Integrated Multimedia City Data (iMCD): A composite survey and sensing approach to understanding urban living and mobility

Piyushimita (Vonu) Thakuriaha,⁎,1, Katarzyna Sila-Nowickab, Jinhyun Hongc, Christina Boididou, Michael Osborne, Catherine Lido, Andrew McHugh

a University of Glasgow, Urban Big Data Centre, 7 Lilybank Gardens, Glasgow, G12 8RZ, UK
b University of Auckland, School of Environment, Auckland, New Zealand
c University of Glasgow, School of Social and Political Sciences, Urban Big Data Centre, Glasgow, UK
d University of Glasgow, Urban Big Data Centre, Glasgow, UK
e University of Glasgow, School of Education, Glasgow, UK

ARTICLE INFO

Keywords:
Wearable sensors
Image data
Urban metabolism
Travel behavior
Social media
Smart cities

ABSTRACT

We describe the Integrated Multimedia City Data (iMCD), a data platform involving detailed person-level self-reported and sensed information, with additional Internet, remote sensing, crowdsourced and environmental data sources that measure the wider social, economic and physical context of the participant. Selected aspects of the platform, which covers the Glasgow, UK, city-region, are available to other researchers, and allows knowledge discovery on critical urban living themes, for example in transportation, lifelong learning, sustainable behavior, social cohesion, ways of being in a digital age, and other topics. It further allows research into the technological and methodological aspects of emerging forms of urban data. Key highlights of the platform include a multi-topic household and person-level survey; travel and activity diaries; a privacy and personal device sensitivity survey; a rich set of GPS trajectory data; accelerometer, light intensity and other personal environment sensor data from wearable devices; an image data collection at approximately 5-second resolution of participants’ daily lives; multiple forms of text-based and multimedia Internet data; high resolution satellite and LiDAR data; and data from transportation, weather and air quality sensors. We demonstrate the power of the platform in understanding personal behavior and urban patterns by means of three examples: an examination of the links between mobility and literacy/learning using the household survey, a social media analysis of urban activity patterns, and finally, the degree of physical isolation levels using deep learning algorithms on image data. The analysis highlights the importance of purposefully designed multi-construct and multi-instrument data collection approaches that are driven by theoretical frameworks underpinning complex urban challenges, and the need to link to policy frameworks (e.g., Smart Cities, Future Cities, UNESCO Learning Cities agendas) that have the potential to translate data to impactful decision-making.

1. Introduction

There has been a great deal of recent interest on high-fidelity and timely measurements of the social, economic and functional characteristics of cities. Quantitative urban research traditionally uses surveys, census and other structured forms of data. The explosive growth in digital devices and Internet-based services have fueled the collection of new forms of data. Infrastructure-based or mobile sensors dedicated to collecting specific streams of transportation, environmental or energy data, or new technologies supporting crowdsourcing and user-generated text, image, video, audio content, have either extended the ability to address existing research questions in new ways. For example, while earlier research utilized vehicle-based GPS to monitor traffic and transportation networks (Sen, Thakuriah, Zhu, & Karr, 1997; Sethi, Bhandari, Koppelman, & Schofer, 1995), the proliferation of cellphones and wearable devices that individuals can carry with them, allow passive monitoring of daily activities in minute detail leading to new research questions in mobility, health, and wellbeing (Appelboom et al.,

⁎ Corresponding author.
E-mail addresses: p.thakuriah@rutgers.edu (P.V. Thakuriah), katarzyna.sila-nowicka@auckland.ac.nz (K. Sila-Nowicka), jinhyun.hong@glasgow.ac.uk (J. Hong), christina.boididou@glasgow.ac.uk (C. Boididou), michael.osborne@glasgow.ac.uk (M. Osborne), catherine.lido@glasgow.ac.uk (C. Lido), andrew.mchugh@glasgow.ac.uk (A. McHugh).
1 Present Address: Edward J Bloustein School of Planning and Public Policy, Rutgers University, 33 Livingston Ave, New Brunswick, NJ 08901, USA.
New forms of data have also opened up entirely new research agendas. One interesting strand of wearable device-generated data is automated photographic images produced through lifelogging devices. Whereas large numbers of cities around the world are now extensively instrumented with CCTV cameras yielding video data for traffic management or crime prevention, social media sites such as Flickr, Instagram and even Google Street View which are populated with image user-generated content or street-based photos have opened up new research questions in the areas of event detection, information propagation, mobility and flow detection, and cultural aspects of cities (Chen & Roy, 2009; Hochman & Manovich, 2013; Sun, Fan, Bakillah, & Zipf, 2015; Yin, Soliman, Yin, & Wang, 2017). Recent developments with small portable cameras either integrated in smartphones or in other devices, or as stand-alone portable/wearable devices, generate a wealth of visual lifelogs consisting of unstructured image data on people’s daily lives, that allow detailed analytics of behaviors and movements (Bolanos, Dimiccoli, & Radeva, 2017; Doherty et al., 2011). A key source of data to study urban phenomena in recent years is online social networks, blogs or social media streams (Marti, Serrano-Estrada, & Nolasco-Crugeda, 2017; Paule, Sun, & Thakuriah, 2018; Quercia & Saez, 2014; Stock, 2018; P. Thakuriah, Sila-Nowicka, & Paule, 2016). Among many other applications, social media data has been used for traffic incident detection (D’Andrea, Ducange, Lazerini, & Marcelloni, 2015), crisis management (Rand, Herrmann, Schein, & Vodopivec, 2015), activity patterns (Hasan & Ukkusuri, 2015), and measurement of labor market (Antenucci, Cafarella, Levenstein, Ré, & Shapiro, 2014).

The combination of survey research and new forms of data (multi-mode data) opens up exciting possibilities of exploring a myriad of complex urban phenomena. In some cases, multi-mode data systems may be designed ground-up to collect data on specific constructs, but by using different data gathering techniques, instruments or devices. One example of multi-mode data is household travel surveys and travel diaries using CAPI or similar techniques, where participants also undertake a survey of their travel using GPS devices (Bricka, Sen, Paleti, & Bhat, 2012; Cottrill et al., 2013; Stopher, FitzGerald, & Xu, 2007). Another example is surveys on household and building characteristics that affect energy use and detailed energy use data from smart meters (Kavousian, Rajagopal, & Fischer, 2013). The Understanding Society longitudinal dataset, consisting of a core annual survey on multiple social, economic and lifecycle constructs, as well as a unique biomarker dataset on survey respondents consisting of blood samples and other physical measurements, as well as a short questionnaire on health, medications and so on, is a multi-mode survey allowing novel research on social determinants of health (Buck & McFarll, 2012).

There are several reasons for using a multi-mode approach to studying urban living. First, analyzing multidimensional social, economic or behavioral problems from multiple perspectives have the potential to allow research questions to be raised – and answered - beyond what is possible by analyzing each dataset separately, and to yield a richer, perhaps more complete view of the phenomena under study (Lahat, Adali, & Jutten, 2015). Many new forms of data are also available at a finer spatial resolution, and are timely and available in real-time or close to real-time, as they are continuously and passively generated by sensor systems or actively generated by citizen participants. Such data can potentially provide information on the wider social and economic context to survey responses. The benefits of mixed-mode survey data collection are many, including cost-effectiveness and reduction in coverage bias and nonresponse error (de Leeuw, 2005), and some of these benefits also potentially accrue to situations where data are collected using entirely different systems, not just with different modes in the same survey process. Second, it helps bring together methodological expertise on data from different disciplinary and practitioner communities, e.g., communities with specialized knowledge in social surveys, sensor systems, GIS, social media or other highly unstructured data, which tend to be in different disciplines. This leads to an interdisciplinary approach to data collection and to data that are not too specific to each application. Third, urban data systems tend to be highly siloed as they are held by researchers within specific disciplines or government administrators having specific responsibilities in transportation, housing, environment and so on. Depending on the purpose of the data collection, a multi-mode approach can help address data fragmentation thereby allowing a more comprehensive, broad-based approach to addressing urban challenges relating to sustainability, social justice, resilience and other end goals. Fourth, having different sources of data on the same construct allows going beyond description, monitoring and hypothesis generation, to explanation, inference and understanding of causality in urban phenomenon. Finally, a major benefit is that there is a possibility of collecting ground truth data, to be able to compare and validate data quality in different data streams. While there is no universal definition of ground truth data, and such data are very specific to applications, it is generally considered to be a dataset that allows verifications of the level of accuracy of other datasets or validation of model estimates.

The last point is especially important because there are many challenges to using some of the newer forms of data, a key issue being to understand the nature of measurement errors, selection and other biases, as well as patterns of missingness and sources of uncertainty in new forms of naturally occurring data. While these questions have been extensively studied in the context of survey, administrative and even some new forms of data, how they occur and understanding their implications in the urban context is relatively recent. These new forms of data not only raise the possibility of timely and publicly accessible data that can be used for critical urban functions, but also raise multiple concerns regarding reliability, bias and coverage.

In this paper, we describe the Integrated Multimedia City Data (iMCD), a multi-strand data platform that could be used to examine complex and multidimensional urban issues described above. The design of the iMCD was motivated by research questions in sustainable transportation, healthy cities, lifelong learning, and their interrelationships. Further motivations arose from technological and methodological questions in data access, sharing and analytics. The objective of this paper is to describe the different strands of the dataset and to present three illustrative examples which highlight the potential of the data for urban social and economic research, as well as for urban informatics research.

The paper is organised as follows: in Section 2, we review relevant literature on data collection, and the conceptual and empirical motivations underpinning the data infrastructure. In Section 3, we describe the iMCD infrastructure and provide details on its multiple strands, and a description of how the strands connect to each other. In Section 4 and sub-sections, we provide the four examples: Section 4.1 gives an example of personal travel and literacy; Section 4.2 demonstrates the power of social media data to understand urban dynamics; Section 4.3 utilizes wearable sensor data to identify how much we move indoors as a part of our daily living; and finally, Section 4.4 provides a unique measure of social isolation detected from image data. Conclusions and comments on future research agenda are given in Section 5.

2. The iMCD framework and data infrastructure

The iMCD is motivated by developments in complex multi-strand and multi-modal data collection, to bring together in one data infrastructure, a number of data streams collected through multiple instruments. The data system enables a comprehensive look at transportation and mobility behaviors; education and lifelong learning; sustainable resource consumption and behaviors; social cohesion, cultural values and political preferences; health and wellbeing; and behaviors around use of ICT and digitalisation of our daily lives.

The basic idea of the iMCD draws from the concept of Digital
Mobility Information Infrastructure (DMII) in urban areas proposed by Thakuriah and Geers (2013), with multiple levels of data, from the personal to wider social, economic and infrastructural. The overarching framework is as given by the UN Sustainable Development Goals, particularly Sustainable Cities (United Nations Development Programme, 2018) and developments in the areas of urban data and smart cities (Aguilera, Peña, Belmonte, & López-de-Ipiña, 2017; Thakuriah, Tilahun, & Zellner, 2017). Within this framework, the core academic disciplines – urban transportation, education, computer science, urban economics, and geography and geoinformatics - brought their own conceptual directions to the project. For example, in the transportation component of the iMCD work, we used as a reference point theories and concepts from the travel behavior literature, by bringing in well-known concepts from economic utility theory that explains travel behavior from a rational decision-making perspective (Conlisk, 1988), as well as notions of bounded rationality (Bonsall, Shires, Maule, Matthews, & Beale, 2007; Mahmassani, Jou, Garling, Laitia, & Westin, 1996). Further, perspectives from psychological and behavioral theories such as Theory of Planned Behavior/Reasoned Action, Technology Acceptance Model, Values-Beliefs-Norm Theory (Ajzen, 1991; Davis, 1989; Madden, Ellen, & Ajzen, 1992), helped inform travel survey questions on attitudes and reasons underpinning travel decisions. One motivation was to understand the role of ICT on travel, and literacy, and on environmental, health and other behaviors, as well as for information supporting the development of persuasive technologies, capitolty, augmented mindfulness and quantified self (Fogg, 1998).

In the education and lifelong learning component of the iMCD work, we used as a reference point the framework of UNESCO's Global Learning Cities Network (GLCN), which argues for the mobilisation of resources to promote education across all sectors and environments (UNESCO, 2017), harnessing lifelong learning to promote more equitable and inclusive societies in line with the Sustainable Development Goals (2015). Using theoretical concepts of Social Identity Theory (Hector & Turner Jordan Christopher, 1985) and Social Capital (Bourdieu & Wacuant, 1992), the concepts measured by the iMCD allows us to situate social inclusion within explanatory frameworks which interpret marginalisation in groups, places and in less tangible domains, such as informal learning. Social arguments for embedding learning at the centre of cities connect to the SDG 4, which calls for lifelong learning to be harnessed as a critical factor in the promotion of social cohesion, in increasingly diverse communities.

### 3. iMCD infrastructure strands

The Greater Glasgow region in Scotland including the City of Glasgow is the third largest urban area in the UK, with a population of 1.209 million (599,855 for City of Glasgow), and with eight local authorities. The iMCD platform covers this region, and consists of the following strands of data:

**Strand 1: Participant survey** A primary data collection effort between March 2015 and November 2015 of a stratified random sample of 1511 households in Glasgow and household members (2095 persons who were older than 16 years) who participated in extensive questionnaire-based surveys as well as a personal sensor experiment.

- **Questionnaire-based Household Survey** (HS) – including questions on (1) transportation preferences, (2) energy use and sustainable consumption patterns, (3) ICT use patterns, (4) attitudes and personal preferences, (5) sociodemographic, health, economic and labor market factors, (6) education, cognitive and literacy levels as measured by specific skills-related instruments, and (7) political preferences, civic participation and citizen engagement behaviors;
- **Diary Survey** (DS) – including (1) Multi-day travel patterns captured through a Travel Diary (TD), and (2) daily activities recorded by participants in an Activity Diary (AcD);

**Strand 2: Internet Information Retrieval** (SM): Information retrieval for a period of a year of a variety of internet-based data covering the period of time over which the social survey and the sensor survey were collected; this strand consists of text-based social media data from Twitter and Foursquare and multimedia data such as flickr and images from news sources;

**Strand 3: Remote Sensing Data** (RS): Very high-resolution satellite data and LiDAR to construct a dynamic digital terrain model of Greater Glasgow;

**Strand 4: Sensor Network Data** (SN): Data captured from a wide range of urban sensors, e.g. transportation, emissions, and weather;

**Strand 5: Manually Annotated Database** (DS): Database of manual annotation of image data, consisting of 27,126 images on 12 features.

**Strand 6: Background data from related projects** Analysis using the above iMCD system strands can be augmented by several other data collections held by the Urban Big Data Centre, or by other, externally held data, for which there are special permissions in place. The main sources of background data are:

- **a. ScotEx Ed data**: Special permission was sought from the HS respondents to link their survey data to administrative data held by the Scottish Government;
- **b. Specialized Private Sector Datasets**: e.g. Strava cyclist GPS data, Zoopla housing sale and rental transaction data, and other such data, that could be used with the iMCD system for greater context in studying specific topics;
- **c. Spatial Urban Data System**: A UK-wide synthetic data system was developed by specific members of the project team using a cloud-based GIS approach, of which most relevant to the case of the iMCD in terms of studying transportation access are measures of public transportation availability and quality (Anjejonu et al., 2019);
- **d. Administrative and Open Data**: consisting of a number of existing data sources that can help in benchmarking and quality assessment, including the census, other specialised government-sponsored surveys, as well as administrative data being published by the City of Glasgow through its Open Data Portal.

Details on Strands 1 through 5 are given in Table 1. Much care was taken to ensure that the data are anonymized so as to minimize the potential of re-identifying participants. The data are encrypted and pseudonymized (with identifying data fields suppressed) for general internal usage, and access to the raw data is restricted to a limited number of authorized employees of the data controlling institution, or the institution to which the data collection team belong. Researchers who are a part of the data collection team and appropriately vetted data services employees are able to access the data through a special secure data service to link different strands of the data infrastructure together, using specific pieces of information such as common personal identifiers, as well as geographic location and time.

Linking of data components by external researchers (those outside the authorized team) is a complex process reflecting data protection requirements, safeguarding against potential personal re-identification, and unique licensing agreements associated with some of the data strands. Due to this reason, there is not a one-size-fits-all approach to data linkage among the different strands. The anonymized household survey and its components are available to external researchers upon request. Linkage on the basis of personal identifiers or even timestamps.
that would need to be contacted to meet this target was calculated using interviews in each DZ to be identified. The total number of households were sorted by postcode and a systematic selection of 34 collected with probability proportionate to size, and within PSUs, selected by quintiles of the Scottish Index of Multiple Deprivation (SIMD) which Local Authority (LA), and within each LA, proportionate stratification stage sampling design was used with proportionate stratification by neighbouring DZ until all PSUs had at least 100 households. A multi-
sident household population of less than 100 were merged with a 500 and 1000 household residents. Prior to selection, DZs with a re-
cycling – Strava Metro – are available through the Urban Big Data Centre) identifiers)

may require the involvement of the data services team, and external

Twitter’s own API to retrieve the full content. Results would typically lieu of the tweets themselves. These can then be used in queries to Twitter’s own API to retrieve the full content. Results would typically lieu of the tweets themselves. These can then be used in queries to taken to Twitter’s own API to retrieve the full content. Results would typically be provided in JSON format.

The primary sampling unit (PSU) for the Household Survey was datazones (DZs), small-area statistical geography in Scotland that are groups of 2001 Census output areas and have populations of between 500 and 1000 household residents. Prior to selection, DZs with a resident household population of less than 100 were merged with a neighbouring DZ until all PSUs had at least 100 households. A multi-stage sampling design was used with proportionate stratification by Local Authority (LA), and within each LA, proportionate stratification by quintiles of the Scottish Index of Multiple Deprivation (SIMD) which scores and ranks every small statistical area in Scotland according to a number of measures that are then combined to form an overall rank and measure of deprivation for the area. Within SIMD strata, PSUs are selected with probability proportionate to size, and within PSUs, addresses were sorted by postcode and a systematic selection of 34 addresses from a random start point, yielding a target number of interviews in each DZ to be identified. The total number of households that would need to be contacted to meet this target was calculated using historical data from the Scottish Household Survey, an annual statistical survey of households in Scotland. This allowed the likely response rate for each PSU to be estimated along with the proportion of households likely to be ineligible. The survey was carried out using Computer Aided Personal Interviewing (CAPI).

The questionnaire had three categories of question: Questions asked of the first household to be interviewed (the Highest Income Householder or their spouse or partner, was asked questions about the composition and characteristics of the household.); Questions asked of all respondents (majority of questions were asked of all respondents); and Questions asked of one random adult in the household (a small number of questions – on cultural and civic activities, such as volunteering, political participation and engagement in social activities) - were allocated to only one adult who was randomly selected from all household adults). Before the start of the main fieldwork phase, a week-long pilot of the survey took place, consisting of interviews at 20 households and the CAPI script to be timed and tested in field and to check that all questions were clear and easily understood by respondents. Interviewers also obtained feedback from respondents on the communication materials to be used in the main fieldwork phase, such as the invitation letter to households. The overall household response rate was 51 per cent.

At the end of the survey, all respondents were asked if they would be willing to be re-contacted about any follow-up research, including for linkage of their data in administrative datasets to their survey record; contact details of those that agreed were recorded. About 75 % of people recruited to take part in the questionnaire-based household survey agreed to be re-contacted for any further research-related projects.

Table 1

| Data Strands                        | Time periods and geography                                                                 | Participants & data points                                                                 |
|-------------------------------------|-------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------|
| **Strand 1: Participant Survey**    |                                                                                            |                                                                                            |
| a. Questionnaire-based household survey for Greater Glasgow region | All Strand 1 data collected over the period March 2015 to November 2015 | 2095 persons older than 16 years; 1511 households |
| b. Dairy Surveys                     |                                                                                            |                                                                                            |
| b(1) Travel diary                   | One-day diary                                                                            | 1509 participants                                                                         |
| b(2) Activity diary                 | First day of GPS survey                                                                  | 297 participants who also participated in sensor survey                                    |
| c. Personal sensor survey           |                                                                                            |                                                                                            |
| c(1) GPS device                     | 7-day period; 5-second interval                                                          | 6,433,150 data points from 333 participants                                               |
| c(2) Lifelogging device             | 2-day period; 5-sec average interval                                                     | 470,484 images with location and timestamp                                                 |
| Image data                          | 2-day period; 5-sec average interval                                                     | 470,484 images with location and timestamp                                                 |
| Accelerometer                       | 2-day period; 5-sec interval                                                             | 470,484 observations                                                                     |
| Magnetometer                        | 2-day period; 5-sec interval                                                             | 470,484 observations                                                                     |
| Colour sensor                       | 2-day period; 5-sec interval                                                             | 470,484 observations                                                                     |
| Temperature                         | 2-day period; 5-sec interval                                                             | 470,484 observations                                                                     |
| PIR (Motion detector)               | 2-day period; 5-sec interval                                                             | 470,484 observations                                                                     |
| d. Device preference survey         | End of personal sensor data collection                                                    | 180 participants of all sensing survey participants                                         |
| **Strand 2: Internet Information Retrieval** |                                                                                           |                                                                                            |
| Textual and multimedia data         |                                                                                            |                                                                                            |
| Twitter                             | July 2015 – Nov 2015                                                                     | 65 million geo-enhanced tweets                                                            |
| Foursquare (central Glasgow only)   | Sept 2013 – Oct 2015                                                                    | 456,894 checkins                                                                          |
| Flickr and Facebook public posts on select groups and pages | Sept 2015 – Nov 2015                                                                    | Limited collection                                                                        |
| News in the form of RSS feeds       | Throughout survey period from multiple sources (no sub-regional geographic identifiers) | Region-level data                                                                          |
| **Strand 3: Remote Sensing**        |                                                                                            |                                                                                            |
| Very High-resolution satellite data | Whole Glasgow region (40 km × 40 km) for 2012 only                                        | Region-level data                                                                          |
| LIDAR data                          | Only for 2003                                                                            | Region-level data                                                                          |
| **Strand 4: Sensor Network Data**   |                                                                                            |                                                                                            |
| Air quality monitoring              | Throughout survey period from Scottish Air Quality Monitoring                           | Site and region-level data                                                                  |
| Transportation data                 | Traffic count and incident data (other Glasgow data on public transportation and cycling – Strava Metro - are available through the Urban Big Data Centre) | Site and region-level data                                                                  |
| Weather data                        | Throughout survey period from theWorldWeatherOnline.com, Met Office 3-h forecasts        | Site-level data                                                                            |
| **Strand 5: Manually Annotated Database** | 2-day period; 5-sec average interval                                                     | 27,126 images on 12 features                                                               |
The travel diary, which asked about travel behaviors on the day previous to that of the interview, was administered to all adults. From that, 1509 participants travelled on the day before filling in the Questionnaire-based Household Survey therefore only this number of people filled the travel diary in. It should be noted that the sensing survey, which consisted of providing a subset of respondents with the GPS and lifelogging devices, were not administered on the same day as the travel diary recall day, which is a limitation of the dataset, as, ideally, researchers should be able to compare the self-reported travel movements and activities in the diary with the measured GPS and lifelogging data. Nevertheless, to partially overcome this limitation, the participants were asked to fill in an additional activity diary on the first day of the GPS data collection which may serve as ground truth.

Invitation letters for the sensor survey were sent out to all of the participants. Next, the GPS devices and lifeloggers were delivered to those who agreed to take part in a week-week long survey. GPS movement data were collected using the Transmit 747 ProS GPS tracker, for seven consecutive days with an interval of 5 s. The total number of GPS data collection participants was 333 individuals, generating 6,433,150 GPS data points in total. A total of 223 participants additionally carried the Autographer lifelogging devices in order to collect images over two days with an average time interval of 5 s. Participants were instructed to wear these devices clipped to the front of their chest, as this strategy proved to have the best angle for all types of photo conditions, while allowing the device to be responsive to body movements identifying the direction person was heading.

There are a total of 470,484 images in the iMCD collection. With each of these images is a set of associated sensor readings, because aside from the camera, the lifelogging device has an accelerometer (measuring linear acceleration in X, Y and Z direction including the gravitational force in g), motion detector (detecting movements using ultrared light), magnetometer (ambient geomagnetic field in uT micro Tesla), thermometer (ambient temperature), GPS sensor (location) and a brightness (light intensity and luminosity) detector. The main objective of the sensors is to autonomously determine when the camera should take a picture, but of course the data enables a number of analytics (such as being able to detect whether a person is indoors or outdoors, indoor orientation and other such factors).

The sensing survey participants were also administered a Device Survey (DS) when the fieldwork team collected the devices. The DS offers the possibility of in-depth behavioral research on privacy preferences. This survey queried about participant’s experience in terms of ease and convenience in using the devices, different behaviors by the participant as they knew their actions were being recorded, their privacy concerns in different scenarios, as well as questions on the reactions of others to the participant’s device (for example, did anyone showed an interest and asked questions or asked for the device to be turned off).

A number of Internet-based text and multimedia user-generated data were captured to provide information on events and background context. Among others, Twitter, Foursquare as well as newsfeed data were collected. We gathered 65 million tweets (4.1 Terabytes) in the period from July 2014- November 2015. An academic licence was used to allow us to maximise the number of collected Tweets. All the geo-located tweets in Glasgow were captured using a polygon around Glasgow (bounding box: (-4.3932, 55.7953; -4.0913, 55.9212)). Furthermore, we collected data from certain users such as: @BBCWestScot, @policescotland, as well as Tweets from certain terms or hashtags such as: #gglasgow, #glasgow2014. Total number of 456,894 Foursquare check-ins collected for the project covered slightly different time range: September 2013 – October 2015 and due to the limitations in scraping these data, a smaller bounding box was chosen. Text and multimedia from various news sources such as the BBC, The Scottish Sun or Daily Record were collected for the same period of time.

Additionally, the iMCD platform covers high-resolution satellite and LIDAR data for Glasgow. Remote sensing data covered two types of data: Worldview-1 50 cm panchromatic stereo pair from May 2012 covering Glasgow to create a 3d model of the area (to obtain a 3d model of high accuracy the images have to be taken when there is a cloudless sky) and 1 m LiDAR DSM & DTM data for 263 sq km area covering Glasgow. The majority of data was captured in 2003, with some from 2010. LIDAR DSM/DTM was supplied by Bluesky. We obtained weather data as background information for iMCD project come from two sources. One was the WorldWeatherOnline.com where the information about Glasgow’s weather is in hourly intervals. The data included are: information about temperature, humidity, precipitation, wind direction, pressure etc.). The second source was the Met Office data with interval varying from 0.5 to 1 h and the data were downloaded not only for Glasgow but also for surrounding meteorological stations.

An important derived dataset to help with machine vision algorithms for behavioral modelling of the image data is a significant Manually Annotated Database (MAD). This database consists of manual annotations of 27,126 images on 12 features, with the total number of images being 470,484. In order to determine a representative sample of participants for the manual annotation, we applied a k-mode clustering algorithm on the whole dataset from the social survey (2095 individuals from which 1509 filled in a travel diary). The optimal number of clusters derived via analyzing a dendrogram was 9, representing therefore 9 groups of people. Using this number, we chose representatives for each of the clusters. Details of this process are given in Sila-Nowicka and Thakuriah (2019). The number of participants from a cluster varied from 1 to 2 participants depending on the number of collected in total images.

The images in the MAD had 12 annotated features for each of the images: mode (e.g. driving on a bus, on a train, etc.), activity mode (standing, sitting, walking, etc.), location (home, car, etc.), location category and group (from points of interest categories from Ordnance Survey POI database), number of people, number of faces, number of people in close proximity, number of faces in close proximity, whether a person is indoor or outdoor, type of activity (cooking etc.) and use of ICT. Moreover, these images have a set of the aforementioned sensor readings linked to them.

4. iMCD research agenda

The iMCD opens an interdisciplinary research agenda. First, it supports urban planning and policy research, as well as social, behavioral and economic research on a number of urban living themes. Second, it supports research into the technological and methodological challenges associated with novel forms of data, and the analytics involved in urban informatics and knowledge discovery. The research agenda is given in Table 2.

4.1. Social and economic policy and urban planning research on urban living themes

The iMCD is intended to support urban policy, planning, and operations research, as well as social and economic research on multiple urban living themes (column 1 of Table 2, expanded in Table 3). We will briefly discuss two of these urban living themes next to give a flavour of the potential use of iMCD research. Others are elaborated to a greater degree in the three detailed examples in Section 4.

Urban Contexts and Neighborhoods: The system allows empirical investigation of social, economic, built environment and behavioral characteristics of neighborhoods, and their relationships to employment, health, education, and other outcomes. Neighborhoods can be characterized in terms of the types of activities undertaken by residents, their perceptions and sentiments, degree of access to services, and social cohesion and sense of belonging. These factors can be examined by analyzing data from several strands of the iMCD: the household survey (HS), travel diary (TD), activity diary (AcD), social media (SM), GPS and lifelogging (LL) data sources.
While the travel diary, activity diary, GPS and lifelogging (primarily image) data allow inferring actual behaviors, the household survey has a large number of questions relating to neighbourhood perceptions (for example, ratings of the neighbourhood as a place to live in, likes and dislikes about the neighborhood). Other research topics that are of potential interest are the level of social cohesion, and safety and sense of belonging primarily based on the HS data, but also through novel uses of SM, GPS and LL data. In contrast to the travel and activity diaries where respondents have self-reported how or when they traveled or which activities they have participated in, the GPS data are simply a report of timestamps and locations from which movements, activities and so on can be inferred using additional processing model. Further, the image data can be mined to, for example, help understand actual behaviors. It should be noted that whereas GPS data allows us to capture primarily outdoor activity, the images and the associated sensor readings, if appropriately modelled, would allow understanding indoor behaviors.

Additionally, social media (for example, Twitter data) allows topic modelling of concepts captured in the Twitter stream generated within neighborhood or about a neighborhood, as well as sentiment analysis of people’s positive, negative or neutral attitudes and opinions about places and neighbourhoods. In the iMCD, the built and physical environment of neighborhoods can be determined from remote sensing data (RS), sensor networks (SN, like traffic flow measurements), LL image data, as well as through Points of Interest databases). Further, resource availability within a neighborhood and access to services can be obtained through the Spatial Urban Data System (SUDS), administrative data (AD), and the activity and travel diary, where the purpose of the activity or trip are self-reported, pointing to the availability of social and economic opportunities in places.

Civic participation, urban engagement, volunteering: The iMCD supports research on urban engagement, and civic and cultural orientations. For example, the household survey items on the extent and type of volunteering and civic activity, together with sociodemographic information, can be used to derive indicators of community activism, and prosocial orientations. We have also listed under the theme “Political preferences and electoral participation” that the HS asks respondents to self-report extent of political awareness and extent and source of knowledge of politics – these survey items could also be potentially useful in determining political engagement levels. Also, as listed under “Sustainable Behaviors and Resource Consumption”, the HS queries on respondents’ awareness of sustainability issues, resource use impacts, and in addition, asks questions about biodiversity and measures environmental literacy. Together, these themes can help determine the extent of engagement on sustainability and community wellbeing, as well as the degree and extent of face-to-face interactions with neighbors. At an aggregate level, opinions and attitudes regarding urban issues, and engagement in civic or political issues can be determined from Twitter and other social and newsfeed data. The survey further asks questions on the extent and type of caregiving. Furthermore, the activity diary picks up on caregiving locations, durations and timing. The type and extent of personal services such as being with a child, or chauffeuring children, or other activities towards the care of children, for example, can also potentially be detected from image data, provided the appropriate training models can be developed.

4.2. Analytics research

A key motivator to the iMCD is to provide a platform that allows in-depth investigation into data science research questions. Examples are: (1) Urban Informatics and Knowledge Discovery; (2) Data governance and sharing; (3) Information management; (4) Information Retrieval; (5) Image Processing; (6) Privacy and security research; (7) GISCience; (8) Human Computer Interaction; (9) Visualisation; (10) Data Standards; (11) Understanding epistemological and political economy challenges of data. Here we touch on two of these topics, with further examples in the case study section.

Urban Informatics and Knowledge Discovery: While there are many definitions of Urban Informatics, one perspective highlights data-driven approaches using machine learning and other data science methods on novel forms of data and focusing to a greater degree on certain urban themes. Several strands of iMCD are particularly well-suited to be tagged as interesting for UI and KE research. Examples are the GPS, image and image sensor data, social media and sensor network data, towards: (1) dynamic resource management: developing data-driven strategies to manage scarce urban resources effectively and efficiently and often supporting real-time decisions regarding competitive use of transportation, energy, environmental and related resources; (2) Knowledge discovery and understanding: discovering patterns and relationships in urban rhythms, interactions and disruptions; mobility and travel; digital economy and use of ICT, and developing explanations for such trends; (3) Urban engagement and civic participation: developing practices, technologies and other processes needed for an informed citi- zenship and for their effective involvement in social and civic life of cities; (4) Urban planning and policy analysis: developing robust data-driven approaches for urban planning, service delivery, policy evaluation and reform, and also for the infrastructure and urban design decisions(Thakuriah et al., 2017). While many of these research themes have been explored with traditional forms of data, the informatics approaches tend to focus on methods that address the novelty of the data and the situational contexts.

Privacy research: The highly personal nature of the data streams requires special consideration of privacy and anonymization strategies, both as individual data strands, as well as when the strands are linked.
| Table 3: Grid of Urban Living Themes and Research. |
|--------------------------------------------------|
| **Urban Context and Neighborhoods** | **Sustainable Behaviours and Resource Consumption** |
| **Neighbourhood activity & perceptions** | **Activities patterns** | **Energy efficiency** | **Energy costs** | **Energy use** |
| **Built environment** | **Social cohesion** | **Sustainable behaviours and resource consumption** | **Family, household assets, living conditions** | **Civic Participation, Urban Engagement, Volunteering** |
| **Safety & sense of belonging** | **Resource availability and use** | **Socio-demographics, race, ethnicity, immigration** | **Physical status, physical activity levels, health & wellbeing** | **Income, Wages, Benefits and Assistance** |
| **Neighbourhood activity & perceptions** | **Travel behavior** | **Labour Market, Employment & work conditions** | **Time use** | **Healthcare perceptions** |
| **Type of device used** | **Time spent on different activities** | **Cultural and leisure activities** | **Physical status, physical activity levels, health & wellbeing** | **Family, household assets, living conditions** |
| **Planning and movement of flows** | **Income from sources other than work** | **Country of origin** | **Income, Wages, Benefits and Assistance** | **Healthcare perceptions** |
| **Latitude and use of space** | **Income** | **Age, gender** | **Time use** | **Healthcare perceptions** |
| **IMD Information Strands Supporting Research** | **Income** | **Self-reported effects of disability** | **Dependency in activities of daily living** | **Healthcare perceptions** |
| **Socio-demographics, race, ethnicity, immigration** | **Labour Market, Employment & work conditions** | **Country of origin** | **Social status, physical activity levels, health & wellbeing** | **Healthcare perceptions** |
| **Physical status, physical activity levels, health & wellbeing** | **Time use** | **Income from sources other than work** | **Healthcare perceptions** | **Family, household assets, living conditions** |
| **Income, Wages, Benefits and Assistance** | **Planning and movement of flows** | **Family, household assets, living conditions** | **Healthcare perceptions** | **Family, household assets, living conditions** |
| **Healthcare perceptions** | **Labour Market, Employment & work conditions** | **Country of origin** | **Social status, physical activity levels, health & wellbeing** | **Family, household assets, living conditions** |
| **Time use** | **Planning and movement of flows** | **Family, household assets, living conditions** | **Healthcare perceptions** | **Family, household assets, living conditions** |
| **Healthcare perceptions** | **Labour Market, Employment & work conditions** | **Country of origin** | **Social status, physical activity levels, health & wellbeing** | **Family, household assets, living conditions** |
| **Time use** | **Planning and movement of flows** | **Family, household assets, living conditions** | **Healthcare perceptions** | **Family, household assets, living conditions** |
| **Healthcare perceptions** | **Labour Market, Employment & work conditions** | **Country of origin** | **Social status, physical activity levels, health & wellbeing** | **Family, household assets, living conditions** |
| **Time use** | **Planning and movement of flows** | **Family, household assets, living conditions** | **Healthcare perceptions** | **Family, household assets, living conditions** |
| **Healthcare perceptions** | **Labour Market, Employment & work conditions** | **Country of origin** | **Social status, physical activity levels, health & wellbeing** | **Family, household assets, living conditions** |
| **Time use** | **Planning and movement of flows** | **Family, household assets, living conditions** | **Healthcare perceptions** | **Family, household assets, living conditions** |
| **Healthcare perceptions** | **Labour Market, Employment & work conditions** | **Country of origin** | **Social status, physical activity levels, health & wellbeing** | **Family, household assets, living conditions** |
| **Time use** | **Planning and movement of flows** | **Family, household assets, living conditions** | **Healthcare perceptions** | **Family, household assets, living conditions** |
| **Healthcare perceptions** | **Labour Market, Employment & work conditions** | **Country of origin** | **Social status, physical activity levels, health & wellbeing** | **Family, household assets, living conditions** |
| **Time use** | **Planning and movement of flows** | **Family, household assets, living conditions** | **Healthcare perceptions** | **Family, household assets, living conditions** |
| **Healthcare perceptions** | **Labour Market, Employment & work conditions** | **Country of origin** | **Social status, physical activity levels, health & wellbeing** | **Family, household assets, living conditions** |
| **Time use** | **Planning and movement of flows** | **Family, household assets, living conditions** | **Healthcare perceptions** | **Family, household assets, living conditions** |
| **Healthcare perceptions** | **Labour Market, Employment & work conditions** | **Country of origin** | **Social status, physical activity levels, health & wellbeing** | **Family, household assets, living conditions** |
| **Time use** | **Planning and movement of flows** | **Family, household assets, living conditions** | **Healthcare perceptions** | **Family, household assets, living conditions** |
| **Healthcare perceptions** | **Labour Market, Employment & work conditions** | **Country of origin** | **Social status, physical activity levels, health & wellbeing** | **Family, household assets, living conditions** |
| **Time use** | **Planning and movement of flows** | **Family, household assets, living conditions** | **Healthcare perceptions** | **Family, household assets, living conditions** |
| **Healthcare perceptions** | **Labour Market, Employment & work conditions** | **Country of origin** | **Social status, physical activity levels, health & wellbeing** | **Family, household assets, living conditions** |
| **Time use** | **Planning and movement of flows** | **Family, household assets, living conditions** | **Healthcare perceptions** | **Family, household assets, living conditions** |
| **Healthcare perceptions** | **Labour Market, Employment & work conditions** | **Country of origin** | **Social status, physical activity levels, health & wellbeing** | **Family, household assets, living conditions** |
| **Time use** | **Planning and movement of flows** | **Family, household assets, living conditions** | **Healthcare perceptions** | **Family, household assets, living conditions** |
| **Healthcare perceptions** | **Labour Market, Employment & work conditions** | **Country of origin** | **Social status, physical activity levels, health & wellbeing** | **Family, household assets, living conditions** |
| **Time use** | **Planning and movement of flows** | **Family, household assets, living conditions** | **Healthcare perceptions** | **Family, household assets, living conditions** |
| **Healthcare perceptions** | **Labour Market, Employment & work conditions** | **Country of origin** | **Social status, physical activity levels, health & wellbeing** | **Family, household assets, living conditions** |
| **Time use** | **Planning and movement of flows** | **Family, household assets, living conditions** | **Healthcare perceptions** | **Family, household assets, living conditions** |
| **Healthcare perceptions** | **Labour Market, Employment & work conditions** | **Country of origin** | **Social status, physical activity levels, health & wellbeing** | **Family, household assets, living conditions** |
| **Time use** | **Planning and movement of flows** | **Family, household assets, living conditions** | **Healthcare perceptions** | **Family, household assets, living conditions** |
| **Healthcare perceptions** | **Labour Market, Employment & work conditions** | **Country of origin** | **Social status, physical activity levels, health & wellbeing** | **Family, household assets, living conditions** |
| **Time use** | **Planning and movement of flows** | **Family, household assets, living conditions** | **Healthcare perceptions** | **Family, household assets, living conditions** |
| **Healthcare perceptions** | **Labour Market, Employment & work conditions** | **Country of origin** | **Social status, physical activity levels, health & wellbeing** | **Family, household assets, living conditions** |
| **Time use** | **Planning and movement of flows** | **Family, household assets, living conditions** | **Healthcare perceptions** | **Family, household assets, living conditions** |
| **Healthcare perceptions** | **Labour Market, Employment & work conditions** | **Country of origin** | **Social status, physical activity levels, health & wellbeing** | **Family, household assets, living conditions** |
| **Time use** | **Planning and movement of flows** | **Family, household assets, living conditions** | **Healthcare perceptions** | **Family, household assets, living conditions** |
| **Healthcare perceptions** | **Labour Market, Employment & work conditions** | **Country of origin** | **Social status, physical activity levels, health & wellbeing** | **Family, household assets, living conditions** |
| **Time use** | **Planning and movement of flows** | **Family, household assets, living conditions** | **Healthcare perceptions** | **Family, household assets, living conditions** |
| **Healthcare perceptions** | **Labour Market, Employment & work conditions** | **Country of origin** | **Social status, physical activity levels, health & wellbeing** | **Family, household assets, living conditions** |
| **Time use** | **Planning and movement of flows** | **Family, household assets, living conditions** | **Healthcare perceptions** | **Family, household assets, living conditions** |
| **Healthcare perceptions** | **Labour Market, Employment & work conditions** | **Country of origin** | **Social status, physical activity levels, health & wellbeing** | **Family, household assets, living conditions** |
| **Time use** | **Planning and movement of flows** | **Family, household assets, living conditions** | **Healthcare perceptions** | **Family, household assets, living conditions** |
| **Healthcare perceptions** | **Labour Market, Employment & work conditions** | **Country of origin** | **Social status, physical activity levels, health & wellbeing** | **Family, household assets, living conditions** |
| **Time use** | **Planning and movement of flows** | **Family, household assets, living conditions** | **Healthcare perceptions** | **Family, household assets, living conditions** |
| **Healthcare perceptions** | **Labour Market, Employment & work conditions** | **Country of origin** | **Social status, physical activity levels, health & wellbeing** | **Family, household assets, living conditions** |
| **Time use** | **Planning and movement of flows** | **Family, household assets, living conditions** | **Healthcare perceptions** | **Family, household assets, living conditions** |
| **Healthcare perceptions** | **Labour Market, Employment & work conditions** | **Country of origin** | **Social status, physical activity levels, health & wellbeing** | **Family, household assets, living conditions** |
| **Time use** | **Planning and movement of flows** | **Family, household assets, living conditions** | **Healthcare perceptions** | **Family, household assets, living conditions** |
| **Healthcare perceptions** | **Labour Market, Employment & work conditions** | **Country of origin** | **Social status, physical activity levels, health & wellbeing** | **Family, household assets, living conditions** |
| **Time use** | **Planning and movement of flows** | **Family, household assets, living conditions** | **Healthcare perceptions** | **Family, household assets, living conditions** |
| **Healthcare perceptions** | **Labour Market, Employment & work conditions** | **Country of origin** | **Social status, physical activity levels, health & wellbeing** | **Family, household assets, living conditions** |
| **Time use** | **Planning and movement of flows** | **Family, household assets, living conditions** | **Healthcare perceptions** | **Family, household assets, living conditions** |
| **Healthcare perceptions** | **Labour Market, Employment & work conditions** | **Country of origin** | **Social status, physical activity levels, health & wellbeing** | **Family, household assets, living conditions** |
| **Time use** | **Planning and movement of flows** | **Family, household assets, living conditions** | **Healthcare perceptions** | **Family, household assets, living conditions** |
| **Healthcare perceptions** | **Labour Market, Employment & work conditions** | **Country of origin** | **Social status, physical activity levels, health & wellbeing** | **Family, household assets, living conditions** |
| **Time use** | **Planning and movement of flows** | **Family, household assets, living conditions** | **Healthcare perceptions** | **Family, household assets, living conditions** |
Many of the data strands individually pose the risk of de-identification even from anonymized data. For example, blurring the faces of participants in an image may help to anonymize persons in the image data, and face detection for the purpose of blurring is an active area of research. One of our approaches to face detection using Convolutional Neural Network (CNN) is described in Section 5.3; other approaches to face detection that is under testing include Haar Cascade Classifiers (Viola & Jones, 2001). At the very least, human faces and license plate numbers should be anonymized by default if the data sets are to be disclosed to the third parties, which agrees with Google Street View’s default privacy policy completely. Our privacy work to date has focused on this problem. Yet, a focus group held by the researchers revealed that the characteristics of the home or office where the image is taken, together with the clothing that the person is wearing, or perhaps even the presence of a pet, can increase the possibility of uniquely re-identifying a specific individual. Hence a decision was made that even with face and license plate blurring, applications for access to the image data by external researchers would require stringent assessment by an independent assessor’s group, and the images themselves must under all circumstances be available only from the secure data service without the possibility of making copies.

As data across Strands 1 and 2 are linked, the risk of re-identification increases, as described in Sila-Nowicka and Thakuriah (2016). For example, non-spatial quasi-identifiers such as ethnicity, job description or disability status that are available from the household survey (which, say, is anonymized with personal identifiers such as participant ID, name, and residential address removed), when combined with the characteristics of travel and locations frequently visited, as inferred from GPS trajectories, could increase the risk of uniquely identifying individuals. An example would be: if one knew from the survey that the ethnicity of a person is Asian, and that her/his age is, say, 20–25 years, and we can infer from her GPS trajectories where they live or which type of Asian grocery store they frequently visit, then one may be able to uniquely identify a person. In such a case, it may become possible to add additional preferences regarding their shopping habits and other movement patterns, which is information that the survey participant did not originally consent to providing data on.

From a data publication perspective and for the potential of analytics using individual data, we have conducted experiments to present a trade-off between data privacy and the resolution of the GPS-based movement data. GPS trajectory data undergo spatial cloaking of significant locations (i.e., locations classified as home, significant third places, and other locations), and grid-based masking with cell sizes that are a minimum of 100 m which gave good results for the anonymization of GPS data and preservation of spatial patterns of travel.

Aside from the above discussion, which is about privacy and data publication, the dataset offers many ways to explore fundamental questions in privacy-preserving algorithms. The GPS movement data and the lifelogging data, particularly the images, provide interesting avenues for privacy research, particularly the study of techniques to prevent the unique re-identification of persons in one dataset based on available information on those persons in other datasets. One example is the need to design ways to avoid re-identification of socio-demographic information that would usually be anonymized in the survey data as per the Data Protection Act, based on trajectory data. An example is the potential identification of the sexual orientation of respondents based on identification of Points of Interest and places visited, as inferred from the GPS trajectory or image data. The GPS data archive requires the use of a number of data privacy techniques to ensure that the home work, or other significant locations of respondents cannot be re-identified. The image data archive provides a rich collection of images in highly noisy environments, with implications for image detection research. Indoor images in particular pose specific challenges especially when persons are in proximity of other objects that may uniquely identify them, even if their faces or bodies in the images are blurred to anonymize them.
5. Illustrative examples of use

In this section, we present three examples of the use of the iMCD system towards the research agenda discussed in Section 4. In Section 5.1, using the household survey, we discuss how literacy and numeracy affects people’s travel behavior choices. In Section 5.2, using social media and other data, we provide an assessment of urban metabolism in the City of Glasgow and spatiotemporal patterns in the predominant use of space. In Section 5.3, we present the results of machine vision algorithms to determine the degree of physical isolation and the links to people’s living conditions.

5.1. Mobility and Literacy: utilizing the iMCD household survey

The iMCD offers different strands of data to study transportation behavior, and the effect of learning and literacy on travel. The household survey, which covers these concepts, allows external researchers to explore these questions without any linkage. It has been well documented that the lack of basic skills such as literacy and numeracy is associated with the high level of unemployment as well as social exclusion (Bynner & Parsons, 2001). In addition, people with a lower level of basic skills are more likely to have semi-skilled or unskilled jobs (National Research and Development Centre for adult literacy and numeracy, 2005). Social exclusion is a complicated and multi-dimensional phenomenon, and several empirical studies have examined its impacts on travel behavior. For example, Stanley and Stanley (2014) showed that mobility (e.g., number of trips) is directly associated with the risk of social exclusion and indirectly with well-being. That means socially excluded people will have a narrower travel horizon and fewer daily trips than those who are socially included, having a lower level of well-being. Moreover, Pro Bono Economics (2014) indicated that the lack of numeracy is related to health outcomes and personal and social skills (e.g., self-esteem and confidence), which are closely linked with travel behavior (Ellaway, Macintyre, Hiscock, & Kearns, 2003).

As shown, previous literature has implied the potential link between basic skills and travel behavior. However, empirical studies are scarce, mainly due to data limitation. The iMCD survey includes several questions about literacy, numeracy, and financial literacy. These questions are constructed based on the definitions from the International Adult Education Survey and OECD practices. For example, the survey asks how confident the interviewee is in using maths in everyday life. The answer was measured at a 4-likert scale, anchored by not at all confident and very confident. Financial literacy involves actual calculations about financial issues. Specifically, two calculation questions about the savings with interest rates and inflation were asked to all interviewees. In addition, a question about the risk management (i.e., “Which is the riskier asset to invest in?”) was asked.

In this example, we focused on numeracy and financial literacy because: 1) both numeracy and financial literacy involve calculation and are linked closely; and 2) those skills influence travel decisions such as trip-making, auto-ownership, scheduling and time/fare calculations. Specifically, we examined how numeracy and financial literacy are correlated with total travel distances calculated based on a one-day travel diary included in the iMCD survey. Based on the results from previous relevant literature, we assumed that people who have higher levels of numeracy and financial literacy will have longer travel distances. Results from the linear regression are shown in Table 4.

Since there are only a small number of observations who said “not at all confident” for the numeracy question (only 36 people among 1426 observations), we combined interviews who said “not at all confident” or “not very confident”. Moreover, we took a log-transformation of the total travel distance variable because it is highly skewed. The control variables (i.e., socio-demographic factors) show the consistent results with previous studies. Our model shows that people who are very confident in using math have longer total travel distances than those who are not at all or not very confident. This result confirms our hypothesis. However, this association is only marginally significant (p-value: 0.099). Financial literacy shows positive and significant associations with total travel distances. Specifically, people who answered correctly for two among the three questions have longer travel distances (i.e., 27 %) than those who did not answer correctly for all three questions. Moreover, those who have all correct answers have longer total travel distances (i.e., 44 %) than those who did not answer correctly for all questions, and its association is significant at the 0.01 level of significance. This implies that people who have higher levels of calculation and financial literacy travel longer distances than those who lack financial literacy. This result shows the importance of education in numeracy and financial literacy for young school-age children to improve their travel horizon and potentially, economic well-being.

### Table 4

| Parameter | Estimate |
|-----------|----------|
| Intercept | 1.498 *** |
| Socio-demographics | | |
| Age | −0.004 ** |
| Gender (Male = 1) | −0.049 |
| Household size | 0.002 |
| Annual income (£) | 0.000 * |
| Education (Higher education = 1) | 0.192 ** |
| Work (work = 1) | 0.431 *** |
| Numeracy (Ref: Not at all + not very confident) | | |
| Quite confident | 0.100 |
| Very confident | 0.200 * |
| Financial literacy (Ref: no correct answer) | | |
| One correct answer | 0.048 |
| Two correct answers | 0.269 ** |
| Three correct answers | 0.444 *** |
| Adjusted-R² | 0.086 |
| Sample size | 1426 |

*** Significant at .01 level.  
** Significant at .05 level.  
* Significant at .10 level.

5.2. Analyzing urban metabolism through social media and GPS data

Monitoring neighbourhood characteristics is important for urban governance. Citizen concerns around safety, neighbourhood satisfaction, gentrification, sentiment towards public resources, as well changes in land-use, are only some of the changes for which up-to-date information is useful. Urban indicators are one approach to track changes at the local level (Albino, Berardi, & Dangelico, 2015; Berardi, 2013; Huovila, Airaksinen, Pinto-Seppi, Piira, & Penttinen, 2016). The goal of such indicators, particularly at the neighbourhood level, are to facilitate performance monitoring, assess trends over time, set future targets and even to support inter-city comparisons. They also have the potential to inform urban planning and operations, various decision-making regarding urban management, raise awareness on critical issues, encourage political interventions and citizen activism, and to support strategies for health behavior and public engagement, and to improve communication among stakeholders working in urban sectoral silos.

Emerging forms of data can be utilized to detect urban spatio-temporal structures in terms of the type of uses of urban spaces, and to track changes to such structures. In an exploratory analysis of spatio-temporal use, Fig. 1 shows different uses of the central area of the City of Glasgow, in terms of degree of social and functional use, as determined by a mix of data. We define urban social space to mean physical spaces exhibiting spatial and temporal patterns reflecting out-of-home social interaction built on networks or other face-to-face interactions, and activities associated with recreation, entertainment and cultural events, while we use the term urban functional space to mean...
Fig. 1. Social and functional activities from Twitter, Foursquare and commuting patterns.
physical spaces with a high degree of work-related, economic and utilitarian use sources (Lathi, Thierstein, & Goebel, 2010; Zhang & Sun, 2016).

As discussed in Thakuriah et al. (2016), we determine where in the social-functional spectrum an area within central Glasgow lies, by combining information from Twitter, Foursquare and commuting patterns from the 2011 UK census. Subject to the constraints around sharing Twitter data as discussed previously, this type of analytics should be possible by external researchers. Higher levels on Twitter and Foursquare and lower levels of census commuting inflows lead to grids being tagged as purely social (Cluster A. High Social–Low Functional (H–L)) and higher levels of inflows and low levels of social activity lead to grids being tagged as purely functional (Cluster D. Low Social–High Functional (L–H)). Ranking high on both dimensions, i.e., both social media as well as inflows, lead to a categorization of Cluster B (High Social–High Functional or H–H), whereas ranking low in both dimensions led to C (Low Social-Low Functional or L–L).

The map shows that areas are predominantly mixed (C) or residential or empty – there are fewer places that are detected as being purely functional or social. Mixed development patterns overlap between functional space and social space, and except for purely residential areas, there are few areas that are purely mono-social or mono-functional. Further, mixed land-use has been an important urban planning goal, and areas which are largely residential have been noted to exhibit social uses. Hence there is increased likelihood of functional and social activities occurring in residential areas, implying that multiple uses occur throughout the day.

Fig. 2 shows this temporal dimension to spaces and is a time-varying extension of Fig. 1, showing the degree of social and functional uses throughout the course of the day. There are parts of the city which remain dominantly residential/empty, functional or mixed throughout the course of the day. Yet, there are areas within the city which are primarily functional or residential in the earlier part of the day, but which turns more social during the evening hours. These activity transition zones are primarily in the centre and eastern parts of the city, as predominant activities change from work or home-based activities to leisure and social activities. Moreover, more areas are labelled as being social in the later hours of the day compared to morning and midday periods.

The activity diary in the iMCD is a rich set of annotations of what people consider to be “social” and “recreational” throughout the day and serves to contextualize the social media analysis. The activity diary data agrees with the aforementioned social data that the overall volume of social activities peak in the evening hours (after 3 PM). Further, annotations of what respondents consider to be social activity during the daytime and in the post-6pm time periods are different. For example, a greater share of social activities during the morning and daytime are social visits to friends and family, the majority (61%) of which take place in outside public and commercial venues, but many (33%) also take place in private, out-of-home places such care homes, schools, churches, and so on. Visits as a share of social activities decline after 3 pm, when social activities are undertaken to a higher degree within the home or in public or commercial places. This explains to a certain extent the increase in social uses seen in Fig. 2 in areas that are primarily residential during the day. The activity diary also shows that a small proportion of activities after 6 PM, which persist to late hours, are work activities. Fig. 2 shows that the 24/7 economy is predominant throughout the day in the west and central parts of the city.

Emerging sources of data therefore provide a way to detect patterns in cities that are important for the efficient management of urban resources. The patterns which we focused on have implications for future resource consumption needs, and in the provision of services. These are primarily spatial patterns in urban form and structure in terms of where and how people socialize, and how and when they work, as well as time-varying diurnal patterns in human concentrations and activity patterns during the course of the day.

5.3. Detecting isolation from personal image data

Our objective with the third case study is to demonstrate the potential of the iMCD image archive to derive high-fidelity understanding of people’s daily lives in ways that are not possible with survey, GPS or other data. The specific example we show here is a unique measure of isolation, which gives the extent to which a person is alone during the course of a day. Social isolation has been noted to lead to mental health challenges among younger people (Matthews et al., 2015) as well as among seniors (Courtin & Knapp, 2017), and also among specific occupational groups (for example, Apostolopoulos, Sönmez, Hege, & Lemke, 2016, for the case of long-haul truckers). Various notions of isolation and loneliness in urban areas (“urban isolation”) are acknowledged to be risk factors for mental and physical health, and have been linked to powerlessness, inequality, alienation, and urban reduced quality of life (Klinenberg, 2001; Leigh-Hunt et al., 2017; Tigges, Browne, & Green, 1998).

While there are surveys and scales designed to measure social isolation, and while time-use diaries, and activity and travel diaries query respondents to self-report how much time a participant is spending alone, self-reports can be biased as respondents may be unable to correctly recall how much time they were alone or the degree of isolation in terms of the level of contact with others. Further, the true extent of isolation, when there is not only lack of interaction with others such as family, friends, acquaintances and co-workers, but when they are in physically isolated situation, and not surrounded by anyone else, not even by strangers, may be very difficult to report.

In this section, we describe a measure of isolation that gives the degree of visible interaction with others either in one-on-one or in a group. The measure fundamentally depends on counting the number of persons and faces in the image, with less socially isolated people having few people or faces in the image and more social persons having greater counts of persons/faces. We then cluster participants to physical isolation groups that are ranked according to the degree of isolation determined by co-detection of persons and faces.

Fig. 3 shows several images from indoor settings. We determine isolation group membership for participants based on such images, using deep learning machine vision algorithms, Convolutional Neural Networks (Krizhevsky, Sutskever, & Hinton, 2017) and TensorFlow (Abadi et al., 2015). While machine vision problems have been extensively studied, and examples of routine applications abound, the unique characteristics of the iMCD image data are the highly variable quality due to a mix of sensor and participant-related or external data quality factors.

Among sensor-related factors, the determination of when an image should be captured is based autonomously by sensor readings on orientation and lighting quality, resulting in uneven time intervals between images or no images taken when light levels are lower than the sensor's threshold values. Further, the device may turn itself off due to low battery, similar to the case of GPS devices. Participant factors in terms of the respondents actions can lead to variable quality in a number of ways: a far from complete set of examples are camera “dead time” when the participant has taken off the lifelogging device, or when they wore the device incorrectly or not as instructed resulting in the same field of view in consecutive images, and due to short recording duration over the course of a day because the participant turned off the device due to privacy concerns, and forgot to turn it back on.

Among external factors, indoor environments pose a particular problem with the image detection algorithms, particularly in situations where there are many other conflicting objects. In the outdoor context, weather particularly precipitation is a unique problem. Because the PIR sensors are very sensitive, high-quality images are difficult at nighttime, and one limitation of our dataset is that there are limited nighttime images. Using k-means clustering, we clustered participants on the basis of three sets of features: (1) percentage of images with 0, 1, 2, 3 or more faces; (2) percentage of images with 0, 1, 2, 3 or more persons;
Fig. 2. Temporal extension of social and functional activities.
and (3) the mean per person of the ratio of number of faces per to the number of people per image. External researchers may be able to access the processed isolation indices in terms of group membership of individual participants, subject to research approval.

Based on the survey data, there does not appear to be gender differences among the three categories, but the most isolated or least social persons are definitely younger, compared to less isolated or more social people. Fig. 4 gives the percent of respondents with varying degrees of isolation by different work status categories as available from the survey. About 55% of those looking after the home or family are classified as being least social according to our approach, and the unemployed and those seeking work are also more likely to be least social in contrast to being most or somewhat social. Of those working, the type of occupation also matters, as shown in the figure on the right. Those in routine or semi-routine manual and service occupations are more likely to be least social compared to senior managers and administrators, or those in technical and craft occupations.

Among other findings not shown in the figure, dog-owners who regularly walk their household dogs three or more times a week are less likely to be in the least social category. Finally, a larger percent of those walking to work or school as a main mode of transportation were likely categorised as most social compared to car drivers.

6. Conclusions

New forms of data open up significant new avenues for urban policy and analytics research. Many of these data sources provide an opportunity to look at complex urban problems from multiple perspectives, or at a finer spatial or temporal resolution than data sources traditionally used for urban research. Yet there are critical questions regarding representativeness, ground truth and other issues around validity and reliability. Due to these reasons, being able to place knowledge inferred from such data in the context of data gathering through well-understood or more “controlled” approaches such as survey methods, is important in the early days of using such data on a routine basis towards decision-making.

The data platform described in this paper, the iMCD, is a multi-mode, multi-construct data system to support urban analytics. The data system enables a comprehensive look at transportation and mobility behaviors; education and lifelong learning; sustainable resource consumption and behaviors; social cohesion, cultural values and political preferences; health and wellbeing; and behaviors around use of ICT and digitalisation of our daily lives. The design of the iMCD was motivated by research questions in sustainable transportation, healthy cities, lifelong learning, and their interrelationships. Further motivations arose from technological and methodological questions in data access, sharing and analytics.

Fig. 3. Indoor images from lifelogging sensors.

Fig. 4. Isolation index derived by machine vision algorithms and respondent work status and occupational category.
Key lessons learned from the project is that the need to form interdisciplinary groups which bring different skills and perspectives into the design, collection and processing stages. Having one or more theoretical frameworks is essential so as to give complex, multi-mode data focus and clarity. In our case, data instruments were designed to capture multiple interweaving ideas resulting from multiple theoretical frameworks underpinning sustainable development goals, smart and connected cities and urban livability, transportation and travel behavior, inclusive societies and identity theory, and education and lifelong learning. Further, as multiple strands are being collected at the same time, having a framework to track whether all instruments or data collection modes are operating as designed is a significant responsibility that requires the development of a comprehensive framework.

The extent of participant comprehension and cooperation in the use of multiple modes of complex questionnaires and devices require strategies that are well thought through. The acquisition of some secondary contextual data requiring special negotiations or purchases, and the limits in the use of such data require both legal as well as technical knowledge. Finally, much of the data in the iMCD will require specialist processing for further use particularly for social and economic research. Other considerations are privacy and information security in the use of sensitive data, which will require anonymization of personal data not only within the mode the data were collected, but also in eventual linkage of such data to other data on the said strands.

Declarations of Competing Interest

None.

Acknowledgement

This work was supported by the Economic and Social Research Council, United Kingdom grant number ES/I011921/1.

References

Abadi, M., et al. (2015). TensorFlow: Large-scale machine learning on heterogeneous distributed systems. Preliminary White Paper, Google Research.

Aguilera, U., Peña, O., Belmonte, O., & López-de-Ipiña, D. (2017). Citizen-centric data sensitive data, which will require anonymization of personal data not for further use particularly for social and economic research. Other considerations are privacy and information security in the use of sensitive data, which will require anonymization of personal data not only within the mode the data were collected, but also in eventual linkage of such data to other data on the said strands.

Declarations of Competing Interest

None.

Acknowledgement

This work was supported by the Economic and Social Research Council, United Kingdom grant number ES/I011921/1.

References

Abadi, M., et al. (2015). TensorFlow: Large-scale machine learning on heterogeneous distributed systems. Preliminary White Paper, Google Research.

Aguilera, U., Peña, O., Belmonte, O., & López-de-Ipiña, D. (2017). Citizen-centric data sensitive data, which will require anonymization of personal data not only within the mode the data were collected, but also in eventual linkage of such data to other data on the said strands.

Declarations of Competing Interest

None.

Acknowledgement

This work was supported by the Economic and Social Research Council, United Kingdom grant number ES/I011921/1.

References

Abadi, M., et al. (2015). TensorFlow: Large-scale machine learning on heterogeneous distributed systems. Preliminary White Paper, Google Research.

Aguilera, U., Peña, O., Belmonte, O., & López-de-Ipiña, D. (2017). Citizen-centric data sensitive data, which will require anonymization of personal data not only within the mode the data were collected, but also in eventual linkage of such data to other data on the said strands.

Declarations of Competing Interest

None.

Acknowledgement

This work was supported by the Economic and Social Research Council, United Kingdom grant number ES/I011921/1.

References

Abadi, M., et al. (2015). TensorFlow: Large-scale machine learning on heterogeneous distributed systems. Preliminary White Paper, Google Research.

Aguilera, U., Peña, O., Belmonte, O., & López-de-Ipiña, D. (2017). Citizen-centric data sensitive data, which will require anonymization of personal data not only within the mode the data were collected, but also in eventual linkage of such data to other data on the said strands.

Declarations of Competing Interest

None.

Acknowledgement

This work was supported by the Economic and Social Research Council, United Kingdom grant number ES/I011921/1.

References

Abadi, M., et al. (2015). TensorFlow: Large-scale machine learning on heterogeneous distributed systems. Preliminary White Paper, Google Research.

Aguilera, U., Peña, O., Belmonte, O., & López-de-Ipiña, D. (2017). Citizen-centric data sensitive data, which will require anonymization of personal data not only within the mode the data were collected, but also in eventual linkage of such data to other data on the said strands.

Declarations of Competing Interest

None.

Acknowledgement

This work was supported by the Economic and Social Research Council, United Kingdom grant number ES/I011921/1.

References

Abadi, M., et al. (2015). TensorFlow: Large-scale machine learning on heterogeneous distributed systems. Preliminary White Paper, Google Research.

Aguilera, U., Peña, O., Belmonte, O., & López-de-Ipiña, D. (2017). Citizen-centric data sensitive data, which will require anonymization of personal data not only within the mode the data were collected, but also in eventual linkage of such data to other data on the said strands.

Declarations of Competing Interest

None.

Acknowledgement

This work was supported by the Economic and Social Research Council, United Kingdom grant number ES/I011921/1.

References

Abadi, M., et al. (2015). TensorFlow: Large-scale machine learning on heterogeneous distributed systems. Preliminary White Paper, Google Research.

Aguilera, U., Peña, O., Belmonte, O., & López-de-Ipiña, D. (2017). Citizen-centric data sensitive data, which will require anonymization of personal data not only within the mode the data were collected, but also in eventual linkage of such data to other data on the said strands.

Declarations of Competing Interest

None.

Acknowledgement

This work was supported by the Economic and Social Research Council, United Kingdom grant number ES/I011921/1.

References

Abadi, M., et al. (2015). TensorFlow: Large-scale machine learning on heterogeneous distributed systems. Preliminary White Paper, Google Research.

Aguilera, U., Peña, O., Belmonte, O., & López-de-Ipiña, D. (2017). Citizen-centric data sensitive data, which will require anonymization of personal data not only within the mode the data were collected, but also in eventual linkage of such data to other data on the said strands.

Declarations of Competing Interest

None.
with wearable computing for urban videographic research. *Journal of Geography in Higher Education, 39*(2), 275–287. https://doi.org/10.1080/03098265.2015.1010143.

Paule, J. G., Sun, Y., & Thakuriah, P. V. (2018). Beyond geo-tagged tweets: Exploring the geo-localization of tweets for transportation applications. In C. Ukkasri, Satish, & Yang (Eds.). Transportation analytics (pp. 1–21). Springer.

Quercia, D., & Saez, D. (2014). Mining urban deprivation from foursquare: Implicit crowdsourcing of city land use. *IEEE Pervasive Computing, 13*(2), 30–36. https://doi.org/10.1109/MPRV.2014.31.

Rand, W., Herrmann, J., Schein, B., & Vodopivec, N. (2015). An agent-based model of urgent diffusion in social media. *Journal of Artificial Societies and Social Simulation, 18*(2). https://doi.org/10.18564/jasss.2616.

Sen, A., Thakuriah, P. (Vonu), Zhu, X.-Q., & Karr, A. (1997). Frequency of probe reports and variance of travel time estimates. *Journal of Transportation Engineering, 123*(4), 290–297. https://doi.org/10.1061/(ASCE)0733-947X(1997)123:4(290).

Sethi, V., Bhandari, N., Koppelman, F. S., & Schofer, J. L. (1995). Arterial incident detection using fixed detector and probe vehicle data. *Transportation Research Part C: Emerging Technologies, 3*(2), 99–112. https://doi.org/10.1016/0968-090X(94)00017-Y.

Siła-Nowicka, K., & Thakuriah, P. (2019). Multi-sensor movement analysis for transport safety and health applications. *PloS One, 14*(1), e0210090. https://doi.org/10.1371/journal.pone.0210090.

Siła-Nowicka, K., & Thakuriah, P. (2016). The trade-off between privacy and geographic data resolution. A case of GPS trajectories combined with the social survey results. *XXIII ISPRS Congress* (pp. 535–542). https://doi.org/10.5194/isprs-archives-XLI-82-535-2016.

Siła-Nowicka, K., Vandrol, J., Oshan, T., Long, J. A., Demšar, U., & Forthingham, A. S. (2016). Analysis of human mobility patterns from GPS trajectories and contextual information. *International Journal of Geographical Information Science, 30*(5), 881–906. https://doi.org/10.1080/13658816.2015.1100731.

Stanley, & Stanley (2014). In T. Gärling, D. Ettema, & M. Friman (Eds.), *Handbook of sustainable travel*.Bördrecht: Springer Netherlands. https://doi.org/10.1007/978-94-007-7034-8.

Stock, K. (2018). Mining location from social media: A systematic review. *Computers, Environment and Urban Systems, 71*, 209–240. https://doi.org/10.1016/j.compenvurbsys.2018.05.007.

Stopher, P., Fitzgerald, C., & Xu, M. (2007). Assessing the accuracy of the sydney household travel survey with GPS. *Transportation, 34*(6), 723–741. https://doi.org/10.1007/s11116-007-9126-8.

Sun, Y., Fan, H., Bakillah, M., & Zipf, A. (2015). Road-based travel recommendation using geo-tagged images. *Computers, Environment and Urban Systems, 53*, 110–122. https://doi.org/10.1016/j.compenvurbsys.2013.07.006.

Tiğges, I. M., Browne, I., & Green, G. P. (1998). Social isolation of the urban poor: Race, class, and neighborhood effects on social resources. *The Sociological Quarterly, 39*(1), 53–77.

Thakuriah, P., & Geers, D. G. (2013). Technology systems for transportation system management and personal use. *SpringerBriefs in computer science* 35–71. https://doi.org/10.1007/978-1-4614-7129-5_3.

Thakuriah, P., Siła-Nowicka, K., & Paule, J. G. (2016). Sensing spatiotemporal patterns in urban areas: Analytics and visualizations using the integrated multimedia city data platform. *Built Environment, 42*(3).

Thakuriah, P., Tilahun, N. Y., & Zellner, M. (2017). *Big data and urban informatics: Innovations and challenges to urban planning and knowledge discovery*. 11–45. https://doi.org/10.1007/978-3-319-40902-3_2.

UNESCO (2017). *Learning cities and the SDGc A guide to action*. Retrieved from https://unesdoc.unesco.org/ark:/48223/pf0000260442.

United Nations Development Programme (2018). No title. Retrieved January 18, 2019, from https://www.undp.org/content/undp/en/home/sustainable-development-goals/goal-11-sustainable-cities-and-communities.html.

Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2001: 1*, (pp. 1–511).

Yin, J., Soliman, A., Yin, D., & Wang, S. (2017). Depicting urban boundaries from a mobility network of spatial interactions: A case study of Great Britain with geo-located Twitter data. *International Journal of Geographical Information Science, 1–21*. https://doi.org/10.1080/13658816.2017.1282615.

Zhang, S., & Sun, H. (2016). A thought about formation of economic function area based on competition. In: *Proc. ICEM 2014: 16th International Conference on Economics and Management, 428–436.*