Improving the Veracity of Open and Real-Time Urban Data

GAVIN MCARDLE and ROB KITCHIN

Data are an integral part of the smart city and are used as input for decision-making, policy formation, and to inform citizens and businesses. Reflecting on our experience of developing software applications which rely on urban data, this article examines the veracity of such data (their authenticity and the extent to which they accurately (in terms of precision) and faithfully (in terms of fidelity, reliability) represent what they are meant to) and how this can be assessed. Open data are often provided with no guarantee about their veracity, continuity or lineage (in terms of documentation that establishes provenance). This allows data providers to share data with undocumented errors, absences and biases. These quality issues can propagate through systems and lead to poor software applications and unreliable 'evidence-based' decisions. In this article, we highlight the janitorial role carried out by data scientists and developers to ensure that data are cleaned, parsed, validated and transformed for use. This process requires effort, knowledge and skill but is rarely shared. We propose the inclusion of crowdsourcing mechanisms to record user observations and fixes for improving the quality of data within open government portals.

The availability of open data detailing various aspects of cities continues to grow. This is driven by pressure on local and national governments and public organizations to release their data into the public domain for use and reuse for civic and commercial purposes, to create transparency in city operations, and as a way of benchmarking a city's performance (Pollock, 2006; Janssen, 2012; Open Knowledge Foundation, 2012). Opening up data in this way, it is argued, will foster innovation, provide the raw material for monitoring tools, allow comparison between jurisdictions, inform decision-making, and ultimately lead to a sustainable, resilient and democratic city (Bates, 2012; Kitchin *et al*., 2015). For example, McKinsey (2013) estimates that as much as $5 trillion a year could be added to enterprises as a result of open data. As the race to open datasets advances, there are risks that the checks and balances necessary to ensure the veracity of the data or to inform users of potential quality issues are not performed. Failing to communicate these risks to data consumers or end users of applications will lead to poor-quality derived data, buggy applications and ultimately to poor decisions. Of course, there are challenges facing data providers in detecting various data issues and describing their veracity. For example, measuring data quality usually requires an understanding of the intended purpose, which may not be known by the data producer when sharing the data. It also requires significant overhead in resourcing to produce and share relevant metadata. In the case of real-time data, the velocity and exhaustiveness of the data pose particular challenges. Nonetheless, failing to tackle data veracity issues would be a retrograde position for the

open data movement, with open data sites potentially seen as little more than untrusted, unverified and uncurated data dumps.

This paper describes two smart city applications which rely on a variety of data sources: first, a real-time dashboard, which uses data generated by city authorities and government agencies to provide an interface showing what is happening in Dublin; second, an application which uses data from the Irish Census and city authorities to simulate and model traffic in Dublin City. The dashboard represents a state-of-the-art application to package real-time streamed data in a form that informed users can quickly grasp (see Gray and O'Brien, 2016, this issue) while the model synthesizes space-time data at the finest spatial level of the individual trip maker to enable rich and detailed predictions (see Batty, 2016, this issue). Both applications involve what is now commonly called 'big data'.

The paper focuses on examining the quality of the data used in both applications, which include real-time urban data relating to transportation and environment. With no guidance on the veracity of the data, except for limited lineage metadata and the reputation of the data providers, we needed to validate each dataset using a combination of domain knowledge and analysis. This paper discusses the steps we took to validate and repair problematic data and presents our interactions with the data providers when errors were discovered. Typically, data cleansing undertaken by data intermediaries (such as dashboard builders) are 'black-boxed' and hidden from end users and the original data producers. This paper discusses the need to inform application users about this process so they can trust the analysis, cleaning, parsing and validating processes and make informed decisions about the data. Despite this being a known issue, and there being examples of veracity metrics and international standards for reporting data quality, open data portals typically do not use them. While our experience shows there is willingness for data providers to engage with data consumers, the resources are not necessarily available to achieve this in a meaningful and large-scale way. In the absence of this, we discuss the possibility of borrowing techniques from crowdsourced open data as a method to curate and report the quality of urban data so that the steps taken by others, and the errors, problems and uses of the data, are shared in the same spirit of Volunteered Geographic Information (VGI). The information revealed through this process can be used by the providers to fix data and also utilized by other data consumers when making a judgement on the veracity of urban data.

In the next section, we present several guidelines and standards which are related to the quality of the data used in our applications. In third section, two case studies are presented to highlight the typical validation process which data consumers apply. In fourth, we examine the possibility of using the wisdom of the crowd and a technical solution to report data quality and usage. In the final section, some conclusions and directions for future work are presented.

Data Veracity Metrics

There have been several guidelines and measures proposed to provide a common platform for describing data quality measures (Batini *et al*., 2009) and the importance of reporting data quality has been recently recognized through several ISO standards, such as ISO 19115-1:2014. These set minimum and mandatory metadata fields that should accompany spatial data, with ISO 19157:2013 a dedicated standard for describing components and principles for the quality of spatial data. These standards do not indicate acceptable thresholds for quality data, but rather mandate the metadata that needs be generated with respect to data veracity in order to receive the standard. Here, we concentrate on some of the most relevant measures for spatial and transport data, the focus of our two case studies, and discuss their application to open data sites. Shi *et al*. (2003) review

the determination and handling of spatial data quality, building on the work of the International Cartographic Association (ICA) who identified seven key metrics related to spatial data accuracy (Guptill and Morrisson, 1995):

- *Lineage*. The history of the data including details of the source material and any tranformations or processes applied in order to produce the final data.

- *Positional Accuracy*. An indication of the horizontal and vertical accuracy of the coordinates used in the data, both to absolute and relative locations. It must account for the processes applied to the data which are described by the lineage.

- *Attribute Accuracy*. The accuracy of the quantitative and qualitative data attached to the spatial data.

- *Completeness*. The degree to which spatial and attribute data are included or omitted from the datasets. It also describes how the sample is derived from the full population and presents the spatial boundaries of the data.

- *Logical Consistency*. The dependability of relationships within the spatial data.

- *Semantic Accuracy*. The quality with which geographical objects are described in accordance with the selected model. Semantic accuracy refers to the pertinence of the meaning of the geographical object rather than its geometry.

- *Temporal Data*. The date of observation, the type of update and the validity period for the data.

Likewise, the transport science community has defined similar measures for reporting the quality of traffic data. Turner (2004) carried out a study of data veracity measures and concluded that there are six core measures required to describe the accuracy of traffic data:

- *Accuracy*. How closely the data collected match actual conditions.

- *Completeness*. The degree to which data values are present in the attributes that require them.

- *Validity*. The degree to which data values satisfy acceptance requirements within the domain.

- *Timeliness*. The degree to which data are provided at the time required.

- *Coverage*. The degree to which data values accurately represent the whole of that which is measured.

- *Accessibility*. The relative ease with which data can be retrieved and manipulated by data consumers.

Additionally, it was recommended that data quality reports are presented in the metadata alongside the datasets. Including metadata about the quality and veracity of data allows data consumers to assign an internalized confidence score to the various aspects of the data. This will influence how the data are used and how the results are interpreted. In the United Kingdom, the Department of Transport publish comprehensive guidelines for conducting traffic modelling which includes instructions related to data quality and maintaining an uncertainty log which lists all assumptions about the input data (WebTag. 2016). Moreover, the Environmental Protection Agency in the United States has developed a set of four questions (EPA, 2006) to which answers should be published alongside environmental data in order to allow data consumers to assess its quality and determine if it is fit for their specific purpose.

1. Can a decision (or estimate) be made with

the desired level of certainty, given the quality of the data?

2. How well did the sampling design perform?

3. If the same sampling design strategy is used again for a similar study, would the data be expected to support the same intended use with the desired level of certainty?

4. Is it likely that sufficient samples were taken to enable the reviewer to see an effect if it was really present?

Combined, the four questions allow data consumers to make informed decisions about using the data for their requirements and also provides a guide for how to interpret the results correctly and the weight to place on the results in a decision-making process.

In contrast, it has been argued by some that big data initiatives utilizing real-time data do not need the same standards of data quality, veracity and lineage because the exhaustive nature of the dataset removes sampling biases and more than compensates for any errors or gaps or inconsistencies in the data or weakness in fidelity (Mayer-Schonberger and Cukier, 2013). The argument for such a view is that 'with less error from sampling we can accept more measurement error' (p. 13) and 'tolerate inexactitude' (p. 16). Nonetheless, the warning 'garbage in, garbage out' still holds and issues of accuracy, completeness, validity, timeliness, coverage and accessibility remain important. For example, real-time data can be biased due to the demographic being sampled (e.g. not everybody uses social media platforms) or the data might be gamed or faked through false accounts or hacking (e.g. there are hundreds of thousands of fake Twitter accounts seeking to influence trending and direct click stream trails) (Bollier, 2010; Crampton *et al*., 2013). Moreover, the technology being used and their working parameters can affect the nature of the data. For example, the quality of a pollution or sound sensor can affect the 'noisiness' of the data generated (Choi *et al*., 2009); which posts on social media are most read or shared are strongly affected by ranking algorithms not simply interest (Baym, 2013). Similarly, APIs (Application Programming Interfaces) structure what data are extracted, for example in Twitter only capturing specific hashtags associated with an event rather than all relevant tweets (Bruns, 2013), with González-Bailón *et al*. (2012) finding that different methods of accessing Twitter data – search APIs versus streaming APIs – produced quite different sets of results. As a consequence, there is no guarantee that two teams of researchers attempting to gather the same data at the same time will end up with identical datasets (Bruns, 2013). There is now a plethora of smart city data standards being developed aimed at improving and aligning the data being generated (see ANSSC, 2015 for an overview).

While these general metrics and associated metadata are applicable to all data, including those held within open data portals, at present, metrics applied to open data are generally more concerned with measuring the nature of the data included or the value of the data portal rather than the quality or veracity of the data contained within. For example, Berners-Lee (2006) presents a star rating for open data and awards the highest quality grade to machine readable and linked open data while data in unformatted pdf files, which are still open data but are not as useable as machine structured data such as CSV (Comma-Separated Values) and JSON (JavaScript Object Notation) file formats, receive a lower grade. Martín *et al*. (2015) focus on studying the usability, functionality and data formats of thirty-six Open Government Data portals. While accuracy of the portals is considered, it is merely a check as to whether the data description matches the data. Similarly, Umbrich *et al*. (2015) apply the core metrics for assessing data quality described by Batini *et al*. (2009) to the metadata provided in data portals, but do

not consider the veracity of the data to which the portal provides access. The Open Data Institute (ODI) has developed a certificate which data producers can use to add credibility to their data. The certification is self-assigned and is obtained by the provider by answering a series of questions about their data. A description of the quality control process needs to be presented alongside the data in order to become accredited (ODI, 2015). Similarly, the EU INSPIRE Directive requires spatial data quality and lineage to be reported alongside the data (Inspire, 2015).

Despite these guidelines, recommendations, certificates and standards for reporting data quality, open data portals typically do not report enough metadata to enable consumers to make a reliable judgement call regarding the quality of the data. A review of open data portals for the urban areas of London (http://data.london.gov.uk/), Paris (http://opendata.paris.fr/) and Dublin (http://www.dublinked.ie/), and the World Council of City Data (which reports data for 253 cities in eighty countries; http://open.dataforcities.org/) reveals that neither general nor specific measures of data quality are reported. While data lineage, such as the age of the data (timeliness) and name of data provider, are generally given, the transformation process from the raw to finished product is not described. Similarly, the spatial and temporal extent is given, but the accuracy and precision measurements are not provided. Although our case studies show how fundamental errors were detected, there are potentially technical, political and financial pressures preventing data providers from delivering this information to consumers. Given the potentially infinite uses of different classes of urban data, it is also difficult for data producers to give reliable veracity and quality scores for each domain. Nonetheless, there is a need for much better analysis and sharing of data quality. In the fourth section, we discuss the possibility of using a crowdsourced approach to rate the quality of data in different domains. The approach would use the 'create, discuss and edit' paradigm used for collecting and curating open data on platforms such as Wikipedia and Open Street Map (OSM).

Case Studies

In this section, we discuss the process which we used to validate and clean urban data for two projects. The description presented is representative of our experience of working with a variety of urban data during the development of the Dublin Dashboard (Kitchin *et al*., 2015) and in human mobility-urban traffic projects (McArdle *et al*., 2012; 2014*a*).

Visualizing Real-Time Traffic Data on the Dublin Dashboard

The Dublin Dashboard (http://www.dublindashboard.ie/) provides citizens, government workers and companies with real-time information, urban indicator and benchmarking data, and other forms of data about all aspects of the city through a series of interactive graphs, maps and applications. It aims to enable users to gain detailed, up-to-date intelligence about the city that will help foster smart decision-making and smart citizens. The data are sourced from a variety of Irish data providers including the Central Statistics Office, the Department of Environment, Community and Local Government, the Environmental Protection Agency, Dublinked, and the four local authorities in Dublin.

The use of urban data is a major component of the Dublin Dashboard. This can be seen through the eleven modules which constitute the dashboard, as we show in figure 1. For example, the *Overview* module shows the current weather, noise levels, air quality, traffic conditions, house prices, rent prices, unemployment and employment rates, plus the number of patients waiting to be admitted to hospitals and the crime rate in Dublin. The *How's Dublin Doing* module consists of a suite of indicators which plot trends describing Dublin in terms of house prices, planning permissions, poverty rates, demographics, educa-

tion levels, crime rates, economic output, etc. Tools to compare data from Dublin to data from other cities are also provided. *Dublin Mapped* provides detailed interactive maps showing the results of the two most recent Irish Censuses. The data are visualized and mapped at a small area level (80–100 households) enabling a detailed analysis. *Dublin Real-Time* consists of interactive maps showing live information on the traffic, travel and environmental conditions in the city. The *Dublin Bay Dashboard* module also provides interactive maps and tools describing the Dublin Bay area and the marine biosphere.

In addition to the modules which were developed specifically for the Dublin Dashboard, the site also acts as a portal to applications and tools developed by others. For example, the *Dublin Planning* and *Dublin Near to Me* modules contain links to location-based services, land zoning data and transport services. The *Dublin Housing* and *Dublin Reporting* modules provide maps showing house prices and tools to allow citizens to report problems such as littering, poor street lighting, and damaged infrastructure. The *Data Stores* module provides access to online repositories of data used in the Dashboard.

Figure 1. The Homepage of the Dublin Dashboard.

One goal of the Dublin Dashboard is to answer questions related to what is happening in the city right now. To achieve this, the dashboard collates, analyses and presents real-time data. This is materialized via real-time maps which show the locations where data are recorded and indicate the current value of the variable being measured. The real-time data streams relate to transport (e.g. how many bikes/spaces are in bike stands, road speeds on different segments, the number of spaces in car parks, general CCTV footage), and environment (e.g. air traffic, air quality, pollution readings, water levels, sound levels, current weather). The data are collected from a variety of sources. In some cases the data are provided via an API which allows a developer to query the data and obtain results in a machine readable format, generally in JSON format; the results can then be digested, presented and used in applications. Other data providers use file formats such as CSV or XML (Extensible Mark-UP Language). In these cases, the data consumer or developer needs to parse and process the files in order to select those data which are relevant to their application domain.

The real-time travel map, shown in figure 2, is one of the most frequently viewed tools in the Dublin Dashboard. The map shows the predicted travel time, by car, on all major artery routes into and out of Dublin City. The data are provided by Dublin City Council (DCC) and are obtained using TRIPS (Travel-time Reporting and Integrated Performance System) which predicts travel time based on data generated by on-street traffic detection technologies (e.g. transduction loops). The data are published by DCC every minute via a CSV file which is available to download

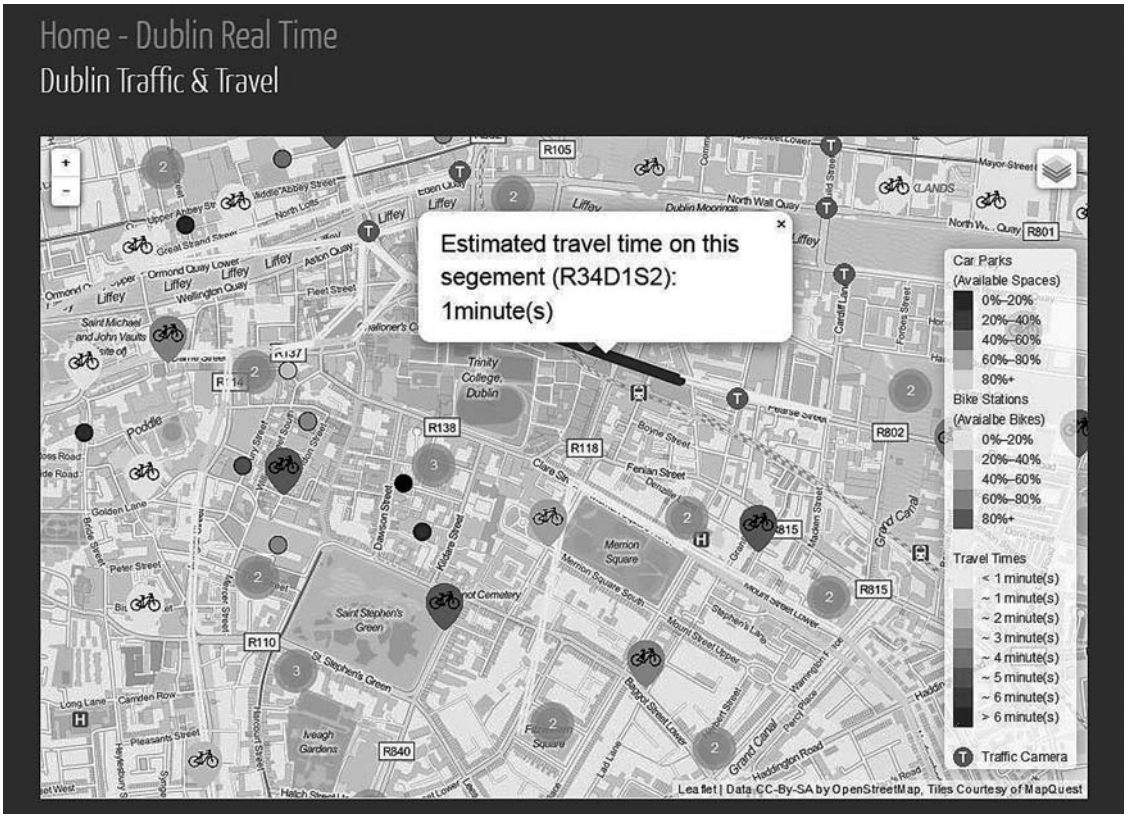

Figure 2. The real-time travel map on the Dublin Dashboard.

Table 1. Sample of journey times from the DCC TRIPS dataset.

From	*To*	*Travel Time* (seconds)
SWORDS RD SHANOWEN RD	INCHICORE ROAD OLD KILMAINHAM LANE	36
DORSET ST GARDINER ST	BLESSINGTON ST DORSET ST	42
CONSTITUTION HILL WESTERN WAY FLATS	PHIBSBORO ROAD NORTH CIRCULAR ROAD DOYLES CORNER	73

from the Dublinked website (Dublinked, 2015*a*). Dublinked is a data portal used by the four Dublin local authorities to share data with the public and organizations in order to promote entrepreneurship and innovation. DCC reserves the right to cancel access or permission for data use and will not be held liable for any losses arising from their use, or from the use of other information based on these data. There is no indication of the data veracity supplied with the dataset.

The CSV travel time data needs to be used in conjunction with other static files which describe the road network. These files are also available to download from Dublinked. The network consists of a list of nodes (named road intersections) and links (roads connecting intersections). A unique ID for each node and link is used to match the travel time data. This allows the travel time for each road segment to be reported and used in other applications and software. After downloading the relevant CSV files, a developer can produce a matrix representation of the travel times. Table 1 shows an example of travel times extracted from the dataset.

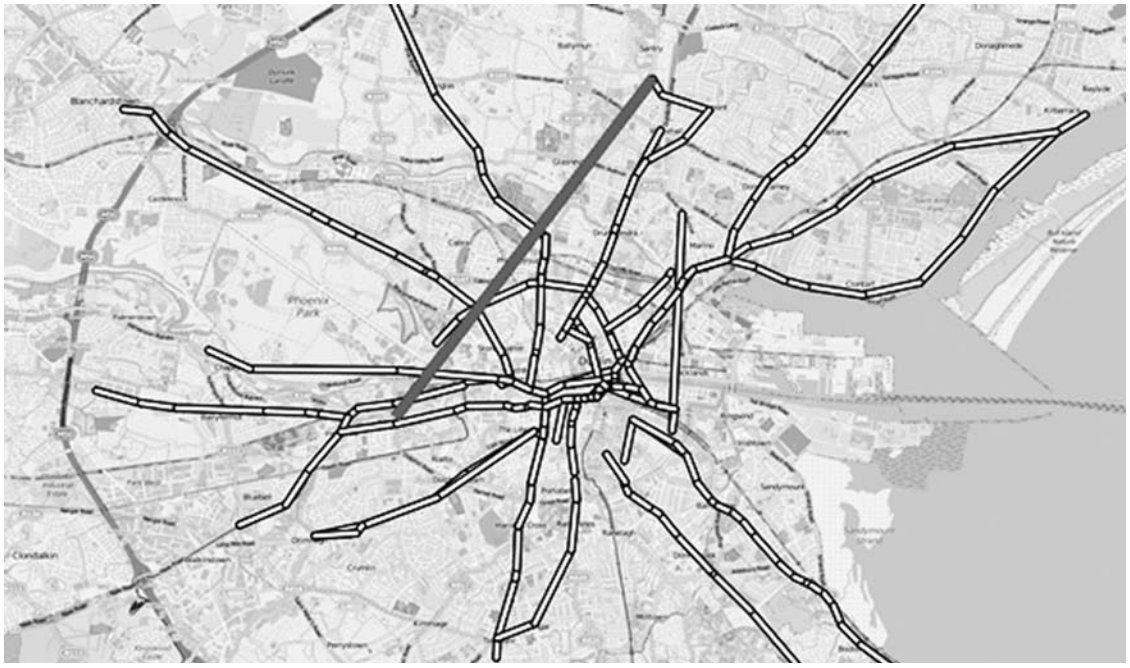

Figure 3. The sections of road for which travel-times are given in the TRIPS dataset. The solid black line is a false road segment over 6 km in length and has a reported travel time of 36 seconds.

It is relatively easy to process the data provided to get to this stage of development. The data can then feed into other applications such as route planning, journey time or traffic analysis software or a travel map like that in the Dublin Dashboard. Without domain knowledge of the geography of Dublin, or mapping the data to add context, the errors in the dataset are not visible. However, creating a map of the segments and times reveals a number of issues. For example, figure 3 shows data in table 1 revealing the impossible journey time of 36 seconds to travel over 6 kilometres due to the inclusion of a false road segment. The dataset contains several examples of impossible journey times like this.

The lineage metadata shows the data originate from a reliable source (DCC) but does not contain processing information, and despite the ease with which we were able to detect the errors without using any specialist tools, the data providers do not report an error. As developers, our solution was to remove the road segments which contained impossible travel times from the dataset and to make no claims regarding the accuracy of the data displayed in the dashboard. While the solution was adequate for our application, we do not report the errors to users of the dashboard nor do we report the techniques we used to identify and fix them. In part, this is because once fixed, they are no longer a problem, but also because we have no way of verifying the data, beyond spotting obvious flaws, without working directly with the data provider or deploying some form of ground truthing for which we have no resources.

Building a Traffic Simulation for Dublin

The second urban data project considered, builds an agent-based traffic simulation for Dublin City (McArdle *et al*., 2014*b*). The model simulates the travel patterns for private vehicles in the Greater Dublin Region and attempts to minimize the travel time for individual vehicles by rerouting commuters on the road network and adjusting departure times through many iterations of the simulation. The simulation completes when equilibrium is achieved and further alterations to routes and travel times will not improve the overall system wide travel time. The project uses a variety of urban data sourced from POWSCAR (Place of Work and School – Census of Anonymized Records) and SCATS (Sydney Coordinated Adaptive Traffic System) as input to the simulation and as a means of validating the simulation results and output.

To create the simulation, a transportation model called *MATSim* (Rieser, 2010) was used. *MATSim* is a multi-agent micro-simulation tool in which each individual of the population is considered an agent with a plan (e.g. travel from home to work) and the tool will provide routing through the network to achieve the plans. As with other transport simulation tools such as *SUMO* (Behrisch *et al*., 2011), *MITSim* (Ben-Akiva *et al*., 2010), *VISSIM* (Fellendorf, 1994) and Megaffic (Suzumura *et al*., 2015), *MATSim* requires a demand to be placed on the road network. For this project the initial demand consists of the home and work locations of individuals organized into an origin–destination matrix augmented with the mode of transport and departure times. This simulation only considers individuals who live or work in Dublin and commute by driving a private car. The demand data were obtained from POWSCAR, a subset of the Irish National Census, which is conducted every 5 years. POWSCAR provides the home, work, school, and college location of individuals; the mode of transport used to commute; the time at which individuals leave their home in the morning along with other variables such as age, socioeconomic grouping, household size and travel time to work, school or college. The home location is anonymized by describing it at a Small Area level which is a geographic area consisting of 80 to 100 households. The work location is presented at a 250-metre grid level. The time of departure

is represented by discrete 30-minute intervals for the morning period, and several transit modes are encoded in the means of transport while travel time is described in minutes. When only drivers who commute to or from County Dublin were considered, a dataset of approximately 300,000 individuals was obtained.

Prior to running the simulation, data verification was carried out to test the validity of the POWSCAR input data for the traffic modelling. Direct ground truth was not an option so data analysis was carried out. The analysis focused on the self-declared journey time parameter, which is an estimate made by individuals regarding their commuting time. The distance between the centroid of the home and work small areas was calculated and used with the journey time to estimate an average speed for the commuting trip. The speed of each commuter is shown in figure 4. The graph shows that many commuters achieved impossible average speeds. The lower horizontal line shows the mean speed of 35 km/h, while the upper horizontal line shows a speed of 120 km/h which is the legal speed limit in Ireland but is not achievable in Dublin during the commuting hours. Further analysis reveals that over 10,000 commuters had an average commuting speed of over 120km/h (3 per cent of the dataset), 7,000 commuters had an estimated average speed of over 200km/h, while over 1,000 commuters had a speed greater than 1,000km/h. This analysis shows inconsistencies in the dataset which were not reported in the documentation. A filter was developed to remove individuals with an impossible travel time from the input matrix before using the data in the traffic simulation process.

The cleaned commuting data was combined with data about non-commuting drivers (unemployed and retired individuals) to give a total of approximately 600,000 drivers. For efficiency reasons a 25 per cent sample was used in the simulation. *MATSim* runs in an iterative fashion. Typically, the population (of drivers) is routed through the road network and travel times, which are dependent on road capacity, are calculated. After each iteration, the travel plans of a sample of the population are altered (e.g. 10 per cent might have their route changed and 10 per cent might have departure time changed). The simulation is run again to determine if there is an improvement in overall travel time for the system. This process continues until further changes have no effect on travel times and the system converges to equilibrium. The time taken to run such simulations is depend-

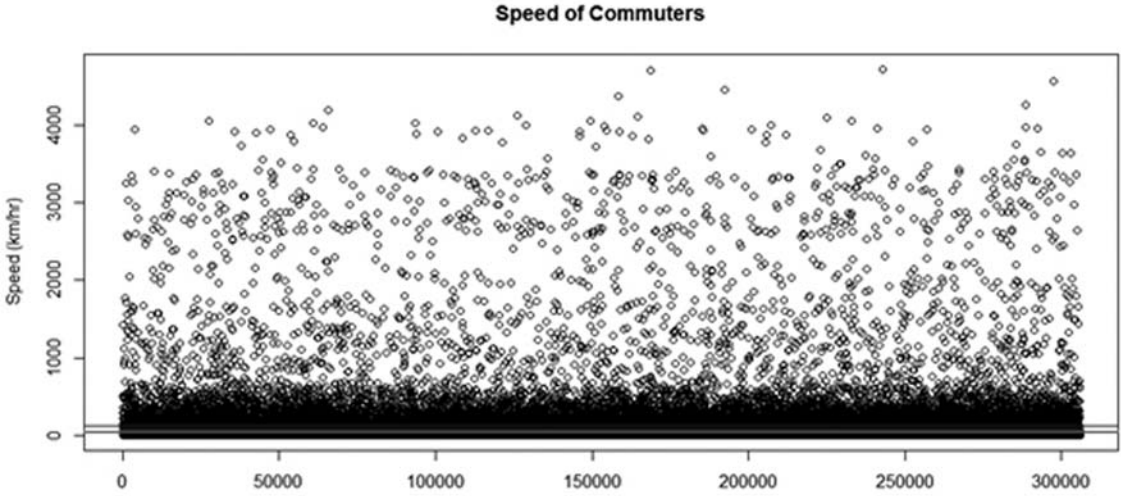

Figure 4. The speed distribution of commuters in the POWSCAR dataset.

Table 2. An example of the CSV file describing the road intersections.

streetSegID	*armNumber*	*armAngle*	*Lato*	*longo*	*latd*	*longd*
681	1	0	53.33981	–6.24184	53.3398	–6.24175
160	1	0	53.34437	–6.26286	53.34435	–6.26276
1396	1	0	53.34513	–6.23838	53.34512	–6.23828
862	1	0	53.34564	–6.24899	53.34563	–6.24889

ent on the complexity and size of the road network, the size of the population and the processing power of the machine used to run the simulation. For the Dublin scenario, 300 iterations were required which took several days to complete. Full details of the Dublin Scenario, including details of how non-commuting trips were included in the simulation can be found in McArdle *et al*. (2014*b*).

The output of the simulation is an hourly count for the number of vehicles using each road segment in Dublin. This enables a 24-hour profile to be produced across the city. In order to validate the effectiveness of the simulation techniques, these data are usually benchmarked against ground truth for the same road segment. The ground truth can be obtained using a manual observation count or by using count data from moveable or embedded sensors in the road surface. Given the cost and resources required to conduct a manual count and the limited coverage of the city that such a count can achieve we opted to use data obtained from SCATS for Dublin. SCATS is a technology which is used to optimize traffic flow by counting cars passing through an intersection and using this data to control the traffic light sequence. While the data are collected in near real-time, a sample of the data in CSV files is made available by DCC via Dublinked (Dublinked, 2015*b*). The data were first provided in January 2012 and updated in April 2012 and made available under a PSI licence. The data are aggregated over 5 minute intervals for each approach to an intersection for each day and data are provided for the period 1 Jan 2012 to 30 April 2012 (8–12 months after the POWSCAR data were generated). Each sensor is described in the CSV file as a street segment ID, arm number and angle which describe the approach road (latd, lond) and the centroid coordinates of the intersection. An example of this is shown in table 2.

The challenge is to map the SCATS sensors to the road network used in the traffic simulation. Initial spatial queries showed there was no direct technique to map the sensor coordinates reliably to a road segment. Geovisual analysis, as shown in figure 5 highlights the problem. The coordinates indicating the location of the SCATS sensors are inaccurate by some unknown offset. Attempts to use translations and transformations to align the sensors with the road network failed and no consistent offset could be determined. Although in some cases, the offset appears to be minor, it is still impossible to determine automatically which road segment on a junction a sensor corresponds to. This is especially true in the city centre where there are many junctions in close proximity. This prevents automatic or manual validation of the count data returned by the simulation. Dialogue was entered into with DCC representatives to remedy the situation. A visualization of the problem as shown in figure 5 was provided. The engineers in DCC were very responsive and conducted their own analysis on the raw data, which describes their sensor network, but no satisfactory solution was found. Finally, we were informed that the location data was only indicative of the location of the sensors. It was therefore impossible to match automatically road seg-

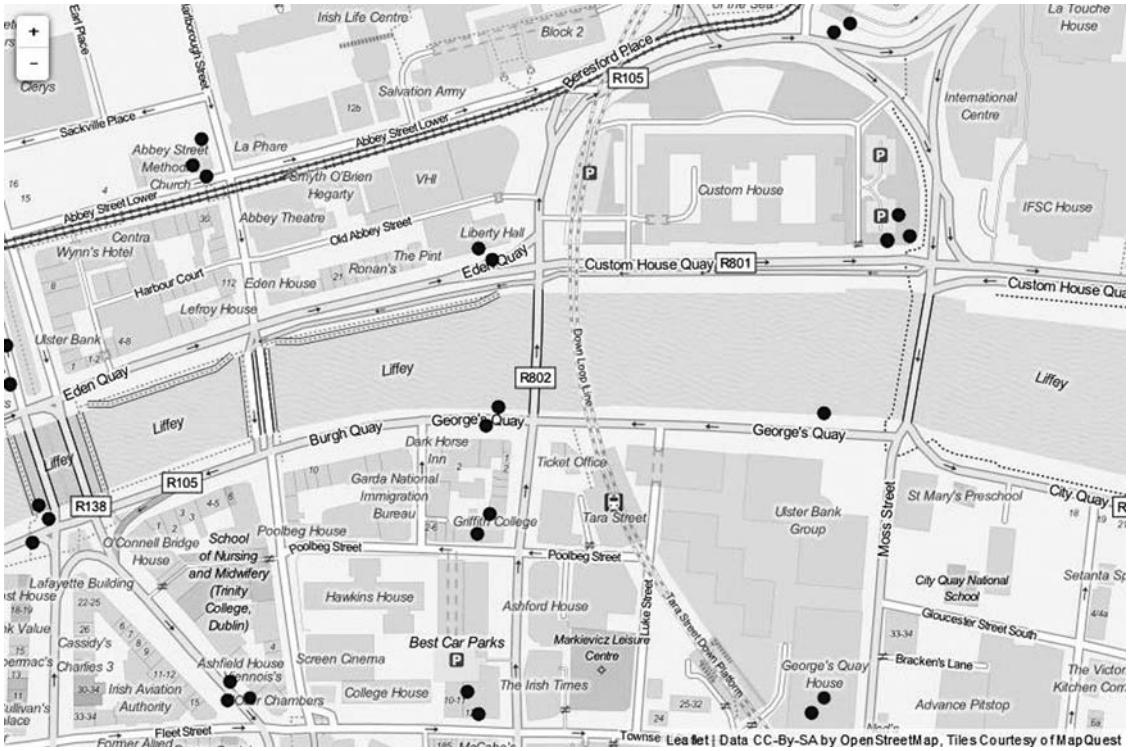

Figure 5. The locations of the SCATS sensors do not correspond to the road intersections.

ments to SCATS sensors with any accuracy. As it was not feasible to recode manually the coordinates of each sensor, this method of validation was abandoned. Instead, data produced by the National Roads Authority in Ireland, which shows the hourly traffic volume between intersections on motorways around Dublin were used to validate the output of the traffic simulation.

As with the first case study project, we spent time and effort examining the data and applying different techniques to validate and repair the data and to determine if it matched our requirements. This effort and its results are not recorded or reported alongside the data or within the applications. Furthermore, despite informing the data providers of the error in the SCATS data, no action was taken by the provider to issue a data quality statement. This means that future data consumers will need to replicate the work that we carried out before finding the errors and will need to determine if the data are fit for their specific purpose.

Discussion

The problem of reporting data quality is recognized and well understood. As the discussion of data veracity metrics above highlights, there are a variety of standards available for reporting data quality and other relevant metadata (e.g. calibration). While some measures are domain specific, several core quality metrics have emerged. From our experience, data producers are not using these measures, or are not reporting them, and instead provide data with no commentary about their veracity and only offer scant details of their lineage. This is through ignorance, lack of resources, indifference, or a lack of expertise. Unfortunately, producing urban data which contain undocumented errors is a retrograde step for the open data movement. As such

veracity issues come to light they have the potential to fuel accusations that open government data portals are untrusted, unverified and uncurated data dumps. While the value of open data for the economy and for business has been well documented (McKinsey, 2013), the cost to business of using poor quality data is also recognized. While some might argue that, in the case of big data, more trumps better, the reality is that poor veracity reduces the validity of analysis and interpretation.

The preferred solution to the lack of documented data veracity is for data producers to document more diligently and extensively such issues in their metadata, along with user guides as to how best to address or compensate for them when using the data for different purposes. However, if the *status quo* remains, the onus falls to data consumers and developers to determine whether they are satisfied that the urban data they are using are reliable and fit for their intended purpose. In our case we were developing two urban applications which used a mix of open administrative census data and real-time travel data which did not have quality measurements in the form of metadata. We therefore applied domain knowledge and various analysis techniques to validate the data. Three different datasets were considered and in each case, errors in the data were discovered. In case study section we documented the process which was used to test the data for our requirements and also described the steps used to clean and repair the data. In one case, we interacted with the data producer but the problem was not resolved nor documented as metadata by the data producer.

The analysis and validation which we carried out required a certain level of expertise, effort and time. Although this effort pays off in the form of a working application or improved data quality for the traffic simulation tool, the process, our findings and our fixes are not recorded or reported which means that this type of effort will need to be replicated by each new consumer of the data. To reduce this, we propose a mechanism for crowdsourcing metadata about the quality of datasets, similar to the collection of Volunteered Geographic Information (Goodchild, 2007). Using the wisdom of the data user crowd could create a more curated form of urban data and encourage greater engagement between data providers and consumers and enhance the reputation of open data portals.

The proposed approach mimics the ethos of Wikipedia and OSM in which users of these websites can contribute and edit content. However rather than directly edit and contribute datasets, we propose that users can contribute and edit metadata to describe the veracity of a dataset and provide feedback about any processing that was applied to validate the data. This could be done using many of the recognized domain specific standards like those outlined in our discussion of data veracity metrics. The open data portal should provide the tools to facilitate and support this crowdsourcing of data veracity, along with a forum to discuss the data and give examples of where they have been used. Some urban data portals such as the Paris data portal do provide a means for discussing datasets, however there is scope to extend its functionality as a reporting and sharing interface. The approach has been successful for OSM and Wikipedia and the editing of data is self-policed by members who form a community so that false or misleading information becomes rare. The approach is akin to the idea of civic hacking in which citizens want to improve services for all (Coleman and Golub, 2008; Perng and Kitchin, 2015).

While there are arguments for not sharing data veracity and processing experiences, such as gaining a competitive advantage, this has not been the case in the open data community. For example, the ODI have members who volunteer time to process open data to improve its usability by translating it into machine-readable formats. Generally

within the crowdsourcing community, individuals do not receive monetary reward for their efforts but receive recognition that their contribution is helping others while also increasing contributors' profile as experts. This proposed approach echoes the more general move towards using crowdsourced data, collected both actively (volunteered) and passively, as a way of creating new official data and official statistics and improving existing government data (Goodchild, 2007; Lauriault and Mooney, 2014). Like Wikipedia, OSM and other crowd-sourced geographic data, which are inherently an unfinished product (Dodge and Kitchin, 2013), determining the veracity of a dataset will be an ongoing task as there are always novel and innovative uses of data for which new quality and veracity metrics will be required.

Following our analysis and using the proposed approach, we would contribute the knowledge that we discovered about the inaccurate positioning of the SCATS sensors (using the ICA data quality methodology), the inaccurate travel times for the TRIPS data (using the Transport Science metrics) and the invalid speeds achieved seen in the POWSCAR dataset. This would benefit future users of these datasets and may lead to a revised dataset being made available by the data producers. It will also allow other data users to update their applications based on this new information.

Conclusion

Our experience in the case studies documented here and other data intensive projects (Gleeson *et al*., 2015; Kitchin *et al*., 2013; McArdle *et al*., 2014*b*; Calabrese *et al*., 2015) highlights several challenges related to the use of urban data regarding its validity, veracity, and reliability. Our experience is typical and shows how errors are handled, or not, by both data producers and consumers. While there are metrics, methodologies and guidelines, and increasingly standards and certificates for measuring the quality and accuracy of data, our experience shows that these are not being widely used in urban open data portals. Data producers seem happy to provide data 'as is', without any guarantee regarding their quality or accuracy either due to laxities in their efforts or to avoid liability for inaccurate data. Doing this on a wide scale in open data portals is potentially dangerous and may lead to the urban data portals being regarded as unreliable by data consumers and critics. Moreover, it potentially jeopardizes the intended economic and civic engagement benefits which are often the goals of such portals. While our examples are not exhaustive in terms of the types of veracity and quality issues for data, they illustrate how issues with accuracy and consistency were detected using analysis and highlight the need for developers to do such checking when no veracity or lineage metadata accompanies urban data.

The reasons why data producers do not carry out such analysis or provide the full lineage of the data is an open question and a further study is required. Issues related to resources, expertise, skills, time and a risk of liability are likely to be cited as causes. Further, data are often provided without knowledge of all the possible end uses and so it is difficult to express data veracity across an exhaustive range of domains.

In the absence of data providers carrying out quality analysis and providing detailed metadata and lineage information, developers must assess data quality and accuracy for their specific needs. Our experience shows the effort required to carry out this process can be great, and is typically lost and becomes black-boxed or encoded in the resulting application or tool. The problem is likely to increase in the era of big data, with many providers such as local government departments being unable to maintain veracity metadata for quickly transitioning data. To resolve this issue, we propose that a data veracity community be developed around the use of open government data, including real-time data. This community can then curate the data by

providing the metadata about veracity, the processing that they carried out in developing applications, and have discussions about the data with other consumers and the producer. This crowdsourcing approach would build on the spirit of sharing seen in the open data community and mimic that seen in the Wikipedia and OSM. This should lead to a greater trust in urban open data portals and result in improved smarter city applications and smarter evidence-based decisions. The next step is to design and integrate this proposed approach with an open data portal and we are exploring the possible implementation of such an endeavour with stakeholders.

REFERENCES

ANSSC (2015) *Directory of Smart and Sustainable Cities Standardization Initiatives and Related Activities*. Washington DC: American National Standards Institute Network on Smart and Sustainable Cities. Available at: https://share.ansi.org/ANSI%20Network%20on%20Smart%20and%20Sustainable%20Cities/ANSSC-Directory-of-Initiatives.pdf.

Bates, J. (2012) This is what modern deregulation looks like: co-optation and contestation In the shaping of the UK's Open Government Data Initiative. *The Journal of Community Informatics*, **8**(2). Available at: http://www.ci-journal.net/index.php/ciej/article/view/845/916.

Batini, C., Cappiello, C., Francalanci, C. and Maurino, A. (2009) Methodologies for data quality assessment and improvement. *ACM Computing Surveys* (CSUR), **41**(3). Article No. 16, doi.10.1145/1541880.1541883.

Batty, M. (2016) Big data and the city. *Built Environment*, this issue.

Baym, N.K. (2013) Data not seen: the uses and shortcomings of social media metrics. *First Monday*, **18**(10). Available at: http://firstmonday.org/ojs/index.php/fm/article/view/4873/3752.

Behrisch, M., Bieker, L., Erdmann, J. and Krajzewicz, D. (2011) Sumo-simulation of urban mobility-an overview, in *Proceedings of SIMUL 2011, the Third International Conference on Advances in System Simulation*, Barcelona, pp. 55–60.

Ben-Akiva, H., Koutsopoulos, N., Toledo, T., Yang, Q., Choudhury, C.G., Antoniou, C. and Balakrishna, R. (2010) Traffic simulation with MITSIMLab, in Barceló, J. (ed.) *Fundamentals of Traffic Simulation*. New York: Springer, pp. 233–268.

Berners-Lee. T. (2006) Linked data: design issues. Available at:http://www.w3.org/DesignIssues/LinkedData.html.

Bollier, D. (2010) *The Promise and Peril of Big Data*. Washington DC: The Aspen Institute. Available at: http://www.aspeninstitute.org/sites/default/files/content/docs/pubs/The_Promise_and_Peril_of_Big_Data.pdf.

Bruns, A. (2013) Faster than the speed of print: reconciling 'big data' social media analysis and academic scholarship. *First Monday* **18**(10). Avaiable at: http://firstmonday.org/ojs/index.php/fm/article/view/4879/3756.

Calabrese, F., Di Lorenzo, G., McArdle, G., Pinelli, F. and Van Lierde, E. (2015) Real-time social event analytics, in *Proceedings of NetMob 2015*, MIT Media Lab, Boston, pp. 126–128.

Choi, S., Kim, N., Cha, H. and Ha, R. (2009) Micro sensor node for air pollutant monitoring: hardware and software issues. *Sensors*, **9**(10), pp. 7970–7987.

Coleman, G. and Golub, A. (2008) Hacker practice: moral genres and the cultural articulation of liberalism. *Anthropological Theory*, **8**(3), 255–277.

Crampton, J.W., Graham, M., Poorthuis, A., Shelton, T., Stephens, M., Wilson, M.W. and Zook, M. (2013) Beyond the Geotag: situating 'big data' and leveraging the potential of the Geoweb. *Cartography and Geographic Information Science*, **40**(2), pp. 130–139.

Dodge, M. and Kitchin, R. (2013) Crowdsourced cartography: mapping experience and knowledge. *Environment and Planning A*, **45**(1), pp. 19–36.

Dublinked (2015*a*) TRIPS Dataset. Available at: https://data.dublinked.ie/dataset/journey-times-across-dublin-city-from-dublin-city-council-traffic-departments-trips-system.

Dublinked (2015*b*) SCATS Dataset. Available at: https://data.dublinked.ie/dataset/volume-data-for-dublin-city-from-dublin-city-council-traffic-departments-scats-system.

EPA (Environmental Protection Agency) (2006) *Data Quality Assessment: A Reviewer's Guide, EPA QA/G-9R, EPA/240/B-06/002*. Washington DC: EPA. Available at: https://www.epa.gov/sites/production/files/2015-08/documents/g9r-final.pdf.

Fellendorf, M. (1994) VISSIM: A microscopic simulation tool to evaluate actuated signal

control including bus priority, in *Proceedings of the 64th Institute of Transportation Engineers Annual Meeting*, Dallas, pp.1–9.

Gleeson, J., Walsh, A.J., Van Egeraat, C., Daly, G., Kitchin, R., Boyle, M., McClelland, A., Foley, R., Haase, T. and Pratschke, J. (2015) *The Atlas of the Island of Ireland*. Available at: http://airo.maynoothuniversity.ie/files/downloads/AtlasoftheIslandofIreland_2015.pdf.

González-Bailón, S., Wang, N., Rivero, A., Borge-Holtoefer, J. and Moreno, Y. (2012) Assessing the bias in communication networks sampled from Twitter. *Social Networks*, **38**(1), pp. 16–27. Available at: http://arxiv.org/abs/1212.1684.

Goodchild, M.F. (2007) Citizens as sensors: the world of volunteered geography. *GeoJournal*, **69**(4), pp. 211–221.

Gray, S. and O'Brien, O. (2016) City dashboards. *Built Environment*, this issue.

Guptill, S.C. and Morrison, J.L. (eds.) (1995) *Elements of Spatial Data Quality*. Oxford: Elsevier.

Inspire (2015) *EU INSPIRE Directive for Spatial Data*. Available at: http://inspire.ec.europa.eu.

Janssen, K. (2012) Open government data: right to information 2.0 or its rollback version? *ICRI Working Paper 8/2012.* Available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2152566.

Kitchin, R., Gleeson, J. and Dodge, M. (2013) Unfolding mapping practices: a new epistemology for cartography. *Transactions of the Institute of British Geographers*, **38**(3), pp. 480–496.

Kitchin, R., Lauriault, T.P. and McArdle, G. (2015) Knowing and governing cities through urban indicators, city benchmarking and real-time dashboards. *Regional Studies, Regional Science*, **2**(1), pp. 6–28.

Lauriault, T.P. and Mooney, P. (2014) Crowdsourcing: a geographic approach to public engagement. *Programmable City Working Paper 6*. Available: SSRN 2518233.

Martín, A.S., De Rosario, A.H. and Pérez, M.D.C.C. (2015) An international analysis of the quality of open government data portals. *Social Science Computer Review*, **34**(3), pp. 298–311.

Mayer-Schonberger, V. and Cukier, K. (2013) *Big Data: A Revolution That Will Change How We Live, Work and Think*. London: John Murray.

McArdle, G., Demšar, U., van der Spek, S. and McLoone, S. (2014*a*) Classifying pedestrian movement behaviour from GPS trajectories using visualization and clustering. *Annals of GIS*, **20**(2), pp. 85–98.

McArdle, G., Furey, E., Lawlor, A. and Pozdnoukhov, A. (2014*b*) Using digital footprints for a city-scale traffic simulation. ACM Transactions on Intelligent Systems and Technology (TIST), **5**(3), article 41.

McArdle, G., Lawlor, A., Furey, E. and Pozdnoukhov, A. (2012) City-Scale traffic simulation from digital footprints, in *Proceedings of the ACM SIGKDD International Workshop on Urban Computing*, pp. 47–54.

McKinsey (2013) *Open Data: Unlocking Innovation and Performance with Liquid Information*. Avilable at: http://www.mckinsey.com/insights/business_technology/open_data_unlocking_innovation_and_performance_with_liquid_information.

ODI (Open Data Institute) (2015) Open Data Certificate. London: ODI. Available at: https://certificates.theodi.org/.

Open Knowledge Foundation (2012) *Open Data Handbook Documentation*, 1.0.0. Available at: http://opendatahandbook.org/.

Perng, S.Y. and Kitchin, R. (2015) Solutions, strategies and frictions in civic hacking. *Programmable City Working Paper 10*. Available at: SSRN 2606939.

Pollock, R. (2006) *The Value of The Public Domain*. London: IPPR. Available at: http://www.ippr.org/publication/55/1526/the-value-of-the-public-domain.

Rieser, M. (2010) Adding Transit to an Agent-Based Transportation Simulation: Concepts and Implementation. PhD thesis, VSP, TU Berlin. Available at: http://svn.vsp.tu-berlin.de/repos/public-svn/publications/vspwp/2010/10-05/20100610_phdthesis_mrieser.pdf.

Shi, W., Fisher, P. and Goodchild, M.F. (2003) *Spatial Data Quality*. London: CRC Press.

Suzumura, T., McArdle, G. and Hiroki, K. (2015) A high performance multi-modal traffic simulation platform and its case study with the Dublin City. *Proceedings of the 2015 Winter Simulation Conference.* Piscataway, NJ: IEEE Press, pp. 767–778.

Turner, S. (2004) Defining and measuring traffic data quality: white paper on recommended approaches. *Transportation Research Record*, No. 1870, pp. 62–69.

Umbrich, J., Neumaier, S. and Polleres, A. (2015) Towards assessing the quality evolution of open data portals, in *ODQ2015: Proceedings of the Open Data Quality: from Theory to Practice Workshop*. Munich. Available at: https://ai.wu.ac.at/~polleres/publications/umbr-etal-2015ODQ.pdf.

WebTAG (2016) UK Department of Transport, *Transport Analysis Guidance*. London: Department of Available at: https://www.gov.uk/guidance/transport-analysis-guidance-webtag.

ACKNOWLEDGEMENTS

The research for this article was funded by a European Research Council Advanced Investigator award (ERC-2012-AdG-323636-SOFTCITY) and Science Foundation Ireland.