Accepted Manuscript

Spatial urban data system: A cloud-enabled big data infrastructure for social and economic urban analytics

Obinna C.D. Anejionu, Piyushimita (Vonu) Thakuriah, Andrew McHugh, Yeran Sun, David McArthur, Phil Mason, Rod Walpole

PII: S0167-739X(18)31904-6
DOI: https://doi.org/10.1016/j.future.2019.03.052
Reference: FUTURE 4877

To appear in: *Future Generation Computer Systems*

Received date : 13 August 2018
Revised date : 6 March 2019
Accepted date : 27 March 2019

Please cite this article as: O.C.D. Anejionu, P.(V. Thakuriah, A. McHugh et al., Spatial urban data system: A cloud-enabled big data infrastructure for social and economic urban analytics, *Future Generation Computer Systems* (2019), https://doi.org/10.1016/j.future.2019.03.052

Spatial Urban Data System: A Cloud-enabled Big Data Infrastructure for Social and Economic Urban Analytics

Obinna C.D. Anejionu*[1], Piyushimita (Vonu) Thakuriah, Andrew McHugh, Yuan Sun, David McArthur, Phil Mason, Rod Walpole

Urban Big Data Centre, 7 Lilybank Gardens, University of Glasgow, Glasgow, UK G12 8RZ

Abstract

The Spatial Urban Data System (SUDS) is a spatial big data infrastructure to support UK-wide analytics of the social and economic aspects of cities and city regions. It utilises data generated from traditional as well as new and emerging sources of urban data. The SUDS deploys geospatial technology, synthetic small area urban metrics, and cloud computing to enable urban analytics, and geovisualization with the goal of deriving actionable knowledge for better urban management and data-driven urban decision making. At the core of the system is a programme of urban indicators generated by using novel forms of data and urban modelling and simulation programme. SUDS differs from other similar systems by its emphasis on the generation and use of regularly updated spatially activated urban area metrics from real or near-real time data sources, to enhance understanding of intra-city interactions and dynamics. By deploying public transport, labour market accessibility and housing advertisement data in the system, we were able to identify spatial variations of key urban services at intra-city levels as well as social and economically-marginalised output areas in major cities across the UK. This paper discusses the design and implementation of SUDS, the challenges and limitations encountered, and considerations made during its development. The innovative approach adopted in the design of SUDS will enable it to support research and analysis of urban areas, policy and city administration, business decision-making, private sector innovation, and public engagement.

[1] Corresponding author: Obinna C.D. Anejionu. obinna.anejionu@glasgow.ac.uk

Having been tested with housing, transport and employment metrics, efforts are ongoing to integrate information from other sources such as IoT, and User Generated Content into the system to enable urban predictive analytics.

Keywords: Urban Big Data Infrastructure; Urban Analytics; Spatial Urban Indicators; Small Area Assessment; Spatial Big Data

1. Introduction

Cities play critical role in society and have increasingly become the focal points for the economy, with the current trend towards increasingly knowledge-intensive economies (European Commission, 2013). At the same time, increasing population concentration in urban areas put pressure on the use of limited city resources and services such as energy, transportation, water, buildings and public spaces (European Commission, 2013b). Cities also account for over 70% of current global CO_2 emissions (OECD, 2012), posing serious challenges arising from environmental pollution, congestion, waste management, and the need for urban sustainability. As a result, cities have been recognised as one of the key elements for future decision-making (Albino et al., 2016; Mori and Christodoulou, 2012).

The transformation of urban areas to smart cities has resulted in the continuous generation of enormous volumes and varieties of data from different sources. Thakuriah et al (2017) noted that the sources of urban data are many, including sensor systems monitoring different aspects of the city, user-generated content such as social media, private business data collected from transaction and customer usage records, as well as traditional sources such as those held by government agencies (registrations, statistics, and archives) and non-government actors (e.g., housing sale and rental data from property agents, and energy usage from energy companies). Together, these have given rise to the urban big data phenomenon. However, for a city to be efficiently managed, data from these disparate sources need to be efficiently integrated, in order

to enable a holistic understanding of the interactions between various city subsystems. Based on the fact that most of the data obtained from cities are spatially-referenced, the interactions between the various city components will be better understood through the deployment of geospatial techniques. In the past, the integration and analysis of large volumes of data presented an enormous task, but with advances in big data analytics, cloud computing and geospatial technology, intra-city interactions can now be monitored and assessed in real time or near-real time, feeding into Urban Informatics, or the utilisation of novel sources of urban data for knowledge discovery, public engagement and business innovations.

In this paper, we describe the Spatial Urban Data System (SUDS), a multi-component system that serves data on multiple social and economic aspects of urban living. SUDS captures key economic and social data of interest and integrates such measurements to generate small-area data in a timely fashion. This approach helps derive new insights that are useful for smart city management. Key capabilities of the system include: automatic acquisition and processing of data from heterogeneous sources, generation of relevant science-based small-area synthetic metrics from acquired data that could potentially be used to generate intra-city indicators for monitoring and assessing the performance of relevant urban area aspects (subsystems); cloud computing infrastructure for the storage, integration and manipulation of urban big data from different sources; robust tools to support spatial urban big data analytics, policy and business decisions tools, public engagement; scenario/predictive modelling and analytics based on generated intra-city metrics, and visualisation tools that will support understanding of the spatial dimensions of the sub-city interactions.

The novelty of this research is fourfold. Firstly, the use of non-traditional sources of data for the generation of synthetic metrics enables the tracking of urban dynamics across an entire country on a regular basis. Secondly, the spatial disaggregation of the metrics (small-area) allows unprecedented insights into sub-city interactions of the various aspects of the urban

area, with an emphasis on assessments of status, needs, disparities and well-being. Thirdly, the spatial big data system developed allows the integration and processing of data from varying sources, with complex geospatial processing, and modern cloud computing systems capable of handling big data. Fourthly, the research developed series of strategies to process and utilised various socioeconomic variables, to understand and manage urban area dynamics.

Section 2 provides an overview of this project – its purpose, significance and contributions. Section 3 reviews related concepts and works that have been undertaken on smart city performance and urban informatics and similar systems that have been proposed to support smart city implementation and management. Section 4 provides a discussion on the design and development of the SUDS, while Section 5 explores some ongoing application of the SUDS. Section 6 discusses certain limitations, constraints and issues encountered and a conclusion is presented in Section 7.

2. Purpose, Significance and Contributions of the Research

SUDS infrastructure is part of the Urban Big Data Centre (UBDC), funded by the Big Data Phase 2 of the UK Research and Innovation's Economic and Social Research Council. The UBDC is a nationwide data service that provides access to urban data to academic researchers, local governments and businesses. The uniqueness of the data service lies in its data collections sourced from a variety of public, private and internet sources including: Zoopla, Experian, Registers of Scotland, Survey, BGS, Met office, Springboard, Twitter, and Facebook; which are used create a big data infrastructure to study dynamic resource management, transport, housing, economic development, migration, lifelong learning, productivity and other social and economic aspects of urban living. The SUDS integrates geospatial data from multiple subprojects to these urban living themes and serves as a capstone project that links these projects to the spatial data infrastructure (SDI).

The key objectives of SUDS are:

Research, knowledge discovery and evaluation: The first and foremost objective of SUDS is to bring together, in one platform, geospatial data on a number of urban living themes, with the ambition of facilitating research and knowledge discovery of social and economic conditions, as well as cross-theme analysis (eg, between economic and health factors, social and environmental factors). By building a platform for the entire UK, SUDS provides the ability to understand regional variations in social and economic factors, and to conduct detailed analysis of how these factors affect poverty, regional deprivation, productivity and other issues of relevance to quality of life and sustainable urban living. Through specially-constructed urban indicators (more details in Section 4.2), we enable research to utilise comprehensive information from multiple sources that utilise novel sources of data, which puts together into composite measures, a number of social and economic variables.

Policy implementation, evaluation and urban operations and service delivery: A second ambition is to support urban policy implementation and evaluation. For instance, aiding in the identification of areas that need attention, improving infrastructure to access jobs, or for better rental housing conditions. Where should policy action be taken and investments made to promote educational outcomes, and for better connection between graduates and local labour markets? Furthermore, national and regional policies often have local effects. For example, cuts in local government funding have critically affected public bus services across England and Wales, especially in deprived areas, thereby limiting peoples access to jobs and education (Topham, 2018). At the same time, decision-makers from specific areas may wish to understand how policies implemented in their areas led to outcomes at the local level, compared to other areas in where such policies were not implemented.

An ambition of SUDS is to provide a framework for longitudinal, over-time content that allows tracking of key measures, changes to which can lead to a determination of the effect of policies and plans. This necessarily implies that data are captured and archived over long periods, under a stable governance model, for which a persistent research platform is needed to ensure research continuity and to deliver persistent services. This in fact is a major motivation of SUDS — to facilitate improved temporal analysis, through the creation of longitudinal synthetic data, by tapping into historical data or by archiving real-time data feeds over time. Such synthetic temporal data will, for instance, enable social scientists to study the dynamics of patterns of interest and link them to changing behaviours and outcomes. They will also help analysts monitor risks to urban areas and the resilience of urban areas to policy and natural interventions (e.g., changes in economic or welfare policy, episodes of extreme weather). Additionally, local administrators are increasingly interested in how to operate improved city services using data-driven practices. SUDS provides a data-driven framework with which to monitor how services could be improved, and offers mechanisms to bring in new types of data that are relevant to the operational problems at hand.

Urban Indicators: A central aspect of SUDS is the utilisation of novel forms of data to generate small-area urban indicators. We discuss this aspect in greater detail in Section 4. The goal of such indicators is typically to facilitate performance monitoring, assess trends over time, set future targets and support inter-city comparisons. They also inform urban planning, operations and a variety of decision-making regarding urban management, raise awareness on critical issues, encourage political interventions and citizen activism, support strategies for health behaviours and well-being, promote public engagement and civic participation, and improve communication among stakeholders working in urban sectoral siloes. However, city-level indicators can mask important variations in performance and well-being within specific neighbourhoods and local areas within a city. This is a critical gap since such indicators can

provide essential information for local community-level action in poorly performing parts of the city. Our focus in SUDS is entirely on creating small-area synthetic data on key policy-relevant factors by drawing on multiple sources of information to enable appropriate place-based decision-making.

Open source development: A key objective of the SUDS platform is to use open source technology as a backbone so that the platform can be replicated elsewhere. The general benefits of open source SDI and extensive growth of open source geospatial technologies have been extensively noted elsewhere (e.g., Hu et al. 2017; Brovelli et al., 2016; Steiniger and Bocher, 2008) and will not be repeated here. Here, our objective is to demonstrate, through the selection of technology components and the configurations employed, how novel forms of urban big data can be offered for use through an open geospatial platform, or replicated by local governments, smart cities SMEs, SDI in less-developed nations, or even how they can form the basis for SDI with other themes as a focus (e.g., health, the environment). However, we also note that with new forms of data, many of which are privately held or are confidential administrative records, not all data services can be open, and there is a need for SDI to be able to support delivery of confidential and private-sector business data. The SUDS platform brings together processes offering security and access control technologies that ensure that data can be accessed and that analytics can be conducted in the safeguarded environment that is obligatory for the processing of such private data.

Larger infrastructure and data acquisition: SUDS is part of a larger data infrastructure (the UBDC), which grows organically with new users, data and technology, and with new government or business initiatives. These characteristics result in SUDS being not a well-defined system (Vandenbroucke, et al., 2013), but rather a "complex, multi-faceted and dynamic environment" that is responsive to new forms of data and stakeholders that enter into the work processes. SUDS benefits from processes in place within the wider infrastructure to

proactively engage with private and government data owners, towards supporting UK industrial strategy. A part of this engagement process leads to new data acquisition from stakeholders. More broadly, the system will play a central role in our stakeholder engagement activities, particularly with policy-makers, businesses and non-profit organisations.

3. Related Works

In this section, we review two strands of literature pertinent to our work – performance monitoring and assessment in small cities, and data systems and infrastructure to support smart city analytics.

3.1 Smart city performance monitoring and assessment

Due to the increasing importance of cities to society, and the need to create a sustainable urban environment, there is a growing interest in robust and efficient methods of monitoring and measuring policy impacts, infrastructure developments, socio-economic factors, resource use, environmental pollution and other processes that contribute to and benefit from the city's metabolism, prosperity and quality of life (European Commission, 2015). Hence, urban metrics/indicators are increasingly important in smart city performance monitoring and assessment, trend assessment over time, and future target-setting (Albino et al, 2016; Airaksinen, 2016; Berardi, 2013). Although a wide range of available indicators (Huovila, 2016; Albino et al., 2016; European Commission, 2015) is being used to monitor smart city performance, most of the indicators are calculated at the national, regional, or city levels. This is because the goals of such indicators are mainly to facilitate performance monitoring, assess trends over time, set future targets and support inter-city comparisons. However, they can mask important intra-city variations (in performance and well-being within specific neighbourhoods and local areas within a city). Furthermore, the indicators are not regularly updated as most tend to be produced from data acquired during censuses. Hence, a major strength of SUDS is

the capability of creating and using small-area synthetic metrics of key policy-relevant factors, based on the data obtained from the various aspects of the city to facilitate small-area analyses that will shed light on underlying city dynamics and inform local and community-level action for poorly performing parts of a city.

Indicators for smart city performance monitoring are classified in different ways (Airaksinen, 2016; European Commission, 2015). The Canadian International Development Agency, (2012) identified three broad categories of indicator: social, economic and environmental (Figure 1). The SUDS programme focuses mainly on social and economic indicators, with less emphasis on environmental aspects, which have received considerable attention from researchers (Shen et al, 2011; Lynch, et al 2011).

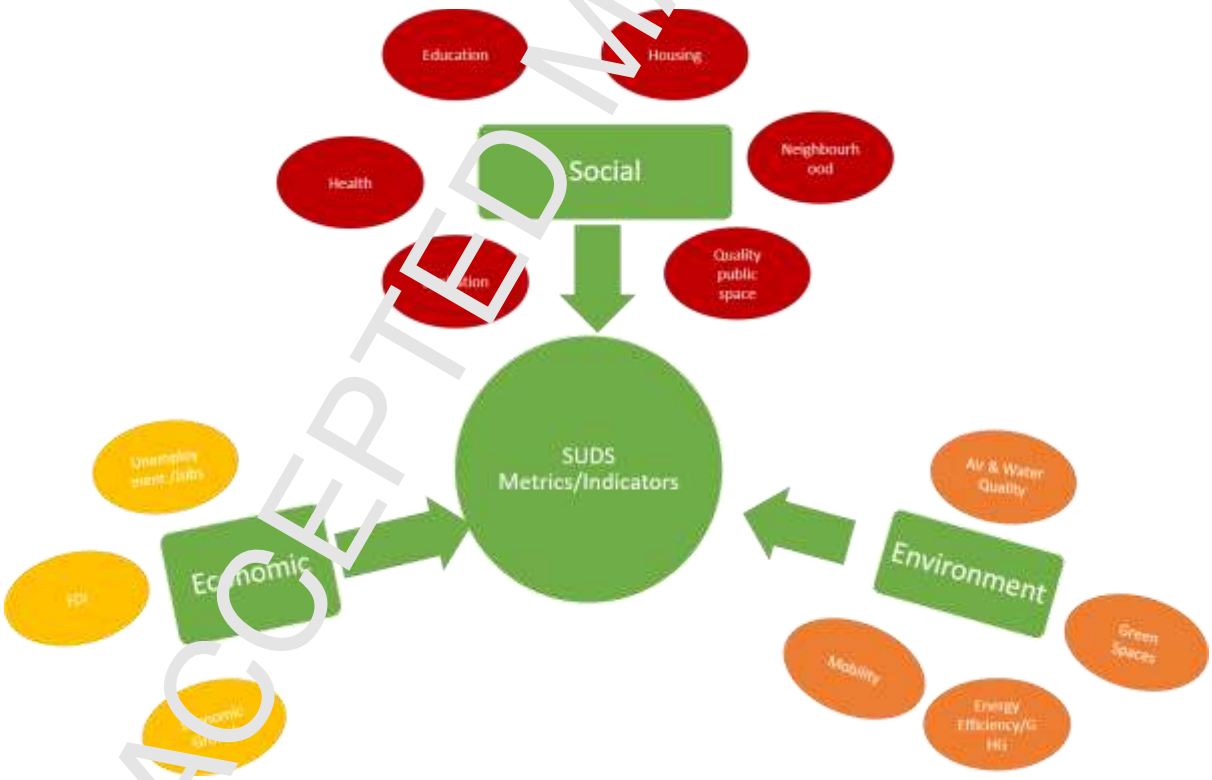


Figure 1. Urban area subsystems and key urban area indicators targeted by SUDS.

3.2 Smart city data infrastructure

In the last two decades, the concept of smart cities has generated great interest within and beyond the research community. With the advent of big data and supporting technologies, urban area and smart city-related studies are becoming prevalent. Different aspects of the smart city are being studied and relevant theoretical and practical steps explored. As a result, a number of authors have proposed various ways smart city could be implemented. However, the use of granular spatially referenced small-area metrics to drive urban area or smart city analytics is still at the nascent stage and the data have not been extensively explored. This gap is among the things compelling the development of the SUDS

To some extent, SUDS could be perceived as a spatial data infrastructure (SDI) for urban area analytics. SDI has been defined by Hu et al. (2017) as the technology, policies, standards, and human resources necessary to acquire, process, store, distribute, and improve utilisation of geospatial data, services, and other digital resources. This definition is in line with the aim of SDIs as noted by several authors (Grus et al., 2011; Crompvoets et al., 2008), which essentially is to facilitate the exchange and sharing of spatial data between stakeholders in the spatial data community. However, SUDS differs from conventional SDIs by not fundamentally focusing on the storage and dissemination of geospatial data, but rather focusing on the combination of spatial and non-spatial data to generate metrics with the aim of providing new insights. In this sense, even though SUDS has storage capability, it mainly serves as an analytics platform that draws data from multiple sources. Hence, SUDS combines the storage capabilities of SDIs with the data processing and analytics capabilities of conventional smart city data infrastructures.

There have been a number of studies on smart city-related infrastructure, most of which have focused on the deployment of internet of things (IoT) to facilitate smart city implementations. Some authors have developed smart city platforms essentially to collect data from sensors without focusing so much on the analysis of the collected data (Bain, 2014, Murty et al., 2008). Zanella et al. (2014) proposed a general reference framework for the design of an urban IoT

that will be based on a centralised architecture through which a set of web services can be exposed. The proposed system was tested with a proof of concept (Padova Smart City) project, which comprises a system for the collection of environmental data (CO level, air temperature and humidity, vibrations, noise, etc.) and monitoring of public street lighting (light intensity) via wireless nodes.

Very recently, Lv et al. (2018), deployed 3D GIS and cloud computing to develop a government affairs service platform for facilitating and handling smart city planning. Soille et al. (2018) proposed a data-intensive computing platform for retrieving information from big geospatial data from earth observation satellites. The platform will facilitate the storage, processing, analysis, and visualization of the satellite images, essentially for applications in agriculture, forestry, environment, disaster risk management, development, health, and energy. For their part, Cicirelli et al. (2017) proposed the iSapiens platform for Smart City applications. This platform operates as an agent-based distributed IoT platform where the bulk of the computations are executed at the edge (instead of within the data core) of the network of computing nodes spread over a city area by agents residing in each node, while all the others, such as computationally demanding tasks, are executed in the cloud. Other previous works, such as the SmartSantander project, have focused on the development of smart city infrastructure with extensive networks for the monitoring of environment pollution (air quality, noise and luminosity levels), outdoor parking, and automated irrigation systems (Sanchez et al., 2014).

Khan et al. (2015; 2013) proposed the development of a cloud-based analysis service that could be used to generate information intelligence and support decision-making for smart future cities management. This system is similar to SUDS, other than in terms of its lesser concern for spatial aspects. Similarly, Babar and Arif (2017) proposed a smart city architecture, based on big data analytics that will comprise a data acquisition and aggregation module (which will

collect varied and diverse data related to city services), a data computation and processing module (which will perform normalization, filtration, processing and data analysis), and an application and decision module (which will formulate decisions and initiate events) to support solutions for smart urban planning and decision making. This system is similar to the SUDS in many respects in the sense that it incorporates data acquisition, processing and analysis components, and is based on big data analytics. However, whereas its central aim is to improve the data processing efficacy to facilitate real-time decision-making, SUDS' main focus is on the rapid or frequent generation of synthetic small-area metrics from a variety of data sources over the long term, and on integrating these metrics to derive new urban area insights and knowledge.

Other studies, such as the IES Cities project, focus on exploiting a combination of open Government data, network sensors and user-supplied data to develop user-centric mobile services constructed around the IoT as a means of supporting smart city applications (Aguilera et al., 2017). Gaur et al. (2015) proposed a Multi-Level Smart City architecture based on semantic web technologies and Dempster-Shafer uncertainty theory to support smart city applications by facilitating the interaction between wireless sensor networks and ICT.

SUDS differs from already existing spatially enabled smart city analytics infrastructure, such as those proposed by Lv et al. (2018) and Khan et al. (2015) by focusing largely on the generation and use of small-area socioeconomic metrics on a countrywide basis collected at regular intervals. Previous indicators and metrics used in studying urban area dynamics are at a higher spatial scale such as regional or national levels. Those that are at smaller scales are limited in extent as they focused on specific areas. However, the small-area metrics generated in this project are at smaller scale (higher spatial detail), higher temporality and covers an entire country. Hence, comparisons can be made at various spatial levels from neighbourhoods, through city-, regional- and national-levels. This facilitates the understanding of intra-city

dynamics and provides "urban health checks" with an emphasis on assessments of status, needs, disparities and well-being. Potential information from IoT sensors forms only part of the data sources for computing urban area metrics, unlike in other systems (Cicirelli et al., 2017; Sanchez et al., 2014), in which IoT forms the core of the infrastructure. The SUDS is designed to be compatible to any modern cloud computing systems such as Snowflake Computing system, Azure SQL Data Warehouse, Amazon Redshift, Oracle Data Warehouse with advanced capabilities for handling big data. The development of the SUDS is informed by multiple global efforts aimed at smart city performance monitoring and comparison. However, SUDs focuses on the generation of synthetic metrics that can be deployed to understand underlying dynamics and to derive deeper insights into sub-city interactions, and which could be extended to generate relevant indicators for urban area monitoring and assessment.

With regards to security, the system was designed to ensure that critical information are protected from unauthorised access and deletion, theft, and data leakage. Modern data warehouses such as those used in SUDS are built to safeguard datasets stored in them. For instance, the Snowflake Data Warehouse uses a comprehensive set of features (IP whitelisting, multi-factor authentication, federated authentication, role-controlled access, automatic encryption of data, maintenance of historical data) that help protect data stored in it against human error, malicious acts, software or hardware failure and ensures data recoverability (Continuous Data Protection – CDP). Another consideration was the choice data centre. The European Union (EU) regulation requires cloud-hosted data to be physically stored within the continent, hence, the cloud system used has secured data centre in two locations (Dublin and Frankfurt) in the EU. This differs from that used by Khan et al. (2015), which was essentially Hadoop-based cloud infrastructure hosted on a server. However, similar security considerations as was made in SUDS were made by Soile et al., 2018, which used Kerberos

authentication and a specific access control list (ACL) mechanism to ensure multi-user environment data security.

4. The SUDS Platform Design and Development - Methods and Approach

SUDS comprises four main components: the Urban Indicator (UI) programme, geospatial processes and analytics, web visualisation (BI and geovisualization dashboards) and cloud computing (Figure 2). The system was designed to use a range of open source and commercial software and tools, including: Extraction Transformation and Loading (ETL) tools (FME and Talend), a spatial database (PostgreSQL/PostGIS), a webmap publishing tool (Geoserver), a cloud-based data warehouse (Snowflake), and business intelligence tools (Tableau/PowerBI). The system can be deployed for medium-scale analytics as currently implemented with countrywide synthetic small-area datasets, and can be scaled up to handle big real-time data when the data inflow increases (e.g., from city sensors or other IoT infrastructure).

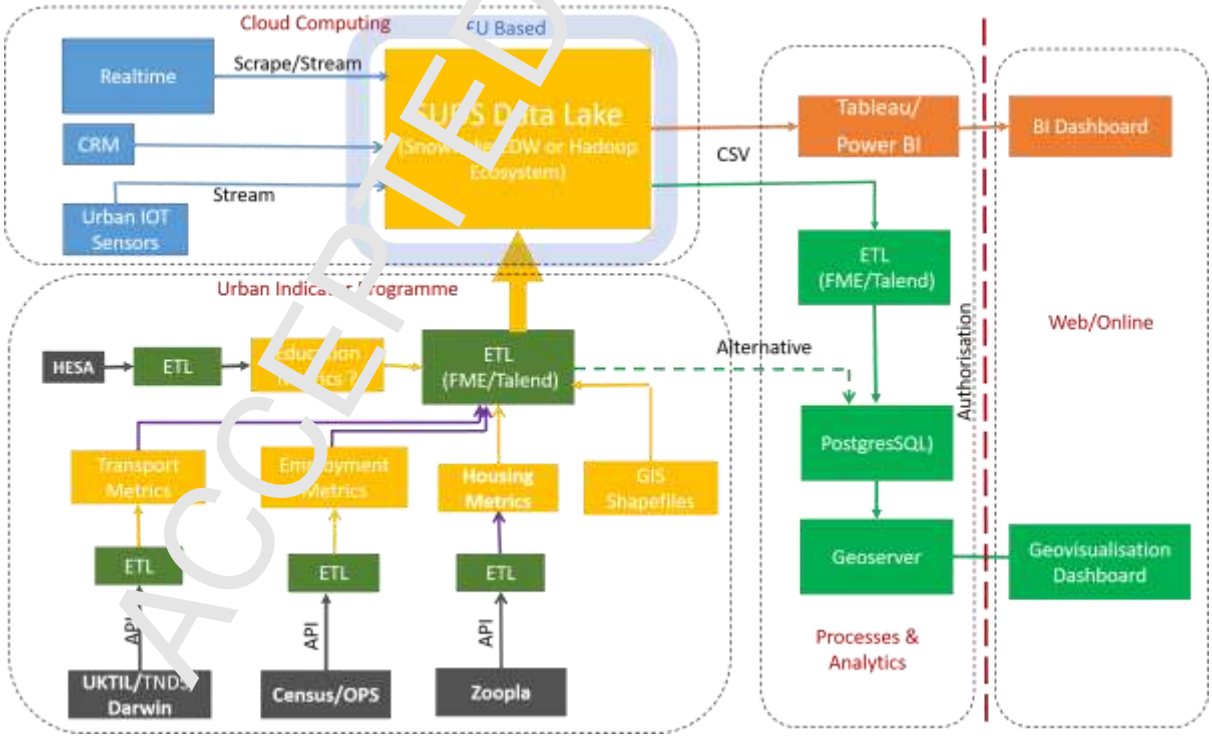


Figure 2. Graphical representation of the SUD system, showing the various components and how they connect to each other

4.1 Main features, functions and processes

The following subsections provide a brief description of the main components of the SUDS.

4.1.1 SUDS geospatial and geovisualization components

SUDS interactive geospatial processing and visualisation components were designed to be a self-service business intelligence system for insight generation and planning. They comprise the geospatial processing and analytics, and the web visualisation (BI and geovisualization dashboards) components of the system. Supported by a powerful geographic database, it has multiple sub-components including: a backend Geographic Information System/spatial processing and analytics; an online web-mapping platform that gives users the ability to have an interactive mapping experience and conduct on-the-fly spatial analytics; a business intelligence dashboard that shows insights; and other specialised tools that enable users to interact and interrogate underlying data in the database. Users can engage with the platform by querying the underlying datasets or conducting multi-metric analyses to gain better insights into multiple dimensions of the city. Figure 3 illustrates the SUDS geospatial processing and visualisation architecture.

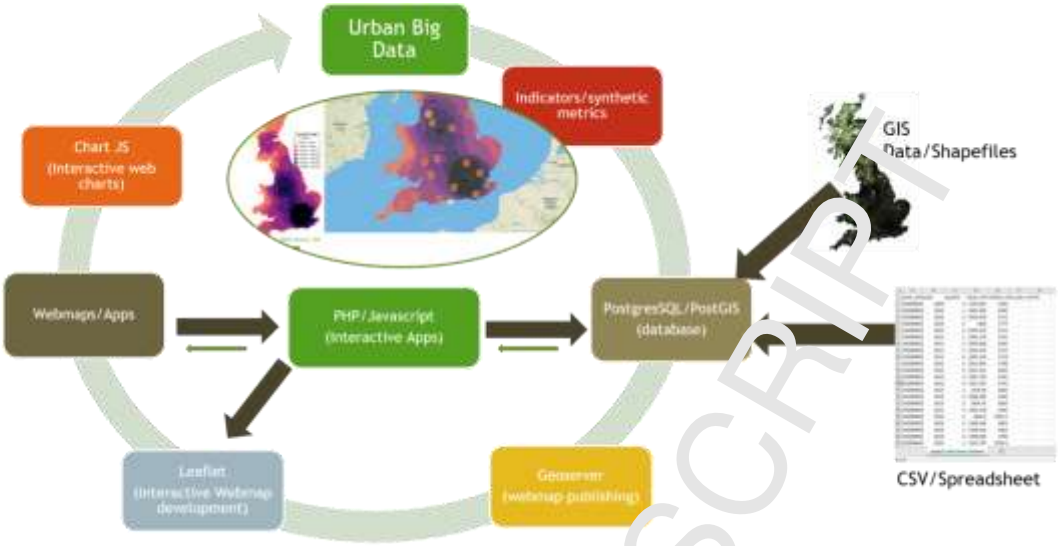


Figure 3. SUDS geospatial component architecture showing how the relevant subcomponents are integrated

The following strategies were adopted in the development of the various subcomponents.

4.1.1.1 Spatial database

PostgreSQL, an open-source object relational database management system, with a powerful spatial database extension (PostGIS), was used to create the SUDS spatial database. This was based on the fact that it is a robust object relational database with advanced spatial analysis functionalities, that can seamlessly connect to: web map publishing tools, most ETL systems such as FME, Talend etc., prominent data warehouses/lakes such as Snowflake computing, Amazon Redshift, Azure SQL Data Warehouse etc., and supports interactive online spatial analytics (dynamic spatial querying of underlying datasets). The spatial database primarily serves as an initial repository for relevant spatially referenced urban area data as well as synthetic or simulated small-area data and derived urban indicators. It is also used as geospatial processing platform considering the fact that most cloud data warehouses capable of dealing with structured and unstructured big data have limited spatial data processing capabilities. This is in line with the proposal of Shaojun et al. (2017), which suggests for a NoSQL database such as (MongoDB, Neo4J, OrientDB etc) to be used as a spatial big data warehouse and a traditional relational spatial database such as PostgreSQL and SQLite used as the application server.

However, in view of progress that have been made in developing open-source geospatial big databases such as GeoMesa (GeoMesa, 2018), GeoWave (LocationTech, 2018) and OmniSci (OminiSci Inc., 2018), we are currently testing the integration of GeoMesa or Geowave in the system.

4.1.1.2 Map publishing tool

As the synthetic metrics were spatially referenced, there was a need for them to be published as web maps. Hence, Geoserver, an open-source server for publishing online geospatial data was deployed as the SUDS web map serving tool. The SUDS spatial database was connected to the Geoserver using the appropriate tools, from which the interactive maps were published. The Geoserver component also served as a link between user web map interactions and the spatial database. Users' queries are sent to the database and results in the form of published maps are returned to them from the database through the Geoserver.

4.1.1.3 Interactive web interface

The public-facing online interface of SUDS was developed with a number of standard web development tools (HTML, PHP, JavaScript and CSS). The web tools drive the web interactive capabilities of the system. The interactive mapping components of the interface were developed with Leaflet, a leading open-source JavaScript library for user-friendly online interactive maps, PHP, and JavaScript codes.

4.1.1.4 Online spatial analytics

Web analytics tools that allow online spatial queries were implemented on the SUDS platform with a combination of JavaScript, PHP and geospatial analytics, to enable users to interact with the underlying datasets. These tools were designed to be simple and easy to use mostly for drilling down into or aggregating information from one or more aspects of the urban area using the indicators/datasets. More complex queries (Multi Criteria Analysis) through which users can integrate information from multiple indicators or sectors of the urban areas were developed to enable a wider understanding of causes and effects of particular outcomes or changes.

4.1.2 Business intelligence and visualisation tools

Chart JS, a flexible JavaScript charting library was initially used alongside PHP and JavaScript codes that queried the database to produce dynamic charts that illustrate BI insights. Through the BI dashboard users can quickly gain insights about the relationship between the spatial query results and other urban area information, such as demography, economic outlook, etc. We are currently revising and testing a new implementation of the BI dashboard with Power BI and Tableau.

4.2 The urban metrics/indicator (UmI) component

The urban metrics/indicator (UmI) is a prominent component of SUDS, whose goal is to develop a range of synthetic metrics that summarise and highlight relationships among multiple dimensions of functional urban sectors. The UmI component comprises a range of spatiotemporal-synthetic or simulated small-area metrics describing diverse aspects of the social, economic, natural, built-environment and physical infrastructure aspects of urban areas that were generated from various datasets. Data used for the UmI component were accessed through a variety of data acquisition and retrieval techniques (APIs and ETL), and processed and formatted using specialised data management methods such as Python and R. These tools together with ETL tools were used to load and wrangle (cleanse, process and transform) the data into suitable formats/standard and transforming them to the same spatial units. Positional information in the raw datasets were converted to coordinates that enabled them to be spatially linked to other spatial datasets. This spatial linkage enabled the processing of the datasets at varying spatial scales such as at intra-neighbourhood-levels (lower super output area – LSOA, and Middle Layer Super Output Areas – MSOA), county-or regional-levels. Subsequently, spatially-activated synthetic data were created from the datasets using a complex set of specialist urban models and simulations, data science and GIS methods. The Spatial ETL tool Feature Manipulation Engine (FME) was used to extract, transform and load spatially-referenced data into the data lake, while non-spatial datasets were handled with Talend

integration software, which has capabilities for data quality and preparation, data integration and management, big data manipulation, cloud storage, and master data management. The synthetic data were post-processed (when possible) in many ways to create simple summaries or composites of information through a process of indicator generation, to yield urban indicators that will help monitor performance. This spatially indexed synthetic data, generated from the UmI programme forms the core of the SUDS database. Some of the metrics covering key city subsystems currently deployed in SUDS include transport availability metrics (TAM), housing affordability metrics (HAM), employment-accessibility metrics (EAM), and education-related metrics (ERM).

4.3 Cloud computing component

The cloud computing component comprises a data warehouse (data lake) in which information from the various components is collated and processed. In addition to serving as a central storage and data processing system, a key purpose of the cloud system is to facilitate real-time information streaming from sensor network gateways and integration and processing of such data with other metrics. In this way, information from urban IoT sensors can be integrated with other urban area information to generate new insights in real time. We are currently testing the development of the data warehouse with Snowflake Computing, which is one of the most promising enterprise data warehouses for big data analytics. Snowflake was chosen because of its relatively high performance, scaling capabilities, speed of computing, simplicity in handling big data and unlimited concurrency support.

5. Application and Results

This section presents an application that identifies UK-wide areas with low levels of public transport quality, labour market accessibility, housing quality and educational barriers. It first describes how we capture, clean and curate the data from multiple novel sources, using a

variety of technological and simulation approaches. We then identify how the different aspects of SUDS allow the areas of interest to be identified.

5.1 Transport Availability Metrics (TAM)

Public transport service data were obtained from the UK Traveline Information Limited (UKTIL), which offers schedule (timetable) data for bus, light rail, tram and ferry services in England, Wales and Scotland (Traveline National Dataset, TNDS [http://www.travelinedata.org.uk/traveline-open-data/traveline-national-dataset/]). Train service schedule data for the entire country was obtained from UK Rail Delivery Group (www.gbrail.info). The public transport schedule data obtained from the UKTIL were in TransXchange format for bus, light rail, tram and ferry services, and in CIF format for train services (Rail Delivery Group, 2016). They were subsequently transformed to the General Transit Feed Specification (GTFS) format, using a modified version of a Python conversion tool (Mooney, 2016). In total, data from 329,314 bus stops/17,880 bus routes, 2,514 rail stations/5,770 rail routes, 1,325 tram stations/93 tram routes, and 306 ferry stations/139 ferry routes in operation in Great Britain (England, Wales, and Scotland) were obtained.

The acquired timetables and locations of stops/stations were used to compute the service levels (frequency of service) at these locations. These were subsequently used to generate useful public transport availability metrics, including average hourly frequency (AHF), density of stops (DOS), density of nighttime stops (DONS), and Density of Routes (DOR) for the whole of Great Britain at LSOA and MSOA levels, which were chosen as the lowest spatial levels of aggregation for SUDS.

The average hourly frequency (AHF) at the stop/station-level was computed as:

$$AHF(i) = \frac{1}{5}\sum_{t \in T} cnt\_trip(i,t) \quad (1)$$

where i is a stop/station, cnt_trip (i,t) is the total count of trips passing through the station (stop) i within a one-hour time slot t on five working days (Monday to Friday); T is the set of one-hour time slots. We focused on working days as a representative of public transport availability because the vast majority of the trips to basic destinations such as workplaces and schools occur mainly on such days. Thus, the public transport availability indicators calculated on working days reflect the extent to which public transport can serve people and support their basic activities.

We used the AHF in conjunction with proximity to compute the transport-availability metrics at the LSOA level. Previous studies measured public transport availability using proximity (walking distance) to stations/stops and service frequency (Minocha et al., 2008; Currie, 2010; Delbosc and Currie, 2011). To measure public transport availability accurately for each LSOA, we took into account the service areas (the area within which people are willing to walk to the station/stop) and service levels (hourly service frequency). The willingness of people to walk to a station decreases as the walking distance to a bus stop increases (Langford et al., 2012). Some studies have suggested 400m (for bus and tram stops) and 800m (for rail and ferry stations) as acceptable maximum walking distances for the different public transport modes (Currie, 2010; Delbosc and Currie, 2011; Langford et al., 2012). These are based on distances that 75 - 80% of people would walk to access a stop/station according to a travel survey (Kittelson and Associates et al., 2003).

The service areas of stations/stops were delineated using spatial buffering. A circular buffer centred on a station/stop is conventionally used to represent the service area of the station/stop. The buffer represents the area where walking distance to a station/stop along the road network is within the acceptable maximum walking distances. The delineated service areas for the

stops/stations were subsequently overlapped with the LSOAs and any stop/station that intersected with an LSOA is allocated to that LSOA, which it is assumed to serve. For each LSOA, the stop-level AHFs for all the allocated stops/stations were aggregated. The LSOA-level AHF is subsequently computed as a combined measure of service level (aggregated AHF) and walking distance using the following:

$$AHF(a) = \sum_{i \in S(a)} AHF(i) * \frac{Area\ (i \cap a)}{Area\ (a)} \quad (2)$$

where a is the LSOA of interest, i is stations/stop, and S(a) is the set of stations/stops whose buffers intersect a. Area (i∩a) represents the overlapping area between i and a; and Area (a) is the area of a. In addition to AHF, two other metrics, density of stops/stations (DOS) and density of nighttime stops/stations (DONS - services between 6pm and 5am) serving an LSOA, were also computed as measures of public transport availability. The DOS for an LSOS (a) was calculated according to the following:

$$DOS(a) = \frac{NOS(a)}{Area\ (a)} \quad (3)$$

Where NOS(a) is the number of stations/stops serving a, and Area(a) is the area of a. The DONS was calculated as:

$$DONS(a) = \frac{NONS(a)}{Area\ (a)} \quad (4)$$

where NONS(a) is the number of nighttime stations/stops serving a, and Area(a) is the area of a.

The computed indicators were subsequently loaded into SUDS data lake for integration with other data using a series of appropriate ETL tools (see Figure 4). With these metrics, public

transport service in various census output areas, counties and regions could be evaluated, compared and ranked. For instance, using global and local spatial clustering approaches (Theil indices – generalized entropy, and Multidirectional Optimal Ecotope-Based Algorithm (AMOEBA) implement via ClusterPy), the TAM was used to identify and levels of spatial inequalities in public transport availability at intra-city, city- and regional-levels across the county. These were subsequently used to identify areas of low PTA at local and global scales; and populations/neighbourhoods at risk of transport poverty. Further details of this process are provided in another report currently under review.

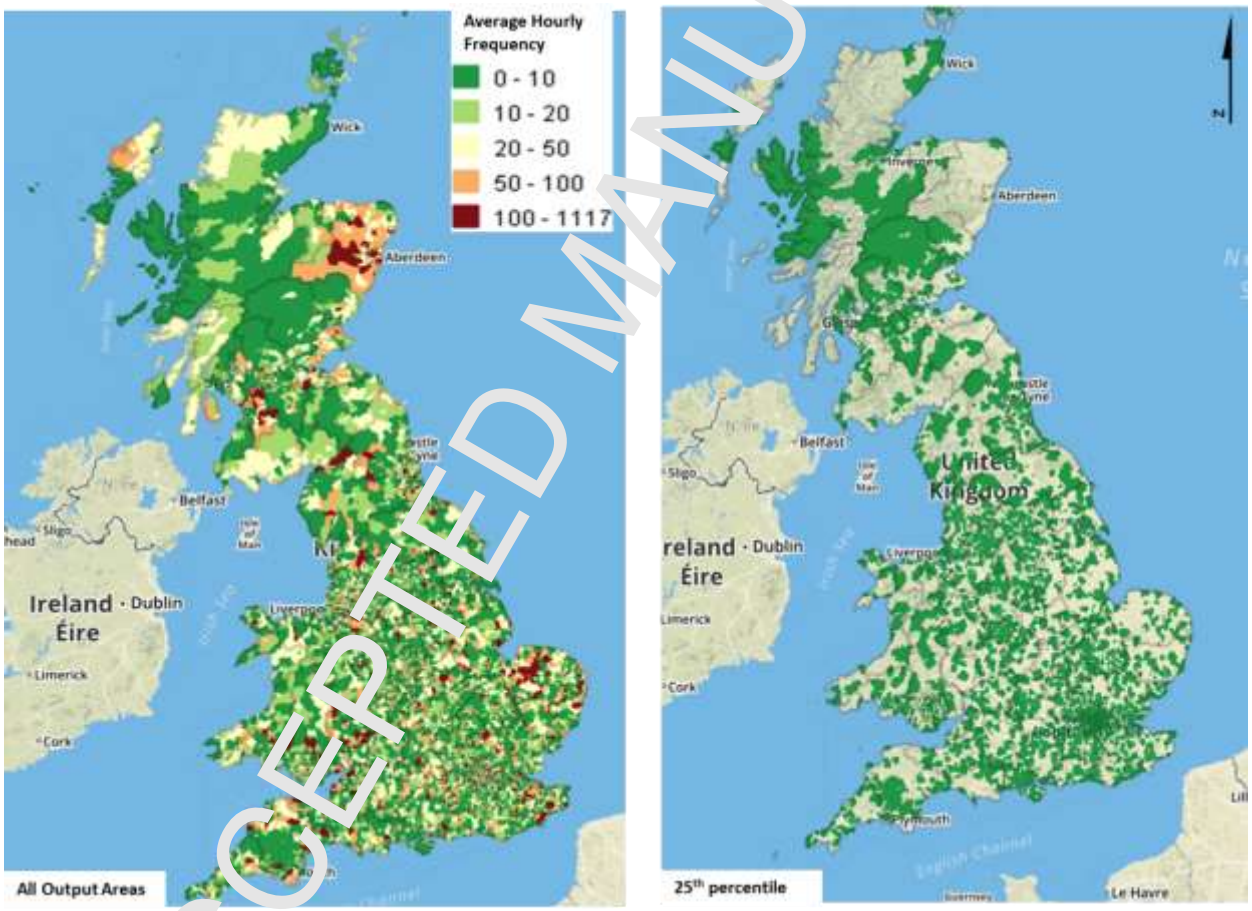


Figure 4. Maps showing one of the Transport Availability Metrics (average hourly frequency - AHF) for all output areas across the UK and output areas with AHF less than the 25th percentile of the countrywide values, displayed on SUDS interface.

5.2 Housing Affordability Metrics (HAM)

Housing indicators are used to highlight the most important features of housing markets (Flood, 1997). Some prominent housing indicators include: house price-to-income ratio, house rent-to-income ratio, floor area per person, mortgage-to-credit ratio, housing investment, household income distribution, housing tenure type, mortgage affordability (Flood, 1997). Computation of the indicators depends on an accurate knowledge of housing dynamics. Currently, there is a considerable knowledge gap concerning the scale and nature of housing dynamics, such as the UK private-rented sector, as most of the available information comes from Census data that are updated only every 10 years. This undermines a clear understanding of changes and associated issues by local authorities, central government and academic researchers. However, to undertake continuous monitoring of the sector over time, housing market information has to be obtained from alternative sources.

Data from the house listings aggregation service Zoopla (https://developer.zoopla.co.uk/) was considered a suitable alternative source for this crucial information. Housing data from properties advertised for sale or rent across Great Britain, from 2010 till present, were acquired under licence, and complemented by price paid data from the Land Registry of England and Wales and Registers of Scotland. Zoopla has over 27 million residential property records in their archive. Access to active and historical property listings is allowed via an Application Programming Interface (API), made available to developers by Zoopla. The UBDC has a licence to access this API with an agreement to download data for the United Kingdom as part of the Centre's housing data catalogue.

Baseline property listings (which contain various types of important historical information about properties) comprising 8 million property records (5 million advertised for sale and 3 million for rent) across Great Britain were initially generated via the Zoopla API with FME data extraction tools (Figure 4), and continuously updated as more properties left the market (closed listings).

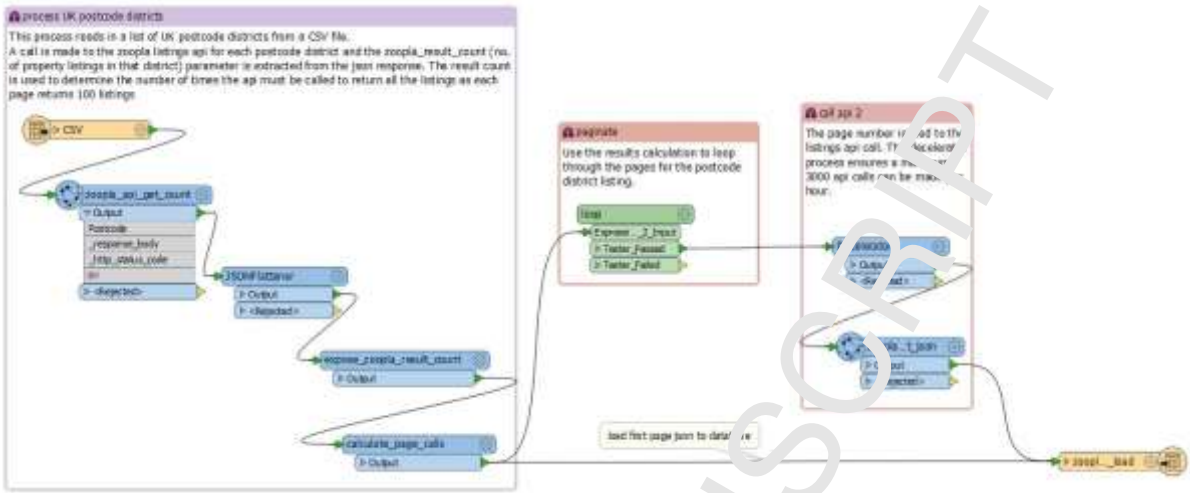


Figure 5. Representation of the workflow of housing data acquisition and transformation with FME

To generate the housing indicators, relevant housing attributes such as property IDs, address, price, description, date of advert, category, number of floors, were extracted from the Zoopla dataset. The data were linked to the LSOA spatial boundaries through the postcodes. Following this, aggregate data for key statistics (mean, median, maximum price, minimum for the rent and sale prices) of the properties, were computed at LSOA level. These were subsequently combined with demographic data at the LSOA to generate further synthetic metrics (Figure 6).

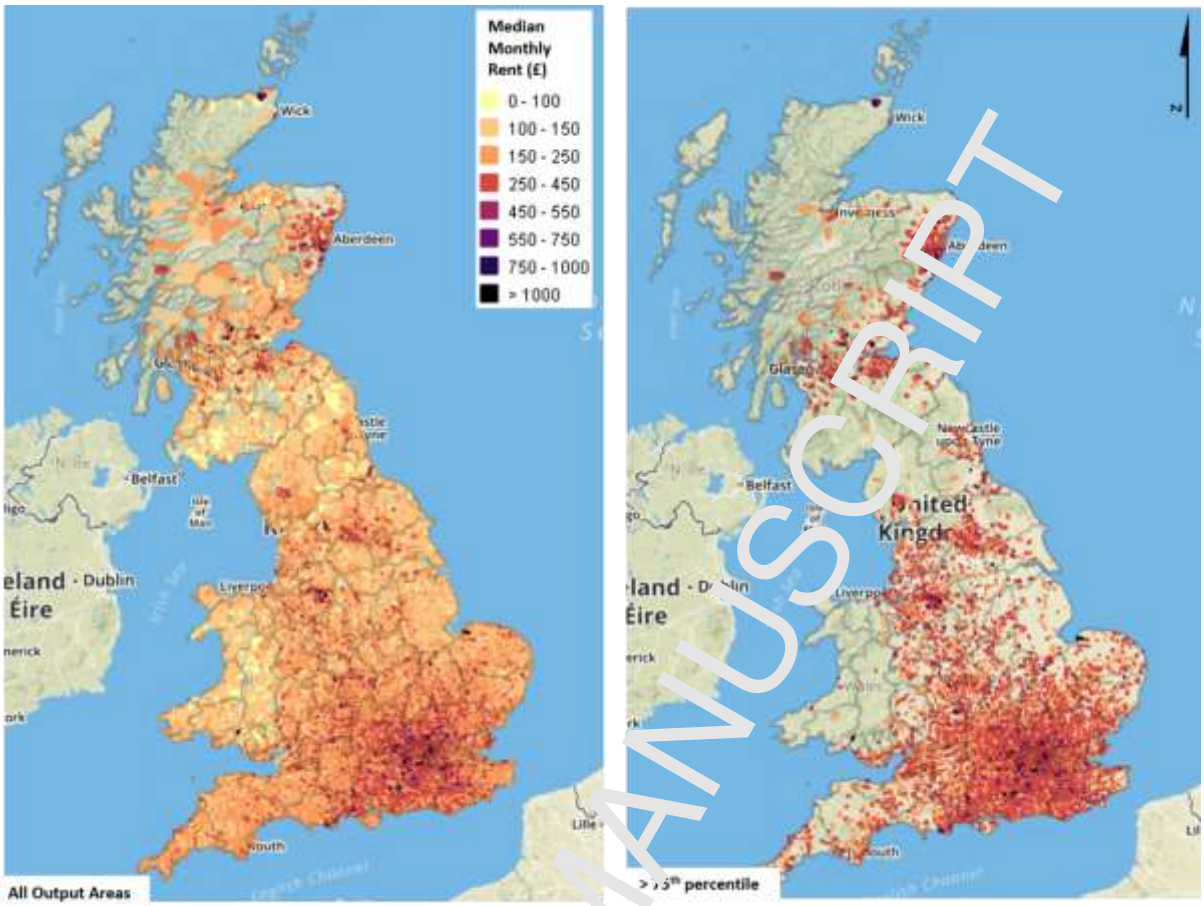


Figure 6. Maps showing monthly median rent prices for all output areas across the UK and output areas with median rent greater than the 75th percentile of the entire country.

5.3 Employment Accessibility Metrics (EAM)

The generation of employment accessibility indicators is driven by the need to continuously obtain more detailed geographical estimates of jobs and locations of workers at small-area levels such as postcodes or output area levels over time (quarterly, annually), rather than using those currently available from the census or the ONS, which are either disclosed at fairly highly aggregated levels or are available only once every 10 years. This is expected to enhance the understanding of the performance of different types of jobs (e.g., low-wage jobs or those in the service sector), as the economy goes through expansions, recessions or stagnation, by breaking down estimates of jobs and workers into different categories of interest. Thus, the metrics are designed to measure the structure and conditions of the local economy and labour markets at

intra-city levels. In addition, the metrics could be extended to become composite synthetic measures of the links between the economy and infrastructure.

Travel to work data from the 2011 census, obtained from the UK Data Service's Flow Data portal was used to determine the number of people reporting that they worked in each output area. This was used as a proxy for employment. Table WF03UK_oa (https://wicid.ukdataservice.ac.uk/), which provides the location of people's residence and work (excluding quasi-workplaces) at the level of output area for the UK, was used. The level of employment in each output area was proxied by aggregating the data by workplace output area. These employment data, combined with travel time information derived from the OpenStreetMap, were used to generate a number of labour market accessibility measures (Figure 7), using the gravity-based measure of potential accessibility developed by Hansen (1959).

To calculate these, a measure of the cost of travelling between each pair of origins and destinations was required. Distance along the road network was used as the measure of travel cost. The road network was represented using OpenStreetMap. An all-pairs shortest-path algorithm was then used to estimate a distance matrix.

Many different methods have been developed to measure accessibility. A popular one, which we used here, is the gravity-based measure of potential accessibility developed by Hansen (1959). This is generally represented as:

$$A_i = \Sigma D_j f(c_{ij}) \qquad (5)$$

where A_i is the accessibility index for zone i, D_j is a measure of the opportunities available at destination i, c_{ij} is the cost of travel between zones i and j, and f() is a cost deterrence function which captures how distance affects the accessibility of opportunities. For our purposes, D was

used to represent the number of people stating they worked in each output area and c_{ij} will be the network distance between output areas i and j.

The deterrence function also has to be defined. Many options are available, but we opted for a simple threshold function of the form:

$$f(c_{ij}) = \begin{cases} 1 \; if \; c_{ij} \leq \tau \\ 0 \; if \; c_{ij} > \tau \end{cases} \quad (6)$$

We evaluated the function for different levels of the parameter τ. The accessibility measure gives the number of employment opportunities that can be reached within a given distance. One advantage of this measure is that it is easy to interpret. Further details of this are not within the scope of the current paper, but are covered in another report.

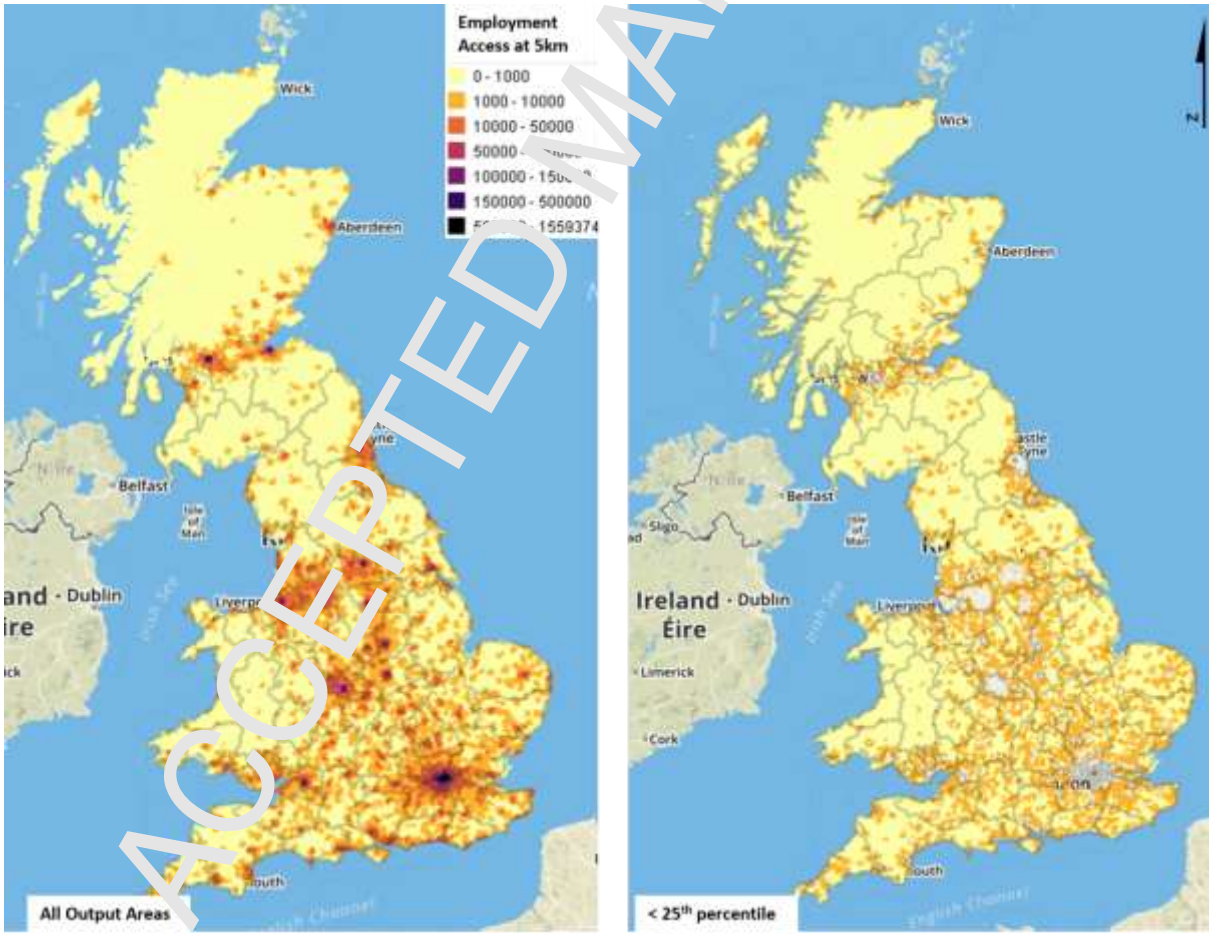


Figure 7. Maps showing employment opportunities within 5km (access 5km) of for all output areas across the UK and output areas with access 5km less than the 25th percentile of the entire country.

5.4 Education-Related Metrics

The creation of education-related metrics (ERM) has been prompted by the desire to examine small area-based drivers of inequalities in educational outcomes (Bell, 2003; Kerr et al, 2014), from Secondary School, through Further and Higher Education and into employment, and against the background of Scotland's Attainment Challenge (https://education.gov.scot/improvement/learning-resources/Scottish Attainment Challenge), which was launched by the Scottish Government in 2015 to achieve equity of educational opportunity and thereby reduce the poverty-related "attainment gap".

Secondary school data were obtained from the Scottish Exchange of Data (ScotXed - http://www.gov.scot/Topics/Statistics/ScotXed), covering the eight local authorities comprising the Glasgow City Region (Glasgow City, East and West Dunbartonshire, North and South Lanarkshire, Renfrewshire, East Renfrewshire, and Inverclyde). The datasets feature individual student-level data from the pupil census and data on all 31 publicly funded secondary schools for the academic years 2007/8 to 2015/16. Pupil data consisted of age, gender, nationality and ethnic background, level of English, receipt of Gaelic education, attendance, and post-school destinations. Educational attainment was measured for all units and courses at levels S4-S6 (senior secondary education, typically of those aged 14-17 years). Schools data cover staffing levels, and proportions of pupils' speaking particular languages at home. The linked pupil and school datasets are extended with other derived and administrative data: the distance (Euclidean) travelled by students between home and school and accessibility to different types of greenspace from the home and school neighbourhoods were calculated from postcode centroids, and home and school locations were linked at datazone level to assign measures of deprivation and rurality.

A broadly similar Higher Education dataset was developed from data supplied by the Higher Education Statistics Agency (HESA- https://www.hesa.ac.uk/). This is a secured data obtained through electronic Data Research and Innovation Service (eDRIS) special licencing arrangement (Safe Haven). It contains approximately 44.7m records for all students attending a Higher Education institution in the UK between 2000/1 and 2015/16, comprising personal characteristics (including home location at postcode sector level), and subject, level and mode of study of courses pursued, level and classification of qualification, and post-HE destination.

The various datasets are currently being used to develop appropriate spatiotemporal indicators of student- and institution-based educational disadvantage at these stages of the educational career. New insights are expected to be derived via the linkage of ERM with other metrics in SUDS, such as the EAM (synergising labour market dynamics with quality of education) and using TAM to provide information about journeys between home and educational institution. These will give a richer understanding of the urban basis of educational inequalities, generating more flexible and locally tailored policy-relevant information and, thereby, solutions to these inequalities.

5.5 Urban Analytics

It is expected that SUDS will be used by policy-makers to undertake several projects that will enhance urban sustainability and smart city management. Some of the potential applications of SUDS include small-area multi-criteria evaluation, where the various metrics can be interactively integrated and explored to understand the dynamics of underlying relationships and locational variability of various city components.

Figure 8 illustrates the SUDS web interface, showing the results obtained from the combination of three SUDS metrics (transport, housing prices and access to employment). The figure shows the spatial distribution of output areas of low liveability (high rent, poor transport services and

low employment opportunities). The following thresholds were used: rent price greater than the 75 percentile, average hourly frequency (AHF) of transport services less than 25 percentile and available jobs within 5 km, less than 25 percentile of countrywide values.

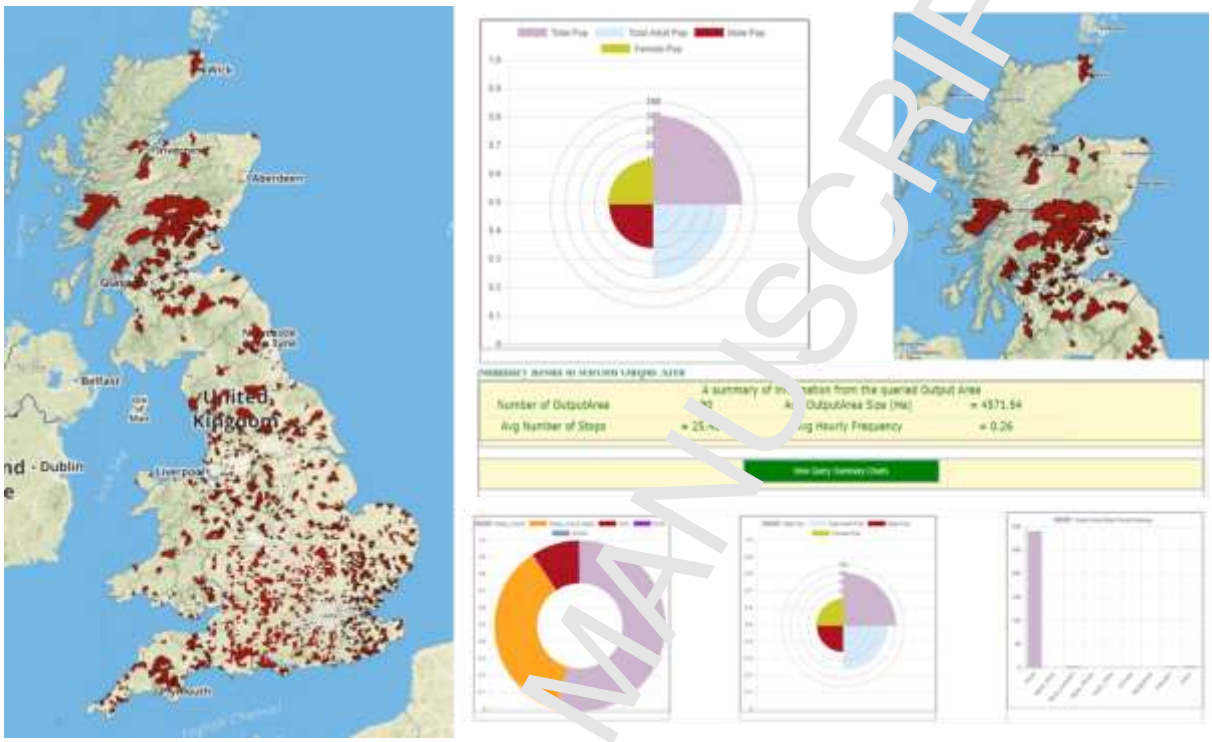


Figure 8. SUDS was used to identify output areas with high monthly rents (>75 percentile), poor access to jobs (<25 percentile) and transport (<25 percentile).

It highlights sub-city variability of these metrics across the country (see Figure 9), which is often masked in other similar systems. For instance, with the exception of London where no output area was identified as having low liveability, others such as Manchester, Glasgow, Aberdeen and Cardiff had few output areas in the low liveability category. This buttresses one of the important aspects of SUDS – identifying intra-city variations that would have otherwise been missed. The interactive nature of SUDS also ensures that users such as city administrators or researchers can set or test different thresholds or use alternative criteria to explore particular aspects of the urban area.

Figure 9. Sub-city variations of low liveability in selected UK cities

This example demonstrates how SUDS can be applied to understand and manage various aspects of the urban area. Although its primary focus is on social and economic aspects, its use could be extended to include environmental attributes. For instance, social scientists could use SUDS to tap into a variety of contextual neighbourhood-level factors that partly explain economic, social, behavioural and attitudinal outcomes of individuals, firms, markets or other institutions and organisations, without which their analysis would potentially suffer from various endogeneity or omitted-variable biases, among many other methodological limitations. For example, suppose a researcher wishes to analyse labour market outcomes such as hours worked or wages earned by low-income single mothers living in urbanised areas in the UK. Aside from the usual sociodemographic, human and social capital factors, SUDS enables the analyst to control for background factors such as transport access, general labour market and industry conditions in the area, and broader economic trends in the region to be introduced into the analysis, thereby facilitating a more complete analysis of the outcomes of interest. For environmental applications, a researcher may be interested in analysing public health outcomes for which SUDS may be able to provide small-area estimates of the characteristics of the built

and physical environments in which people live, work or go to school, such as housing density and accessibility, alternatives to driving such as walking or cycling infrastructure, and access to high-quality food outlets, green space, clean air and clean water.

SUDS was also designed to inform policy-making, invite public, private and citizen action to address challenges in urban transport, housing, the environment, education, land-use, urban design, labour markets and employment conditions, public health, social care, and other policy areas. In this respect, SUDS will enable the public to engage with academic outputs relevant to the understanding of urban areas. The goal is to stimulate a range of civic and business innovations with the adoption of SUDS by urban digital infomediaries (Thakuriah et al, 2017). It is the aspiration that debates stimulated by SUDS will lead to improved services and wellbeing of people, places and infrastructure, and facilitate communication and exchange of information among stakeholders towards these objectives.

The system will also serve as a tool that will support data-related engagement with data owners, and encourage data owners to contribute data. More broadly, the system is intended to play a central role in stakeholder engagement activities, particularly with policy-makers, businesses, data providers and non-profit organisations. Hence, an important component of SUDS is the development of visualisation and interactive mapping components that provides a unique opportunity for intentional, meaningful interactions on city life that provide opportunities for mutual learning between urban researchers and members of the public. “Mutual learning” here refers not just to the acquisition of knowledge, but also to the increased familiarity with a breadth of perspectives, frames, and worldviews (American Association for the Advancement of Science, 2018) and helps to “empower people, broaden attitudes and ensure that the work of universities and research institutes is relevant to society and wider social concerns” (RCUK, 2018).

Specifically, the employment availability metrics will be useful for job-accessibility studies that involve matching workers to jobs. For example, those involving non-standard shift-work, which are often low-wage jobs, may not be available to workers who are dependent on public transport, or who have no car, or who have difficulty running a car during periods of high fuel prices, or to other transport users if otherwise suitable jobs are located in areas that are not well-served by transit during off-peak periods. Similarly, manufacturing enterprises located in areas with high levels of congestion may be affected by the inability of just-in-time freight delivery during certain hours of the day, or even the ability to attract employment for non-standard work shifts. Non-standard shifts may further affect the quality of access to local services and social activities in the absence of reliable transport.

These reasons underline the need to estimate the spatio-temporal locations of workers and jobs in terms of precise estimates of the geographical location of policy-relevant categories of worker residences and jobs, as well as the temporal shifts of those jobs.

The housing metrics could be used to gauge the effect of certain changes in policy or industrial activities. For instance, the effects of disruptors or accelerators in a society, such as the establishment of new industries or the collapse of existing ones, could be measured in terms of changes in house sales and rent prices. It is important in smart city management for these sorts of dynamics to be picked up as quickly as possibly at very detailed spatial levels without having to rely on Census statistics, which are gathered much less frequently, in order to use this information to quickly cushion the adverse effects of utility-associated benefits.

Another future application of SUDS is in the area of urban predictive modelling and analysis, where machine learning could be used in conjunction with the metrics generated by SUDS to gain deeper insights about urban area dynamics such as those associated with: predicting future outcomes from structural, infrastructural, commercial and industrial changes and impacts on

urban dwellers such as where job losses might happen, or where house prices might rise or fall; gauging urban area emotions or reactions to policy changes; and predicting the location of events, such as riots and various types of crime, through the use of existing datasets. Incorporating data from urban IoT would facilitate real-time monitoring of environmental quality using SUDS. Hourly air pollution data automatically generated by monitoring networks could be streamed into the SUDS cloud data lake to facilitate real-time monitoring of air quality across cities. This could subsequently be integrated with other datasets to generate further insights, for example, by combining active travel data (e.g., from Strava) to dynamically monitor exposures to pollution. In the same vein, available data from smart meters and street lights could be used to gain a detailed understanding of energy usage over time across cities. These data could also be used to derive metrics and indicators for measuring socioeconomic factors such as household poverty.

6. Challenges, Limitations, and Issues

There are certain challenges, limitations and issues encountered in the development of the SUDS infrastructure, especially relating to data governance, data acquisition, information management, and system reproducibility, which are briefly discussed below.

6.1 UBDC Data Services and Data Governance

SUDS is a manifestation of UBDC's data service. Unlike comparable data platforms, SUDS uses not only open data and derived data products, but also data licensed by UBDC under more restrictive agreements. This demands additional controls and governance mechanisms but offers opportunities to achieve broader, higher spatial resolution insights, reflecting a broader, growing emphasis on data sharing, versus open data. As a public good, open data is highly desirable but many factors, often related to privacy or commercial sensitivities, limit the feasibility that all potentially useful data can be made available under open licences. The

benefits of data sharing for doing research work have been widely discussed (Chatham House Data Sharing Advisory Group, 2016), albeit offset with concerns – that wealthier stakeholders are best positioned to benefit, at the cost of poorer communities, or that data subjects' privacy may be at risk because of the practice (van Panhuis et al., 2014). UBDC's data service aims to minimise barriers to the use of data in the resolution of urban challenges. Broadening access means providing a service that is free at the point of use, and negotiating with data owners to agree terms for data sharing that are as unrestrictive as possible, while protecting the interests of individuals and organisations affected. UBDC partly achieves this by offering data owners reassurances through its policies for managing data access.

UBDC datasets are grouped into one of three categories and members of each are candidates for publication within the SUDS platform. The first is the Centre's open data collection – typically licensed under Open Government or Creative Commons data licences, these datasets are accessible via a public portal to any prospective user. They can likewise be published on the SUDS platform with few limitations. The second category, which involves additional restrictions, is UBDC's safeguarded data collection. This comprises of datasets that have associated bespoke licensing and data sharing arrangements. End users wishing to access these data must agree to the relevant terms and the permitted uses of such data, and the nature of permitted outputs are more strictly limited. The limitations imposed, and the possibilities for platforms like SUDS are specific to each data sharing agreement. The third category is controlled data, those datasets with additional restrictions related primarily to the sensitivity of their content. These are mostly individual level data, such as administrative health or social care datasets, where there is an onus on protecting individuals' privacy. In such cases, physical access is restricted to within secure safe environments. Outputs are subject to formal approval processes (particularly to ensure that risks of disclosure are managed). In many cases UBDC's role with respect to controlled data is to broker access between third parties (typically data

owners, users and administrators of safe indexing, access and analytics environments) with no custodial role.

UBDC has infrastructure and governance controls in place to support users wishing to access datasets across each collection, with data released via SUDS subject to the same processes and limitations. Informing licensing, ingest and data processing, UBDC's data accessioning policy defines seven primary stages. These are 1) negotiation of dataset licensing, where data sharing agreements and end user licensing arrangements are agreed and formalised; 2) physical acquisition of data, where data and associated metadata are physically transferred and received; 3) dataset assessment, where datasets are evaluated and additional processing requirements identified; 4) dataset processing, where applicable processing is undertaken; 5) data documentation, where accompanying documentation is created, validated and standardised; 6) dataset definition, where one or more agreed data packages are defined and their manifests recorded; and 7) dataset publication, where data is published to one or more delivery platforms. Several stages operate iteratively with new data products defined, produced and published in response to emerging researcher requirements or data additions/changes. In terms of the user experience, UBDC's end user delivery policy controls access to data within UBDC's collections. This establishes several stages whereby end users' purposes are defined and compared with relevant data sharing policy(ies); sub-licensing documentation is exchanged, completed and stored, and data is securely transferred or made accessible to authorised, authenticated users through a secure platform. For the most sensitive controlled data that UBDC facilitates access to (e.g. individual-level health data) additional governance processes require prospective users to satisfy an independent committee of the scientific and public benefit impacts of their proposed work, and of the appropriate mitigation of associated risks.

Predictability, negotiating these policies and processes is much simpler for acquiring and sharing open data, than, for example, commercially sensitive business data.

UBDC approaches the accessioning of a given dataset with its safe accessibility of foremost importance. Agreements with data owners may not permit widespread sharing of raw data to general audiences but it may be possible to negotiate rights to publish derived aggregate data products instead. Data requirements vary by projects and circumstances – for instance, although one community of users may require access to individual level higher education attainment data another may benefit just as much from aggregate, rounded summary data (particularly if accessible with few practical restrictions). Similarly, synthetic data offers opportunities to create widely shareable resources that are more credible if produced with reference to real-world, but highly controlled, datasets. Furthermore, although SUDS is available online and built primarily using open source technology, it is by no means a wholly open data platform. Limitations to data availability are supported, and end user licensing constraints can be enforced to ensure that only authorised, authenticated users may access particularly datasets or higher resolution data content.

6.2 Data acquisition, processing and software integration issues

Some of the challenges and limitations encountered in the development of SUDS revolve around data acquisition, licensing and protection, as well as the choice of software to be used for the various components of SUDS. Access to some of the data from commercial vendors through APIs is usually subject to certain conditions, which must be considered when designing the workflows for data retrieval.

Another issue is the choice of the appropriate level of spatial and temporal resolution of the metrics that should be made publicly available. SUDS aims to calculate and display the urban

area metrics at highly granular levels, in finer detail than has been achieved with previously computed metrics/indicators. However, this is also subject to data licensing agreements and the need to preserve anonymity, especially with the implementation of the General Data Protection Regulation (GDPR) in Europe in May 2018. This informs the use of Census output areas as the base spatial scale for SUDS.

6.3 Managing dataset licensing and associated sensitivities

One of the principle non-technical challenges in delivering the SUDS architecture is rationalising the terms and conditions of usage and the varying sensitivities of datasets originating from many sources. The goal of the CRDC, when negotiating data sharing agreements, as part of its data service responsibility, is to be able to support broad accessibility and utility of data, with the fewest possible constraints. Predictably, this rarely happens without compromises, which in turn leads to restrictions or responsibilities bespoke to each agreement. These are often limits on the permitted types of users and usage (e.g., academic researchers only), requirements for physically accessing data (e.g., via secure centralised data stores) or in terms of what can be published following research activities. They extend to aspects of data protection law, the scale and scope of liabilities and aspects of academic freedom. Pricing models for provision of data to third parties are also variable.

Although this paper does not specifically cover legal interoperability issues, our ambitions for SUDS to combine disparate sources and support analysis based on parameters from multiple datasets establishes it as a challenging consideration. Considerable related work has focused on interoperability of open research data licenses (see for example RDA-CODATA, 2016) or issues associated with deploying open data within business and government contexts (Morando, 2013). The compatibility of free and open source licenses within a software context are also well explored (Rosen, 2004). Given the increasing emphasis being placed on the value

of shared data, acknowledging the limits of what can be made wholly open, there remains uncertainty as to what can be done when combining multiple, more restricted sources. Within SUDS we approach this issue in a bottom-up manner by adopting a cautious approach to information sharing, establishing a licensing process as a gateway to data access and enforcing limits on accessibility to the platform as well as individually presented datasets. Increasing use of synthetic data may offer a means of bypassing particularly restrictive terms and conditions (although the feasibility of this approach may depend on a number of factors, not least the terms and conditions of a given license). Convincing data owners of the value of contributing to a shared pool of data with a view to them realising benefits from accessing the whole remains a significant objective.

In addition to the constraints associated with licensing terms and conditions, the use of individual, person-level records present further challenges. Several datasets in use within SUDS, such as the HESA and ScotXed education data present specific personal data and privacy issues, as covered by legislation such as the GDPR. Access to these data is tightly regulated and corresponding data-sharing agreements impose demands regarding the environments within which they can be accessed, and the permitted outputs that may emerge. At present, the use of information based on these types of sources requires significant manual intervention to produce aggregate outputs within a secure data access environment. Outputs are subject to statistical disclosure control prior to their incorporation within SUDS. The development of solutions to facilitate the safe integration of personal data sources remains a key objective. The risk of statistical disclosure and compromising of privacy is an additional important motivation for the generation of synthetic populations.

These are salient issues that must be thoroughly considered while developing a system, like SUDS, that is intended to be publicly available at high levels of spatial and temporal resolution.

6.3 System Reproducibility

With regards to choice of software, there is currently a wide range of commercial and open-source software that could be deployed to perform some of the tasks in SUDS. Despite the many benefits of availability of a wide range of technologies, this on its own presents a challenge, especially regarding how to determine appropriate sets of software to be deployed. Software applications, even those developed to perform similar tasks, have different performance capabilities in certain respects. This calls for careful consideration and the challenges they present must be cautiously navigated when developing an infrastructure like SUDS. Wherever convenient, SUDS' first choice is the deployment of open-source tools and software. Robustness, speed and ease of usability of the software were also considered. Combining different software into a system also presents a challenge. We have overcome some of these data and software integration problems in SUDS by using spatial and non-spatial ETL tools to drive the workflow.

7. Conclusions and Future Work

In this paper, we have described the Spatial Urban Data System (SUDS), a part of the UK ESRC-funded Urban Big Data Centre (UBDC). SUDS is a small-area geospatial big data system that delivers complex data analytics at the scale of a country, allowing regional comparisons and sub-area analysis, on a variety of social and economic attributes of urban living. At the core of the system is a programme of urban indicators generated by using novel forms of data and an urban modelling and simulation programme. Using public transport, labour market accessibility and housing advertisement data, we were able to show areas that are deprived of certain urban services in the UK. One of the key objectives of the system is to disseminate the technology to local governments, small businesses and other users such as NGOs in less-developed nations. For this reason, the technology used is open-source and

replicable elsewhere. The system grows organically with new policies, stakeholders and data opportunities. A robust user base is recruited using a recruitment and communications plan.

The SUDS differs from existing spatially enabled smart city analytics infrastructures in that it focuses largely on the generation and use of spatially enabled socioeconomic metrics collected countrywide at regular intervals to facilitate the understanding of intra-city dynamics and to provide "urban health checks." Researchers have noted the greater efforts being made to measure and monitor environmental aspects than those made to represent social, economic and governance aspects (Shen et al, 2011; Lynch, et al 2011). This informs SUDS' focus on social and economic, health and well-being conditions to enable a more comprehensive assessment of urban living, in line with sustainable development goals. SUDS provides a quantitative multidimensional foundation for comprehensive urban quality of life assessment.

It can be deployed for smart city performance monitoring and assessment at an intra-city level in a timely manner. Other application areas include high-resolution urban area indicator generation that could drive city comparison and ranking; urban area predictive analytics for forecasting future outcomes and impacts of policies and changes; multi-criteria evaluation of impacts of urban area accelerators, disrupters and policies; and real-time monitoring of urban area dynamics. Furthermore, through the cloud computing component, data streams from urban IoT sensor networks could be processed and integrated with other datasets, such as historical data from various facets of the urban environment to derive new insights. Key unique selling points of SUDS include:

- integration and processing of spatially-activated big data from varying sources, with complex geospatial processing, and modern cloud computing systems, to derive deeper insights into sub-city interactions,

- generation and use of frequently updated small-area socioeconomic synthetic metrics on a countrywide basis,
- facilitation of the understanding of intra-city dynamics through the integration of data from various aspects of the urban area,
- development of series of strategies to process and utilised various socioeconomic variables, to understand and manage urban area dynamics,
- compatibility with modern cloud computing systems such as Snowflake Computing system, Azure SQL Data Warehouse, Amazon Redshift, Oracle Data Warehouse with advanced capabilities for handling big data.

Ongoing work includes the development of additional metrics from other aspects of the urban area, including health and wellbeing, environmental, and user-generated contents such as those from social media (Twitter, Facebook, Reddit, etc.) or transactional data. Data on athletic activities of city residents that could be used to gauge city lifestyle have been acquired from Strava under a licence. The Strava data contain spatially referenced information on various activities including cycling, running, and walking that could be integrated into SUDS. The Strava dataset comprises millions of anonymised and aggregated data of rides and runs uploaded regularly by Strava users via their mobile phones or GPS devices. Relevant metrics are currently being generated from the data. In addition, automation of the system through the use of APIs and ETL tools to obtain real-time travel data from sources such as Darwin and NextBus are being tested and optimised. Various components of SUDS are also being optimised for more efficient integration, processing and analysis of data, and for the visualisation of outputs. Advanced open-source geospatial big databases such as GeoMesa and GeoWave are currently being explored for possible incorporation with SUDS.

Finally, a training and capacity-building programme is underway to ensure that a wide base of potential users have the skills in GIS, software programming and related areas and are also familiar with the data to use the system as a part of their programmes.

Future work planned for SUDS will help develop it into a leading spatial big data platform with fully functional big data analytics capabilities, with a machine-learning component that will drive urban area predictive modelling and analytics, and real-time analytic tools to enable integration with the urban IoT.

Acknowledgements

We acknowledge the Economic and Social Research Council (ESRC) who funded the Urban Big Data Centre (UBDC) to undertake this project as part of the Big Data Phase 2 of the UK Research and Innovation's.

Data Sources

Zoopla

Citation: Zoopla Limited. Economic and Social Research Council. Zoopla Property Data [data collection]. University of Glasgow - Urban Big Data Centre.
Mandatory attribution: Zoopla Limited, © 2015

Experian

Citation: Goad Plan Data. Economic and Social Research Council. Goad Plan Data (Experian), 2018 [data collection]. University of Glasgow - Urban Big Data Centre.
Mandatory attribution: Experian's services are not intended to be used as the sole basis for any business decision, and are based upon data which is provided by third parties, the accuracy and/or completeness of which it would not be possible and/or economically viable for Experian to guarantee. Experian's services also involve models and techniques based on statistical analysis, probability and predictive behaviour. Experian is therefore not able to accept any liability for any inaccuracy, incompleteness or other error in the Experian data which arises as a result of data provided or any failure to achieve a particular result.

Registers of Scotland

Citation: Registers of Scotland. Economic and Social Research Council. Registers of Scotland All Sales Data [data collection]. University of Glasgow - Urban Big Data Centre.
Mandatory attribution: © Crown copyright. Material is reproduced with the permission of the Keeper of the Registers of Scotland.

Strava

Citation: Strava Inc. Economic and Social Research Council. Strava Metro data - Scotland, Glasgow, Manchester, Tyne and Wear [data collection]. University of Glasgow - Urban Big Data Centre.
Mandatory attribution: Data Licensed by Strava

BGS

Citation: NERC. British Geological Survey. Economic and Social Research Council. British Geological Survey Data [data collection]. University of Glasgow - Urban Big Data Centre.

Met office

Citation:
Met Office. Economic and Social Research Council. Met Office: Forecasts Data [data collection]. University of Glasgow - Urban Big Data Centre.
Met Office. Economic and Social Research Council. Met Office: Observations Data [data collection]. University of Glasgow - Urban Big Data Centre.
Mandatory attribution: Contains public sector information licensed under the Open Government Licence

Springboard

Citation: Springboard. Economic and Social Research Council. Springboard's Footfall Benchmark Data, 2018 [data collection]. University of Glasgow - Urban Big Data Centre.

Twitter

Citation: Urban Big Data Centre. Economic and Social Research Council. iMCD Project: Twitter Data, 2015 [data collection]. University of Glasgow - Urban Big Data Centre.

Facebook

Citation: Urban Big Data Centre. Economic and Social Research Council. iMCD Project: Facebook Data, 2015 [data collection]. University of Glasgow - Urban Big Data Centre.

References

Aguilera, U., Peña, O., Belmonte, O., López-de-Ipiña, D. (2017). Citizen-centric data services for smarter cities. *Future Generation Computer Systems*, 76 234–247.

Airaksinen, M. (2016). Smart cities, can the performance be measured? http://www.vttresearch.com/Impulse/Pages/Smart-cities,-can-the-performance-be-measured.aspx (accessed on 12th March, 2018).

Albino,V., Berardi, U. and Dangelico, R.M. (2015). Smart Cities: Definitions, Dimensions, Performance, and Initiatives. *Journal of Urban Technology*, 22(1), 3-21, DOI: 10.1080/10630732.2014.942092.

American Association for the Advancement of Science. (2018). Why Public Engagement Matters. http://www.aaas.org/pes/what-public-engagement (accessed on May 21, 2018).

Bain, M. Sentilo - Sensor and Actuator Platform for Smart Cities, 2014. https://joinup.ec.europa.eu/community/eupl/document/sentilo-sensor-andactuator-platform-smart-cities (Accessed on 20th May 2018).

Babar, M. and Arif, F. (2017). Smart urban planning using Big Data analytics to contend with the interoperability in Internet of Things. *Future Generation Computer Systems*, 77 (2017) 65–76.

Bell, J. (2003). Beyond the school gates: The influence of neighbourhood on the relative progress of pupils. *Oxford Review of Education*, 29, 485-502.

Berardi, U. (2013). "Sustainability Assessments of urban Communities through Rating Systems," *Environment, Development and Sustainability*, 15 (6), 1573–1591.

Brovelli, M.A., Minghini, M., Moreno-Sanchez, R. and Oliveira, R. (2018). Free and open source software for geospatial applications (FOSS4G) to support Future Earth. 10, (4), 386-404.

Canadian International Development Agency. (2012). Indicators for Sustainability How cities are monitoring and evaluating their success. https://sustainablecities.net/wp-content/uploads/2015/10/indicators-for-sustainability-in-case-studies-final.pdf (accessed on 17th February, 2018).

Chatham House Data Sharing Advisory Group. (2016). Public health surveillance: a call to share data. International Association of National Public Health Institutes.

Crompvoets, J., Rajabifard, A., van Loenen, B. & Delgado Fernandez, T. (2008). A multi-view framework to assess spatial data infrastructures. Melbourne: Digital Print Centre, The University of Melbourne, Australia.

Currie, G., 2010. Quantifying spatial gaps in public transport supply based on social needs. J. Transp. Geogr. 18, 31-41.

Delbosc, A., Currie, G., 2011. Using Lorenz curves to assess public transport equity. J. Transp. Geogr. 19(6), 1252-1259.

Dirks, S., and Keeling, M. (2009). A Vision of Smarter Cities: How Cities Can Lead the Way into a Prosperous and Sustainable Future. https://www-03.ibm.com/press/attachments/IBV_Smarter_Cities_-_Final.pdf (accessed on 17th February, 2018).

European Commission (2013a). Report for the European Parliament: Mapping Smart Cities in the EU. *IP/A/ITRE/ST/2013-02*.

European Commission. (2013b). EIP SCC, European Innovation Partnership on Smart Cities and Communities, Strategic Implementation Plan, 14.10.2013, http://ec.europa.eu/eip/smartcities/ (accessed on 17th February, 2018).

European Commission. (2015). Indicators for Sustainable Cities. Science for Environment Policy report. http://ec.europa.eu/environment/integration/research/newsalert/pdf/indicators_for_sustainable_cities_IR12_en.pdf (accessed on 17th February, 2018).

Flood, J. (1997). Urban and Housing Indicators. Urban Studies, 34, (10), 1635-1665.

Gaur, A., Scotney, B., Parr, G., and McClean, S. (2015). Smart City Architecture and its Applications based on IoT. Presented at the 5th International Symposium on Internet of Ubiquitous and Pervasive Things (IUPT 2015). *Procedia Computer Science*, 52, 1089 – 1094. doi: 10.1016/j.procs.2015.05.122.

GeoMesa. (2018). GeoMesa. https://www.geomesa.org/#downloads. (accessed on 6th November, 2018).

Giffinger, R., and Gudrun, H. (2010). "Smart Cities Ranking: An Effective Instrument for the Positioning of Cities?" ACE Architecture," *City and Environment*, 4 (12), 7–25.

Grus, L., Castelein, W., Crompvoets, J., Overduin, T., van Loenen, B., van Groenestijn, A., Rajabifard, A., and Bregt, A.K. (2011). An assessment view to evaluate whether Spatial Data Infrastructures meet their goals. *Computers, Environment and Urban Systems*, 35, 217–229.

Hu, Y. and Li, W. (2017). Spatial Data Infrastructures. in (ed.) John P. Wilson .*The Geographic Information Science & Technology Body of Knowledge*. http://dx.doi.org/10.22224/gistbok/2017.2.1

Huovila, A., Airaksinen, M., Pinto-Seppä, I., Piira, K. and Morl Penttinen, T. (2016). Smart city performance measurement system. *Paper presented at 41st IAHS WORLD CONGRESS Sustainability and Innovation for the Future 13-16th September 2016 Albufeira, Algarve, Portugal.*

Kerr, K., Dyson, A., Raffo, C. (2014). Education, Disadvantage and Place: Making the Local Matter, Bristol: Policy Press.

Khan, Z., Anjum, A., Soomro, K., Tahir, M.A. (2015). Towards cloud based big data analytics for smart future cities, J. Cloud Comput. 4 (1) 2.

Khan, Z., Anjum, A., Kiani, S.L. (2013). Cloud based big data analytics for smart future cities, in: Proceedings of the 2013 IEEE/ACM 6th International Conference on Utility and Cloud Computing, IEEE Computer Society, 2013, 381–386.

Li, S., Yang, H., Huang, Y. and Zhou, C. (2017). Geo-spatial Big Data Storage Based on NoSQL Database. Wuhan Daxue Xuebao (Xinxi Kexue Ban)/Geomatics and Information Science of Wuhan University 42(2):163-169. DOI: 10.13203/j.whugis20140774.

LocationTech. (2018). About GeoWave. https://locationtech.github.io/geowave/ (accessed on 6th November, 2018).

Lv, Z., Li,X., Wang, W., Zhang,B., Hu, J., Feng, S. (2018). Government affairs service platform for smart city. *Future Generation Computer Systems* 81,443–451

Minocha, I., Sriraj, P.S., Metaxatos, P., Thakuriah, V., 2008. Analysis of transit quality of service and employment accessibility for the greater Chicago, Illinois, region. Transp. Res. Record 2042, 20-29.

Mooney, A. (2016). Python tool to convert TransXChange data to GTFS. https://github.com/adamhukemooney/txc2gtfs. (accessed on 17th February, 2018).

Morando, F. (2013). Legal interoperability: Making open government data compatible with businesses and communities. JLIS.it, 4(1), 441. doi:http://dx.doi.org/10.4403/jlis.it-5461

Mori, K., and Christodoulou, A. (2012). "Review of Sustainability Indices and Indicators: Towards a
Murty, R.N., Mainland, G., Rose, I., Chowdhury, A.R., Gosain, A. , Bers, J. , Welsh, M. (2008). CitySense: An urban-scale wireless sensor network and testbed, in: Technologies for Homeland Security, 2008 IEEE Conference, 583–588. http://dx.doi.org/10.1109/THS.2008.4534518.

OECD, 2012, OECD Environmental Outlook to 2050, OECD Publishing.

OminiSci Inc. (2018). Get Started with OmniSci. https://www.omnisci.com/platform/downloads/. (accessed on 6th November, 2018).

RDA-CODATA Legal Interoperability Interest Group. (2016, October 20). Legal Interoperability of Research Data: Principles and Implementation Guidelines. Zenodo. http://doi.org/10.5281/zenodo.162241

Research Councils UK. (2018). Research council partners and public engagement with research.https://www.ukri.org/public-engagement/research-council-partners-and-public-engagement-with-research/.

Lawrence Rosen (2004). Open Source Licensing: Software Freedom and Intellectual Property Law. Prentice Hall PTR, Upper Saddle River, NJ, USA.

Sanchez, L., Muñoz, L., Galache, J.A., Sotres, P. , Santana, J.R., Gutierrez, V. , Ramdhany, R. , Gluhak, A., Krco, S., Theodoridis, E., Pfisterer, D. (2014). SmartSantander: IoT experimentation over a smart city testbed, Comput. Netw. 61 (C), 217–238. http://dx.doi.org/10.1016/j.bjp.2013.12.020. (accessed on 17th February, 2018).

Soille, P., Burger, A., De Marchi, D., Kempeneers, P., Rodriguez, D., Syrris, V. , and Vasilev, V. (2018). A versatile data-intensive computing platform for information retrieval from big geospatial data. *Future Generation Computer Systems*, 81, 30–40.

Steiniger, S. and Bocher, E. (2008). An overview on current free and open source desktop GIS developments. International Journal of Geographical Information Science, 23, (10), 1345-1370.

Sun, Y. and Thakuriah, P. (2018). An Assessment of Inequalities in Public Transport Availability Using General Transit Feed Specification Data in England and Wales. *Journal of Transport Geography*. (under review).

Thakuriah, P., Dirks, L. and Keita, Y. (2017) Digital infomediaries and civic hacking in emerging urban data initiatives. In: Thakuriah, P., Tilahun, N. and Zellner, M. (eds.) *Seeing Cities Through Big Data: Research Methods and Applications in Urban Informatics*. Series: Springer geography. Springer, NY, pp. 189-207.

Topham, G. (2018). Bus services in 'crisis' as councils cut funding, campaigners warn. https://www.theguardian.com/uk-news/2018/jul/02/bus-services-in-crisis-as-councils-cut-funding-campaigners-warn.

UN - United Nations (2008). World Urbanization Prospects: The 2007 Revision Population Database.http://www.un.org/esa/population/publications/wup2007/2007WUP_Highlights_web.pdf. (accessed on 17th February, 2018).

van Panhuis WG, Paul P, Emerson C, et al. (2014). A systematic review of barriers to data sharing in public health. BMC Public Health, 14, 1144. doi:10.1186/1471-2458-14-1144 pmid: 25377061

Cicirelli, F., Guerrieri, A., Spezzano, G., and Vinci, A. (2017). An edge-based platform for dynamic Smart City applications. *Future Generation Computer Systems*, 76, 106–118.

Zanella, A., Bui, N., Castellani, A. , Vangelista, L., Zorzi, M. (2014). Internet of Things for Smart Cities, IEEE Internet of Things J. 1 (1) , 22–32. http://dx.doi.org/ 10.1109/JIOT.2014.2306328.

Obinna Anejionu is a Scientific Computing Officer/Spatial Data Scientists at the Urban Big Data Centre, with his primary responsibility focusing on the development of the Spatial Urban Data System (SUDS). He obtained B.Sc in Geoinformatics and Surveying from the University of Nigeria Nsukka, M.Sc GIS and Remote Sensing from the University of Greenwich, London, and PhD in Geography (Remote Sensing and GIS applications in environmental monitoring) from the University of Lancaster.

Prior to his current role, he worked at various universities in the UK including Imperial College London as well as at the University of Nigeria. Obinna's current research interest focuses on the integration of geospatial technologies (Geomatics, GIS and remote sensing), big data analytics and cloud technology in solving environmental, socio-political and socioeconomic challenges.

His career has progressed into diverse areas with spokes into GIS, Environmental Remote Sensing, Spatial Analytics, Web Mapping, Cloud Technology, Big Data, Data Science, Machine learning. He has published several findings of his research in various high impact international journals.

Piyushimita (Vonu) Thakuriah is Distinguished Professor of Rutgers University in the Greater New York City Area. Her research focuses on the connections among transportation, society and technology with the view that sustainable and socially-just mobility in cities of the future will require planners to jointly consider social and technological challenges. She is interested in "smart" public transportation, bicycle and pedestrian active transportation, as well as in collaborative/shared transportation systems. She is interested in the potential of technologies to improve livability, learning and engagement in cities and on the effective use of Big Data for urban planning, policy and business innovations.

Prior to her position at the University of Glasgow, she was a professor in the University of Illinois at Chicago and a postdoctoral fellow at the National Institute of Statistical Sciences, both in the US. Vonu has published over 165 journal papers, conference proceedings and technical reports. She has been Principal Investigator of grants of approximately £14.89 million (about $25 million) and has been involved in research grants of over £3.5 million in total (as PI, Co-I or Investigator) in the US and the UK. She served as the Principal Investigator of a series of research grants to study mobility and accessibility outcomes experienced by low-wage workers, persons with disabilities and seniors.

Andrew McHugh is Senior IT and Data Services Manager at the Urban Big Data Centre (UBDC) at the University of Glasgow and is responsible for the development, management and implementation of the data services, data collections and IT strategy of the Centre. Prior to joining UBDC Andrew was the Centre Manager for CREATe, the RCUK Centre for Copyright and New Business Models in the Creative Economy, providing leadership across seven University partners, overseeing several innovations in data development and visualisation, online media and communications and internal project management. Prior to that he established specialisms in issues associated with research data management, digital curation and digital preservation with high profile roles in projects such as the Digital Curation Centre (funded by JISC), 3D Coform (EU FP7) and Planets (EU FP6). Among his most notable contributions were the development of metrics and standards to support the audit and certification of data repositories, including DRAMBORA and ISO 16363. Andrew was awarded an LLB/Hons (Scots Law) in 2000 and MSc in Information Technology in 2001. He completed his PhD in Computing Science in 2016, developing a risk based ontology supporting the preservation of data. Although driven by an interest in technology, Andrew is extremely sensitive to the importance of legal, social and cultural issues that inform IT systems and data, and this is reflected in his career to date and his approach to technical problems.

David Philip McArthur is Lecturer in Transport Studies in the Urban Big Data Centre at the University of Glasgow. He completed his PhD at the Norwegian School of Economics and has since held posts at

The Western Norway University of Applied Sciences, The University of Oslo and the University of Glasgow. He has published 19 peer-reviewed publications papers as journal articles or book chapters. His research interests focus on using data to better manage urban transport; particularly in relation to active travel.

Yeran Sun is a postdoctoral research associate in urban analytics (GIS) at the Urban Big Data Centre, University of Glasgow, United Kingdom. His research interests are GIS, geospatial big data analysis, transport and health. Before he joined the Urban Big Data Centre, Dr Sun completed his PhD in Geography (GIScience) at Heidelberg University, Germany. He received his B.S. degree from Northwest University, China, and his M.S. degree from University of Chinese Academy of Sciences. Both his B.S. and M.S. are in GIS.

Phil Mason is a Research Fellow in the University of Glasgow's School of Education, based in the Urban Big Data Centre, where he has been part of the Educational Disadvantage and Place research team since 2017, researching neighbourhood effects on educational outcomes of Scottish school, further and higher education students. He is a quantitative researcher with broad interests in neighbourhoods and the built environment and their influence on inequalities in life outcomes, education and employment, physical and mental health and positive wellbeing, and health behaviours (especially physical activity). He has a background in pure and applied Evolutionary Biology (PhD from University of East Anglia, Norwich, UK; Postdoctoral Fellow at Universidad Autónoma de Madrid, Spain; Higher Scientific Officer at Ministry of Agriculture, Fisheries and Food, Slough, UK) and Decision Sciences (MSc from University of Westminster, London, UK), but since 2004 he has worked in policy-relevant areas of Social Science research, most notably as the lead statistician on the GoWell Research and Learning Programme, based in Urban Studies at the University of Glasgow.

Rod Walpole is a Scientific Computing Officer at UBDC specialising in GIS and spatial data analysis. His current focus is on helping to set up the core Data Centre and ensure that spatial data sets it contains are managed efficiently and to develop an intuitive web mapping front end to display these resources.

He has been working in this area for the past 20 years on spatial data requirements for a range of organisations including Defra, the Met Office, Cable & Wireless, Natural England, the Environment Agency, European Commission Joint Research Centre and the European Space Agency. He has worked extensively with satellite and aerial remote sensing data in the past and he is an Associate Fellow of the Remote Sensing and Photogrammetry Society.

Authors' Photos

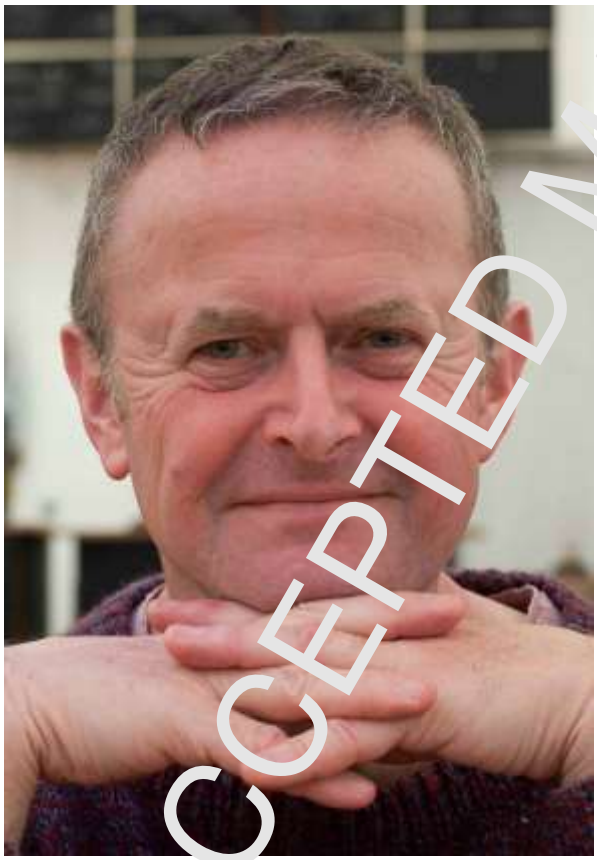
Obinna Anejionu

Piyushimita (Vonu) Thakuriah

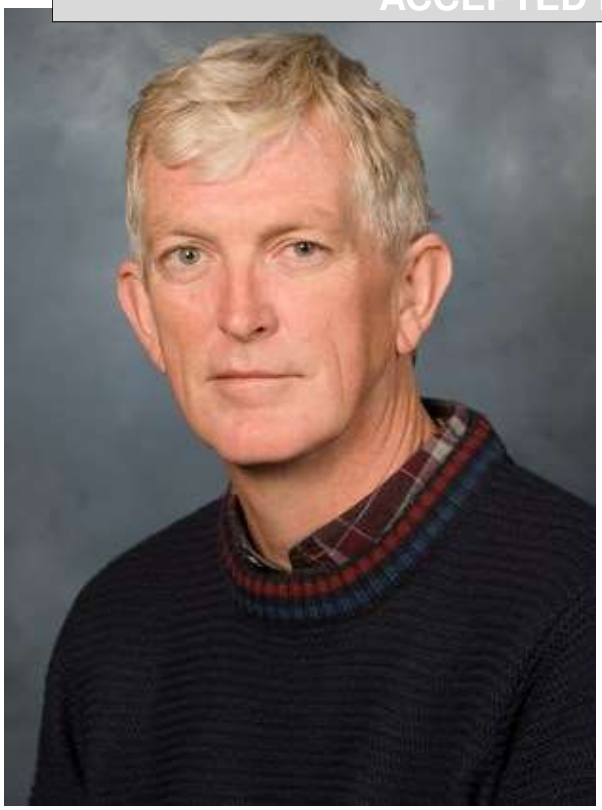
Andrew McHugh

David McArthur

Yeran Sun



Philip Mason



Rod Walpole

Highlights

- A new system for country-wide urban small-area analytics
- Cloud-enabled spatial big data infrastructure for research and policy
- Country-wide social and economic urban analytics capability
- Generation and deployment of unique spatially-activated urban area metrics
- Identification of intra-city variations in key urban services