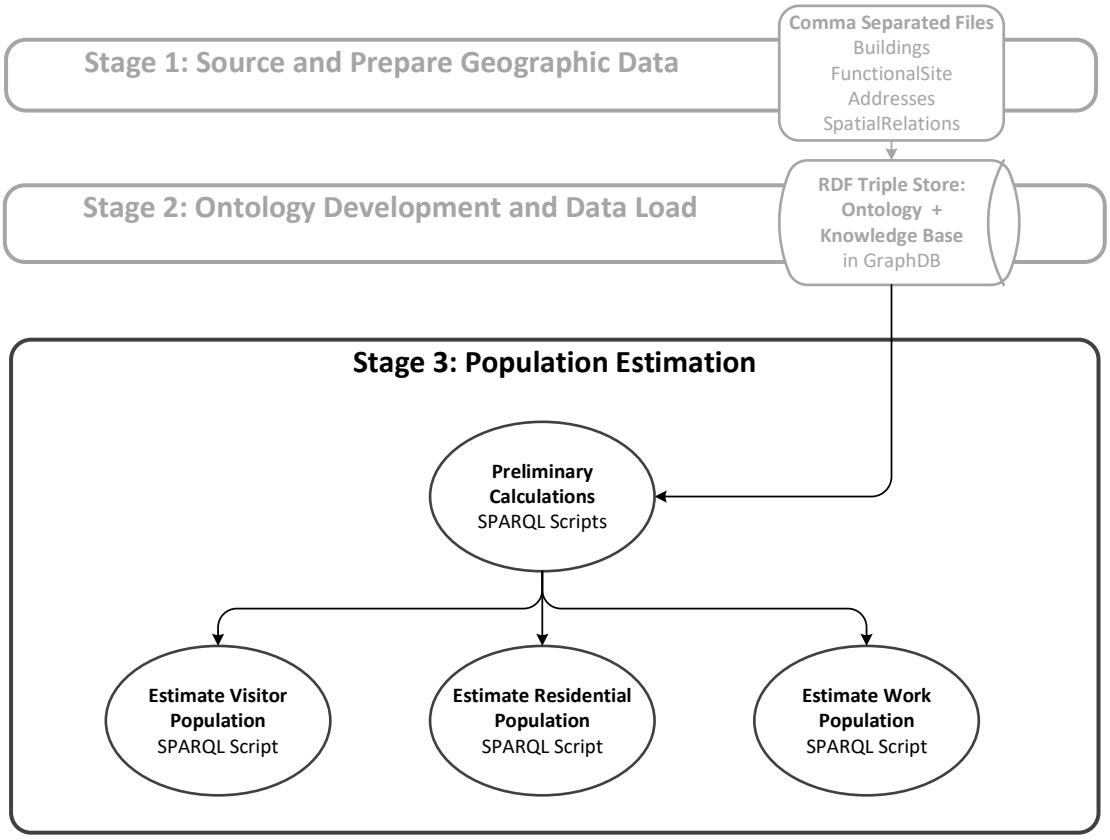


Figure 35: Population Estimation in the context of the modelling framework: the stges involed in Poulation Estimation

7.1 Proof of Concept Requirements

It is essential that the model developed in Chapter 6 can be queried effectively in order to estimate population. The requirements of the estimation algorithms are set out in Section 6.1 on page 165. To summarise, these are focused on the ability to select subsets of addresses based on: their classification; the regions with which they are associated; numbers of other addresses of each type within those regions; properties of the buildings with which the addresses are associated, including the number of other addresses within the building.

It is also necessary to be able to select the TS, and properties of the relevant interval based on the time supplied, as well as to aggregate calculated values to return a population estimation for the selected addresses in the AOI.

7.2 Population Estimation

Population is estimated for an AOI rather than for individual addresses. An AOI could be defined as a set of addresses added to the RDF triple store prior to the time of query. However, for the purpose of proof of concept, a simpler approach is to use existing regions as an AOI. In this case, WZs are used as the AOI so that performance can be assessed for AOIs with different, well-defined, characterisations (from COWZ).

Each address has three maximum capacity properties. One for each of residential, work and visitor populations. The maximum capacity is calculated for both residential and commercial addresses based on the disaggregation processes outlined in Figure 33, and is sourced from the census small area statistics. For visitor populations, the maximum capacity is estimated based on building size and the number of addresses within the buildings, unless recorded visitor numbers are available. For example, the hospital visitor data are from an official NHS source and the magnitude of visitor numbers is known, whereas the TSs derived from GPT have no magnitude, so this must be estimated based on the proportion of the building that the address occupies, along with its estimated floor area. It is possible that some industry data could provide these population magnitude data; however, these will often be considered commercially sensitive and will therefore not be made available in the public domain.

Population is estimated for addresses using a decision rule set and the maximum capacity of the address is represented as a floating-point number, so the population estimation is a floating-point calculation.

It is important to note that the estimated population is not a building attribute, nor is it an address attribute, but is a value that changes over time and is calculated only for the specified query time. In the modelling framework as implemented, it is never recorded as a per-address value, but is estimated for a set of addresses comprising an AOI. The population attribute stored for individual address is the maximum capacity, broken down by activity type. This is the most people that would be expected to be present at a single address at *any* time.

7.2.1 Handling Time

The key to estimating the population is the ability to select the correct TS interval and query the occupancy rate property of the interval, so that it is possible to apply the correct occupancy rate to the maximum capacity for each individual address. The TS chosen is based on the functional class of the address and the activity being modelled. Each activity is modelled separately, and each activity can have its own TS for either a SIC Group, or an individual address (it is also possible

to add a functional class TSs but there is no current data for these, so they are not included in the estimation scripts). These different types of addresses are handled in separate sub-queries within the SPARQL scripts.

For commercial addresses, the data repository is queried using graph pattern matching in SPARQL (as described in Section 2.2.4.5) using the following patterns:

1. Select addresses of interest: commercial addresses
2. Select TS for those addresses
3. Reduce TS selection to TSs for the correct day type
4. Reduce TS selection to TSs for the correct activity type
5. Select the relevant interval for each of the TSs

This returns the correct interval for each address, based on its TS, from which the occupancy rate can be retrieved. The relevant interval for each TS is identified using the following SPARQL formulation:

```
1     { SELECT (MAX(?start) as ?maxStart)
2         WHERE{
3              ?interval popont:hasStartTime ?start .
4             FILTER (?start <=
              "1899-12-31T14:00:00.000Z"
              ^^xsd:dateTime) .
5         }
6    }
```

Line 3 selects all intervals that have a start time less than that indicated in Line 4. Line 1 selects the interval with the maximum time from that selection. This ensures that the time specified returns the appropriate interval, from which the occupancy rate can be queried. Note the use of xsd:dateTime (XSD is XML Schema Definition and provides the datatypes used by SPARQL) data definition in Line 4. This is used as GraphDB does not use xsd:Time. In practical terms, this means that the progression from the interval before midnight on one day to the interval after midnight on the next day can be specified. However, in this script, the date is left unused for simplicity, and the date is effectively null. It is also possible to select the interval with the maximum start time because xsd:dateTime can be ordered using string literals, so this provides a good solution to the absence of xsd:Time in GraphDB.

7.2.2 The SPARQL Scripts

There are three SPARQL scripts used to estimate population, which can be found in Appendix D. There is a script for each of Working, Residential and Visitor population estimation. The cases that each handles are presented in Table 27 on page 166, and broken down by the activity script they

are contained within in Table 34. Each address case is handled in a separate sub-query within the commercial, residential or visitor query. These cases represent the activity being modelled (workers, visitors or residential), whether the address is inside a FS (yes or no) and whether a maximum capacity of the address needs to be estimated from building size and topology (yes or no).

There are several points of note related to this particular implementation of the modelling framework. Firstly, it is a straightforward task to increase the sophistication of these queries as sub-queries can be used to model additional address cases. This implementation is deliberately simplified so that only the necessary features of the query language are demonstrated.

Secondly, the general approach is to identify those addresses with specific TSs, estimate the population at those addresses, and then to use the SIC Groups for the commercial addresses without TSs. Given that in the available data, the SIC Groups relate only to commercial worker activities, and the address specific TSs relate only to commercial visitor activities, this has not been implemented in the scripts, but it would be a straightforward exercise to incorporate this when new data are available.

Finally, this implementation explicitly states the relationships between addresses and address-specific TSs. It is possible to model this in a different way, using inferred relationships. In this case, the required differences in the scripts would be minor.

Table 34: Summary of the population estimation scripts, which shows the different address scenarios calculated in each script

Activity	**Algorithm Details**	**Notes**
Residential	1. Select all residential addresses in WZs (AOIs)	Demonstrates: Final arithmetic Filtering by addresses with different membership of FS Handling time
	2. Filter by not in FS	
	3. Select residential TS and intervals for correct day type and time	
	4. Retrieve occupancy from TS and capacity from address and calculate population estimate	
	5. Group (sum) all population estimates by WZ	
Commercial Work	1. Select all commercial addresses in WZs (AOIs)	Demonstrates: Final arithmetic Handling time
	2. Select SIC Group TS (present only for Work activity)	

Activity	Algorithm Details	Notes
	3. Select correct interval for each TS by day type and time	*Notes:* *FSs handled by the data – due to use of address capacity rather than building capacity* *All work capacities are sourced from the census data.*
	4. Retrieve occupancy from TS and capacity from address and calculate population estimate	
	5. Group (sum) all population estimates by WZ	
Commercial Visitor	1. Select all commercial addresses in WZs (AOIs)	<u>Demonstrates:</u> Final arithmetic Filtering by addresses with different membership of FSs Use of different algorithms for different types of address Handling time *Notes:* *Non-FS addresses use address capacity (specified or estimated)* *FS addresses use site aggregated building floor area for capacity.*
	2. Select visitor TS and intervals for correct day type and time	
	3. Select and calculate estimated population for addresses • not in FS with specified capacity • not in FS without specified capacity • in FS without specified capacity • in FS without specified capacity	
	4. Calculate sum of all the estimated populations	

7.2.3 Functional Sites

FSs are used within the scripts for calculating an aggregated floor area for all of the buildings within a FS. This is then used in the same way as a single building floor area where the building capacity is not available.

That population estimates are calculated for addresses rather than buildings means that there is no requirement to model FSs in greater spatial detail when the estimates are calculated at the scale of WZs. However, should there be a requirement for estimated population counts at a finer scale, such as individual buildings in a FS, or rooms within a building, the algorithms used here would require some development. The same techniques already demonstrated can be used for redistributing populations across buildings or groups of buildings within a FS, with different treatments for different addresses within the site where necessary.

7.3 Population Estimation Validation Data

Validation of time-specific population estimates is recognised as a challenging exercise (Martin, Cockings & Leung 2015) as there are no ground-truth data or techniques that are adequate for

comparison with model outputs. One approach taken for evaluating whether the outputs from a model are as expected, and reflecting the varying patterns of population over time, is to use a proxy for time-specific population. In this case, the proxy is mobile network data.

These data are generated by a passive network-monitoring tool installed on devices that records signal strength when a user is in, or out of network coverage and uploads these to the network when back in network coverage. Users opt in to the use of this tool. Location data are based on GPS signals. As a result, there is a high degree of selection bias in the proxy data. This is bias towards subscribers to the mobile network, those who have opted in to use of the application and those who have turned on GPS on their device(s).

What this means for the validation data, is that it is unlikely to represent certain demographic groups well, including the very young and very old, and that it is possible that it under-represents populations at home addresses as GPS may be less likely to be active at these locations. There are other issues with mobile network data, described by Jacques (2018) and already discussed. These include the lack of one-to-one relationship between the devices, producing ownership bias (*ibid.*).

These mobile network data have a fine spatial and temporal granularity, which is not present in mobile site or cell based data. They have been sourced as a count of number of unique devices spatially aggregated to three decimal places in WGS84 (approximately 60m x 120m intervals), and temporally aggregated to 15 minute time slots. These have been further aggregated, to calculate the sum of maximum numbers of devices within each WZ, for each hour over a three-month period covering 1st February to 30th April 2017. Finally, they have been aggregated again to calculate the average of maximum number of devices for each hour in the three Day Types represented in the TSs for working populations (and therefore also the residential TSs). These are derived from the day types in the Time Use Survey as part of Population247NRT (Cockings et al. 2017): Day Type 1 is Sunday, Day Type 4 is weekday (averaged Monday to Friday) and Day Type 7 is Saturday. The averaging of the three-month time period covered by the proxy data means that noise in the TSs from non-average weekdays, such as the first day of a school holiday, or a bank holiday, will be reduced. The result is a time series of averaged numbers of devices on the network. While these data will not indicate actual numbers of individuals at different places at different times, it is anticipated that general patterns of population will be represented within the time-series.

7.4 Results

The model output is generated for a one-week period, using the same day types. The model needs to be run three times for each hour that is represented in the time series. That is 168 times for each of working, residential and visitor populations. The weekday values are averaged for each hour of the day. The result is a time-series representing Sunday (Day Type 1), Monday to Friday (Day Type 4) and Saturday (Day Type 7) for each of the activities for each WZ in the smaller study areas.

Because of the lack of accurate magnitude data, the proxy time-series and the modelled time-series are both normalised using z-scores so that they can be compared. The results of these comparisons are presented graphically in Figure 36 to Figure 41, with each figure representing one WZ. The WZs presented here are those for which site-specific data have been loaded into the model, along with two predominantly residential WZs as identified from COWZ. The WZs are described with an accompanying contextual map, and a second graph indicating the proportion of estimated population engaged in the three different activities. RMSE is presented in Table 35.

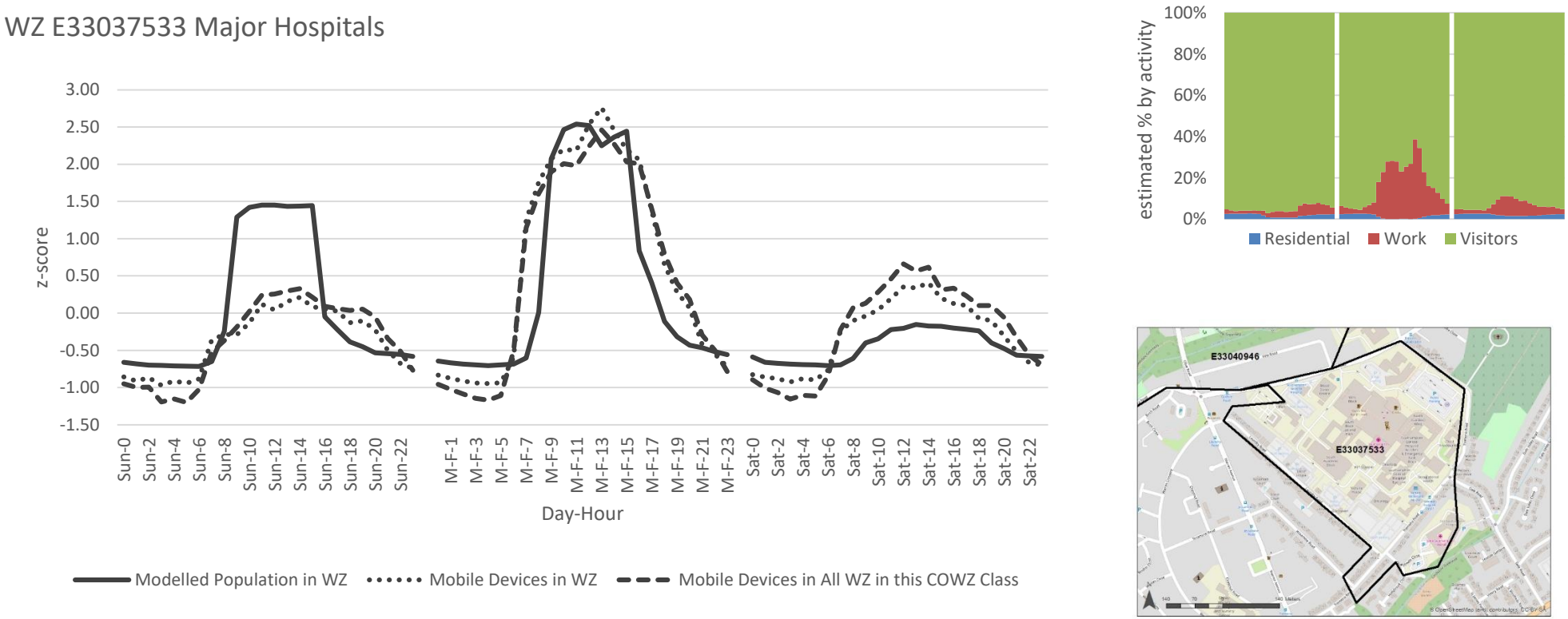


Figure 36: Results of Model Validation for WZ containing Southampton General Hospital. Left: a comparison of the standardised model outputs against the mobile network data for the WZ and for all WZ with this COWZ class. Top Right: the percent of population present, for the three activities modelled (work, residential and visitor). Bottom Right, contextual map of the WZ. There are 10 mobile network data points in this WZ.

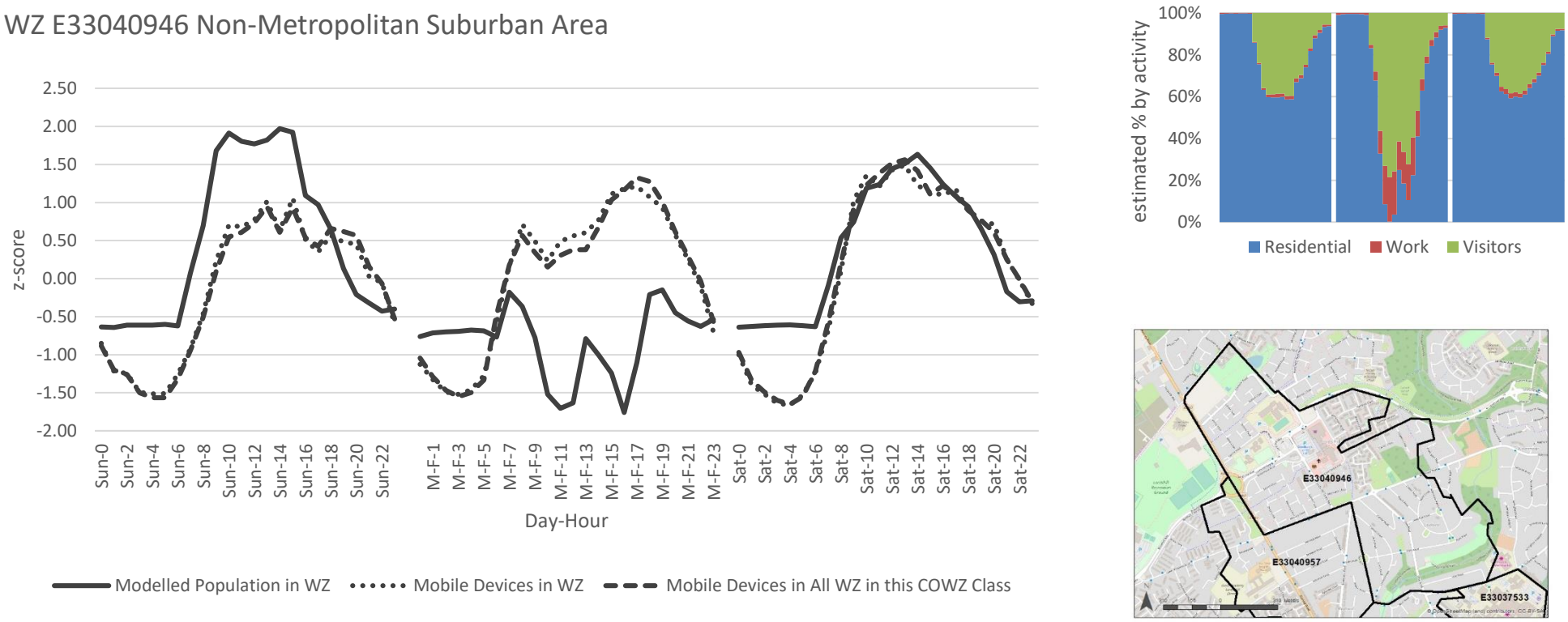


Figure 37: Results of Model Validation for WZ containing the Lordshill addresses with individual TSs. Left: a comparison of the standardised model outputs against the mobile network data for the WZ and for all WZ with this COWZ class. Top Right: the percent of population present, for the three activities modelled (work, residential and visitor). Bottom Right, contextual map of the WZ. There are 75 mobile network data points in this WZ.

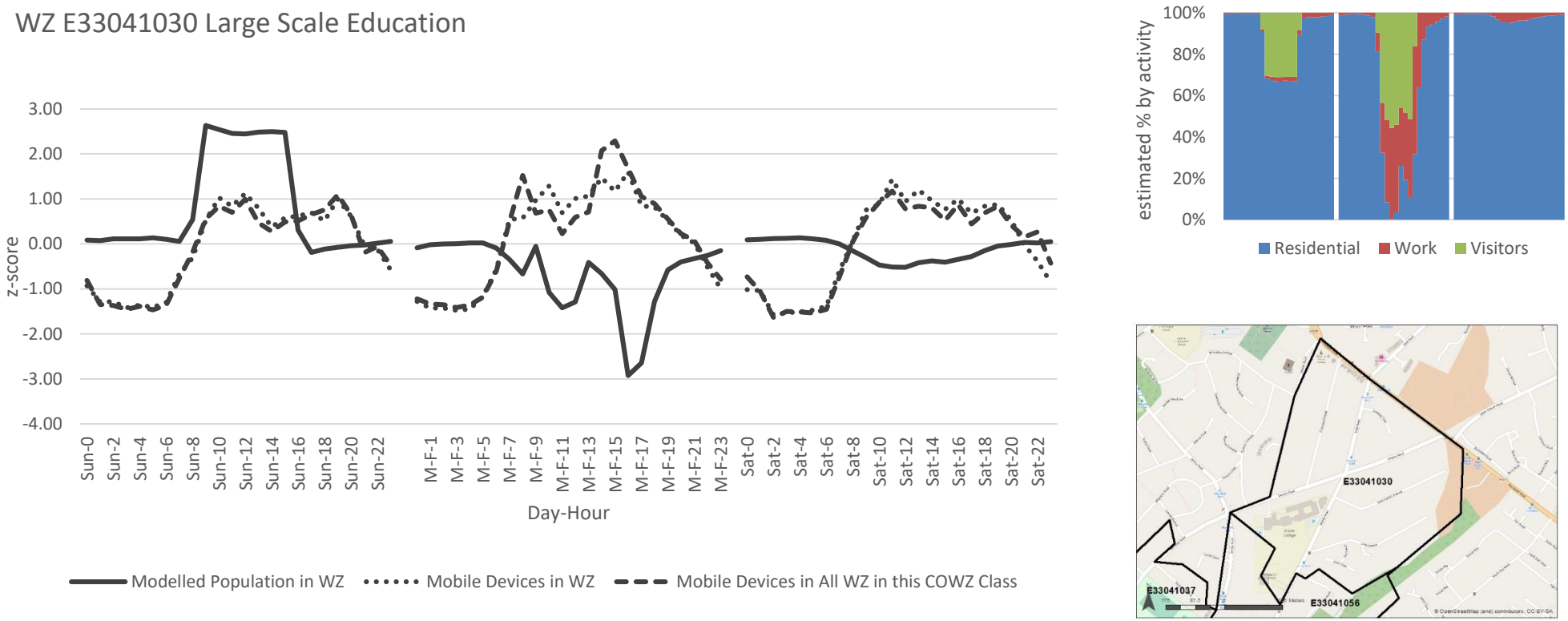


Figure 38: Results of Model Validation for WZ containing a college for 16-18 year olds. Left: a comparison of the standardised model outputs against the mobile network data for the WZ and for all WZ with this COWZ class. Top Right: the percent of population present, for the three activities modelled (work, residential and visitor). Bottom Right, contextual map of the WZ. There are 25 mobile network data points in this WZ.

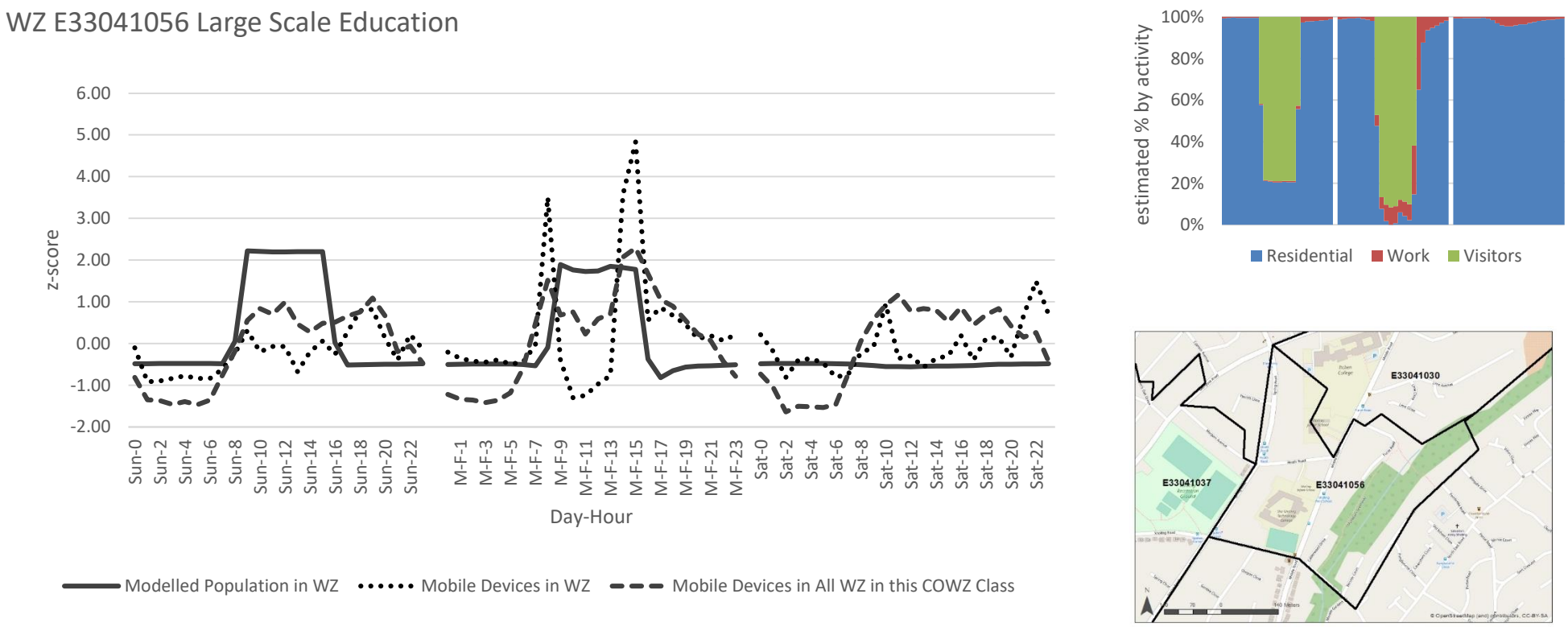


Figure 39: Results of Model Validation for WZ which contains primary schools. Left: a comparison of the standardised model outputs against the mobile network data for the WZ and for all WZ with this COWZ class. Top Right: the percent of population present, for the three activities modelled (work, residential and visitor). Bottom Right, contextual map of the WZ. There are 20 mobile network data points in this WZ.

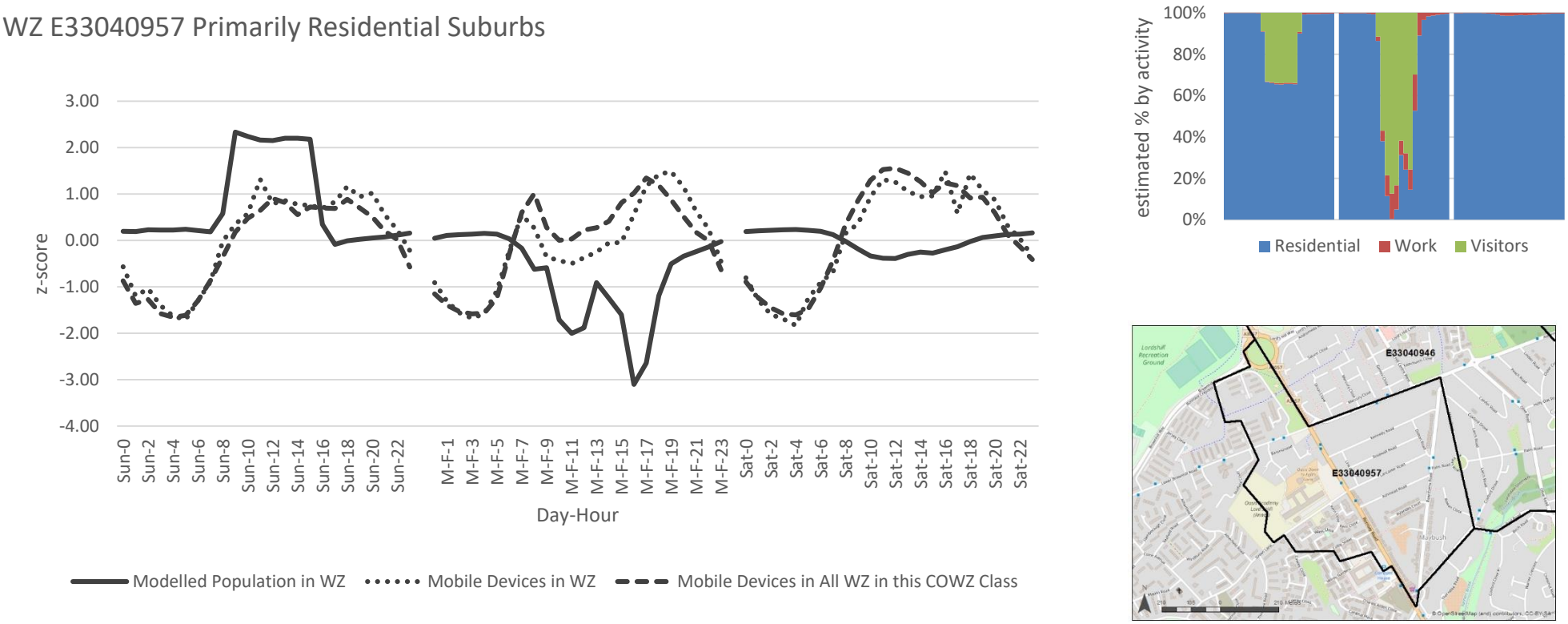


Figure 40: Results of Model Validation for WZ containing primarily residential suburbs in the Shirley study area. Left: a comparison of the standardised model outputs against the mobile network data for the WZ and for all WZ with this COWZ class. Top Right: the percent of population present, for the three activities modelled (work, residential and visitor). Bottom Right, contextual map of the WZ There are 39 mobile network data points in this WZ.

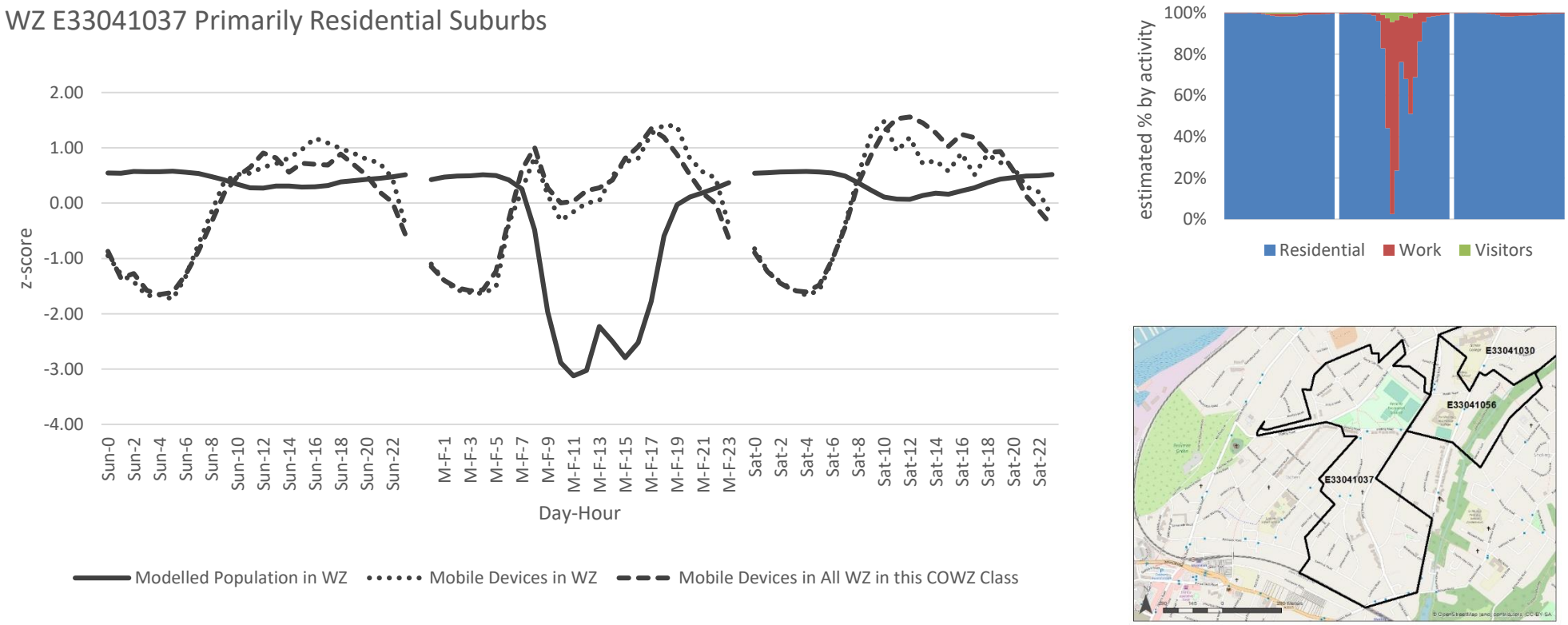


Figure 41: Results of Model Validation for WZ containing primarily residential suburbs in the Itchen study area. Left: a comparison of the standardised model outputs against the mobile network data for the WZ and for all WZ with this COWZ class. Top Right: the percent of population present, for the three activities modelled (work, residential and visitor). Bottom Right, contextual map of the WZ. There are 48 mobile network data points in this WZ.

Table 35: Model Output Results RMSE Workplace Zones and their classification.

Workplace Zone	COWZ	Defining Characteristics	RMSE
E33037533	Major Hospitals	Southampton General Hospital, which has the most detailed visitor population data	0.624
E33040946	Non-metropolitan suburban area	Lordshill, including a large supermarket, bingo hall, library, public house and health centre, all with temporal profiles derived from GPT.	0.991
E33041030	Large scale Education	Itchen College with students aged 16-18. Uses additional schools data for number of pupils.	1.480
E33041056	Large scale Education	Itchen Schools including primary schools for 4-11 year olds (escorted children). Uses additional schools data for number of pupils.	1.332
E33040957	Primarily residential suburbs	Shirley residential, no additional data available	1.410
E33041037	Primarily residential suburbs	Itchen residential, no additional data available	1.569

The results of running the model clearly show some agreement with the proxy population data. Only one of these WZs has a major road running through it that could significantly affect the results by adding in-transit devices to the validation data. Notwithstanding known issues with using mobile phone data for validation, some of the broad patterns can be seen to be representative and there are a few points of note:

With regard to the WZ containing the major hospital, in Figure 36, the model output closely follows the broad patterns of occupation. On Sundays the model appears to over-estimate the population, and on Saturdays to slightly under-estimate the population. However, this can be accounted for by considering the methods used for estimating visitor numbers at the hospital when the TS was generated. The pragmatic approach that assumed each visitor had an accompanying visitor, and the even distribution of outpatients throughout the days of the week may explain these issues. The population breakdown in the smaller graph indicates that visitors make up the majority of the modelled population on all days.

In Figure 37, the results for the WZ containing the Lordshill addresses whose Visitor TSs were sourced from GPT, the weekend days reflect similar patterns for the model and the proxy data.

The weekdays, however reflect almost an opposite pattern. This is most likely to be caused by the process used to generate residential TSs. In this implementation, this has been created as a single TS to apply across the entire study area which will generalise the results across the entire study area. In general, one would expect the residential population to fall in a residential area during the day. However, this will be impacted by the socio-economic characteristics of the population and the industries in which people work. The breakdown of industries that people work in, their socio-economic groups and the demographic breakdown are published per OA, meaning that it would be possible to generate residential TSs for each individual OA, which better reflects the working patterns of its residents. In this particular WZ, the model produces a dip in population, whereas the proxy data do not, indicating that in this residential area, there are more people present than anticipated by the model.

Similar dips in modelled population are indicated in the two primarily residential suburb WZs in Figure 40 and Figure 41. Here, there are more devices observed in the proxy data within the WZ during the day than the model suggests. Again, this can be accounted for by the method used to generate the residential TS, that assumes this is an inverse of the work activity TS. It does not account for people who remain at home during the day, for work, or because they do not work through choice, retirement or unemployment. The data and TS in this model are also not broken down demographically, by socio-economic group or employment status. As the results of the model reflect the data that are utilised, such a breakdown would lead to an improvement in the results and this represents a clear opportunity for further research.

In the case of the two large-scale education WZs represented in Figure 38 and Figure 39, the case for including TSs for different demographic groups is clear. In the WZ containing primary schools for 4-11 year olds, there is a clear peak in the proxy data at the start and the end of the school day, as parents deliver and collect their children. During school hours, the proxy data displays a drop in numbers, suggesting that children in these schools do not have mobile devices. As already stated, this is a result of the proxy data not being suitable as a proxy for population, as it excludes the very young and very old (this is a clear demonstration of the need for an integrative model). The model however, correctly indicates the presence of the schoolchildren in the WZ. In this WZ, the additional variability at weekends may be caused by the surrounding residential population, who, as already discussed, are not best represented in the TSs.

In the WZ that includes a sixth form college, aimed at 16-18 year olds, the weekday patterns are quite different. In the proxy data, there is a smaller peak at the start and end of the day, potentially resulting from students arriving at college, switching off their phones and then switching them back on at the end of the day (although clearly not all students switch their

phones off). The model, however, does not indicate a very large increase in population during the school day, but a dip. The cause of this is the sparseness of the input data. In this case, the schools capacity data does not include this college, although the address is present in the ABP data with the correct classification. This demonstrates the need to know the provenance of the data to understand why some facts may be missing.

In summary, the results of the model reflect the data that are used in the modelling process, and there is a clear opportunity for increasing the granularity of these data by introducing breakdowns by sub-populations, by e.g. age, employment status or industry in which residents are employed. This is particularly true for the residential population estimation.

7.5 Chapter Review

This chapter has presented the methods used to estimate population using the modelling framework, and the results achieved using these techniques. The aim was to demonstrate that the semantic web technologies that are used are suitable for such analysis. The population estimation algorithm is intentionally simplified for this purpose and the results reveal that the model is, in fact, functioning as expected and accounting for different population profiles at different addresses, whether these are site-specific for visitors, or generic, based on SIC Groups for workers, or residential and therefore equal across all WZs and addresses within them.

Chapter 8 provides a discussion of all three parts of the modelling framework, as well as the overall model.

Chapter 8 Discussion

This thesis is a response to the lack of flexible population models capable of integrating increasingly diverse data sources to estimate population at very fine spatial and temporal scales.

This need for increased spatio-temporal has been identified following a review of applications of population data, and current population modelling approaches based on available data. There is no current method for estimating population at very fine temporal *and* spatial scales. Added to this requirement for more detail, it is possible to access an increasing number of data sources that could be utilised in population models. These range from well-established geographic, administrative, survey, and statistical data, to new data that are becoming available from sensors, big data analytics and social media. These data may represent site-specific or area-specific population at different times. They may also represent different common activities including work, residential and leisure activities. The data environment is constantly evolving, with improvements and extensions to the well-established data sets, as well as the new data coming on line.

The modelling framework focuses on laying the foundations for an alternative approach to address these issues, using semantic web technologies in an ontological approach. The aim has been to develop the framework in such a way as to ensure that it can be used with diverse, alternative and constantly evolving data sources, and that it can generalise to use in different data environments. As such, the research presented here has been a proof of concept, ensuring that the technologies employed are flexible enough to meet this need. The data that have been modelled are deliberately restricted to a small number of sites, so that the focus remains with the functioning of the model itself and not with data development. The data environment that the modelling framework has been developed for is GB, with access to some of the best commercial data available. The framework, however, is capable of using alternative data sources due to the simplicity of the modelling technologies. These data could be alternatives available within GB, or they could be from an entirely different data environment.

This chapter is divided into four main sections. The first three sections discuss parts one, two and three of the modelling framework, presented in Chapters 4 to 7. This comprises a review of the objectives laid out in Chapter 1, in turn, with a discussion of whether, and how they have been met and further discussion of the wider implications. Section four sets out how, together, these three parts of the framework meet the research aims, also set out in Chapter 1. A discussion of the limitations of the framework, and the contribution that such an ontology based modelling framework can make in the population estimation domain are also presented in this section.

8.1 Stage One: Data Preparation

Stage One of the modelling framework is concerned with data preparation and integration. Core geographic datasets were sourced, linked via spatial or database means, and prepared for import into the RDF triple store. Additional attributes were also calculated using spatial analysis, counting addresses with different primary class in each building.

> *Objective 1: Prepare appropriate core datasets so that they can be utilised within the model.*

This objective has been met. While there is no formal validation of this stage, all data were carefully checked after they were processed to ensure that the processing techniques were appropriate.

In this stage, all data have been tested for their suitability, and successfully linked, producing suitable input to Stage Two of the framework, the ontology. Some issues with the data were highlighted during the process of linking and preparing the data. The primary issue was that of unexpectedly inconsistent classification quality in the ABP data, which led to the definition of Objective 2, below.

The need for GIS use in Stage One is reinforced by the need for linking the data inputs. However, the evolution of the data environment is already leading to datasets becoming more open, such as OS MasterMap (Ordnance Survey 2018) and ONS UPRN Directory (ONS 2018b), which explicitly links ABP data to statistical and administrative regions. This trend may, in time, lead to more data becoming more explicitly linked. ABP data are already supplied with cross references to the MasterMap TOIDs with which they are associated, and continuous improvement within the datasets (since the data used in this analysis were sourced) is reducing the number of addresses for which spatial analysis is necessary. In this case, the balance of GIS and ontology employed within the modelling framework will shift towards ontology, as the GIS will no longer be required for linking the data.

Associated with this, is the ability to transfer this modelling framework to different data environments. The geographic data requirements for the ontology can be reduced to building lists, addresses (with a classification and a linkage to the buildings), and statistical regions. The quality of these data will vary in different data environments, but if address data with even a rudimentary classification scheme are available, the modelling environment will still be applicable.

The floor area of buildings is also required for estimating capacity of buildings, so these data need to be from geographic sources, even if they are not processed in the GIS part of the modelling framework. Alternative techniques for more accurately calculating number of storeys in a building, such as those used by Wu, Blunden & Bahaj (2018), could easily be implemented in the data preparation stage of the modelling. It is also feasible that data could be from a more detailed source, such as a Building Information Model (BIM) that indicates the proportions of floor space given over to different activities. This could also be incorporated into the model.

In a future in which geographic and statistical data are fully linked, through attribution or through use of semantic web technologies, there would be no need to include such a data preparation stage to this modelling framework, and instead a data gathering or identification exercise would be required. The potential for reducing the amount of GIS processing is an attractive prospect, as the main issue encountered during this stage was the speed of processing caused by the complexity of the spatial query and size of the datasets being used.

It is worth noting some specifics about the data that have been prepared. Firstly, there was no data preparation required for the statistical attribution of WZs or OAs as it was possible to load these tables directly into the RDF triple store.

Secondly, the data utilised have been gathered for demonstration purposes only, and while they are generally representative of the patterns of occupation for different activities, this data gathering exercise was a pragmatic approach, involving assumptions about, for example, the number of visitors who accompany in-patients or outpatients to hospital.

Thirdly, due to the simple structure of the RDF data model, there is no need for a dataset to be structured in certain ways. For example, a TS does not need to be a structured and ordered list of same length intervals with their attributes, but can instead be disordered, or with varying length attributes as these are specified on data load.

Finally, issues related to temporal mismatches in the data have not been addressed in this research. There is clearly a possibility, using data subject to continuous improvement programmes, that features will appear in one dataset, but not be acknowledged in related data until the first dataset's update is published. For example, where new buildings have been constructed, the features that represent them cannot include cross references to the new address features until those addresses are included in a data release. These issues can be dealt with by data suppliers in production processes. It is not necessary to address these details where they have not been dealt with for the proof of concept.

Some of the developments for this stage were necessary only because this research is focused on proof of concept, rather than on developing a working production system. The identification of FOIs that was carried out during this stage. In a working production system, the extents of the geographic area covered by the data would be larger, so that extracts would not be necessary. Instead, features within the AOI, and with appropriate attribution would be identified during the query stage of the modelling. If, for example, address data were available as a linked dataset with the cross references included in an ontology, then data extraction from address data would not be necessary at this stage, and SPARQL queries could be used to select the live addresses that are also occupiable from the full dataset, a procedure that is carried out in GIS in this research.

> *Objective 2: Supplement the classification of address data where this is required for population estimation purposes.*

Meeting Objective 2 involved developing a process to classify residential addresses. This is required for the application of estimates of average population by dwelling type. Although not strictly necessary for proof of concept, this is useful to demonstrate the application of statistical data to specific feature types, rather than all features within the statistical region. This is a requirement for development of a more sophisticated population estimation algorithm.

The 95% success rate for classification of dwelling type in the PSA suggests that this objective has been well met. Some of the characteristics of the Eastleigh study area led to a reduced success rate, as indicated by the validation in which only 72% of the sampled buildings were correctly classified. There were some clear examples that stood out as potential improvements. These issues include:

1. Building heights are unsuitable for indicating the number of storeys in a building, either because of the form of the building, or because the heights were inaccurate.
2. Buildings linked through trivial boundaries such as garages or entrance porches.
3. Buildings linked by archways that interfere with the neighbourhood calculations.

Where the form of the building is an issue, or where buildings are linked through trivial boundaries or archways a more sophisticated algorithm could be developed to reduce or remove the impact. If these issues are dealt with, in the sample used for validation, the success rate rises to 81%. Data error accounts for nearly 14% of the remaining errors. Eastleigh is an area with several relatively new developments that are included in the small study area, so these data errors will be negated over time.

This high degree of accuracy in the topological analysis was unexpected. The anticipated use of building heights and textual analysis to identify flats was unnecessary as the data linkages (spatial

and database) along with the address classifications clearly indicate the presence of flats by inclusion of more than one address if one of them is residential. This has not been the case in previous incarnations of address data, such as AL2.

One particular topological arrangement provides difficulties for the classification algorithm, that in which a first order neighbour is also a second order neighbour. This issue may not be resolvable in this approach and an alternative algorithm may be necessary. This unexpected issue has not been addressed. That there was only one topological arrangement not yielding accurate dwelling type classification was also unexpected.

It is clear that this approach to identifying dwelling type is dependent on the presence of high quality address data. However, the addresses require only a primary class, indicating whether they are residential or commercial, in order to be suitable for this type of analysis. There is no reason why alternative dwelling types should not be used, for example, detached, or part of a block containing several addresses. Building data at a smaller scale, not subject to licence agreements, could be utilised to this end to add an extra dimension to those data.

Processes such as this could feed valuable information back to data suppliers, such as the LAs who supply the address classification to the ABP product.

This approach to classifying residential addresses brings the methods of Orford & Radcliffe (2007) up to date with the richer data environment that now exists, with the evolution of address data to its current state. As the data environment continues to evolve, these methods will need review, but the principles should hold regardless of the data specifics (so long as the data are broadly similar).

The classification produced by the process is acceptable for the purposes of the modelling framework presented here, allowing the estimated average occupants per dwelling type to be applied to the addresses in the data. Where misclassification occurs, this is likely to have a minor impact as these are based around individual addresses, rather than entire buildings.

Alternative processes could be developed for commercial addresses, including those involving textual analysis, as was originally anticipated for the residential classification, data linkage to POIs, or web scraping.

Overall, the data preparation carried out in Stage One meets the needs of the modelling framework, in that the output from Stage One is appropriate input to Stage Two, the ontology. Should the data environment change and new data sets be produced in the GB or data with a

different structure be used from a different country, this is the only part of the modelling framework that would require significant review.

8.2 Stage Two: Ontology Development

The purpose of developing the ontology is for information system development (rather than for knowledge generation or domain specification). It formalises the spatial and temporal relations and hierarchies from the real world, at the conceptual and operational levels, in a domain-specific ontology. This ontology has been developed for the specific operational task of modelling population at the building level.

> *Objective 3: Develop an ontological model that can utilise core data sources, and that is capable of integrating other data sources, including in the attribute and temporal domains.*

In meeting this objective, all of the relevant features and relationships from the core data sources have been successfully modelled. The success of this exercise has been measured by loading the diverse data sources into the triple store and running queries. The results of the load and queries have been confirmed as accurate using a GIS and spreadsheet approach to make certain that the data load is correct and that the queries can be executed correctly using the technologies employed.

TSs have also been successfully modelled and loaded into the data repository and tested to ensure that the appropriate interval (and associated properties) can be retrieved. The ability to make inferences from the loaded data has also been proved in identifying addresses within FSs using a role inclusion axiom, and using inverse properties for feature selection. This proof was gained using simple and manageable synthetic data to ensure the model was functioning as expected which proved the validity of the ontology itself.

Three main methods of inference are used within this modelling framework. Firstly, the use of a role inclusion axiom that defines a relationship based on a series of classes and relationships, by stepping along the property chains. The *isAssociatedWithFS* relationship between addresses and FSs is defined in this way, as addresses that have an *isAssociatedWithRegion* relationship with a region that has an *isPartOf* relationship with a FS. It would be possible to model these relationships using relational databases, or by writing programs that execute on flat files, but that is an over-complicated approach.

Secondly, inferences can be made using a transitive object property. This is a single property that is inferred to connect two individuals however many steps are taken to connect them thereby operating through generations within the class hierarchy. A typical example is that of a mereonomy; in which parts of a part are also part of the whole, so if A is part of B, and B is part of C, then A is part of C (W3C 2004). This capability is exploited in the *hasProperPart* property on FSs, so that a finer breakdown of buildings could indicate whether, for example, a room that is within a building is also within the FS that the building is associated with. This is also used on the *hasFollowingInterval* property for intervals, so that all intervals that follow a specified interval can be identified. It could also be used in an implementation containing a hierarchy of statistical regions, for example.

Finally, inferences can be made by using inverse properties. In the data repository, relationships are defined only in one direction. The inverse is defined within the ontology so that there is no need to state the inverse explicitly. Where a pair of inverse properties between two classes are defined, only one of these relationships between individuals needs to be expressed, as the inverse can be inferred. For example, addresses are defined with an *isAssociatedWithRegion* relationship with buildings. It is explicitly stated that this relationships inverse is *hasAssociatedPlaces*, so there is no need to state, for every address, that a building contains a set of addresses as this can be inferred. This is useful if additional features are added to the data repository, as explicitly stated relationships do not need to be reviewed.

The main issue in the development of the ontology was the iterative, continuous improvement nature of the process combined with the different technologies employed (Protégé for ontology development, GraphDB Free for data load and testing). However, the simplicity of the resulting model was unexpected. The complexity of the built environment and the data, including the number of data sources led to an expectation of a very complicated ontology. The result of the development is, at the top-level hierarchy, very simple.

There remain limitations in the semantic web technologies. The ontology design has been influenced by the expressive limitations of OWL. The main example of this is with the *isProperPartOf* relationship.

IsProperPartOf and *hasProperPart* should be irreflexive, i.e. they cannot have this relationship with themselves (a proper part is a part that is also not the whole). They should also be transitive: if region A *isProperPartOf* region B, and region B *isProperPartOf* region C, then it stands to reason that region A *isProperPartOf* region C. However, OWL cannot allow both transitivity and irreflexivity on the same property as this would result in undecidability (Horrocks, Kutz & Sattler 2006; W3C OWL Working Group 2012b). This is a clear example of the necessity for a trade-off

between the reasoner capabilities and the complexity of the ontology. In this case, some of the data checking occurs in the GIS part of the modelling framework. Irreflexivity is already checked in spatial analysis during data preparation, with a high level of confidence in the data. The transitivity, however, is required in the ontology as it provides the functionality needed for inference so the modelling decision is to include transitivity rather than irreflexivity.

A second limitation of the reasoner is related to neighbourhood analysis. Limitations of OWL prevent identification of second order neighbours (buildings are identified as second order neighbour to themselves leading to logical inconsistencies and failure of the reasoner). This would be a useful addition to the ontology to reduce the reliance on GIS for such analysis.

The final limitation of the reasoner that led to compromise in the ontology design is related to a partiallyOverlaps relationship, that a region should have with itself. This relationship is required if regions need to be used for specifying an AOI. In the model, the *isPartOf* and *hasPart* relationships are essential to the transitive ability to step through generations to identify membership of regions higher up in a hierarchy. These two relationships are inverse and disjoint, meaning that if *X isPartOf Y*, *Y* cannot have the relationship *isPartOf* with *X*. The reasoner fails due to logical inconsistency if these relationships are also set as disjoint with *partiallyOverlaps*. It is not possible to be a part of something and to partially overlap it, they are two separate relations (as set out in **Error! Reference source not found.**), and so the decision was taken to allow an AOI to be defined only by the relationship it has with addresses. This allows the AOI to be defined by loading an additional RDF file into the triple store, which states that a set of addresses are related to an AOI, without specifying their relationship with any other features within the data repository.

In this instance, such limitations of OWL verify the need for a GIS component that guarantees the data will not have logical inconsistencies. The need for spatial analysis to establish the spatial relationships between features also means that there remains a requirement for GIS. GeoSPARQL is a geographic query language for RDF data. It uses an RDF representation for geographic information based on the General Feature Model (form ISO19109) so that features and geometry can be represented in RDF and subsequently queried using the GeoSPARQL extension (Perry & Herring 2011). This requires further investigation: to establish the role that GeoSPARQL could play in population estimation ontology.

It would be useful to be able to identify parades of shops, where several local stores are connected in a single block, so that these can be treated as a single entity for the population estimation. Ontology offers the capability to handle this, using adjacency. An *externallyConnected* property can be added to buildings to achieve this. This is another area for further investigation and is a different problem to the identification of second order neighbours discussed earlier.

Most importantly, it is now possible to represent these, and the supporting, spatial relationships required for detailed spatio-temporal population estimation in a qualitative manner, in a very simple data model that facilitates the integration of additional knowledge, without changes to the ontology. It is a simple exercise to add to the knowledge contained within the database, and additional knowledge can be gained through inference.

This means that as new data become available, from new sources, they can be easily integrated. In particular, those new datasets coming online from IoT, big data and sensors can indicate patterns of occupation for individual places, sometimes down to the room level. As national datasets are opened up, or alternative open sources become available, existing data could be linked directly, re-using existing data, rather than loaded into a local data store using the structures of the semantic web.

In some circumstances, population data may be available at the building level rather than the address level. This is likely to be the case for FSs. It may be more appropriate to handle these circumstances within the ontology rather than in the data. This would require some re-engineering of the ontology.

8.3 Stage Three: Population Estimation

The population estimation part of the modelling framework is intended only to demonstrate that all of the required components are available for the disaggregation of small area population counts and the allocation of site-specific population to individual addresses.

> *Objective 4: Ensure all arithmetic, aggregation and inference operations, required for population estimation, can be effectively applied using the ontology and semantic web technologies employed, and that it is possible to allocate population to the appropriate geographic features.*

Much of the requirement for the population estimation have been met through appropriate development of the ontology, including the ability to disaggregate small area statistics to individual addresses, and the ability to use more than one TS where required, such as where an address has visitor and work activities. In this case, disaggregation utilises dwelling type statistics for residential addresses and industry type statistics for commercial addresses, using simple disaggregation algorithms. With further development, this could be extended to incorporate demographic breakdowns by age, sex and socio-economic group: the technologies utilised here provide all of the capabilities required to increase the complexity of the algorithms, and the model design facilitates this by allowing more than one TS for individual sites, functional classes,

groups of classes. Furthermore, development of the algorithms could include the ability to apply different TSs to different demographic groups within one activity, e.g. the ability to use different TS for the visitor activities for schoolchildren than for adult populations.

The estimation algorithms need to build on this by treating addresses differently depending on whether they are associated with a FS, and whether they have site-specific TS or need to use a SIC Group TS. The SPARQL queries developed as part of the ontology development (i.e. the prerequisite calculations outlined in Figure 33), demonstrated that appropriate selections can be made, that arithmetic calculations can be carried out and that maximum occupancy rates can be allocated to individual addresses based on the address characteristics. Ontology development also included ensuring that the correct interval can be selected for each TS depending on required time of query.

Many potential enhancements could be made to the population estimation algorithm. For example, enhancements could be made to the development of the residential TS, which in this implementation has been created as a single TS to apply across the entire study area. The breakdown of industries that people work in are published for OAs, meaning that it would be possible to generate residential TSs for each individual OA, which better reflects the working patterns of its residents. The graphs in Figure 36 to Figure 41 clearly show that in WZs with sizeable residential populations the model produces a dip in population present, whereas the proxy data do not. This is a result of the TSs used, not the functioning of the model as a whole, or the scripts. It is worth noting here that sourcing more representative residential TSs, and applying them to the individual addresses within the modelling framework requires only an additional load to the data repository, and no changes to the model itself.

This model is conceptually focused on containers for population (addresses) and their capacity, rather than being a source-destination model that finds the containers into which to place measured population. There is currently no adjustment made should the official data indicate that, e.g. there are 100 residents who live in flats in an OA, but the topological data indicate only two flats present in that OA. A source-destination model would account for this anomaly, but potentially by distributing all 100 residents between two flats.

In the development of SPARQL scripts for the estimation of population at specific times of day, the ability to make inferences, aggregate data and allocate population have been clearly demonstrated. Data aggregation is applied where the GROUP BY clause is used in the SPARQL scripts, to find count, maximum or sum of a selected attribute. Estimated occupancy rates for the selected intervals are applied to maximum occupancy, to calculate estimated population at

specific addresses, which are then summed to provide an estimated population present in the AOI.

Some or all of the prerequisite calculations could be executed within the population estimation scripts, as well as the SPARQL scripts used to aggregate number of addresses within the buildings. This would further reduce the reliance on GIS for these prerequisite tasks.

The queries developed for this objective return good results that are representative of the data. They have been assessed and validated to prove their workings as far as this is possible (using a combination of spreadsheet analysis and GIS analysis).

The results of running the model on real data are set out in Figure 36 to Figure 41. Overall, the results reveal less about the actual population in the WZs represented, and more about the model workings. The model queries are working effectively, utilising all three activity TSs, as expected. The issues with the model revealed by the results are attributable to three main causes: the sparseness of the data loaded to the model; the unsuitability of mobile network data as proxy for population in some places (e.g. primary schools); and the lack of sophistication in the development of residential TSs.

The generation of the SPARQL scripts was unexpectedly straightforward. That the SPARQL language was developed with the intention of being readable to those familiar with SQL has contributed to this, but once the pattern matching is understood, the development is straightforward. This is partly down to the ease of use of the GraphDB Free software used for development.

The queries that have been developed for this proof of concept do not cover every choice that would be made in a production system, or in any data environment. They would need further development if e.g. class TSs were available, FSs were no longer available or additional classes and sub-classes are added to the ontology. They would require only minor edits where attribute names differ.

8.4 Modelling Framework

There is no conceptual limit to the temporal scales represented, as the TS is a continuous variable. The interval lengths could be reduced to seconds without influencing the functioning of the model. Should the spatial scale need to be increased this is possible, a room could be used as a sub-class of address, or as a sub-class of Place, with its own TS. In this way, e.g. a university could be modelled down to the level of offices and lecture rooms, according to timetables that provide

TSs for visitors and workers. If data are available only at a coarser scale than that used in this implementation, these could be modelled as a different type of topographic area, perhaps a "block" class. Again, these changes to the ontology to accommodate alternative data environments has no impact on the concepts that are modelled.

This makes the modelling environment both generalizable and transferable for use with alternative data, including the evolving environment in GB, both for the core geographic data and the ancillary data that contributes to the population estimation. This is the result of the different levels of the class hierarchy being used for different levels of detail. Minor adjustments for the new data, such as alternative attribution, would be required. Alterations to the ontology might also be necessary so that alternative classifications can be utilised, but the core concepts will still be applicable: places (addresses) will continue to be associated with regions (buildings) and temporal signatures. The key dataset here, as with the classification of dwelling types, is addresses with functional class attribute. If the class attribute provides only a distinction between residential and commercial addresses, a scaled back version of the ontology could be used, because of the hierarchical nature of both functional class and TS classes. Alternative classifications could also be modelled by adapting the ontology and the principle relationships would still hold true.

The results of the model implementation are a reflection the data that have been loaded into the RDF triple store. There is great scope for improvement in both the data and the algorithms. The data could be improved by sourcing complete data sets for TSs and maximum capacities. Geographic data could also be improved to include 3D building models or BIM so that floor surface estimates for individual addresses can be more accurately determined. Other potential improvements could be made to the population estimation algorithms, which would need to not only account for those data improvements and additions, but also become more sophisticated with the current data implementation.

The model provides a more spatially detailed population estimate than has previously been possible. It uses the same continuous temporal scales as Population 24/7, but increases the spatial scale of the estimate to any area greater than a building (or smaller if individual rooms are modelled in this way), due to the fine modelling scales. Additional knowledge can be very easily added to the calculations without affecting the model itself.

RDF, the data model on which the semantic web is built facilitates the integration of data. It also makes it a trivial exercise to add to the knowledge base. OWL facilitates the inferences that create new knowledge and SPARQL facilitates the queries that are used for population estimation.

Without these semantic web technologies, such a simple model that meets all the requirements for detailed spatio-temporal population estimation would not be possible.

The investigation of the role of ontology and semantic web technologies for population estimation has resulted in successfully building a modelling framework that produces convincing output, based on the qualitative spatial relationships between geographic features.

In time, as more geographic datasets are opened up and linked, the potential for using these data in this type of model can only grow. If all the required data were stored in RDF and linked, there would be no need to load them into the data repository, and they could be accessed directly. There are indications that data products are moving towards this state, with the partial opening of national datasets, both geographic and statistical. These data are available to users in the public and private sectors and could be combined with whichever other data individual organisations have access to, in order to add to the knowledge within their data repositories. There is potential for a huge amount of data to be available in the knowledge base that could contribute to the population estimation.

This modelling framework, as implemented for this thesis, has outputs subject to licencing constraints. In future, with the opening up of national datasets, even in part, this may not be a restriction and the framework may become more accessible to a greater number of application use cases.

The use of Day Type in this model reflects the current use of cyclical patterns of human behaviour over time. The model therefore, creates output that reflects a day like the day type specified in the query. The more data that is in the repository and utilised by the model, the more a day like this day becomes a better simulacrum of the day itself. Ultimately, this modelling technique, with the data sources that could be incorporated, has the potential to progress towards a now-casting system, a technique for very short-range forecasting such as that used for weather forecasting (Meterological Office 2011). In such a system, the query would not supply a retrospective date, but a date in the near future, that could account for additional factors, including weather (which affects activity patterns). Such heterogeneous data would require a flexible framework such as the one presented here.

8.5 Other Points for Consideration

Besides the limitations of GIS used in Stage One (primarily the complex data models and slow processing times with large datasets), the semantic web technologies employed in Stage Two and

three (OWL limitations) and the SPARQL queries that require greater sophistication, there are several other issues worth considering.

There are conceptual limitations with a model that is concerned with estimating population based on container sizes, rather than distributing known population counts, or tracking individuals. Firstly, there is potential for double counting in the results. Currently all people are estimated to be either at residential or work addresses (but from the address perspective, rather than the individual perspective). This means that all estimated visitors will also be counted as either at work or at home. The inverse calculation that would be necessary to generate residential TSs without double-counting residents at home and as visitors requires assumptions about destinations of residents. This is an area for further investigation and a potential opportunity for Population 24/7 methods to inform this model.

With any development of the algorithms, this needs to be a consideration to avoid such double counting wherever possible. Consequently, this model is not volume preserving. This means that people could be gained, or lost in the model if they are not properly distributed. Development of the redistribution methods could allocate fixed total populations into containers in proportion to the estimates currently produced in the model. In this way, the model in its current state produces a weighting to be used by additional algorithms that are concerned with volume preserving.

Another conceptual limitation relates to the handling of time, which allows a query to estimate population counts in an area at a particular time, but not to indicate how many people may have passed through an area over a time period. This requires a different approach to time handling.

There are several sets of data not explicitly represented within this implementation of the modelling framework. These include, but are not limited to visitors to a region who are neither residents nor workers (students are included in this category). While some models require visiting populations to be explicitly modelled (Smith, Martin & Cockings 2012; Martin, Cockings & Leung 2015), these do not need to be explicitly modelled as they are represented in the capacities of places that are modelled within the framework, such as hotels. A notable exception to this is camping and caravan sites, as these will count as open spaces and will need to be dealt with separately within the model.

Vacant properties are not modelled, but given their own time profile, this would conceptually be no different from any of the other addresses that are modelled. Likewise, transitory addresses are not modelled. This group of features includes, for example, markets that are present only intermittently, and events such as festivals. There is no reason why these could not be modelled

as places, with a TS that reflects their transitory nature. A regular market may only have people attending for one day a week, or a festival for two days a year. This would be also be applicable in a now-casting development of this framework, in which one-off, non-cyclical events could be added to the model.

The evolving nature of data has already been discussed from the perspective of its utility within the model. However, with evolving data, techniques for data refresh need to be available within the systems that use them. Apart from the need to edit the ontology and query when data evolve, a change in background data would need to be handled carefully in a production system. Whether this is best handled in the GIS part of the framework or in the triple stores would depend on the type of change. With the addition of new features, with no structural change to the data, the relationships between elements stored within a triple store would all need to be checked to ensure that changes to subject or object of triples are cascaded through all the related elements.

Regional differences in TS are not considered here. For example, a hotel in a coastal resort may have a very different TS to a hotel in a large city. There is no conceptual reason why time profiles cannot be localised, as with an individual address.

No transportation is included in the modelling framework. It would be feasible to develop the ontology to incorporate linear features, or to utilise the transportation cartographic areas within topographic datasets. Conceptually, this would be very different to the building containers that are currently modelled within this framework, as transportation data represent counts of vehicles or individuals passing through, so the interval length needs to be reduced to a snapshot instance rather than an interval. This would produce another type of TS. This could be tackled in a separate, external system that uses sensor and traffic-type data that could be compatible with this framework.

Chapter 9 Conclusions

This chapter provides a summary of the findings of the thesis. First, the two research aims are evaluated to determine to what extent they have been met. This is followed by a summary of the limitations of the modelling framework, followed by opportunities and recommendations for further research. Finally, the novel contributions of this work to the population estimation domain are presented.

9.1 Evaluation of Research Aims

Two research aims and four specific objectives were outlined in the introductory chapter of this thesis. Chapter 8 provides a discussion of the objectives, and concludes that all objectives have been met. This section addresses whether the aims of the research have been met.

> *The aim of this research is to develop a generalizable modelling framework for population estimation that enables high levels of spatial and temporal detail. The model must also have the ability to integrate data from diverse sources.*

The three-part modelling framework presented here integrates population data in the spatial, temporal and attribute domains to estimate population at very fine spatial and temporal scales. Ontology modelling has facilitated the assignment of temporal attributes to addresses through the modelling of relationships between cartographic and address objects, which has enabled an address level estimation of population at a specific time that can be aggregated for any AOI. The approach involved the preparation of core geographic data sets by explicitly stating the spatial relationships between the objects of interest. The result is that these features do not require a spatial representation within the data model. Explicit definitions of these objects and relationships within the ontology reduces the population estimation activity to a series of relatively straightforward queries that will automatically incorporate additional data as it is loaded to the knowledge base.

Because data preparation is contained within its own stage, the framework is transferrable to other data environments where even rudimentary classification of addresses are available (as it is possible to use less detailed attribute data) or to richer data environments (by using more detailed attribute data and class hierarchies in the ontology, or modelling to sub-building level). In these cases, Stage One of the framework would be altered accordingly.

The ontological model is also generalizable: it could be applied over much larger areas without the need to change the concepts and techniques used. The framework is also able to accommodate changes in the data environment without changing the conceptual approach to the

estimation, such as with the addition of new or alternative data. Adding new attribute data to the triple store would not require any actions other than re-running the model, unless the ontology itself required additional modelling. This would be the case only if new classes were required, such as a different type of place, or a new classification scheme.

Diverse data sources can be integrated into the model: the breakdown of complex data models to the extremely simple RDF data model allows new data to be added very easily, and represents all of the data from spatial, temporal and attribute domains. The key to integration of additional data sources will be in understanding exactly what they represent.

The result is a framework that can integrate data from a variety of sources, including: spatial data, represented qualitatively; temporal data, represented in TSs; and attribute data, and produce a convincing population estimation at any time, down the address level.

Given the proof of concept focus of this research, the three-stage modelling framework has met the first stated aim. The function of addresses, and therefore the activities that occur at them, can be used to estimate population at those addresses, as this varies over time. Neighbourhood statistics can be incorporated into the model, as can site-specific data.

> *The second aim of this research is to investigate the potential role of using ontology and associated Semantic Web technologies for population estimation.*

From the outset, this second aim was central to the model development. Ontology and semantic web technologies can play a pivotal role in the integration of data for population estimation. The ontology facilitates the population estimation, by formalising the connections between the data, specifically the temporal signatures relationship with the addresses, by using group, class, or site-specific data. The data linkages that have proved possible in the ontology-based framework are very difficult to model in a GIS, or relational database environment, because the relationships are between classes of things rather than the instances of the things.

Different levels of detail in all three domains can be handled. In the temporal domain, this is demonstrated by the variable-length intervals that can be handled. In the attribute domain because the specificity of the classes may vary, and in the spatial domain as the representation of the features is qualitative rather than quantitative and data for different spatial resolutions can be utilised (site or statistical region).

Because the modelling framework has been developed using semantic web technologies, it is fundamentally very simple. To prepare data that needs to be loaded into an ontology data store, whose main relationships are between just four super-classes (region, place, TS and interval), and

then to query that data store is a very simple approach to a very complex problem. The detail beneath those super-classes becomes more complex when different data are added, but the fundamental concepts remain the same.

9.2 Limitations

The limitations of this modelling framework fall into three main categories: those related to data, those related to the semantic web technologies and finally conceptual limitations. All three provide opportunity for further research.

As always, the accuracy of model output is limited by the quality and completeness of the input data. However, the results of running the estimation algorithms, presented in Chapter 8 demonstrate that the approach is performing as expected: able to model minute-by-minute changes in population present (with appropriate temporal signatures) at individual addresses, however sparse the input data. This provides population estimates at very high levels of spatial and temporal detail. In this implementation of the modelling framework, the data were kept deliberately sparse. In some situations, the pragmatic approach of not over-baking the data before loading them into the model (e.g. the hospital data) means that the numbers of people present at the addresses may be in the wrong order of magnitude. For proof of concept however, this is adequate as the general patterns of population appear to be reflected in the model output. With the limited site-specific input data used, this model could not currently be used to estimate population counts, only patterns of occupation in the specified AOI.

The second set of limitations is concerned with the technologies employed. There were three specific cases where the limitations of OWL led to a trade-off between the complexity of the model that best represents the real world and the capabilities of the reasoners. First, where externally connected cannot be used to identify neighbourhood. Second where irreflexive and transitive properties cannot both be set on a relationship (so that it was necessary to choose between the two different options) and finally where it was not possible to use the *partiallyOverlaps* relationship at the same time as either *isPartOf* or *hasPart*.

Finally, the conceptual limitations of the bottom-up approach to modelling, and the associated development of the TS for residential addresses (discussed in Section 8.5) that does not take visitor TSs into account leading to the potential for double counting.

9.3 Recommendations for Further Research

Several areas for potential complementary research have been highlighted during analysis of the methods and results. These are laid out below.

9.3.1 Transportation

This model does not include any accounting of people in the various modes of transportation: road, rail, air, water and pedestrian. The presence of transportation routes within every area means that this absence needs to be addressed. While this is already a significant research field, the Population24/7, and its successor, Population247NRT do incorporate a background transportation layer. Opportunities for adding transportation into the modelling framework include avenues such as incorporating linear features in the model with the associated transportation attribution and temporal signatures, or incorporating the appropriate layers from the Population247NRT. The model in its current state however does not handle raster data, so this would require some adjustment.

Adding transportation requires some conceptual adjustment, as the number of people who pass along a road over a period of time, for example, is not the same as the number of people who would be expected on the road at any one time.

9.3.2 Estimation Algorithm

The algorithm that was used in this research was by necessity very crude, and set out only to demonstrate that the required calculations could be made. The development of more sophisticated algorithms for population disaggregation could include many more statistical datasets, for example, those listed as potential candidate datasets in England and Wales, in Section 4.5. Such developments would require consideration of the demographic breakdown in residential, work and visitor activities. Associated temporal signature developments could also be made, to incorporate regional variations. For example, occupation levels of different types of hotel (city break, airport hotels, holiday hotels, those providing conference facilities etc.) may be different, and while these types of hotels may be recognised in a classification, it is possible that temporal activity levels are different even between cities and regions.

9.3.3 Large Scale Implementation

In order to build on the foundations of the modelling framework laid down in this thesis, a large-scale implementation would be a sensible next step. The incorporation of minimal data for the

work undertaken to date should be built upon, with many other datasets being added to make the model more realistic and to test its capabilities thoroughly. Additional data could contribute to TSs include many of the sources already mentioned, such as sensor data, BIM 3D building models, observed data, as well as commercial data should this be available.

Such a large-scale implementation would require a thorough understanding of temporal mismatches in the data: which data are more susceptible to delay in being released; how these relate to the other data in the model and whether the impact is under or over-estimation of population, at different times and different places.

9.3.4 Potential Model Extensions

A large-scale implementation could also extend the model itself, to explore the addition of different types of places: events, and those at different scales, such as rooms within buildings and outdoor spaces. There is also potential for using the remaining qualitative relationships from RCC8, and use the topological adjacency relationship to determine neighbourhood within the ontology, rather than in the GIS.

Taken to its natural conclusion, the ability to harvest real-time population proxy information from sensors, traffic data or mobile network data and integrate them with a large range of less dynamic datasets would not be an unrealistic goal. Ultimately, this could lead to “now casting”, a modelling technique that, while not providing live situational results, allows short-term predictions of the current situation, and can utilise additional factors that influence the location of people, such as weather and traffic incident data.

GeoSPARQL is a geographic query language for RDF data. It uses an RDF representation for geographic information based on the General Feature Model (form ISO19109) so that features and geometry can be represented in RDF and subsequently queried using the GeoSPARQL extension (Perry & Herring 2011). The modelling framework outlined here is based around the ontology and proving the concept of using ontology with the geographic features as the unifying classes. As this is the focus of this research, the use of GeoSPARQL and attempts to represent the geographies within RDF have not been made. Additional improvements to the model could therefore include investigating the potential of GeoSPARQL and whether this could further reduce the need for the use of GIS in Stage One of the modelling framework.

9.4 Novel Contributions

The development of this modelling framework has moved the potential for high-resolution spatiotemporal populations models a step closer to practical implementation. The key innovation is in driving spatial resolution to a high level of detail while also incorporating the continuous temporal detail, with the ability to add more detail for specific sites where data are available, so that individual sites do not have to use averaged population data if measured data are available. It represents a significant change in approach, as compared to previous population models, in that the spatial relationships are modelled in a qualitative manner, rather than quantitatively. The fact that the potential data sources are, and will continue to be evolving highlights the value of the development of such a flexible modelling framework capable of integrating new data.

This modelling framework can integrate the very fine spatial detail of the building models from Ahola et al. (2007) and Greger (2014), while also incorporating the flexibility to extend this to larger areas, due to the ease with which additional address function's temporal signatures can be added to the model, as well as information specific to individual addresses. The temporal scale also provides more flexibility than the model from Zhang, Sunila & Virrantaus (2010), which extends Ahola's Helsinki model to the whole of Finland, but at the expense of the detail in attribution and temporal scale. This framework also has the potential to incorporate sub-address level population models, by adding a Place sub-class that represents rooms. There is no conceptual reason why site-specific models such as Charles-Edwards & Bell's (2013) university campus model and Jochem, Sims, E. A. Bright, et al.'s (2013) airport model, cannot also be incorporated into this framework. Conceptually the model could also handle outdoor spaces such as parks, beaches or sporting grounds, should the temporal data and capacity be available. Again, this would require a new Place sub-class to link with the temporal signature.

The classification of dwelling types to provide supplementary information to the address data has brought the methods of Orford & Radcliffe (2007) up to date in line with developments in the available data. When the original techniques were applied, there was no distinction between residential and commercial addresses. The richer datasets available in GB today enable these alternative methods to be applied. This is the case for any complete address data set, which includes at least a primary classification.

Ontology is not often applied to geography due to the inherent difficulties in defining geographic concepts, and the interdisciplinary nature of the subject and this is the first time that an ontology has been created to handle the data necessary for estimating population at fine spatial and temporal scales.

Adding the detail in the spatial domain to the already detailed temporal population estimation models will be a welcome improvement to the applications that require this higher resolution data. The model does require development with considerable refinement in the algorithms used and many more datasets to be utilised, but this proof of concept lays the foundations for a working model, and shows promise for the implementation of a production system.

As more data become available, this framework will not require re-working, as the concepts and relationships are already well defined. As a result, additional data should be straightforward to incorporate without any further changes. If data are also linked in the sense that they are connected through URIs, this can lead to even smoother incorporation of new datasets, through SPARQL query, with no data load process required.

The multiple practical application areas cited in Chapter 1, such as mobile telecommunications and emergency response and planning activities, could benefit from this approach to modelling population. In the commercial domain, it is feasible that detailed data that is under licence could also inform the model, something that may not be a possibility in the research and public domains. For applications that require an estimate of population within a specific AOI at short notice, this modelling approach is able to meet that need. If developed into a production system, an RDF triple store can be queried in a timely fashion, without the need to build new databases.

Finally, this research has been undertaken in response to the need for greater spatial detail while maintaining the temporal and attribute detail in current models. In particular, this research is closely connected to the Population 24/7 project (Martin, Cockings & Leung 2015). There is scope for this project to inform Population 24/7 to push the spatial dimension to higher levels of detail, and vice versa, as the data that informs Population 24/7 is also appropriate intelligence for this modelling framework.

Appendix A Source Data Preparation

A.1 Buildings and BLPUs

Number of buildings and BLPUs inside buildings for each combination of primary classes within single buildings.

BLPU Primary Class Names	Number of Buildings	Number of BLPUs inside buildings
Commercial, 1	7580	9616
Commercial, Land, 2	124	361
Commercial, Land, Object of Interest, 3	1	6
Commercial, Land, Other (OS Only), 3	14	121
Commercial, Land, Other (OS Only), Parent Shell, 4	3	19
Commercial, Land, Other (OS Only), Parent Shell, Residential, 5	3	148
Commercial, Land, Parent Shell, 3	9	48
Commercial, Land, Parent Shell, Residential, 4	2	359
Commercial, Land, Residential, 3	8	51
Commercial, Object of Interest, 2	29	101
Commercial, Object of Interest, Parent Shell, 3	2	12
Commercial, Object of Interest, Parent Shell, Residential, 4	1	7
Commercial, Object of Interest, Residential, 3	1	8
Commercial, Other (OS Only), 2	152	776
Commercial, Other (OS Only), Parent Shell, 3	85	1391
Commercial, Other (OS Only), Parent Shell, Residential, 4	11	224
Commercial, Other (OS Only), Residential, 3	12	54
Commercial, Parent Shell, 2	381	1829
Commercial, Parent Shell, Residential, 3	237	5205
Commercial, Residential, 2	1563	4461
Dual Use, 1	29	29
Dual Use, Residential, 2	2	4
Land, 1	1690	1795
Land, Object of Interest, 2	1	2
Land, Other (OS Only), 2	18	149
Land, Parent Shell, 2	39	95

Appendix A

BLPU Primary Class Names	Number of Buildings	Number of BLPUs inside buildings
Land, Parent Shell, Residential, 3	6	217
Land, Residential, 2	24	51
Military, 1	11	12
Military, Other (OS Only), 2	1	3
Military, Parent Shell, 2	1	2
Object of Interest, 1	185	191
Object of Interest, Parent Shell, 2	2	5
Object of Interest, Parent Shell, Residential, 3	5	100
Object of Interest, Residential, 2	10	23
Other (OS Only), 1	627	687
Other (OS Only), Parent Shell, 2	59	149
Other (OS Only), Parent Shell, Residential, 3	4	106
Other (OS Only), Residential, 2	12	43
Parent Shell, 1	4773	5001
Parent Shell, Residential, 2	4459	36756
Residential, 1	108486	119001
Grand Total	**130662**	**189218**

A.2 Tertiary Classification

Tertiary classification of AddressBase Premium Data: There are 563 possible classes in ABP. Not all of these addresses are to quaternary level. This table lists all addresses up to the tertiary level. There are 304 of these.

Tertiary Class	Primary_Desc	Secondary_Desc	Tertiary_Desc
C	Commercial		
CA	Commercial	Agricultural	
CA01	Commercial	Agricultural	Farm / Non-Residential Associated Building
CA02	Commercial	Agricultural	Fishery
CA03	Commercial	Agricultural	Horticulture
CA04	Commercial	Agricultural	Slaughter House / Abattoir
CB	Commercial	Ancillary Building	
CC	Commercial	Community Services	
CC02	Commercial	Community Services	Law Court
CC03	Commercial	Community Services	Prison
CC04	Commercial	Community Services	Public / Village Hall / Other Community Facility
CC05	Commercial	Community Services	Public Convenience
CC06	Commercial	Community Services	Cemetery / Crematorium / Graveyard. In Current Use.
CC07	Commercial	Community Services	Church Hall / Religious Meeting Place / Hall
CC08	Commercial	Community Services	Community Service Centre / Office
CC09	Commercial	Community Services	Public Household Waste Recycling Centre (HWRC)
CC10	Commercial	Community Services	Recycling Site
CC11	Commercial	Community Services	CCTV
CC12	Commercial	Community Services	Job Centre
CE	Commercial	Education	
CE01	Commercial	Education	College
CE02	Commercial	Education	Children's Nursery / Creche
CE03	Commercial	Education	Preparatory / First / Primary / Infant / Junior / Middle School
CE04	Commercial	Education	Secondary / High School
CE05	Commercial	Education	University
CE06	Commercial	Education	Special Needs Establishment.

Appendix A

Tertiary Class	Primary_Desc	Secondary_Desc	Tertiary_Desc
CE07	Commercial	Education	Other Educational Establishment
CH	Commercial	Hotel / Motel / Boarding / Guest House	
CH01	Commercial	Hotel / Motel / Boarding / Guest House	Boarding / Guest House / Bed And Breakfast / Youth Hostel
CH02	Commercial	Hotel / Motel / Boarding / Guest House	Holiday Let/Accommodation/Short-Term Let Other Than CH01
CH03	Commercial	Hotel / Motel / Boarding / Guest House	Hotel/Motel
CI	Commercial	Industrial Applicable to manufacturing engineering maintenance storage / wholesale distribution and extraction sites	
CI01	Commercial	Industrial Applicable to manufacturing engineering maintenance storage / wholesale distribution and extraction sites	Factory/Manufacturing
CI02	Commercial	Industrial Applicable to manufacturing engineering maintenance storage / wholesale distribution and extraction sites	Mineral / Ore Working / Quarry / Mine
CI03	Commercial	Industrial Applicable to manufacturing engineering maintenance storage / wholesale distribution and extraction sites	Workshop / Light Industrial
CI04	Commercial	Industrial Applicable to manufacturing engineering maintenance storage / wholesale distribution and extraction sites	Warehouse / Store / Storage Depot
CI05	Commercial	Industrial Applicable to manufacturing engineering maintenance storage / wholesale distribution and extraction sites	Wholesale Distribution
CI06	Commercial	Industrial Applicable to manufacturing engineering maintenance storage / wholesale distribution and extraction sites	Recycling Plant

Tertiary Class	Primary_Desc	Secondary_Desc	Tertiary_Desc
CI07	Commercial	Industrial Applicable to manufacturing engineering maintenance storage / wholesale distribution and extraction sites	Incinerator / Waste Transfer Station
CI08	Commercial	Industrial Applicable to manufacturing engineering maintenance storage / wholesale distribution and extraction sites	Maintenance Depot
CL	Commercial	Leisure - Applicable to recreational sites and enterprises	
CL01	Commercial	Leisure - Applicable to recreational sites and enterprises	Amusements
CL02	Commercial	Leisure - Applicable to recreational sites and enterprises	Holiday / Campsite
CL03	Commercial	Leisure - Applicable to recreational sites and enterprises	Library
CL04	Commercial	Leisure - Applicable to recreational sites and enterprises	Museum / Gallery
CL06	Commercial	Leisure - Applicable to recreational sites and enterprises	Indoor / Outdoor Leisure / Sporting Activity / Centre
CL07	Commercial	Leisure - Applicable to recreational sites and enterprises	Bingo Hall / Cinema / Conference / Exhibition Centre / Theatre / Concert Hall
CL08	Commercial	Leisure - Applicable to recreational sites and enterprises	Zoo / Theme Park
CL09	Commercial	Leisure - Applicable to recreational sites and enterprises	Beach Hut (Recreational Non-Residential Use Only)
CL10	Commercial	Leisure - Applicable to recreational sites and enterprises	Licensed Private Members Club
CL11	Commercial	Leisure - Applicable to recreational sites and enterprises	Arena / Stadium
CM	Commercial	Medical	
CM01	Commercial	Medical	Dentist
CM02	Commercial	Medical	General Practice Surgery / Clinic

Appendix A

Tertiary Class	Primary_Desc	Secondary_Desc	Tertiary_Desc
CM03	Commercial	Medical	Hospital / Hospice
CM04	Commercial	Medical	Medical / Testing / Research Laboratory
CM05	Commercial	Medical	Professional Medical Service
CN	Commercial	Animal Centre	
CN01	Commercial	Animal Centre	Cattery / Kennel
CN02	Commercial	Animal Centre	Animal Services
CN03	Commercial	Animal Centre	Equestrian
CN04	Commercial	Animal Centre	Vet / Animal Medical Treatment
CN05	Commercial	Animal Centre	Animal / Bird / Marine Sanctuary
CO	Commercial	Office	
CO01	Commercial	Office	Office / Work Studio
CO02	Commercial	Office	Broadcasting (TV / Radio)
CR	Commercial	Retail	
CR01	Commercial	Retail	Bank / Financial Service
CR02	Commercial	Retail	Retail Service Agent
CR04	Commercial	Retail	Market (Indoor / Outdoor)
CR05	Commercial	Retail	Petrol Filling Station
CR06	Commercial	Retail	Public House / Bar / Nightclub
CR07	Commercial	Retail	Restaurant / Cafeteria
CR08	Commercial	Retail	Shop / Showroom
CR09	Commercial	Retail	Other Licensed Premise / Vendor
CR10	Commercial	Retail	Fast Food Outlet / Takeaway (Hot / Cold)
CR11	Commercial	Retail	Automated Teller Machine (ATM)
CS	Commercial	Storage Land	
CS01	Commercial	Storage Land	General Storage Land
CS02	Commercial	Storage Land	Builders Yard
CT	Commercial	Transport	
CT01	Commercial	Transport	Airfield / Airstrip / Airport / Air Transport Infrastructure Facility
CT02	Commercial	Transport	Bus Shelter
CT03	Commercial	Transport	Car / Coach / Commercial Vehicle / Taxi Parking / Park And Ride Site
CT04	Commercial	Transport	Goods Freight Handling / Terminal
CT05	Commercial	Transport	Marina
CT06	Commercial	Transport	Mooring

Tertiary Class	Primary_Desc	Secondary_Desc	Tertiary_Desc
CT07	Commercial	Transport	Railway Asset
CT08	Commercial	Transport	Station / Interchange / Terminal / Halt
CT09	Commercial	Transport	Transport Track / Way
CT10	Commercial	Transport	Vehicle Storage
CT11	Commercial	Transport	Transport Related Infrastructure
CT12	Commercial	Transport	Overnight Lorry Park
CT13	Commercial	Transport	Harbour / Port / Dock / Dockyard / Slipway / Landing Stage / Pier / Jetty / Pontoon / Terminal / Berthing / Quay
CU	Commercial	Utility	
CU01	Commercial	Utility	Electricity Sub-Station
CU02	Commercial	Utility	Landfill
CU03	Commercial	Utility	Power Station / Energy Production
CU04	Commercial	Utility	Pump House / Pumping Station / Water Tower
CU06	Commercial	Utility	Telecommunication
CU07	Commercial	Utility	Water / Waste Water / Sewage Treatment Works
CU08	Commercial	Utility	Gas / Oil Storage / Distribution
CU09	Commercial	Utility	Other Utility Use
CU10	Commercial	Utility	Waste Management
CU11	Commercial	Utility	Telephone Box
CU12	Commercial	Utility	Dam
CX	Commercial	Emergency / Rescue Service	
CX01	Commercial	Emergency / Rescue Service	Police / Transport Police / Station
CX02	Commercial	Emergency / Rescue Service	Fire Station
CX03	Commercial	Emergency / Rescue Service	Ambulance Station
CX04	Commercial	Emergency / Rescue Service	Lifeboat Services / Station
CX05	Commercial	Emergency / Rescue Service	Coastguard Rescue / Lookout / Station
CX06	Commercial	Emergency / Rescue Service	Mountain Rescue Station
CX07	Commercial	Emergency / Rescue Service	Lighthouse
CX08	Commercial	Emergency / Rescue Service	Police Box / Kiosk

Appendix A

Tertiary Class	Primary_Desc	Secondary_Desc	Tertiary_Desc
CZ	Commercial	Information	
CZ01	Commercial	Information	Advertising Hoarding
CZ02	Commercial	Information	Tourist Information Signage
CZ03	Commercial	Information	Traffic Information Signage
L	Land		
LA	Land	Agricultural - Applicable to land in farm ownership and not run as a separate business enterprise	
LA01	Land	Agricultural - Applicable to land in farm ownership and not run as a separate business enterprise	Grazing Land
LA02	Land	Agricultural - Applicable to land in farm ownership and not run as a separate business enterprise	Permanent Crop / Crop Rotation
LB	Land	Ancillary Building	
LB99	Land	Ancillary Building	
LC	Land	Burial Ground	
LC01	Land	Burial Ground	Historic / Disused Cemetery / Graveyard
LD	Land	Development	
LD01	Land	Development	Development Site
LF	Land	Forestry	
LF02	Land	Forestry	Forest / Arboretum / Pinetum (Managed / Unmanaged)
LF03	Land	Forestry	Woodland
LL	Land	Allotment	
LM	Land	Amenity - Open areas not attracting visitors	
LM01	Land	Amenity - Open areas not attracting visitors	Landscaped Roundabout
LM02	Land	Amenity - Open areas not attracting visitors	Verge / Central Reservation
LM03	Land	Amenity - Open areas not attracting visitors	Maintained Amenity Land
LM04	Land	Amenity - Open areas not attracting visitors	Maintained Surfaced Area
LO	Land	Open Space	

Tertiary Class	Primary_Desc	Secondary_Desc	Tertiary_Desc
LO01	Land	Open Space	Heath / Moorland
LP	Land	Park	
LP01	Land	Park	Public Park / Garden
LP02	Land	Park	Public Open Space / Nature Reserve
LP03	Land	Park	Playground
LP04	Land	Park	Private Park / Garden
LU	Land	Unused Land	
LU01	Land	Unused Land	Vacant / Derelict Land
LW	Land	Water	
LW01	Land	Water	Lake / Reservoir
LW02	Land	Water	Named Pond
LW03	Land	Water	Waterway
M	Military		
MA	Military	Army	
MA99	Military	Army	
MB	Military	Ancillary Building	
MB99	Military	Ancillary Building	
MF	Military	Air Force	
MF99	Military	Air Force	
MG	Military	Defence Estates	
MN	Military	Navy	
MN99	Military	Navy	
O	Other (OS Only)		
OA	Other (OS Only)	Aid To Navigation	
OA01	Other (OS Only)	Aid To Navigation	Aid To Aeronautical Navigation
OA02	Other (OS Only)	Aid To Navigation	Aid To Nautical Navigation
OA03	Other (OS Only)	Aid To Navigation	Aid To Road Navigation
OC	Other (OS Only)	Coastal Protection / Flood Prevention	
OC01	Other (OS Only)	Coastal Protection / Flood Prevention	Boulder Wall / Sea Wall
OC02	Other (OS Only)	Coastal Protection / Flood Prevention	Flood Gate / Flood Sluice Gate / Flood Valve

Tertiary Class	Primary_Desc	Secondary_Desc	Tertiary_Desc
OC03	Other (OS Only)	Coastal Protection / Flood Prevention	Groyne
OC04	Other (OS Only)	Coastal Protection / Flood Prevention	Rip-Rap
OE	Other (OS Only)	Emergency Support	
OE01	Other (OS Only)	Emergency Support	Beach Office / First Aid Facility
OE02	Other (OS Only)	Emergency Support	Emergency Telephone (Non Motorway)
OE03	Other (OS Only)	Emergency Support	Fire Alarm Structure / Fire Observation Tower / Fire Beater Facility
OE04	Other (OS Only)	Emergency Support	Emergency Equipment Point / Emergency Siren / Warning Flag
OE05	Other (OS Only)	Emergency Support	Lifeguard Facility
OE06	Other (OS Only)	Emergency Support	Life / Belt / Buoy / Float / Jacket / Safety Rope
OF	Other (OS Only)	Street Furniture	
OG	Other (OS Only)	Agricultural Support Objects	
OG01	Other (OS Only)	Agricultural Support Objects	Fish Ladder / Lock / Pen / Trap
OG02	Other (OS Only)	Agricultural Support Objects	Livestock Pen / Dip
OG03	Other (OS Only)	Agricultural Support Objects	Currick
OG04	Other (OS Only)	Agricultural Support Objects	Slurry Bed / Pit
OH	Other (OS Only)	Historical Site / Object	
OH01	Other (OS Only)	Historical Site / Object	Historic Structure / Object
OI	Other (OS Only)	Industrial Support	
OI01	Other (OS Only)	Industrial Support	Adit / Incline / Level
OI02	Other (OS Only)	Industrial Support	Caisson / Dry Dock / Grid
OI03	Other (OS Only)	Industrial Support	Channel / Conveyor / Conduit / Pipe
OI04	Other (OS Only)	Industrial Support	Chimney / Flue
OI05	Other (OS Only)	Industrial Support	Crane / Hoist / Winch / Material Elevator

Tertiary Class	Primary_Desc	Secondary_Desc	Tertiary_Desc
OI06	Other (OS Only)	Industrial Support	Flare Stack
OI07	Other (OS Only)	Industrial Support	Hopper / Silo / Cistern / Tank
OI08	Other (OS Only)	Industrial Support	Grab / Skip / Other Industrial Waste Machinery / Discharging
OI09	Other (OS Only)	Industrial Support	Kiln / Oven / Smelter
OI10	Other (OS Only)	Industrial Support	Manhole / Shaft
OI11	Other (OS Only)	Industrial Support	Industrial Overflow / Sluice / Valve / Valve Housing
OI12	Other (OS Only)	Industrial Support	Cooling Tower
OI13	Other (OS Only)	Industrial Support	Solar Panel / Waterwheel
OI14	Other (OS Only)	Industrial Support	Telephone Pole / Post
OI15	Other (OS Only)	Industrial Support	Electricity Distribution Pole / Pylon
ON	Other (OS Only)	Significant Natural Object	
ON01	Other (OS Only)	Significant Natural Object	Boundary / Significant / Historic Tree / Pollard
ON02	Other (OS Only)	Significant Natural Object	Boundary / Significant Rock / Boulder
ON03	Other (OS Only)	Significant Natural Object	Natural Hole (Blow / Shake / Swallow)
OO	Other (OS Only)	Ornamental / Cultural Object	
OO02	Other (OS Only)	Ornamental / Cultural Object	Mausoleum / Tomb / Grave
OO03	Other (OS Only)	Ornamental / Cultural Object	Simple Ornamental Object
OO04	Other (OS Only)	Ornamental / Cultural Object	Maze
OP	Other (OS Only)	Sport / Leisure Support	
OP01	Other (OS Only)	Sport / Leisure Support	Butt / Hide
OP02	Other (OS Only)	Sport / Leisure Support	Gallop / Ride
OP03	Other (OS Only)	Sport / Leisure Support	Miniature Railway
OR	Other (OS Only)	Royal Mail Infrastructure	

Tertiary Class	Primary_Desc	Secondary_Desc	Tertiary_Desc
OR01	Other (OS Only)	Royal Mail Infrastructure	Postal Box
OR02	Other (OS Only)	Royal Mail Infrastructure	Postal Delivery Box / Pouch
OR03	Other (OS Only)	Royal Mail Infrastructure	PO Box
OR04	Other (OS Only)	Royal Mail Infrastructure	Additional Mail / Packet Addressee
OS	Other (OS Only)	Scientific / Observation Support	
OS01	Other (OS Only)	Scientific / Observation Support	Meteorological Station / Equipment
OS02	Other (OS Only)	Scientific / Observation Support	Radar / Satellite Infrastructure
OS03	Other (OS Only)	Scientific / Observation Support	Telescope / Observation Infrastructure / Astronomy
OT	Other (OS Only)	Transport Support	
OT01	Other (OS Only)	Transport Support	Cattle Grid / Ford
OT02	Other (OS Only)	Transport Support	Elevator / Escalator / Steps
OT03	Other (OS Only)	Transport Support	Footbridge / Walkway
OT04	Other (OS Only)	Transport Support	Pole / Post / Bollard (Restricting Vehicular Access)
OT05	Other (OS Only)	Transport Support	Subway / Underpass
OT06	Other (OS Only)	Transport Support	Customs Inspection Facility
OT07	Other (OS Only)	Transport Support	Lay-By
OT08	Other (OS Only)	Transport Support	Level Crossing
OT09	Other (OS Only)	Transport Support	Mail Pick Up
OT10	Other (OS Only)	Transport Support	Railway Pedestrian Crossing
OT11	Other (OS Only)	Transport Support	Railway Buffer
OT12	Other (OS Only)	Transport Support	Rail Drag
OT13	Other (OS Only)	Transport Support	Rail Infrastructure Services
OT14	Other (OS Only)	Transport Support	Rail Kilometre Distance Marker

Tertiary Class	Primary_Desc	Secondary_Desc	Tertiary_Desc
OT15	Other (OS Only)	Transport Support	Railway Lighting
OT16	Other (OS Only)	Transport Support	Rail Mile Distance Marker
OT17	Other (OS Only)	Transport Support	Railway Turntable
OT18	Other (OS Only)	Transport Support	Rail Weighbridge
OT19	Other (OS Only)	Transport Support	Rail Signalling
OT20	Other (OS Only)	Transport Support	Railway Traverse
OT21	Other (OS Only)	Transport Support	Goods Tramway
OT22	Other (OS Only)	Transport Support	Road Drag
OT23	Other (OS Only)	Transport Support	Vehicle Dip
OT24	Other (OS Only)	Transport Support	Road Turntable
OT25	Other (OS Only)	Transport Support	Road Mile Distance Marker
OT26	Other (OS Only)	Transport Support	Road Kilometre Distance Marker
OT27	Other (OS Only)	Transport Support	Road Infrastructure Services
OU	Other (OS Only)	Unsupported Site	
OU01	Other (OS Only)	Unsupported Site	Cycle Parking Facility
OU04	Other (OS Only)	Unsupported Site	Picnic / Barbeque Site
OU05	Other (OS Only)	Unsupported Site	Travelling Persons Site
OU08	Other (OS Only)	Unsupported Site	Shelter (Not Including Bus Shelter)
P	Parent Shell		
PP	Parent Shell	Property Shell	
PS	Parent Shell	Street Record	
R	Residential		
RB	Residential	Ancillary Building	
RC	Residential	Car Park Space	
RC01	Residential	Car Park Space	Allocated Parking

Appendix A

Tertiary Class	Primary_Desc	Secondary_Desc	Tertiary_Desc
RD	Residential	Dwelling	
RD01	Residential	Dwelling	Caravan
RD02	Residential	Dwelling	Detached
RD03	Residential	Dwelling	Semi-Detached
RD04	Residential	Dwelling	Terraced
RD06	Residential	Dwelling	Self Contained Flat (Includes Maisonette / Apartment)
RD07	Residential	Dwelling	House Boat
RD08	Residential	Dwelling	Sheltered Accommodation
RD10	Residential	Dwelling	Privately Owned Holiday Caravan / Chalet
RG	Residential	Garage	
RG02	Residential	Garage	Lock-Up Garage / Garage Court
RH	Residential	House In Multiple Occupation	
RH01	Residential	House In Multiple Occupation	HMO Parent
RH02	Residential	House In Multiple Occupation	HMO Bedsit / Other Non Self Contained Accommodation
RH03	Residential	House In Multiple Occupation	HMO Not Further Divided
RI	Residential	Residential Institution	
RI01	Residential	Residential Institution	Care / Nursing Home
RI02	Residential	Residential Institution	Communal Residence
RI03	Residential	Residential Institution	Residential Education
U	Unclassified		
UC	Unclassified	Awaiting Classification	
UP	Unclassified	Pending Internal Investigation	
X	Dual Use		
Z	Object of Interest		
ZA	Object of Interest	Archaeological Dig Site	
ZM	Object of Interest	Monument	
ZM01	Object of Interest	Monument	Obelisk / Milestone / Standing Stone
ZM02	Object of Interest	Monument	Memorial / Market Cross

Tertiary Class	Primary_Desc	Secondary_Desc	Tertiary_Desc
ZM03	Object of Interest	Monument	Statue
ZM04	Object of Interest	Monument	Castle / Historic Ruin
ZM05	Object of Interest	Monument	Other Structure
ZS	Object of Interest	Stately Home	
ZU	Object of Interest	Underground Feature	
ZU01	Object of Interest	Underground Feature	Cave
ZU04	Object of Interest	Underground Feature	Pothole / Natural Hole
ZV	Object of Interest	Other Underground Feature	
ZV01	Object of Interest	Other Underground Feature	Cellar
ZV02	Object of Interest	Other Underground Feature	Disused Mine
ZV03	Object of Interest	Other Underground Feature	Well / Spring
ZW	Object of Interest	Place Of Worship	
ZW99	Object of Interest	Place Of Worship	

A.3 Class:SIC Group Lookup

ABPClass	Class_Desc	SIC	SIC_Section	SIC Group Name
C	Commercial	C	Manufacturing	C
CA	Agricultural	A	Agriculture, forestry and fishing	ABDE
CA01	Farm / Non-Residential Associated Building	A	Agriculture, forestry and fishing	ABDE
CA02	Fishery	A	Agriculture, forestry and fishing	ABDE
CA03	Horticulture	A	Agriculture, forestry and fishing	ABDE
CA04	Slaughter House / Abattoir	A	Agriculture, forestry and fishing	ABDE
CB	Ancillary Building	A	Agriculture, forestry and fishing	ABDE
CC	Community Services	R		RSTU
CC02	Law Court	O	Public administration and defence; compulsory social security	OPQ
CC03	Prison	O	Public administration and defence; compulsory social security	OPQ
CC04	Public / Village Hall / Other Community Facility	R	Arts, entertainment and recreation; other service activities	RSTU
CC05	Public Convenience	E	Water supply, sewerage, waste management and remediation activities	ABDE

ABPClass	Class_Desc	SIC	SIC_Section	SIC Group Name
CC06	Cemetery / Crematorium / Graveyard. In Current Use.	R	Arts, entertainment and recreation; other service activities	RSTU
CC07	Church Hall / Religious Meeting Place / Hall	R	Arts, entertainment and recreation; other service activities	RSTU
CC08	Community Service Centre / Office	N	Administrative and support service activities	KLMN
CC09	Public Household Waste Recycling Centre (HWRC)	E	Water supply, sewerage, waste management and remediation activities	ABDE
CC10	Recycling Site	E	Water supply, sewerage, waste management and remediation activities	ABDE
CC11	CCTV			
CC12 (0)	Job Centre	R	Arts, entertainment and recreation; other service activities	RSTU
CE	Education	P	Education	OPQ
CE01	College	P	Education	OPQ
CE02	Children's Nursery / Crèche	P	Education	OPQ
CE03	Preparatory / First / Primary / Infant / Junior / Middle School	P	Education	OPQ
CE04	Secondary / High School	P	Education	OPQ
CE05	University	P	Education	OPQ
CE06	Special Needs Establishment.	P	Education	OPQ
CE07	Other Educational Establishment	P	Education	OPQ

ABPClass	Class_Desc	SIC	SIC_Section	SIC Group Name
CH	Hotel / Motel / Boarding / Guest House	I	Accommodation and food service activities	GI
CH01	Boarding / Guest House / Bed And Breakfast / Youth Hostel	I	Accommodation and food service activities	GI
CH02	Holiday Let/Accomodation/Short-Term Let Other Than CH01	I	Accommodation and food service activities	GI
CH03	Hotel/Motel	I	Accommodation and food service activities	GI
CI	Industrial Applicable to manufacturing, engineering, maintenance, storage / wholesale distribution and extraction sites	C	Manufacturing	C
CI01	Factory/Manufacturing	C	Manufacturing	C
CI02	Mineral / Ore Working / Quarry / Mine	B	Mining and quarrying	ABDE
CI03	Workshop / Light Industrial	C	Manufacturing	C
CI04	Warehouse / Store / Storage Depot	B	Mining and quarrying	ABDE
CI05	Wholesale Distribution	G	Wholesale and retail trade; repair of motor vehicles and motor cycles	GI
CI06	Recycling Plant	E	Water supply, sewerage, waste management and remediation activities	ABDE
CI07	Incinerator / Waste Transfer Station	E	Water supply, sewerage, waste management and remediation activities	ABDE
CI08	Maintenance Depot	H	Transport and storage	HJ

ABPClass	Class_Desc	SIC	SIC_Section	SIC Group Name
CL	Leisure - Applicable to recreational sites and enterprises	R	Arts, entertainment and recreation; other service activities	RSTU
CL01	Amusements	R	Arts, entertainment and recreation; other service activities	RSTU
CL02	Holiday / Campsite	I	Accommodation and food service activities	GI
CL03	Library	R	Arts, entertainment and recreation; other service activities	RSTU
CL04	Museum / Gallery	R	Arts, entertainment and recreation; other service activities	RSTU
CL06	Indoor / Outdoor Leisure / Sporting Activity / Centre	R	Arts, entertainment and recreation; other service activities	RSTU
CL07	Bingo Hall / Cinema / Conference / Exhibition Centre / Theatre / Concert Hall	R	Arts, entertainment and recreation; other service activities	RSTU
CL08	Zoo / Theme Park	R	Arts, entertainment and recreation; other service activities	RSTU
CL09	Beach Hut (Recreational, Non-Residential Use Only)	R	Arts, entertainment and recreation; other service activities	RSTU
CL10	Licensed Private Members' Club	I	Accommodation and food service activities	GI
CL11	Arena / Stadium	R	Arts, entertainment and recreation; other service activities	RSTU
CM	Medical	Q	Human health and social work activities	OPQ
CM01	Dentist	Q	Human health and social work activities	OPQ
CM02	General Practice Surgery / Clinic	Q	Human health and social work activities	OPQ
CM03	Hospital / Hospice	Q	Human health and social work activities	OPQ

ABPClass	Class_Desc	SIC	SIC_Section	SIC Group Name
CM04	Medical / Testing / Research Laboratory	M	Professional, scientific and technical activities	KLMN
CM05	Professional Medical Service	Q	Human health and social work activities	OPQ
CN	Animal Centre	A		ABDE
CN01	Cattery / Kennel	A		ABDE
CN02	Animal Services	M	Professional, scientific and technical activities	KLMN
CN03	Equestrian	R	Arts, entertainment and recreation; other service activities	RSTU
CN04	Vet / Animal Medical Treatment	M	Professional, scientific and technical activities	KLMN
CN05	Animal / Bird / Marine Sanctuary	A		ABDE
CO	Office	N		KLMN
CO01	Office / Work Studio	N		KLMN
CO02	Broadcasting (TV / Radio)	J	Information and communication	HJ
CR	Retail	G	Wholesale and retail trade; repair of motor vehicles and motor cycles	GI
CR01	Bank / Financial Service	K	Financial and insurance activities	KLMN
CR02	Retail Service Agent	G	Wholesale and retail trade; repair of motor vehicles and motor cycles	GI

ABPClass	Class_Desc	SIC	SIC_Section	SIC Group Name
CR04	Market (Indoor / Outdoor)	G	Wholesale and retail trade; repair of motor vehicles and motor cycles	GI
CR05	Petrol Filling Station	G	Wholesale and retail trade; repair of motor vehicles and motor cycles	GI
CR06	Public House / Bar / Nightclub	I	Accommodation and food service activities	GI
CR07	Restaurant / Cafeteria	I	Accommodation and food service activities	GI
CR08	Shop / Showroom	G	Wholesale and retail trade; repair of motor vehicles and motor cycles	GI
CR09	Other Licensed Premise / Vendor	R		RSTU
CR10	Fast Food Outlet / Takeaway (Hot / Cold)	I	Accommodation and food service activities	GI
CR11	Automated Teller Machine (ATM)			
CS	Storage Land			
CS01	General Storage Land	H	Transport and storage	HJ
CS02	Builders' Yard	F		F
CT	Transport	H	Transport and storage	HJ
CT01	Airfield / Airstrip / Airport / Air Transport Infrastructure Facility	H	Transport and storage	HJ
CT02	Bus Shelter	H	Transport and storage	HJ

ABPClass	Class_Desc	SIC	SIC_Section	SIC Group Name
CT03	Car / Coach / Commercial Vehicle / Taxi Parking / Park And Ride Site	H	Transport and storage	HJ
CT04	Goods Freight Handling / Terminal	H	Transport and storage	HJ
CT05	Marina	H	Transport and storage	HJ
CT06	Mooring	H	Transport and storage	HJ
CT07	Railway Asset	H	Transport and storage	HJ
CT08	Station / Interchange / Terminal / Halt	H	Transport and storage	HJ
CT09	Transport Track / Way	H	Transport and storage	HJ
CT10	Vehicle Storage	H	Transport and storage	HJ
CT11	Transport Related Infrastructure	H	Transport and storage	HJ
CT12	Overnight Lorry Park	H	Transport and storage	HJ
CT13	Harbour / Port / Dock / Dockyard / Slipway / Landing Stage / Pier / Jetty / Pontoon / Terminal / Berthing / Quay	H	Transport and storage	HJ
CU	Utility	D	Electricity, gas, steam and air conditioning supply	ABDE
CU01	Electricity Sub-Station	D	Electricity, gas, steam and air conditioning supply	ABDE
CU02	Landfill	E	Water supply, sewerage, waste management and remediation activities	ABDE

ABPClass	Class_Desc	SIC	SIC_Section	SIC Group Name
CU03	Power Station / Energy Production	D	Electricity, gas, steam and air conditioning supply	ABDE
CU04	Pump House / Pumping Station / Water Tower	E	Water supply, sewerage, waste management and remediation activities	ABDE
CU06	Telecommunication	J		HJ
CU07	Water / Waste Water / Sewage Treatment Works	E	Water supply, sewerage, waste management and remediation activities	ABDE
CU08	Gas / Oil Storage / Distribution	D	Electricity, gas, steam and air conditioning supply	ABDE
CU09	Other Utility Use	M		KLMN
CU10	Waste Management	E	Water supply, sewerage, waste management and remediation activities	ABDE
CU11	Telephone Box			
CU12	Dam	E	Water supply, sewerage, waste management and remediation activities	ABDE
CX	Emergency / Rescue Service	O	Public administration and defence; compulsory social security	OPQ
CX01	Police / Transport Police / Station	O	Public administration and defence; compulsory social security	OPQ
CX02	Fire Station	O	Public administration and defence; compulsory social security	OPQ
CX03	Ambulance Station	O	Public administration and defence; compulsory social security	OPQ
CX04	Lifeboat Services / Station	O	Public administration and defence; compulsory social security	OPQ

ABPClass	Class_Desc	SIC	SIC_Section	SIC Group Name
CX05	Coastguard Rescue / Lookout / Station	O	Public administration and defence; compulsory social security	OPQ
CX06	Mountain Rescue Station	O	Public administration and defence; compulsory social security	OPQ
CX07	Lighthouse	O	Public administration and defence; compulsory social security	OPQ
CX08	Police Box / Kiosk	O	Public administration and defence; compulsory social security	OPQ
CZ	Information			
CZ01	Advertising Hoarding			
CZ02	Tourist Information Signage			
CZ03	Traffic Information Signage			
L	Land			
LA	Agricultural - Applicable to land in farm ownership and not run as a separate business enterprise	A		ABDE
LA01	Grazing Land	A	Agriculture, forestry and fishing	ABDE
LA02	Permanent Crop / Crop Rotation	A	Agriculture, forestry and fishing	ABDE
LB	Ancillary Building			
LC	Burial Ground			
LC01	Historic / Disused Cemetery / Graveyard			

ABPClass	Class_Desc	SIC	SIC_Section	SIC Group Name
LD	Development	F	Construction	F
LD01	Development Site	F	Construction	F
LF	Forestry	A	Agriculture, forestry and fishing	ABDE
LF02	Forest / Arboretum / Pinetum (Managed / Unmanaged)	A	Agriculture, forestry and fishing	ABDE
LF03	Woodland			
LL	Allotment			
LM	Amenity - Open areas not attracting visitors			
LM01	Landscaped Roundabout			
LM02	Verge / Central Reservation			
LM03	Maintained Amenity Land			
LM04	Maintained Surfaced Area			
LO	Open Space			
LO01	Heath / Moorland			
LP	Park			
LP01	Public Park / Garden			
LP02	Public Open Space / Nature Reserve			

ABPClass	Class_Desc	SIC	SIC_Section	SIC Group Name
LP03	Playground			
LP04	Private Park / Garden			
LU	Unused Land			
LU01	Vacant / Derelict Land			
LW	Water			
LW01	Lake / Reservoir			
LW02	Named Pond			
LW03	Waterway			
M	Military	O	Public administration and defence; compulsory social security	OPQ
MA	Army	O	Public administration and defence; compulsory social security	OPQ
MB	Ancillary Building	O	Public administration and defence; compulsory social security	OPQ
MF	Air Force	O	Public administration and defence; compulsory social security	OPQ
MG	Defence Estates	O	Public administration and defence; compulsory social security	OPQ
MN	Navy	O	Public administration and defence; compulsory social security	OPQ
O	Other (Ordnance Survey Only)			
OA	Aid To Navigation			

ABPClass	Class_Desc	SIC	SIC_Section	SIC Group Name
OA01	Aid To Aeronautical Navigation			
OA02	Aid To Nautical Navigation			
OA03	Aid To Road Navigation			
OC	Coastal Protection / Flood Prevention			
OC01	Boulder Wall / Sea Wall			
OC02	Flood Gate / Flood Sluice Gate / Flood Valve			
OC03	Groyne			
OC04	Rip-Rap			
OE	Emergency Support			
OE01	Beach Office / First Aid Facility			
OE02	Emergency Telephone (Non Motorway)			
OE03	Fire Alarm Structure / Fire Observation Tower / Fire Beater Facility			
OE04	Emergency Equipment Point / Emergency Siren / Warning Flag			
OE05	Lifeguard Facility			
OE06	LIfe / Belt / Buoy / Float / Jacket / Safety Rope			

ABPClass	Class_Desc	SIC	SIC_Section	SIC Group Name
OF	Street Furniture			
OG	Agricultural Support Objects			
OG01	Fish Ladder / Lock / Pen / Trap			
OG02	Livestock Pen / Dip			
OG03	Currick			
OG04	Slurry Bed / Pit			
OH	Historical Site / Object			
OH01	Historic Structure / Object			
OI	Industrial Support			
OI01	Adit / Incline / Level			
OI02	Caisson / Dry Dock / Grid			
OI03	Channel / Conveyor / Conduit / Pipe			
OI04	Chimney / Flue			
OI05	Crane / Hoist / Winch / Material Elevator			
OI06	Flare Stack			
OI07	Hopper / Silo / Cistern / Tank			

ABPClass	Class_Desc	SIC	SIC_Section	SIC Group Name
OI08	Grab / Skip / Other Industrial Waste Machinery / Discharging			
OI09	Kiln / Oven / Smelter			
OI10	Manhole / Shaft			
OI11	Industrial Overflow / Sluice / Valve / Valve Housing			
OI12	Cooling Tower			
OI13	Solar Panel / Waterwheel			
OI14	Telephone Pole / Post			
OI15	Electricity Distribution Pole / Pylon			
ON	Significant Natural Object			
ON01	Boundary / Significant / Historic Tree / Pollard			
ON02	Boundary / Significant Rock / Boulder			
ON03	Natural Hole (Blow / Shake / Swallow)			
OO	Ornamental / Cultural Object			
OO02	Mausoleum / Tomb / Grave			
OO03	Simple Ornamental Object			
OO04	Maze			

ABPClass	Class_Desc	SIC	SIC_Section	SIC Group Name
OP	Sport / Leisure Support			
OP01	Butt / Hide			
OP02	Gallop / Ride			
OP03	Miniature Railway			
OR	Royal Mail Infrastructure			
OR01	Postal Box			
OR02	Postal Delivery Box / Pouch			
OR03	PO Box			
OR04	Additional Mail / Packet Addressee			
OS	Scientific / Observation Support			
OS01	Meteorological Station / Equipment			
OS02	Radar / Satellite Infrastructure			
OS03	Telescope / Observation Infrastructure / Astronomy			
OT	Transport Support			
OT01	Cattle Grid / Ford			
OT02	Elevator / Escalator / Steps			

ABPClass	Class_Desc	SIC	SIC_Section	SIC Group Name
OT03	Footbridge / Walkway			
OT04	Pole / Post / Bollard (Restricting Vehicular Access)			
OT05	Subway / Underpass			
OT06	Customs Inspection Facility			
OT07	Lay-By			
OT08	Level Crossing			
OT09	Mail Pick Up			
OT10	Railway Pedestrian Crossing			
OT11	Railway Buffer			
OT12	Rail Drag			
OT13	Rail Infrastructure Services			
OT14	Rail Kilometre Distance Marker			
OT15	Railway Lighting			
OT16	Rail Mile Distance Marker			
OT17	Railway Turntable			
OT18	Rail Weighbridge			

ABPClass	Class_Desc	SIC	SIC_Section	SIC Group Name
OT19	Rail Signalling			
OT20	Railway Traverse			
OT21	Goods Tramway			
OT22	Road Drag			
OT23	Vehicle Dip			
OT24	Road Turntable			
OT25	Road Mile Distance Marker			
OT26	Road Kilometre Distance Marker			
OT27	Road Infrastructure Services			
OU	Unsupported Site			
OU01	Cycle Parking Facility			
OU04	Picnic / Barbeque Site			
OU05	Travelling Persons Site			
OU08	Shelter (Not Including Bus Shelter)			
P	Parent Shell			
PP	Property Shell			

ABPClass	Class_Desc	SIC	SIC_Section	SIC Group Name
PS	Street Record			
R	Residential			
RB	Ancillary Building			
RC	Car Park Space			
RC01	Allocated Parking			
RD	Dwelling			
RD01	Caravan			
RD02	Detached			
RD03	Semi-Detached			
RD04	Terraced			
RD06	Self Contained Flat (Includes Maisonette / Apartment)			
RD07	House Boat			
RD08	Sheltered Accommodation			
RD10	Privately Owned Holiday Caravan / Chalet			
RG	Garage			
RG02	Lock-Up Garage / Garage Court			

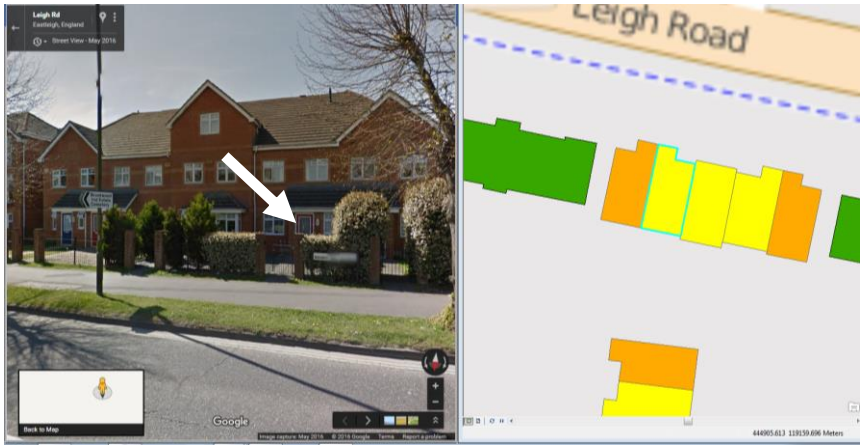
ABPClass	Class_Desc	SIC	SIC_Section	SIC Group Name
RH	House In Multiple Occupation			
RH01	HMO Parent			
RH02	HMO Bedsit / Other Non Self Contained Accommodation			
RH03	HMO Not Further Divided			
RI	Residential Institution	Q	Human health and social work activities	OPQ
RI01	Care / Nursing Home	Q	Human health and social work activities	OPQ
RI02	Communal Residence			
RI03	Residential Education			
U	Unclassified			
UC	Awaiting Classification			
UP	Pending Internal Investigation			
X	Dual Use			
Z	Object of Interest			
ZA	Archaeological Dig Site			
ZM	Monument			
ZM01	Obelisk / Milestone / Standing Stone			

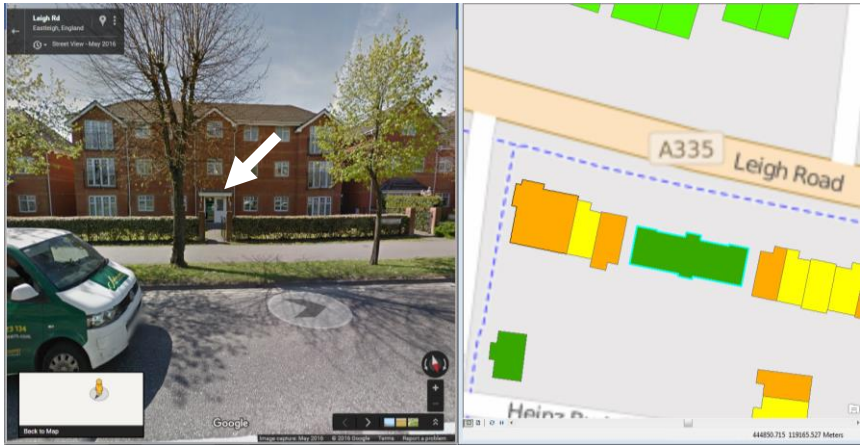
ABPClass	Class_Desc	SIC	SIC_Section	SIC Group Name
ZM02	Memorial / Market Cross			
ZM03	Statue			
ZM04	Castle / Historic Ruin			
ZM05	Other Structure			
ZS	Stately Home			
ZU	Underground Feature			
ZU01	Cave			
ZU04	Pothole / Natural Hole			
ZV	Other Underground Feature			
ZV01	Cellar			
ZV02	Disused Mine			
ZV03	Well / Spring			
ZW	Place Of Worship	R	Arts, entertainment and recreation; other service activities	RSTU

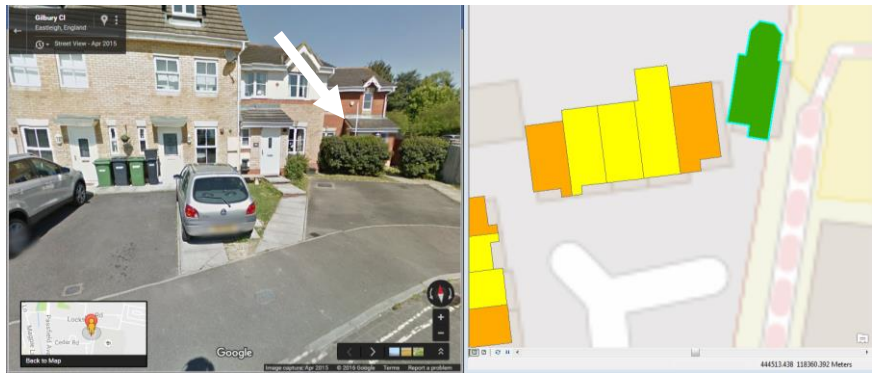
Appendix B Validation of Residential Data Classification

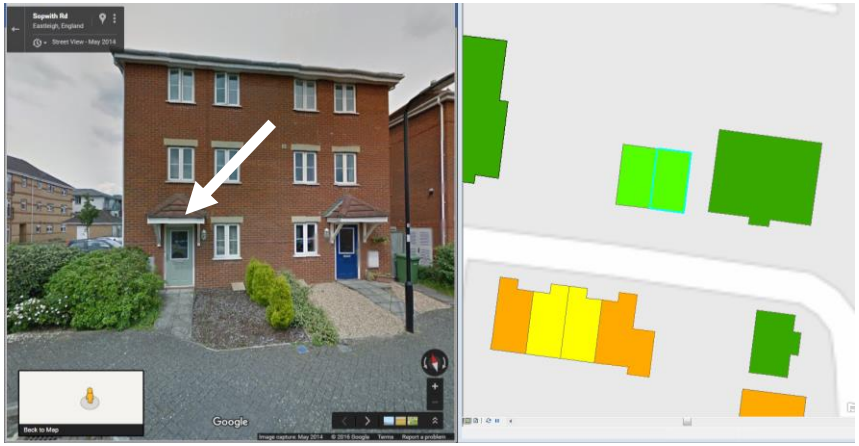
B.1 Individual Building Validation using Google Street View

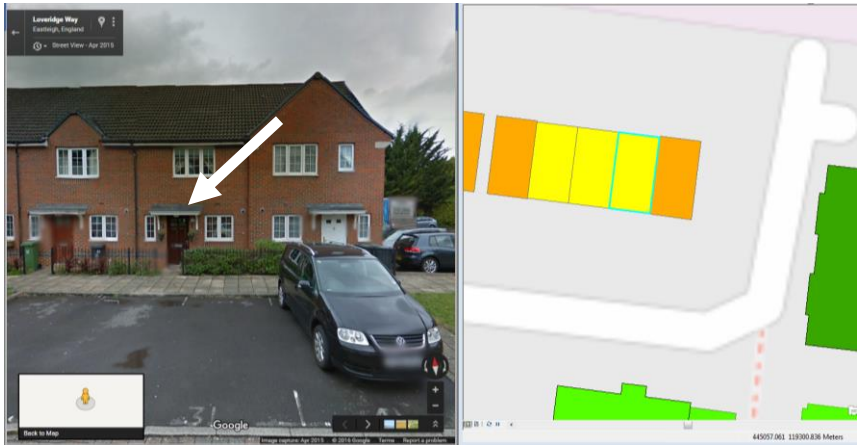
The following table provides the results of a selection of the stratified sample in the Eastleigh Study Area that were assessed for their classification validity. All addresses in the sample have results recorded in the same way.

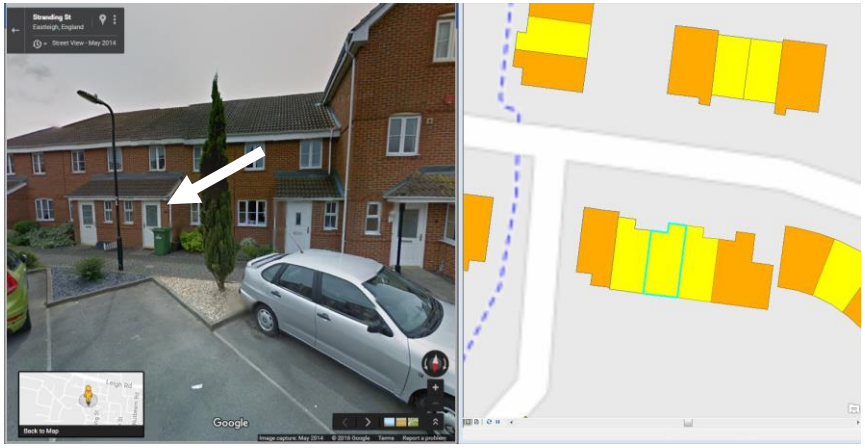
OBJ ID	Bldg hght	BLPU classes	Num Res Addr	Num Addr	Dwelling Type	Valid	Street View Capture Date	Result	GSV/ArcMap Evidence
1	6.6	RD	1	1	Mid-Terrace	Yes	May 2016	Valid	

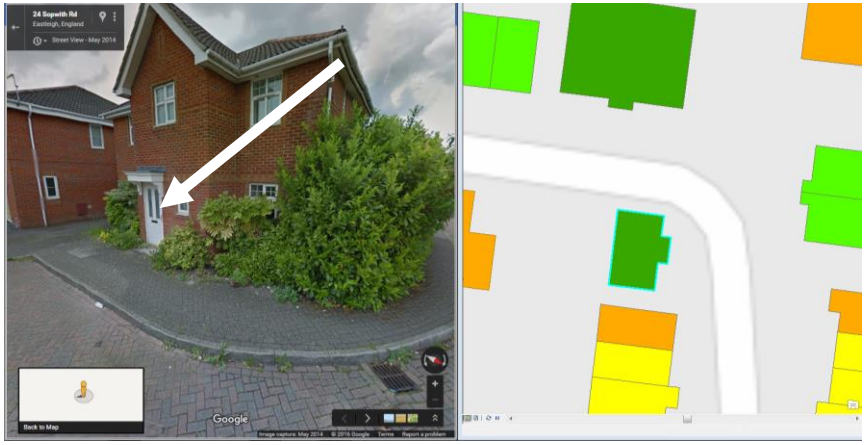
OBJ ID	Bldg hght	BLPU classes	Num Res Addr	Num Addr	Dwelling Type	Valid	Street View Capture Date	Result	GSV/ArcMap Evidence
2	8.2	RD RD RD RD RD RD	6	6	Detached - Flat	Yes	May 2016	Valid	

OBJ ID	Bldg hght	BLPU classes	Num Res Addr	Num Addr	Dwelling Type	Valid	Street View Capture Date	Result	GSV/ArcMap Evidence
3	4.3	RD	1	1	Detached	Yes	April 2015	Valid	
4	5.8	RD	1	1	Detached	Yes	April 2015	Valid	

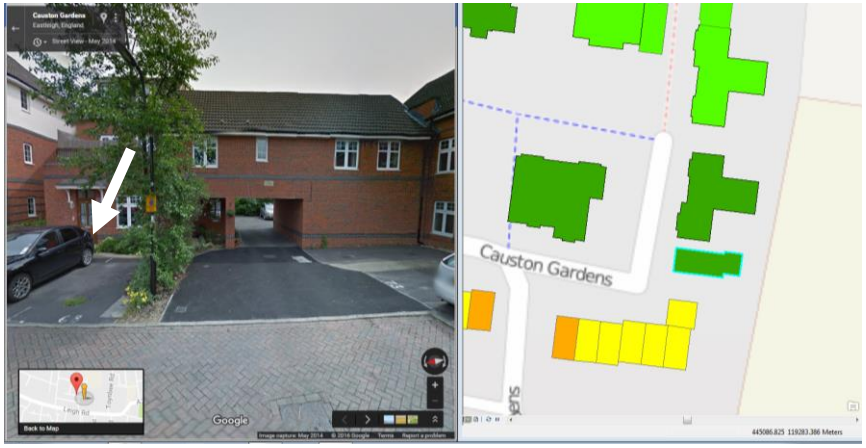
OBJ ID	Bldg hght	BLPU classes	Num Res Addr	Num Addr	Dwelling Type	Valid	Street View Capture Date	Result	GSV/ArcMap Evidence
6	7	RD	1	1	Semi-Detached	Yes	May 2014	Valid	

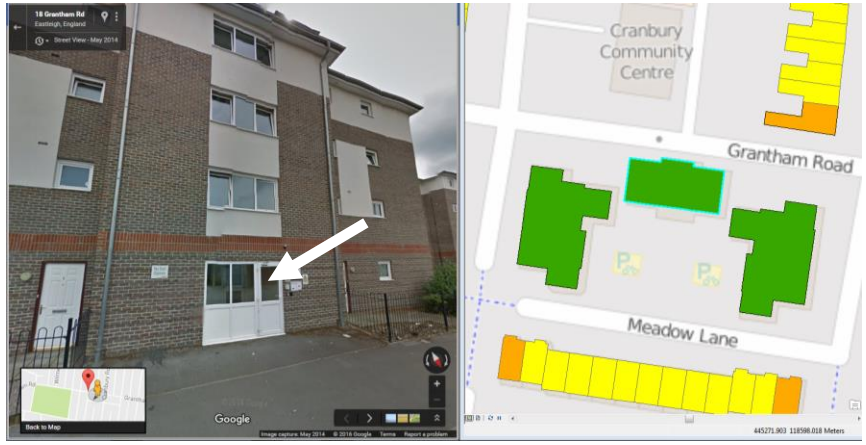
OBJ ID	Bldg hght	BLPU classes	Num Res Addr	Num Addr	Dwelling Type	Valid	Street View Capture Date	Result	GSV/ArcMap Evidence
7	6.1	RD	1	1	Mid-Terrace	Yes	April 2015	Valid	

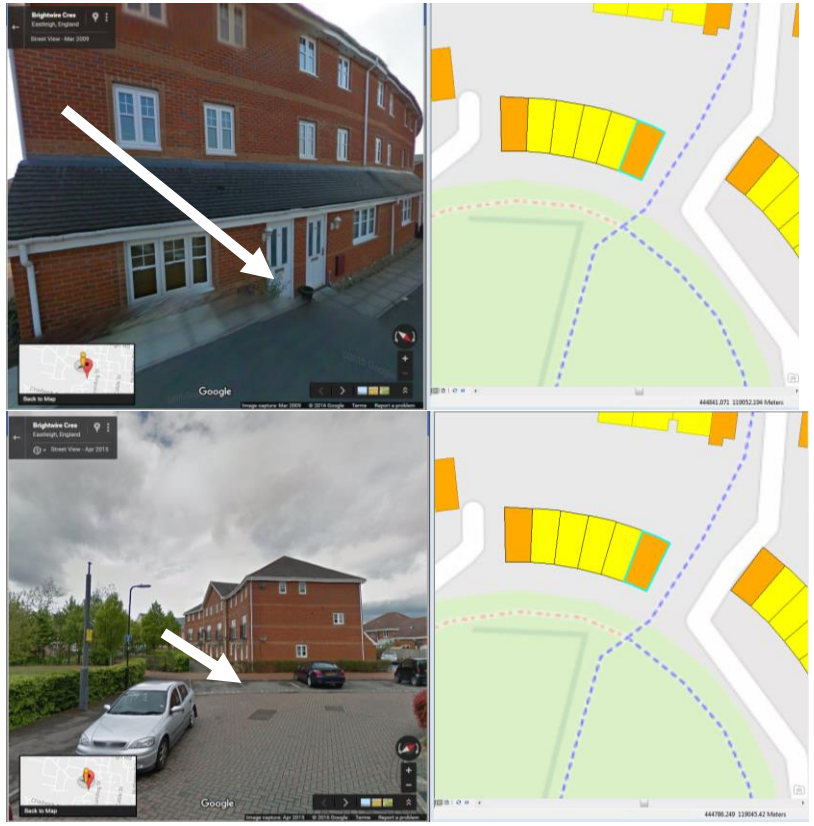
OBJ ID	Bldg hght	BLPU classes	Num Res Addr	Num Addr	Dwelling Type	Valid	Street View Capture Date	Result	GSV/ArcMap Evidence
8	2.3	RD	1	1	Mid-Terrace – Bungalow	No – not a bungalow Building height	May 2014	Data Error (height)	

OBJ ID	Bldg hght	BLPU classes	Num Res Addr	Num Addr	Dwelling Type	Valid	Street View Capture Date	Result	GSV/ArcMap Evidence
15	4.8	RD	1	1	Detached	Yes	May 2014	Valid	

OBJ ID	Bldg hght	BLPU classes	Num Res Addr	Num Addr	Dwelling Type	Valid	Street View Capture Date	Result	GSV/ArcMap Evidence
16	4.8	RD	1	1	End-Terrace	Yes	May 2014	Valid	

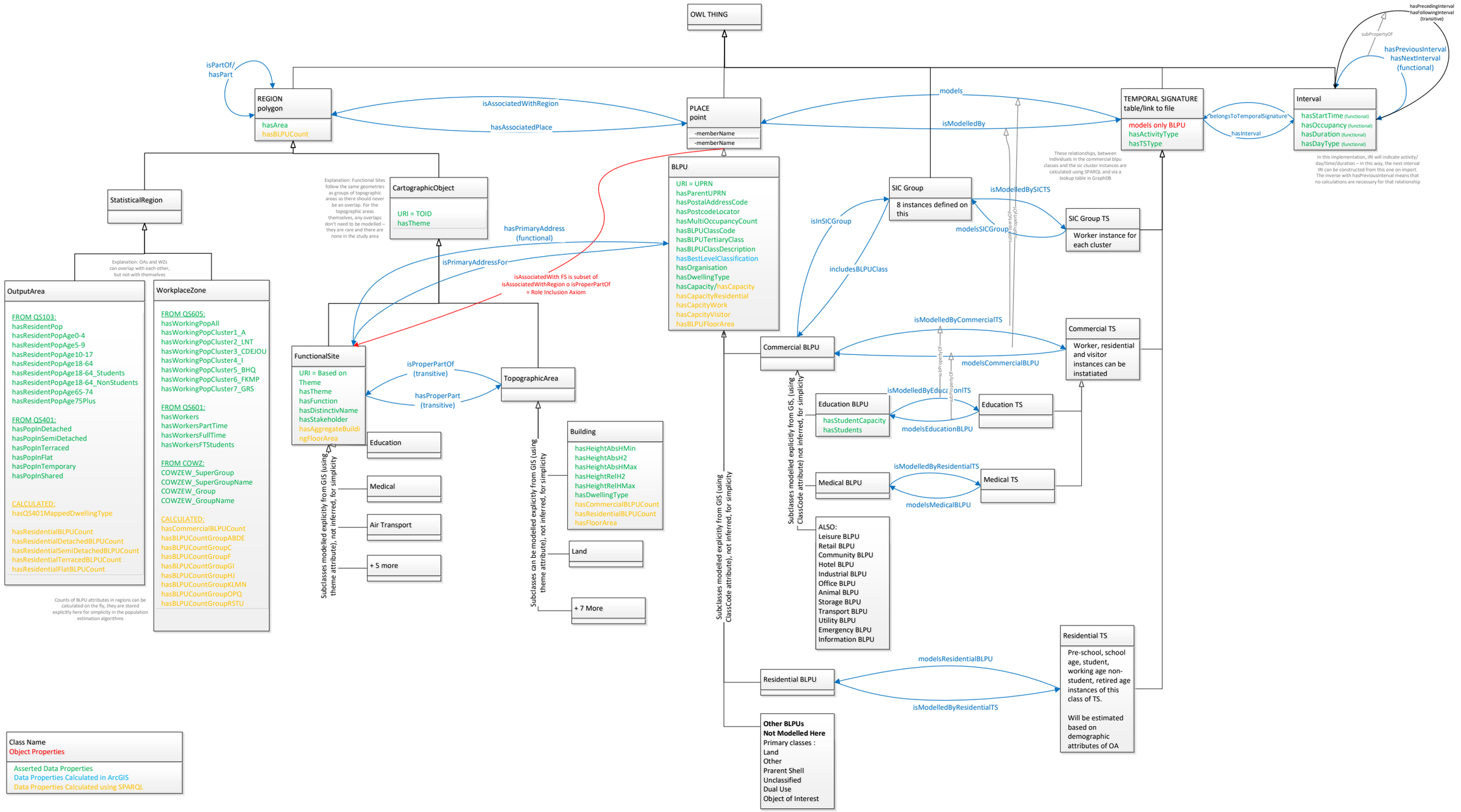
OBJ ID	Bldg hght	BLPU classes	Num Res Addr	Num Addr	Dwelling Type	Valid	Street View Capture Date	Result	GSV/ArcMap Evidence
17	7.8	RD RD	2	2	Detached - Flat	NOT SURE from this –is example of issue with the archway not containing an address.	May 2014	Valid	

OBJ ID	Bldg hght	BLPU classes	Num Res Addr	Num Addr	Dwelling Type	Valid	Street View Capture Date	Result	GSV/ArcMap Evidence
18	10.7	RD06 RD06 RD06 RD06 RD06 RD06 RD06 PP RD06 RD06 RD06 RD06	11	12	Detached - Flat	Yes	May 2014	Valid	

20	6.6	RD RD	2	2	End-Terrace - Flat	Yes – door on front and door on back of building – each is a different property – there's no indication of how the property is split inside. Is two properties within one building as indicated by number of residential addresses inside the building boundary.	May 2009 & April 2015	Valid	

Appendix C Ontology

C.1 Detailed Ontology Diagram



C.2 Example Data Load Scripts

Examples of the data load scripts are below.

C.2.1 LoadBuildingsSP

```
#================================================================
# PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
# BUILDINGSSP IS AN OUTPUT FROM THE GIS PROCESSING.
# CHANGE THE SERVICE TO CHANGE WHICH INPUT FILE IS LOADED:
#
# PREREQUISITES: NONE
#================================================================
#================================================================
# BUILDINGSSP IS ALL BUILDINGS IN THE AREA WHETHER THEY CONTAIN
# AN ADDRESS OR NOT
#================================================================

PREFIX bldgs: <http://example.com/resource/>
PREFIX spif: <http://spinrdf.org/spif#>
PREFIX popont: <http://www.example.org/pop#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

INSERT {
    ?bldgIRI a popont:BuildingsMMTArea ;
} WHERE {
    SERVICE <http://localhost:7200/rdf-bridge/2476318047662> {
        ?row a bldgs:Row ;
            bldgs:mmt_Toid ?mmt_Toid .

            BIND (iri(concat("http://www.example.org/pop#TOID_",
            spif:encodeURL(STR(?mmt_Toid)))) as ?bldgIRI)
    }
}
```

C.2.2 LoadAddresses

```
#================================================================
# LOAD ADDRESSES (BLPUS) WITH CLASSES AND LINKS TO BUILDINGS:
# ADDRESSES IS AN OUTPUT FROM THE GIS PROCESSING.
# CHANGE THE SERVICE TO CHANGE WHICH INPUT FILE IS LOADED:
#
# PREREQUISITES: NONE
#================================================================

PREFIX addr: <http://example.com/resource/>
PREFIX spif: <http://spinrdf.org/spif#>
PREFIX popont: <http://www.example.org/pop#>

INSERT {
  ?UprnIRI a ?Class ;
    popont:isAssociatedWithRegion ?BuildingIRI ;
    popont:hasBLPUClassCode ?abp_CLASSIFICATION_CODE ;
    popont:hasBLPUClassDescription ?abp_Class_Desc ;
    popont:hasDwellingType ?DTWFlats ;
```

```
    popont:hasBLPUTertiaryClass ?tertiary .
} WHERE {
    SERVICE <http://localhost:7200/rdf-bridge/1982290567024>

        ?row a addr:Row ;
         addr:blpu_UPRN ?blpu_UPRN ;
         addr:abp_CLASSIFICATION_CODE ?abp_CLASSIFICATION_CODE ;
         addr:abp_Class_Desc ?abp_Class_Desc ;
         addr:Chosen_TOID ?Chosen_TOID ;
         optional { ?row addr:DTWFlats ?DTWFlats .}

       BIND (iri(concat("http://www.example.org/pop#UPRN_",
             spif:encodeURL(STR(?blpu_UPRN)))) as ?UprnIRI)
       BIND (iri(concat("http://www.example.org/pop#TOID_",
             spif:encodeURL(?Chosen_TOID))) as ?BuildingIRI)
       BIND (IRI(CONCAT("http://www.example.org/pop#",
             STR(?abp_CLASSIFICATION_CODE), "_BLPU")) AS ?Class)
        BIND (SUBSTR(?abp_CLASSIFICATION_CODE,1,4) as ?tertiary)
    }
}
```

C.2.3 LoadSchoolsStudentCapacity

```
#=============================================================
# PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
# ONLY FOR THE STUDY AREA
#
# PREREQUISITES: REQUIRES THE ADDRESSES TO BE PRESENT - SCHOOL
# CAPACITY IS ONLY CREATED FOR EXISTING CE BLPUs
#=============================================================
PREFIX sch: <http://example.com/resource/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX popont: <http://www.example.org/pop#>
PREFIX spif: <http://spinrdf.org/spif#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

INSERT {
    ?UprnIRI popont:hasStudentCapacity ?SchoolCapacity .
    ?UprnIRI popont:hasStudents ?NumberOfPupils .
    ?UprnIRI popont:hasCapacityVisitor ?visitors .
    ?UprnIRI popont:isModelledByEducationTS ?TSIRI .
} WHERE {
    # only addresses that are education:
    ?UprnIRI rdf:type popont:CE_BLPU .

    SERVICE <http://localhost:7200/rdf-bridge/2201011820921> {
        ?row a sch:Row ;
            sch:UPRN ?UPRN ;
            sch:PhaseOfEducation ?PhaseOfEducation ;
            optional { ?row sch:SchoolCapacity
                ?SchoolCapacity . }
            optional { ?row sch:NumberOfPupils
                ?NumberOfPupils . }

        # create uprn iri (subject):
        BIND (iri(concat("http://www.example.org/pop#UPRN_",
                  spif:encodeURL(STR(?UPRN)))) as ?UprnIRI)
```

```
        # create ts iri (object):
        BIND (iri(concat(
             "http://www.example.org/pop#isModelledByEducation",
             "TS")) as ?TSIRI)

        # calculate visitor capacity from number of pupils:
        BIND(
            IF(?NumberOfPupils = 0, ?SchoolCapacity,
            IF(?NumberOfPupils > 0, ?NumberOfPupils, ""))
            AS ?visitors)
    }
}
```

C.2.4 CreateSchoolTS

```
#================================================================
# CREATE TEMPORAL SIGNATURE AND LINK TO ALL SCHOOL INSTANCES
# VISITOR CAPACITY ONLY
#
# PREREQUISITES: REQUIRES THE ADDRESSES TO BE PRESENT - SCHOOL
# CAPACITY IS ONLY  CREATED FOR EXISTING CE BLPUs
#================================================================
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX popont: <http://www.example.org/pop#>
PREFIX schts: <http://example.com/resource/>
PREFIX spif: <http://spinrdf.org/spif#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

INSERT {
    # create a temporal signature for schools (all education
    # blpus):
    ?schools_TS a ?TS_Class ;
        popont:hasActivityType "Visitor" ;
    popont:hasTSType "Class" .

    # link the schools blpus to their temporal signature
    ?sch_addr popont:isModelledByCE_TS ?schools_TS .
} WHERE {
    # all schools:
    ?sch_addr rdf:type popont:CE_BLPU .

    # Schools temporal signature and class IRIs:
    BIND (iri("http://www.example.org/pop#TS_Class_CE_Visitor")
         as ?schools_TS)
    BIND iri("http://www.example.org/pop#
         CommercialEducationTSVisitor") as ?TS_Class)
}
```

Appendix D Population Estimation

D.1 Population Estimation Scripts: Residential Activity

```
#================================================================
# BUILD THE QUERY THAT ESTIMATES POPULATION FOR ALL RESIDENTIAL
# ACTIVITY AT RESIDENTIAL ADDRESSES (NOT INSIDE FS)
#
# TO RUN THIS FOR DIFFERENT DAYS/TIMES, SELECT THE DAY
# TYPE (0 - 6) AND CHANGE THE DATE TIME IN THE FILTER
#================================================================
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX popont: <http://www.example.org/pop#>
PREFIX spif: <http://spinrdf.org/spif#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

# PROOF OF WORKINGS:
#CONSTRUCT {
#    ?address popont:hasPopCount ?numResPeople .
#} WHERE {

SELECT ?WZ (SUM(?numResPeople) as ?sumResPop) WHERE {
    # GET ADDRESSES WITH CORRECT TEMPORAL SIGNATURE:
    # residential addresses:
    ?address rdf:type popont:R_BLPU .
    ?address popont:isAssociatedWithRegion ?WZ .
    ?WZ rdf:type popont:WorkplaceZone .

    # and that are not in functional sites:
    FILTER (!EXISTS {?address popont:isAssociatedWithFS ?FS . }) .

    # sic group of the commercial addresses:
    ?address popont:isModelledByResidentialTS ?resTS .

    # GET ALL INTERVALS IN THE WORK ACTIVITY TS FOR THE SIC GROUP:
    # FOR THE CORRECT DAY TYPE:
    ?resTS popont:hasInterval ?interval .
    ?interval rdf:type popont:Interval .
    ?interval popont:hasDayType 4 .

    # APPLY THE FILTERS: latest time that is <= specified time:
    { SELECT (MAX(?start) as ?maxStart)
        WHERE{
            ?interval popont:hasStartTime ?start .
            FILTER (?start <= "1899-12-
            31T14:39:00.000Z"^^xsd:dateTime) .
        }
    }
    ?interval popont:hasStartTime ?maxStart .
    ?interval popont:hasOccupancy ?occ .
    ?address popont:hasCapacityResidential ?cap .

    BIND (?cap*?occ/100.0 as ?numResPeople)
} GROUP BY ?WZ
```

D.2 Population Estimation Scripts: Work Activity at Commercial Addresses

```
#================================================================
# BUILD THE QUERY THAT ESTIMATES POPULATION FOR ALL WORK
# ACTIVITY AT COMMERCIAL ADDRESSES (INSIDE AND NOT INSIDE FS)
#
# MODELS ALL DAY TYPES IN ONE SCRIPT
# TO RUN THIS FOR DIFFERENT TIMES, SELECT THE DAY
# TYPE (0 – 6) AND CHANGE THE DATE TIME IN THE FILTER
#================================================================
# ALL ADDRESSES INCLUDED REGARDLESS OF WHETHER THEY ARE INSIDE A
# FUNCTIONAL SITE, AS THE MAX CAPACITIES WITHIN ADDRESSES WITHIN
# FUNCTIONAL SITES HAVE ALREADY BEEN DISTRIBUTED
#================================================================
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX popont: <http://www.example.org/pop#>
PREFIX spif: <http://spinrdf.org/spif#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

# PROOF OF WORKINGS:
#CONSTRUCT {
#    ?address popont:hasPopCount ?numPeople .
#} WHERE {

SELECT DISTINCT ?DT ?WZ (SUM(?numPeople) as ?sumComPop) WHERE {
    # GET ADDRESSES WITH CORRECT TEMPORAL SIGNATURE:
    # commercial addresses:
    ?address rdf:type popont:C_BLPU .
    ?address popont:isAssociatedWithRegion ?WZ .
    ?WZ rdf:type popont:WorkplaceZone .

    # sic group of the commercial addresses:
    ?address popont:isInSICGroup ?sicGroup .

    # WORK TS of the sic group:
    ?sicGroup popont:isModelledBySICTS ?TS .
    ?TS popont:hasActivityType "Work" .

    # GET ALL INTERVALS IN THE WORK ACTIVITY TS FOR THE SIC GROUP:
    # FOR THE CORRECT DAY TYPE:
    ?TS popont:hasInterval ?interval .
    ?interval rdf:type popont:Interval .
    ?interval popont:hasDayType ?DT .

    # APPLY THE FILTERS: latest time that is <= specified time:
    { SELECT (MAX(?start) as ?maxStart)
        WHERE{
            ?interval popont:hasStartTime ?start .
            FILTER (?start <= "1899-12-
            31T11:00:00.000Z"^^xsd:dateTime) .
        }
    }
    ?interval popont:hasStartTime ?maxStart .
    ?interval popont:hasOccupancy ?occ .
    ?address popont:hasCapacityWork ?cap .
```

```
    BIND (?cap*?occ/100.0 as ?numPeople)
} GROUP BY ?DT ?WZ ORDER BY ?DT ?WZ
```

D.3 Population Estimation Scripts: Visitor Activity at Commercial Addresses

```
#================================================================
# BUILD THE QUERY THAT ESTIMATES POPULATION FOR ALL VISITOR
# ACTIVITY AT COMMERCIAL ADDRESSES
#
# TO RUN THIS FOR DIFFERENT DAYS/TIMES, SELECT THE DAY
# TYPE (0 – 6) AND CHANGE THE DATE TIME IN THE FILTER
#================================================================
# CALCULATE COMMERCIAL VISITOR ESTIMATES
# to include:
#   for address or class temporal signatures
#   with and without estimate of visitor capacity
#================================================================
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX popont: <http://www.example.org/pop#>
PREFIX spif: <http://spinrdf.org/spif#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

## PROOF OF WORKINGS:
#CONSTRUCT {
#    ?address popont:hasPopCount ?present .
#} WHERE {

SELECT DISTINCT ?WZ (SUM(?present) as ?visitorsPresent) WHERE {

    # GET TEMPORAL SIGNATURE AND INTERVAL TO USE FOR EACH TS:
    #=======================================================
    {    # get the max time for each temporal signature
        SELECT ?TS (MAX(?start) as ?maxStart)
        WHERE {
            ?TS rdf:type popont:TemporalSignature .
            ?TS popont:hasActivityType ?TSActivityType .
            FILTER (?TSActivityType = "Visitor" ) .

            ?TS popont:hasInterval ?interval .
            ?interval popont:hasDayType 0 .
            ?interval popont:hasStartTime ?start .
            FILTER (?start <= "1899-12-
             31T01:00:00.000Z"^^xsd:dateTime) .
    } GROUP BY ?TS
    }

    # now get the interval with the max time for each temporal
    # signature
    ?TSig popont:hasInterval ?int .
    ?TSig popont:hasActivityType ?AT .
    FILTER (?AT = "Visitor")
    ?int popont:hasDayType 4 .

    ?int popont:hasStartTime ?maxStart .

    # ESTIMATE POPULATION AT ADDRESSES:
    #==================================
    # occupancy rate from interval
```

```
    ?int popont:hasOccupancy ?occ .

    # addresses modelled by the TSs
    ?address popont:isModelledByTS ?TSig .
    ?address popont:isAssociatedWithRegion ?WZ .
    ?WZ rdf:type popont:WorkplaceZone .

   {
        # DO THE CALCULATIONS FOR ADDRESSES NOT IN FS & WITH
        # CAPACITY:
        #-------------------------------------------------------
        FILTER (!EXISTS {?address popont:isAssociatedWithFS ?FS .
                ?FS rdf:type popont:FunctionalSite}) .

        ?address popont:hasCapacityVisitor ?cap .

        BIND (?cap*(?occ/100) as ?present )
    }
    UNION
    {
        # DO THE CALCULATIONS FOR ADDRESSES NOT IN FS & WITHOUT
        # CAPACITY:
        #-------------------------------------------------------
        FILTER (!EXISTS {?address popont:isAssociatedWithFS ?FS .
                ?FS rdf:type popont:FunctionalSite}) .

        FILTER(!EXISTS{?address popont:hasCapacityVisitor ?cap .})
        ?address popont:hasBLPUFloorArea ?FA .

        BIND ((?FA/15)*(?occ/100) as ?present )
    }
    UNION
    {
        # DO THE CALCULATIONS FOR ADDRESSES IN FS & WITH CAPACITY:
        #-------------------------------------------------------
        ?address popont:isAssociatedWithFS ?FS .
        ?FS rdf:type popont:FunctionalSite .

        ?address popont:hasCapacityVisitor ?cap .

        BIND (?cap*(?occ/100) as ?present )
    }
    UNION
    {
        # DO THE CALCULATIONS FOR ADDRESSES IN FS & WITHOUT
        # CAPACITY:
        #-------------------------------------------------------
        ?address popont:isAssociatedWithFS ?FS .
        ?FS rdf:type popont:FunctionalSite .

        FILTER(!EXISTS{?address popont:hasCapacityVisitor ?cap .})
        ?FS popont:hasAggregateBuildingFloorArea ?FSFA .

        BIND ((?FSFA/15)*(?occ/100) as ?present )
    }
} GROUP BY ?WZ ORDER BY ?WZ
```

Glossary of Terms

Address: "a means of referencing an object for the purposes of unique identification and location" (Katalysis Limited 2014). In common usage, an address is a means of finding a property, but it can also relate to non-postally addressable objects. Throughout this thesis, addresses are defined as structured textual descriptions of spatial objects, including, but not limited to, those used in the management of postal delivery.

Ambient Population: a temporally averaged population density measure, i.e. representing an average of population throughout a specified period such as a day.

Ancillary Data: Additional datasets that provide further information for a process.

BLPU: Basic Land and Property Unit, is an area of land in uniform property rights. This forms the basis of address features generated from AddressBase Premium.

Cadastral: related to a map or survey results showing property boundaries, used for land registry and legal process.

Cartographic Objects: real world features modelled within a Geographic Information System (GIS), and with geometries represented using polygons, lines and points. Attributes supply additional information about the individual cartographic objects.

Census Output Area (OA): In England and Wales, OAs are the lowest geographical level at which census estimates are provided (ONS 2012b). They were built from clusters of adjacent unit postcodes and were "designed to have similar population sizes and be as socially homogeneous as possible based on tenure of household and dwelling type" *ibid*. For the 2011 Census, the target size for an OA was to contain between 100 and 625 individuals or 40 to 250 households.

Choropleth Map: An areal thematic map. Areas (polygons) are shaded in different patterns or colours depending on the magnitude of the attribute value being depicted.

Classes: in ontology, classes are formal, explicit definitions of concepts.

Communal Establishments: managed residential establishments, including prisons, university halls, care homes, hotels or bed and breakfast accommodation with 10 or more guest beds and army barracks (Department for Communities and Local Government 2012).

Dasymetric Mapping: Dasymetric (equal density) mapping is a cartographic technique that uses ancillary data to guide redistribution of population within the areas for which the population statistics are published, in order to portray the population location more accurately.

Delivery Points: A delivery point is an address to which mail can be delivered.

Functional Classification: Classification based on function rather than form. An addresses function may be, for example, residential, while the form of the building that it is associated with may be semi-detached.

Functional/Inverse Functional Property: In a subject-predicate-object triple, a functional property is property that an individual can only have one relationship of this type (e.g. hasBirthMother), where the individual is the subject. An inverse functional property is one that an individual can only have one relationship of this type if the individual is the object (e.g. isBirthMotherOf).

Functional Site: important sites such as airports, schools, hospitals or ports that are comprised of one or more than one topographic area in the MasterMap Topography layer.

gawk: a programming language used for data extraction.

Geographical Information System (GIS): a system for storing, managing, visualising and analysing data, with the ability to handle spatial data and spatial data models.

Houses in Multiple Occupation: properties that accommodate more than one unit of living accommodation, they share one or more of the basic amenities, such as toilet, washing and cooking facilities, and they can be buildings or parts of buildings, individual self-contained flats, or converted buildings (The National Archives 2004).

Interpolation: This is the process of inserting an estimated value between known or measured values. Areal Interpolation may distribute values for a single area within that area to infer greater detail than is available from the measured values.

Mereology: the theory of part-whole relationships. Mereological relationships can be used in qualitative spatial reasoning, are concerned with containment, and are represented in RCC5.

Mereotopology: a unified theory of mereological relations and topological relations (Casati & Varzi 1999) the relationships additional to those in mereology, are concerned with connectedness are represented in RCC8.

Non-Postally Addressable Object: A real world object that does not receive post. These may be allotments, parks, post boxes, communications masts and many other object types that do not

receive post but that need to be modelled as cartographic objects in a GIS database, because these can be occupied.

Ontology: In computer science, "an explicit and formal specification of a conceptualisation of a domain of interest" (Davies, Studer & Warren 2006), that allows inferences to be made from data.

Open World Assumption: The open world assumption refers to the web's approach to data: we must always assume that more information may come to light (Allemang & Hendler 201110). Database technologies generally used a closed world assumption, which is that something is assumed to be false if it is not held within the database (Hart & Dolbear 2013, p52).

Output Area: the smallest areal unit for publication of residential statistics from the census in England, Wales and Northern Ireland.

Population Density: a count of population per unit area e.g. 20 people per $100m^2$.

Postcode: The postcode system is UK wide and identifies postal delivery areas. It is used by the Office for National Statistics (ONS) for the main geographic reference when collecting data. Postcodes have several, nested components. From largest areal unit to smallest, these are: Postcode Area, Postcode District, Postcode Sector and Unit Postcode.

Property: a parcel of land, building or portion of a building such as a flat.

Properties: In ontology, formal, explicit definitions of the relationships between resources or between resources (object properties) and values such as strings or integers (value properties).

Pycnophylactic: This is the property of mass preservation as defined by (Tobler 1979). A population redistribution process preserves the pycnophylactic property if people are not created or removed as part of the process. Redistributed population, when re-aggregated to the source areal units will sum to the same number as the original source zone population values.

Resource Description Framework (RDF): the standard model for data interchange on the web (W3C 2014b).

Reflexive/Irreflexive Property: In ontology, a reflexive property is a relationship that an individual has with itself (e.g. knows). An irreflexive property is a relationship that an individual cannot have with itself (e.g. isMotherOf).

Semantic Web: "an extension of the web in which information is given well-defined meaning, better enabling computers and people to work in cooperation" (Berners-Lee, Hendler & Lassila 2001).

SIC Group: a grouping of standard industrial classification sections based on their temporal activity patterns.

Sites of Human Activity: places where human activities such as work, leisure or travel occur, this may be an address, a cartographic feature such as a building or an open space, or a collection of cartographic features relating to the same activity.

Spatial Disaggregation: the process of distributing population counts across the areal unit to which they apply.

Taxonomy: A taxonomy is a class hierarchy, with super and sub-classes, and provides a means of inheritance for objects and relationships defined in the ontology.

Temporal Signatures: a description of human activity for a place, at different times of the day, week or season. This is a graph of occupation of a building by per cent of maximum occupation for the modelled time periods.

Thiessen Polygons: Thiessen polygons are proximal zones generated from point geometries.

Topology: The theory of spatial relations of whole objects that do not change when space is distorted. These relations include containment and adjacency.

Transitive Property: A transitive property skips through generations. E.g. if A is inside B, and B is inside C, then A is inside C.

Triple: In ontology, the basic unit of data, defined as subject, predicate and object.

Triple store: In ontology, the database that stores all of the triples that define the ontology and its data.

Unit Postcode: The base unit of postal geography. The unit postcode is a *collection* of addresses that are usually adjacent. Some postcodes are assigned to individual addresses that meet criteria related to volume of mail received. These are very often business addresses. Each unit postcode typically contains 15 addresses, although they may contain more (up to 100) (ONS 2012b). There are approximately 1.7 million postcodes in the UK. Note that unit postcodes are not areas, but are lists of addresses.

Visitors: non-work and non-residential population present at an address.

Workplace Zones: the smallest areal unit for publication of workplace statistics from the census in England, Wales and Northern Ireland.

List of References

Abburu, S., 2012. A Survey on Ontology Reasoners and Comparison. *International Journal of Computer Applications*, 57(17), pp.33–39.

Agarwal, P., 2005. Ontological Considerations in GIScience. *International Journal of Geographical Information Science*, 19(5), pp.501–536.

Ahola, T. et al., 2007. A spatio-temporal population model to support risk assessment and damage analysis for decision-making. *International Journal of Geographical Information Science*, 21(8), pp.935–953.

Alahmadi, M., Atkinson, P. & Martin, D., 2013. Estimating the spatial distribution of the population of Riyadh, Saudi Arabia using remotely sensed built land cover and height data. *Computers, Environment and Urban Systems*, 41, pp.167–176.

Allemang, D. & Hendler, J., 2011. *Semantic Web for the Working Ontologist : effective modeling in RDFS and OWL* Second., Morgan Kaufmann.

Allen, J.F., 1983. Maintaining Knowledge about Temporal Intervals. *Research Contributions*, 26(11), pp.832–843.

Allen, J.F., 1984. Towards a General Theory of Action and Time. *Artifical Intelligence*, 23(2), pp.123–154.

Aubrecht, C. et al., 2013. Multi-level geospatial modeling of human exposure patterns and vulnerability indicators. *Natural Hazards*, 68(1), pp.147–163.

Barth, D., 2009. The bright side of sitting in traffic: Crowdsourcing road congestion data. *Google Official Blog*. Available at: https://googleblog.blogspot.com/2009/08/bright-side-of-sitting-in-traffic.html.

Batista e Silva, F., Gallego, J. & Lavalle, C., 2013. A high-resolution population grid map for Europe. *Journal of Maps*, 9(1), pp.16–28.

Batty, M., 2013. Big data, smart cities and city planning. *Dialogues in Human Geography*, 3(3), pp.274–279.

BBC, 2013. What has caused the European floods? Available at: http://www.bbc.co.uk/news/world-europe-22774962 [Accessed August 6, 2014].

Berners-Lee, T. et al., 1998. RFC 2396: Uniform Resource Identifiers (URI): Generic Syntax. , p.39.

Berners-Lee, T., Hendler, J. & Lassila, O., 2001. The Semantic Web. *Scientific American*, 284(5), pp.35–43.

Bhaduri, B. et al., 2007. LandScan USA: A high-resolution geospatial and temporal modeling approach for population distribution and dynamics. *GeoJournal*, 69(1–2), pp.103–117.

Bhaduri, B., Bright, E. & Coleman, P., 2005. Development of a high resolution population dynamics model. *Geocomputation*.

Brachman, J.R. & Levesque, J.H., 1984. The Tractability of Subsumption in Frame-Based Description Languages. *Proceedings of the National Conference on Artificial Intelligence*, pp.34–37.

Buja, K. & Menza, C., 2013. Sampling Design Tool for ArcGIS - Instruction Manual. , pp.1–16.

Burt, S., 2005. Cloudburst upon Hendraburnick Down: The Boscastle storm of 16 August 2004. *Weather*, 60(8), pp.219–227.

Calder, A., 2009. Building the address register for the 2011 Census. *Population Trends*, (138), pp.22–26.

Casati, R. & Varzi, A.C., 1999. *Parts and places: The structures of spatial representation. Mit Press.*, The MIT Press.

Centre for Ecology and Hydrology, 2008. Land Cover Map 2007 [TIFF geospatial data], Scale 1:250000, Updated: 18 July 2008, Using: EDINA Environment Digimap Service, Downloaded: 2015-12-27. Available at: http://digimap.edina.ac.uk.

Charles-Edwards, E. & Bell, M., 2013. Estimating the Service Population of a Large Metropolitan University Campus. *Applied Spatial Analysis and Policy*, 6(3), pp.209–228.

Cheng, T. & Adepeju, M., 2014. Modifiable temporal unit problem (MTUP) and its effect on space-time cluster detection. *PLoS ONE*, 9(6), pp.1–10.

Cockings, S. et al., 2013. Getting the foundations right: Spatial building blocks for official population statistics. *Environment and Planning A*, 45(6), pp.1403–1420.

Cockings, S. et al., 2017. Population247NRT: Near real-time spatiotemporal population estimates for health, emergency response and national security. Available at: https://gtr.ukri.org/projects?ref=ES%2FP010768%2F1 [Accessed August 24, 2018].

Cockings, S., Martin, D. & Harfoot, A., 2015. A Classification of Workplace Zones for England and Wales (COWZ-EW). , p.14.

Cohn, A.G. et al., 1997. Qualitative Spatial Representation and Reasoning with the Region Connection Calculus. *GeoInformatica*, 1(3), pp.275–316.

Cohn, A.G. & Renz, J., 2008. Qualitative Spatial Representation and Reasoning. In F. van Harmelen, V. Lifschitz, & B. Porter, eds. *Handbook of Knowledge Representation*. Elsevier B.V., pp. 551–584.

Çolak, S., Lima, A. & González, M.C., 2016. Understanding congested travel in urban areas. *Nature Communications*, 7, p.10793.

Çöltekin, A. et al., 2011. Modifiable temporal unit problem. *Progress in Physical Geography*.

Commission for Architecture and the Built Environment, 2005. *Better Neighbourhoods: Making Higher Densities Work*, London. Available at: http://webarchive.nationalarchives.gov.uk/20110118095356/http:/www.cabe.org.uk/files/better-neighbourhoods.pdf.

Cormen, T.H., Leiserson, C.E. & Rivest, R.L., 1990. *Introduction to Algorithms* 1st ed., Cambridge,Mass. ; London: MIT Press.

Cova, T.J. & Goodchild, M.F., 2002. Extending geographical representation to include fields of spatial objects. *International Journal of Geographical Information Science*, 16(6), pp.509–532.

Davies, J., Studer, R. & Warren, P., 2006. *Semantic Web Technologies trends and research in ontology-based systems*, Chichester: John Wiley & Sons Inc.

Department for Communities and Local Government, 2012. Definitions of general housing terms. *Government Guidance*. Available at: https://www.gov.uk/guidance/definitions-of-general-housing-terms [Accessed January 26, 2016].

Department for Transport, 2017. Traffic Counts. Available at: http://www.dft.gov.uk/traffic-counts/ [Accessed April 3, 2016].

Deville, P. et al., 2014. Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences*, 111(45), pp.15888–15893.

Diakakis, M. et al., 2017. Identifying elements that affect the probability of buildings to suffer flooding in urban areas using Google Street View. A case study from Athens metropolitan

area in Greece. *International Journal of Disaster Risk Reduction*, 22(February), pp.1–9.

Dobson, J.E. et al., 2000. LandScan: A global population database for estimating populations at risk. *Photogrammetric Engineering and Remote Sensing*, 66(7), pp.849–857.

DuCharme, B., 2013. *Learning SPARQL* 2nd Editio., Sebastapol: O'Reilly.

Duke-Williams, O. & Rees, P., 1998. Can census offices publish statistics for more than one small area geography? An analysis of the differencing problem in statistical disclosure. *International journal of geographical information science : IJGIS*, 12(6), pp.579–605.

European Commission, 2014. Infrastructure for Spatial Information in Europe Data Specification on Addresses - Technical Guidelines. , (March).

Facebook, 2016. Connecting the world with better maps. Available at: https://code.facebook.com/posts/1676452492623525/connecting-the-world-with-better-maps/ [Accessed February 24, 2016].

FGDC, 2010. United States Thoroughfare, Landmark, and Postal Address Data Standard: Address Data Classification. FGDC Document Number FGDC-STD-016-2011.

Fisher, P., 1997. The pixel: A snare and a delusion. *International Journal of Remote Sensing*, 18(3), pp.679–685.

Fisher, P., Wood, J. & Cheng, T., 2004. Where is Helvellyn? Fuzziness of multi-scale landscape morphometry. *Transactions of the Institute of British Geographers*, 29(1), pp.106–128.

Foley, D.L., 1952. The Daily Movement of Population into Central Business Districts All use subject to JSTOR Terms and Conditions THE DAILY MOVEMENT OF POPULATION INTO CENTRAL BUSINESS DISTRICTS *. , 17(5), pp.538–543.

Fotheringham, A.S. & Wong, D.W.S., 1991. The Modifiable Areal Unit Problem in Multivariate Statistical Analysis. *Environment and Planning A*, 23(7), pp.1025–1044.

Frank, A.U., 1998. Spatial Ontology. In O. Stock, ed. *Spatial and Temporal Reasoning*. Springer Science & Business Media, pp. 135–151.

Freire, S. & Aubrecht, C., 2011. Assessing Spatio-Temporal Population Exposure to Tsunami Hazard in the Lisbon Metropolitan Area. *Proceedings of the 8th International ISCRAM Conference*, (May), pp.1–5.

Freire, S., Aubrecht, C. & Wegscheider, S., 2013. Advancing tsunami risk assessment by improving

spatio-temporal population exposure and evacuation modeling. *Natural Hazards*, 68(3), pp.1311–1324.

Fry, C., 1999. GIS in telecommunications. *Geographic information systems : principles, Techniques, applications and Management*, 2, pp.819–826.

Galton, A., 1998. Space, Time, and Movement. In O. Stock, ed. *Spatial and temporal reasoning*. Springer Science & Business Media, pp. 321–352.

Garrity, T.F., 2008. Getting Smart. *IEEE Power and Energy Magazine*, 6(2), pp.38–45.

GeoData Institute, University of Southampton, WorldPop. Available at: http://www.worldpop.org.uk/ [Accessed March 25, 2016].

Gershuny, J. & Sullivan, O., 2017. United Kingdom Time Use Survey 2014-2015.

Goodchild, M.F., 1992. Geographical Data Modeling. , 18(4).

Goodchild, M.F., 2013. Prospects for a Space–Time GIS. *Annals of the Association of American Geographers*, 103(5), pp.1072–1077.

Goodchild, M.F., Anselin, L. & Deichmann, U., 1993. A framework for the areal interpolation of socioeconomic data. *Environment and Planning A*, 25(1989), pp.383–397.

Goodchild, M.F. & Lam, N.S.-N., 1980. Areal Interpolation: a variant of the traditional spatial problem. *Geo-processing*, 1, pp.197–312.

Goodchild, M.F., Yuan, M. & Cova, T.J., 2007. Towards a general theory of geographic representation in GIS. *International Journal of Geographical Information Science*, 21(3), pp.239–260.

Google, 2016. Google StreetView. Available at: https://mapstreetview.com/#ubwht_-vnn4_7i.a_5e43.

Google, 2018. Popular Times, Wait Times and Visit Duration. Available at: https://support.google.com/business/answer/6263531?hl=en-GB [Accessed August 20, 2018].

Greger, K., 2014. Spatio-Temporal Building Population Estimation for Highly Urbanized Areas Using GIS. *Transactions in GIS*, 19(1), pp.129–150.

Gregory, D. et al., 2011. *Dictionary of Human Geography*, John Wiley & Sons.

Gruber, T.R., 1993. A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2), pp.199–220.

Gruber, T.R., 1995. Toward Principles for the Design of Ontologies Used for Knowledge Sharing. *International Journal of Human-Computer Studies*, 43, pp.907–928.

Hägerstrand, T., 1976. Geography and the study of interaction between nature and society. *Geoforum*, 7, pp.329–334.

Hägerstrand, T., 1975. Space, Time and Human Conditions. In A. Karlqvist, L. Lunqvist, & F. Snickars, eds. *Dynamic allocation of urban space*. Farnborough: Saxon House.

Hägerstrand, T., 1973. The Domain of Human Geography. In R. J. Chorley, ed. *Directions in Geography*. London: Methuen & Co.

Hägerstrand, T., 1970. What About People in Regional Science. *Papers in Regional Science*, 24, pp.7–24.

Hallin, P.O., 1991. New Paths for Time-Geogrpahy. *Geografiska Annaler. Series B, Human Geography*, 73(3), pp.199–207.

Harper, G. & Mayhew, L., 2012. Applications of Population Counts Based on Administrative Data at Local Level. *Applied Spatial Analysis and Policy*, 5(3), pp.183–209.

Hart, G. & Dolbear, C., 2013. *Linked Data: A Geographic Perspective*, CRC Press.

Van Heijst, G., Schreiber, A.T. & Wielinga, B.J., 1997. Using explicit ontologies in KBS development. *International Journal of Human-Computer Studies*, 45, pp.183–292.

Highways England, 2015. TRIS – User Guide Revision 3 Highways England – Data.gov.uk – Journey Time and Traffic Flow Data April 2015 onwards – User Guide. , (April), pp.1–14.

Horrocks, I., Kutz, O. & Sattler, U., 2006. The Even More Irresistible SROIQ. *Proc. of the 10th Int. Conf. on Principles of Knowledge Representation and Reasoning (KR2006)*, pp.57–67.

HSE, 2017. National Population Database. Available at: https://www.hsl.gov.uk/what-we-do/data-analytics/national-population-database [Accessed August 22, 2018].

Huang, Q. & Wong, D.W.S., 2015. Modeling and Visualizing Regular Human Mobility Patterns with Uncertainty : An Example Using Twitter Data Modeling and Visualizing Regular Human Mobility Patterns with Uncertainty : An Example Using Twitter Data. *Annals of the Association of American Geographers*, 105(6)(November), pp.1179–1197.

Intelligent Addressing, 2010. Summary Product Descritpion for the National Land and Property Gazetteer. Available at: http://www.iahub.net/docs/1263829255172.pdf [Accessed July 8, 2014].

Jacques, D.C., 2018. Mobile Phone Metadata for Development. , pp.1–28.

Jacquez, G., 2011. It's about space and time: From the modi- fiable areal unit problem (MAUP) to the modifiable temporal unit problem (MTUP) to the modifiable spatio-temporal unit problem (MSTUP).

Jochem, W.C., Sims, K., Bright, E. a., et al., 2013. Estimating traveler populations at airport and cruise terminals for population distribution and dynamics. *Natural Hazards*, 68(3), pp.1325–1342.

Jochem, W.C., Sims, K., Bright, E.A., et al., 2013. Estimating traveler populations at airport and cruise terminals for population distribution and dynamics. *Natural Hazards*, 68(3), pp.1325–1342.

Katalysis Limited, 2014. An Open National Address Gazetteer. , (January), p.101.

Keßler, C. & McKenzie, G., 2018. A geoprivacy manifesto. *Transactions in GIS*, 22(1), pp.3–19.

Krisp, J.M., 2010. Planning Fire and Rescue Services by Visualizing Mobile Phone Density. *Journal of Urban Technology*, 17(1), pp.61–69.

Kwan, M.-P., 2004. GIS Methods in Time-Geographic Research: Geocomputation and Geovisualization of Human Activity Patterns. *Geografiska Annaler: Series B, Human Geography*, 86(4), pp.267–280.

Kwan, M.-P. & Neutens, T., 2014. Space-time research in GIScience. *International Journal of Geographical Information Science*, 28(5), pp.851–854.

Land Information New Zealand, 2010. Landonline Bulk Data Extract Overview V24.

Langford, M., 2013. An Evaluation of Small Area Population Estimation Techniques Using Open Access Ancillary Data. *Geographical Analysis*, 45(3), pp.324–344.

Langford, M., 2006. Obtaining population estimates in non-census reporting zones: An evaluation of the 3-class dasymetric method. *Computers, Environment and Urban Systems*, 30(2), pp.161–180.

Langford, M. & Unwin, D.J., 1994. Generating and mapping population density surfaces within a

geographical information system. *The Cartographic journal*, 31(1), pp.21–26.

Langran, G., 1992. *Time in Geographic Information Systems*, London: Taylor & Francis Ltd.

Lenntorp, B., 1999. Time-geography - At the end of its beginning. *GeoJournal*, 48(3), pp.155–158.

Linard, C. & Tatem, A.J., 2012. Large-scale spatial population databases in infectious disease research. *International Journal of Health Geographics*, 11(7).

Longley, P.A. et al., 2011. *Geographic Information Systems and Science* Third Edit., John Wiley & Sons.

Malleson, N. & Andresen, M.A., 2016. Exploring the impact of ambient population measures on London crime hotspots. *Journal of Criminal Justice*, 46, pp.52–63.

Martin, B.A., 2016. *RADPOP: A New Modelling Framework for Radiation Protection*. University of Southampton. Available at: https://eprints.soton.ac.uk/412256/1/Becky_Alexis_Martin_PhD_Thesis_final.pdf.

Martin, D., 1996. An assessment of surface and zonal models of population. *International journal of geographical information systems*, 10(8), pp.973–989.

Martin, D. et al., 2000. Refining Population Surface Models : Experiments with Northern Ireland Census Data. , 4(4).

Martin, D., Cockings, S. & Harfoot, A., 2013. Development of a geographical framework for census workplace data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 176(2), pp.585–602.

Martin, D., Cockings, S. & Leung, S., 2015. Developing a Flexible Framework for Spatiotemporal Population Modeling. *Annals of the Association of American Geographers*, 105(4), pp.754–772.

Martin, D., Cockings, S. & Leung, S., 2009. Population 24/7: building time-specific population grid models. In *European Forum for Geostatistics*. pp. 1–11.

Martin, D., Cockings, S. & Smith, A., 2017. *Population24/7 An open gridded population dataset for England and Wales*, Available at: http://pop247.geodata.soton.ac.uk/.

Martinez-Cruz, C., Blanco, I.J. & Vila, M.A., 2012. Ontologies versus relational databases: Are they so different? A comparison. *Artificial Intelligence Review*, 38(4), pp.271–290.

Massey, D., 2016. Space-time, “science” and the relationship between physical geography and

human geography. , 24(3), pp.261–276.

Matheson, J., 2014. The Census and Future Provision of Population Statistics in England and Wales : Recommendation from the National Statistician and Chief Executive of the UK Statistics Authority The Census and Future Provision of Population Statistics in England and Wales. , (March), p.12.

McPherson, T.N. et al., 2006. A Day-Night Population Exchange Model for Better Exposure and Consequence Management Assessments. *Symposium on the Urban Environment,86th AMS Annual Meeting*, p.6.

McPherson, T.N. & Brown, M.J., 2004. Estimating daytime and nighttime population distributions in U.S. cities for emergency response activities. *Bulletin of the American Meteorological Society*, pp.557–566.

Mennis, J., 2003. Generating Surface Models of Population Using Dasymetric Mapping. *The Professional Geographer*, 55(1), pp.31–42.

Meterological Office, 2011. Nowcasting. Available at: https://www.metoffice.gov.uk/learning/making-a-forecast/hours-ahead/nowcasting [Accessed August 20, 2018].

MHCLG, 2016. Private Renting. Available at: https://www.gov.uk/private-renting.

Miller, H.H.J., 2005. What About People in Geographic Information Science. In P. Fisher & D. J. Unwin, eds. *Re-presenting GIS*. John Wiley & Sons., pp. 215–242.

Mitas, L. & Mitasova, H., 1999. Spatial interpolation. *Tutorial*, pp.481–492.

Musen, M.A., 2015. The Protege Project: A look back and a look forward. *AI Matters. Association of Computing Machinery Specific Interest Group in Artificial Intelligence*, 1(4).

National Land and Property Gazetteer, 2007. BS7666 Explained, National Land and Property Gazetteer Fact Sheet. , (February).

National Land and Property Gazetteer, 2014. The National Land and Property Gazetteer. Available at: http://www.nlpg.org.uk/nlpg/link.htm?nwid=19 [Accessed April 1, 2014].

Nomis, 2017a. QS103EW - Age by single year. , (April 2015), pp.2015–2018. Available at: http://www.nomisweb.co.uk/census/2011/qs103ew.pdf.

Nomis, 2017b. QS401EW - Accommodation type. , (April 2015), pp.2010–2011. Available at:

http://www.nomisweb.co.uk/census/2011/qs401ew.pdf.

Nomis, 2016. WP605EW - Industry (Workplace population). , (October). Available at: http://www.nomisweb.co.uk/census/2011/wp605ew.pdf.

Nordbeck, S. & Rystedt, B., 1970. Isarithmic maps and the continuity of reference interval functions. *Geografiska Annaler. Series B, Human Geography*, 52(2), pp.92–123.

O'Sullivan, D. & Unwin, D.J., 2003. *Geographic information analysis*, John Wiley & Sons.

OED, 2016. Geometry. *Oxford English Dictionary*. Available at: http://www.oed.com/view/Entry/77794?redirectedFrom=geometry#eid [Accessed April 3, 2016].

Ofcom, 2012. *4G Coverage Obligation Notice of Compliance Verification Methodology: LTE*, Available at: http://stakeholders.ofcom.org.uk/binaries/consultations/award-800mhz/statement/4GCov-verification.pdf.

ONS, 2014. 2011 Census: Workplace Population Analysis. Available at: http://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/articles/workplacepopulationanalysis/2014-05-23 [Accessed March 3, 2016].

ONS, 2012a. 2011 Census Address 2011 Census Address Register Evaluation Report. , (October), pp.1–42.

ONS, 2015. 2011 census analysis: What does the 2011 census tell us about people living in communal establishments? *Office for National Statistics*, pp.1–32.

ONS, 2012b. A Beginner's Guide to UK Geography. Available at: http://webarchive.nationalarchives.gov.uk/20160105160709/http://www.ons.gov.uk/ons/guide-method/geography/beginner-s-guide/index.html [Accessed January 16, 2014].

ONS, 2018a. Nomis. Available at: https://www.nomisweb.co.uk/ [Accessed August 23, 2018].

ONS, 2018b. ONS UPRN Directory (February 2018). Available at: http://geoportal.statistics.gov.uk/datasets/ce09395b1b674300a12b885b64692c7a [Accessed August 20, 2018].

ONS, 2017. Small area population estimates QMI. , (October), pp.1–14. Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/methodologies/smallareapopulationestimatesqmi.

ONS, 2007. *UK Standard Industrial Classification of Economic Activities 2007*, Available at: file:///H:/Documents/Research/all files/UK Standard Industrial Classification of Economic Activities 2003.pdf.

Ontologydesignpatterns.org, 2010. What is a pattern. Available at: http://ontologydesignpatterns.org/wiki/Odp:WhatIsAPattern [Accessed July 25, 2017].

OntoText, 2016. GraphDB Free Documentation.

Openshaw, S., 1984. The modifiable area unit problem. *Concepts and Techniques in Modern Geography*, 38, pp.1–41.

Ordnance Survey, 2013a. AddressBase Premium - CSV Technical specification V1.3. , p.44.

Ordnance Survey, 2013b. AddressBase products user guide. , pp.1–23. Available at: https://www.ordnancesurvey.co.uk/docs/user-guides/addressbase-products-user-guide.pdf.

Ordnance Survey, 2015a. Building Height Attribute Coverage.

Ordnance Survey, 2018. Open OS MasterMap. Available at: https://www.ordnancesurvey.co.uk/business-and-government/products/open-mastermap.html.

Ordnance Survey, 2011. OS MasterMap Address Layer and Address Layer 2 User guide Contents V1.4. , p.62.

Ordnance Survey, 2013c. OS MasterMap Sites Layer Technical Specification V1.2. , p.28.

Ordnance Survey, 2013d. OS MasterMap Sites Layer User guide V1.0. , p.23.

Ordnance Survey, 2014. OS MasterMap Topography Layer: User guide and technical specification V1.12. *Ordnance Survey*, p.150.

Ordnance Survey, 2015b. Sites – now part of OS MasterMap Topography Layer. Available at: http://www.ordnancesurvey.co.uk/business-and-government/products/sites-layer.html [Accessed April 5, 2016].

Orford, S., 2010. Towards a data-rich infrastructure for housing-market research: Deriving floor-area estimates for individual properties from secondary data sources. *Environment and Planning B: Planning and Design*, 37(2), pp.248–264.

Orford, S. & Radcliffe, J., 2007. Modelling UK residential dwelling types using OS Mastermap data: A comparison to the 2001 census. *Computers, Environment and Urban Systems*, 31(2),

pp.206–227.

Patterson, L. et al., 2007. Assessing spatial and attribute errors in large national datasets for population distribution models : a case study of Philadelphia county schools. , pp.93–102.

Paull, D., 2003. A Geocoded National Address File for Australia : The G-NAF What, Why, Who and When. , pp.1–16.

Penn State College of Earth and Mineral Sciences, 2014. Nature of Geographic Information, 3. MAF/TIGER. Available at: w.e-education.psu.edu/natureofgeoinfo/c4_p3.html [Accessed April 1, 2014].

Perry, M. & Herring, J., 2011. GeoSPARQL - A geographic query language for RDF data. , pp.1–17. Available at: http://www.opengeospatial.org/.

Pesaresi, M. et al., 2013. A global human settlement layer from optical HR/VHR RS data: Concept and first results. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 6(5), pp.2102–2131.

Pesaresi, M., 2015. Global Human Settlement Layer, 1st Urbanization Workshop. Available at: https://ec.europa.eu/jrc/sites/default/files/pesaresi_urbanisation_workshop_28may2015.pdf.

Petrov, A., 2012. One Hundred Years of Dasymetric Mapping: Back to the Origin. *Cartographic Journal*, 49(3), pp.256–264.

Peuquet, D., Smith, B. & Brogaard, B., 1998. The ontology of fields. *Report of a Specialist Meeting held under the auspices of the Varenius Project*, pp.1–42.

Peuquet, D.J. & Duan, N., 1995. An event-based spatiotemporal data model (ESTDM) for temporal analysis of geographical data. *International journal of geographical information systems*, 9(1), pp.7–24.

Powers, S., 2003. *Practical RDF*, O'Reilly.

Pred, A., 1977. The Choreography of Existence: Comments on Hägerstrand's Time-Geography and Its Usefulness. *Economic Geography*, 53(2), pp.207–221.

PSMA Australia Limited, 2014. G-NAF Product Description. , (February), p.70. Available at: http://www.psma.com.au/?product=g-naf.

Pultar, E. et al., 2010. EDGIS: a dynamic GIS based on space time points. *International Journal of*

Geographical Information Science, 24(3), pp.329–346.

Raper, J.., Rhind, D.. & Shepherd, J.., 1992. *Postcodes: The New Geography*, Harlow : Longman.

Raper, J., 2005. Spatio-temporal ontology for digital geographies. In P. Fisher & D. J. Unwin, eds. *Re-presenting GIS*. John Wiley & Sons.

Renner, K. et al., 2018. Spatio-temporal population modelling as improved exposure information for risk assessments tested in the Autonomous Province of Bolzano. *International Journal of Disaster Risk Reduction*, 27(May), pp.470–479.

Renz, J., 2001. *Qualitative Spatial Reasoning with Topological Information*, Springer Science & Business Media.

Richardson, D.B. et al., 2015. Replication of scientific research: addressing geoprivacy, confidentiality, and data sharing challenges in geospatial research. *Annals of GIS*, 21(2), pp.101–110.

Royal Mail, 2014. Postcode Address File End-User Homepage. Available at: http://www.poweredbypaf.com/end-user/products/data-products/paf-raw-data/ [Accessed March 30, 2014].

Schmitt, R.C., 1956. Estimating Daytime Populations. *Journal of the American Institute of Planners*, 22(2), pp.83–85.

Shaw, S.-L. & Wang, D., 2000. Handling Disaggregate Travel Data in GIS. *GeoInformatica*, 4(2), pp.161–178.

Shaw, S., Bombom, L.S. & Yu, H., 2008. A Space-Time GIS Approach to Exploring Large Individual-based Spatiotemporal Datasets. *Transactions in GIS*, 12(4), pp.425–441.

Shuttleworth, I. & Martin, D., 2016. People and places: Understanding geographical accuracy in administrative data from the census and healthcare systems. *Environment and Planning A*, 48(3), pp.594–610.

Skouby, K.E. et al., 2014. Smart Cities and the Aging Population. *OUTLOOK: Visions and research directions for the Wireless World*, 12, pp.1–12.

Sleeter, R. & Wood, N., 2006. Estimating daytime and nighttime population density for coastal communites in oregon. *44th Urban and Regional Information Systems Association Annual Conference, British Columbia.*, pp.1–15.

Smith, A. et al., 2015. Assessing the impact of seasonal population fluctuation on regional flood risk management. *ISPRS International Journal of Geo-Information*, 4(3), pp.1118–1141.

Smith, A., Martin, D. & Cockings, S., 2012. 24 / 7 Population modelling for enhanced assessment of exposure to natural hazards.

Smith, A., Martin, D. & Cockings, S., 2016. Spatio-Temporal Population Modelling for Enhanced Assessment of Urban Exposure to Flood Risk. *Applied Spatial Analysis and Policy*, 9(2), pp.145–163.

Smith, B., 1996. Mereotopology: A theory of parts and boundaries. *Data & Knowledge Engineering*, 20(3), pp.287–303.

Smith, B. & Mark, D.M., 2003. Do mountains exist? Towards an ontology of landforms. *Environment and Planning B: Planning and Design*, 30(3), pp.411–427.

Smith, B. & Mark, D.M., 1998. Ontology and Geographic Kinds. In *International Symposium on Spatial Data Handling*. Vancouver, Canada.

Smith, B. & Munn, K., 2008. *Applied Ontology*,

Smith, G. & Fairburn, J., 2008. RR678 Research report: Updating and improving the National Population Database to National Population Database 2.

Southampton City Council, 2015. Southampton Statistics. Available at: https://www.southampton.gov.uk/council-democracy/council-data/statistics/ [Accessed December 28, 2015].

Stevens, F.R. et al., 2015. Disaggregating Census Data for Population Mapping Using Random Forests with Remotely-Sensed and Ancillary Data. *PLoS ONE*, pp.1–22.

Sutton, P. et al., 2001. Census from Heaven: An estimate of the global human population using night-time satellite imagery. *International Journal of Remote Sensing*, 22(16), pp.3061–3076.

Sutton, P., Elvidge, C. & Obremski, T., 2003. Building and Evaluating Models to Estimate Ambient Population Denstiy. *Photogrammetric Engineering and Remote Sensing*, 69(5), pp.545–553.

Tenerelli, P., Gallego, J.F. & Ehrlich, D., 2015. Population density modelling in support of disaster risk assessment. *International Journal of Disaster Risk Reduction*, 13, pp.334–341.

The National Archives, 2004. *Housing Act 2004, Chapter 34, Section 254-259*, Available at: Legislation.gov.uk.

Thrift, N., 1977. An introduction to time geography. *Concepts and Techniques in Modern Geography*, 13, pp.1–37.

Tobler, W.R., 1970. A Computer Movie Simulation Urban Growth in Detroit Region. *Economic Geography*, 46, pp.234–240.

Tobler, W.R., 1979. Smooth pycnophylactic interpolation for geographical regions. *Journal of the American Statistical Association*, 74(367), pp.519–30.

U.S. Census Bureau, 2013. TIGER / Line Shapefiles Technical Documentation. *Geography*.

U.S. Census Bureau Fact Finder, U.S. Census Bureau. MAF/TIGER Database. Available at: http://factfinder2.census.gov/help/en/glossary/m/maf_tiger_database.htm [Accessed July 9, 2014].

Ulbrich, U. et al., 2003. The central European floods of August 2002 : Part 1 – Rainfall periods and flood development. *Weather*, 58(August 2002), pp.371–377.

United Nations Statistics Division, 2014. Cenusus Clock. Available at: http://unstats.un.org/unsd/demographic/sources/census/2010_PHC/default.htm [Accessed March 4, 2016].

Varzi, A., 2016. Mereology. *The Stanford Encyclopedia of Philosophy*. Available at: http://plato.stanford.edu/archives/spr2016/entries/mereology [Accessed April 3, 2016].

W3C, 2015. Linked Data. Available at: https://www.w3.org/standards/semanticweb/data [Accessed May 22, 2018].

W3C, 2014a. Query Operators Semantics Temporal Operators. Available at: https://www.w3.org/community/rsp/wiki/Query_Operators_Semantics [Accessed March 23, 2016].

W3C, 2014b. Resource Description Framework. *W3C Semantic Web*. Available at: https://www.w3.org/RDF/ [Accessed May 21, 2018].

W3C, 2008. SPARQL Query Language for RDF. Available at: https://www.w3.org/TR/rdf-sparql-query/ [Accessed May 22, 2018].

W3C, 2004. W3C Working Draft. Available at: https://www.w3.org/2001/sw/BestPractices/OEP/SimplePartWhole/ [Accessed July 27, 2017].

W3C, 2003. What is HyperText. *What is HyperText*. Available at: https://www.w3.org/WhatIs.html [Accessed June 2, 2017].

W3C OWL Working Group, 2012a. OWL 2 Web Ontology Language Document Overview (Second Edition). Available at: https://www.w3.org/TR/2012/REC-owl2-overview-20121211/ [Accessed May 1, 2018].

W3C OWL Working Group, 2012b. OWL2 Syntax. Available at: https://www.w3.org/TR/owl2-syntax/#Global_Restrictions_on_Axioms_in_OWL_2_DL [Accessed August 19, 2018].

Wang, D. & Cheng, T., 2001. A spatio-temporal data model for activity-based transport demand modelling. *International Journal of Geographical Information Science*, 15(6), pp.561–585.

Worboys, M.F., 1994. A unified model for spatial and temporal information. *The Computer Journal*, 37(1).

Worboys, M.F. & Duckham, M., 2004. *GIS: A Computing Perspective* Second., CRC Press.

Wright, J.K., 1936. A method of mapping densities of population: With Cape Cod as an example. *Geographical Review*, 26(1), pp.103–110.

Wu, Y., Blunden, L.S. & Bahaj, A.S., 2018. City-wide building height determination using light detection and ranging data. *Environment and Planning B: Urban Analytics and City Science*, 0(0), pp.1–15.

Xie, Y., 1995. The overlaid network algorithms for areal interpolation problem. *Computers, Environment and Urban Systems*, 19(4), pp.287–306.

Yu, H. & Shaw, S., 2008. Exploring potential human activities in physical and virtual spaces: a spatio-temporal GIS approach. *International Journal of Geographical Information …*, 22(4), pp.409–430.

Yuan, M., 2015. Frontiers of GIScience : Evolution , State-of-Art , and Future Pathways. In P. Thenkabail, ed. *Remote Sensing Handbook*. CRC Press.

Yuan, M., 1994. Wildfire conceptual modeling for building GIS space-time models. In *GIS/LIS*. pp. 860–869.

Zhang, C. & Qiu, F., 2011. A Point-Based Intelligent Approach to Areal Interpolation. *The Professional Geographer*, 63(2), pp.262–276.

Zhang, Z., Sunila, R. & Virrantaus, K., 2010. A spatio-temporal population model for alarming,

situational picture and warning system. *Remote Sensing and Spatial Information Sciences*, 38, pp.69–74.