sciendo

Journal of Official Statistics, Vol. 37, No. 1, 2021, pp. 121–147, http://dx.doi.org/10.2478/JOS-2021-0006

Can Smart City Data be Used to Create New Official Statistics?

Rob Kitchin[1] *and Samuel Stehle*[2]

In this article we evaluate the viability of using big data produced by smart city systems for creating new official statistics. We assess sixteen sources of urban transportation and environmental big data that are published as open data or were made available to the project for Dublin, Ireland. These data were systematically explored through a process of data checking and wrangling, building tools to display and analyse the data, and evaluating them with respect to 16 measures of their suitability: access, sustainability and reliability, transparency and interpretability, privacy, fidelity, cleanliness, completeness, spatial granularity, temporal granularity, spatial coverage, coherence, metadata availability, changes over time, standardisation, methodological transparency, and relevance. We assessed how the data could be used to produce key performance indicators and potential new official statistics. Our analysis reveals that, at present, a limited set of smart city data is suitable for creating new official statistics, though others could potentially be made suitable with changes to data management. If these new official statistics are to be realised then National Statistical Institutions need to work closely with those organisations generating the data to try and implement a robust set of procedures and standards that will produce consistent, long-term data sets.

Key words: Big data; transport; environment; data quality; key performance indicators.

1. Introduction

Official statistics are appropriate, reliable, consistent data related to a jurisdiction that adhere to professional and scientific standards and are impartially compiled and produced by official statistical agencies (UNECE 1992). Over the past decade there has been significant attention paid to the potential of big data–that is, data that are produced continuously and have an exhaustive sample within a domain (Kitchin and McArdle 2016) – for compiling official statistics as produced by National Statistical Institutions (NSIs) and other government agencies (Florescu et al. 2014; Kitchin 2015; Struijs et al. 2014). Several big data sources have been examined as to their prospective utility, such as mobile network data for mobility and tourism statistics, web searches and job advertisements for labour statistics, real-estate websites for property statistics, social

[1] Maynooth University Social Sciences Institute, Maynooth University, County Kildare, Ireland. Email: Rob.Kitchin@mu.ie.
[2] National Centre for Geocomputation, Maynooth University, County Kildare, Ireland. Email: samstehle@gmail.com.
Acknowledgments: The research for this article was funded by Science Foundation Ireland, grant number 15/IA/3090. We are also grateful to Aidan Condron, John Dunne, and Konstantinos Giannakouris, who read and provided comments on an initial draft.

media for consumer confidence and well-being statistics, supermarket scanner data for price and household consumption statistics, traffic inductive loops and automatic number plate recognition cameras for transport statistics, smart meters for energy statistics, and satellite imagery for land-use, agricultural and environmental statistics (ESSC 2014). These data, it is hypothesised, hold a number of potential opportunities and benefits for agencies producing official statistics (see Table 1), not least of which are more granular detail (at the level of individuals, address points, transactions), more timely production (Eurostat 2014b), entirely or partially replacing existing statistical sources, improving estimates from statistical sources, and providing additional complementary or entirely new statistical outputs (Florescu et al. 2014). In addition, they are generally direct measures of phenomena (rather than what people say they do), and are more timely being continually generated as a core element of system operation. They obviate survey fatigue, and they can add significant value for marginal cost given they are already being produced (Landefeld 2014; Struijs et al. 2014). However, these opportunities also require new accreditation procedures (Eurostat 2014a), and have their own unique quality and usability issues.

As a consequence, NSIs and related agencies such as Eurostat, the ESS, the United Nations Economic Commission for Europe (UNECE), the United Nations Economic and Social Council (ECOSOC), and the United Nations Statistical Division (UNSD) have

Table 1. Opportunities, challenges and risks of big data for official statistics.

Opportunities	Challenges	Risks
• complement, replace, improve, and add to existing data sets • produce more timely outputs • compensate for survey fatigue of citizens and companies • complement and extend micro-level and small area analysis • improve quality and ground truthing • refine existing statistical composition • easier cross-jurisdictional comparisons • better linking to other data sets • new data analytics producing new and better insights • reduced costs • optimization of working practices and efficiency gains in production • redeployment of staff to higher value tasks • greater collaboration with computational social science, data science, and data industries • greater visibility and use of official statistics	• forming strategic alliances with big data producers • gaining access to data, procurement and licensing • gaining access to associated methodology and metadata • establishing provenance and lineage of data sets • legal and regulatory issues, including intellectual property • establishing suitability for purpose • establishing data set quality with respect to veracity (accuracy, fidelity), uncertainty, error, bias, reliability, and calibration • technological feasibility • methodological feasibility • experimenting and trialing big analytic • institutional change management • ensuring inter-jurisdictional collaboration and common standards	• mission drift • damage to reputation and losing public trust • privacy breaches and data security • inconsistent access and continuity • resistance of big data providers and populace • fragmentation of approaches across jurisdictions • resource constraints and cutbacks • privatisation and competition

Source: Kitchin (2015, 473)

established big data working groups/committees and/or undertaken scoping and pilot research to evaluate the opportunities, challenges and risks of using big data in their work (e.g., ESSC 2014, Eurostat 2014b, UNECE 2014, ECOSOC 2015). The 2017 Bogota Declaration adopted by the 4th UN Global Conference on Big Data for Official Statistics recognised the need for greater collaboration and modernisation in pursuit of sustainable statistical resources, while proposing global platforms and partnerships to facilitate data sharing (UN Big Data Working Group 2017). In the European Union, the Heads of the NSIs signed the Scheveningen Memorandum (ESSC 2013) that committed them to examining and evaluating the use of big data in official statistics. In 2018, the ESSC approved the Bucharest Memorandum (ESSC 2018), providing additional considerations for modernizing official statistics, including skills training, national and international coordination, and responding to General Data Protection Regulations and other privacy concerns. Despite expanded mandates to use big data, there remain good reasons for hesitancy in mainstreaming the use of big data in creating official statistics, as set out in Table 1. Of particular concern are: gaining access to the data and associated metadata on a long-term, on-going basis; establishing and maintaining data quality; gaining and validating methodological transparency and ensuring long-term stability in the measurement and production processes; losing control of key aspects of the statistical system; and minimizing risks with respect to privacy and data security. These challenges and risks are not easy to handle and surmount. While still excited by the potential of using big data and pursing their prospective use, NSIs are still some way from passing judgement on their viability in compiling official statistics fearing that they might compromise the integrity of their products and their position as trusted agencies.

In this article, we explore these issues and the viability of using big data produced by several smart city technologies for creating new official statistics, resulting from a project conducted in partnership with a NSI (Central Statistics Office, Ireland). The smart city agenda seeks to improve urban life through the application of digital technologies to the management and delivery of city services and infrastructures and the solving of urban issues (Karvonen et al. 2018; Townsend 2013). While the term 'smart city' has gained popularity in recent years, there is a relatively long history of digital systems being used to understand and manage city services stretching back to the 1950s (Kitchin 2017). Given advances in digital technology and networking, a relatively broad set of smart city technologies are presently being used across the globe by city administrations, companies and citizens (see Table 2). The majority of these systems are used for operational purposes and generate and use big data to monitor, manage and direct system and infrastructure performance in or near real-time. Collectively, they produce and analyse a wealth of data about cities and several of the sources hold promise for supplementing or creating new official statistics related to transport and travel, environment, energy, waste, emergency services, and home life.

This article expands our understanding of the usability of big data by evaluating several open data sources largely produced outside of the fold of traditional sources of official statistics data (e.g., state survey and administrative data), including private-sector data, and identifying 15 criteria that are necessary to develop usable official statistics based on the suggestions of the Scheveningen and Bucharest Memoranda. Our analysis stems from a project that assessed the characteristics and potential of the big data sources available to

Table 2. Smart city technologies.

Domain	Example technologies
Government	E-government systems; online transactions; city operating systems; performance management systems; urban dashboards
Security and emergency services	Centralised control rooms; digital surveillance; predictive policing; coordinated emergency response
Transport	Intelligent transport systems; integrated ticketing; smart travel cards; bikeshare; real-time passenger information; smart parking; logistics management; transport apps; dynamic road signs; mobility tracking
Energy	Smart grids; smart meters; energy usage apps; smart lighting
Waste	Compactor bins and dynamic routing/collection
Environment	IoT sensor networks (e.g., pollution, noise, weather; land movement; flood management); dynamically responsive interventions (e.g., automated flood defenses)
Buildings	Building management systems; sensor networks
Homes	Smart meters; app controlled smart appliances

Source: Kitchin (2016, 12)

us as open data (see Table 3), and the challenges and risks in using these data for official statistics. The research was undertaken as part of a larger project focused on the creation of city dashboards. In general, a city dashboard uses visual analytics – dynamic and/or interactive graphics and maps to display and communicate information about the performance, structure, pattern and trends of cities (Kitchin et al. 2015). The data can be sourced from operational and administrative systems, generated by municipalities, government departments and NSIs. The operational data are typically big data and accessed using an API. The data are largely understood to be useful for tracing what is happening in the city in real-time, at particular time points, and over time, but are not considered validated official statistics for the city. Frequently there are no quality checks on such data, with APIs fed directly from the data sensor. These are supplemented with more traditional administrative and statistical data, preferably data published on a sub-annual basis (monthly, quarterly, half-yearly).

Our choice of what data to examine was largely defined by access and availability, and we invested time into trying to leverage additional closed data sets for use in the dashboard. All of the forms of data detailed in Table 3 were openly available through APIs and data stores in the two cities for which we have been building dashboards (Dublin and Cork). As can be seen by comparing Tables 2 and 3, it is thus apparent that a very significant barrier to using smart city technologies as a data source for official statistics is that the majority of data are not openly available, especially those that are produced by commercial enterprises. The Bucharest Memorandum and the UN Fundamental Principles of Official Statistics (UNECE 1992) stipulate that integration of official statistics be sustainable, accessible, collaborative, impartial, and transparent, which is not representative of proprietary data in general. Moreover, while some producers of official statistics might have powers to demand access to data, especially NSIs, it will be difficult to force companies to change their data formats, standards and systems to meet demands over commercial imperatives, or to hand over propriety data of commercial value, and

Table 3. The smart city data examined.

Data	Access	Spatial granularity	Spatial coverage	Temporal granularity	Real time	Query archive	What does it represent
Public transport GPS location	Poor. Some data supressed; given access to two month sample	GPS points	National	1 minute	Yes	Yes	Locations of public transit vehicles and expected time until next stop
Public transport at stops (bus, tram and train) – Real-time passenger information (RPTI)	Fair (system is occasionally down for a few hours)	GPS points	National	Seconds	Yes	Yes	Real-time expected time until next service at stop
Travel time/Roadway traversal time	Good national; poor local	Lines – roadway segments	Local and national	2 minutes	Yes	Yes	How long it takes to drive a vehicle along a stretch of roadway
Inductive loop counters	Poor. Access given to a 3 month sample.	GPS points	Local (800 across the city) and national	Unknown but likely seconds	Yes	No	How many vehicles pass over each induction loop
Car park spaces	Fair for local authority. Unavailable private car parks	GPS points	Car parks owned by local authority	2 minutes	Yes	Yes	Car park use and availability
Flight locations	Fair. Need receiver or pay subscription fee	GPS points	Global	$<$ second	Yes	No	Locations of all flights above a city
Flight arrivals/departures	Good	Airport	Airport	1 minute	Yes	Yes	Arrival and departure time of flights from airport
Maritime boat locations	Fair. Need receiver or pay subscription fee	GPS points	Global	1 minute	Yes	No	Locations of all boats in the harbour and coastal areas
Bikeshare	Good	GPS points	Bike stations in city	5 minutes at API level, event-driven at station-level	Yes	Yes	Bike usage and stand availability

Table 3. Continued

Data	Access	Spatial granularity	Spatial coverage	Temporal granularity	Real time	Query archive	What does it represent
CCTV cameras	Poor (feeds are often down for long periods)	GPS points	Local and national networks	5 minutes	Yes	No	Single static image sampled every 5 minutes from camera
Sound levels	Good	GPS points	14 locations in one local authority	5 minutes (aggregation of 3 readings)	Yes	Yes	Ambient sound levels
Air quality	Good	Regional	National	Unknown (minimum daily)	Yes	Yes	Air quality is an amalgamation of several measures of particulate matter and visibility
Air pollution levels	Poor (images but not data points)	GPS points	National (39 stations, not all counties covered)	30 minutes	Yes	No	Measure of pollution levels related to NO^2 and SO^2
Tide level	Good	GPS points	National (32 stations)	Varies by sensor, ~30 mins	Yes	Yes	Water level, wave height, wave period, water temperature, wind speed/wind direction, gust speed/-direction
River level	Good	GPS points	National (456 stations)	Unknown	Yes with delay	Yes	Height of water in waterways as a result of tides and rainfall drainage
Weather	Good	GPS points	National (82 stations)	1 minute	Yes	Yes	Measures of temperature, humidity, windspeed and direction, precipitation, road temperature

such demands are likely to be resisted and subject to legal action. Those data we had open access to are limited to just transport and environment domains and, as we discuss below, even these have significant other access issues, along with limitations around associated metadata.

2. Measures to Assess Smart City Data for Use in Official Statistics

In 2014, UNECE developed a hierarchical quality framework for evaluating big data that consisted of 14 quality dimensions grouped into three hyperdimensions: source (institutional factors); data (quality of the data themselves); and metadata (information about the data and its production) (UNECE 2014). Within the 14 dimensions are "factors to consider," which provide actionable measures to evaluate data against. Eurostat (2018) assigned the quality aspects of this framework to the three phases – input, throughput, output – of the business process of producing statistics to group quality aspects into seven key criteria: coverage, accuracy and selectivity; measurement error; comparability over time; linkability; processing errors; process chain control; and model errors and precision. Our research has sought to apply a modified version of the UNECE and Eurostat frameworks to evaluate the potential utility of creating new official statistics from open smart city data.

Table 4 diagrams the measures that we used to assess our data and rates each data set on a level of "not ready" to "good". A rating of "good" means that the data satisfies the necessary criterion for use in creating official statistics at the moment of evaluation. "Fair" ratings represent problems that users are aware of and accept or are able to correct with modest effort. Fair ratings are inadequacies for sustained data use, but do not preclude official statistics as institutional fixes can be implemented. A "poor" rating represents significant data issues that are largely unsolvable without taking a different approach to data generation, processing and sharing. The denotation "Not ready" indicates problems that can only be solved by data providers, which is a more significant barrier to using such data now or in the future, as it is outside the control of statistics compilers. Importantly, very few data sources are rated as "poor" or "not ready." Many of the problems that we observe in real-time data are solvable issues, rather than exclusionary ones. Ideally, for the data to be considered as suitable for use in official statistics we would want it to score well in relation to all of these criteria since a rating below 'good' in relation to any of them undermines the veracity, validity and utility of the data. We first discuss these measures and the reasons for examining them, before them to evaluate the potential of the 16 data sets to become the basis of official statistics (see Tables 3 and 4).

2.1. Source Factors

2.1.1. Access

Access refers to whether data can be obtained for third party use and analysis. Access to operational data for compilers of official statistics may be inconsistent and hindered by the inability to control the data collection process. Most smart city data remain closed, foreclosing a detailed analysis of their characteristics. The 16 data sets we analyse are the ones for which we could source API or FTP access. These data are generally open as part

Table 4. Assessing the potential utility of smart city data for official statistics

Hyperdimension	Source			Data				Metadata				
Data	Sustainability/ Reliability Status	Transparency and interpretability	Privacy	Fidelity	Cleanliness	Completeness	Coherence	Metadata availability	Changes over time	Standardisation	Methodological transparency	Relevance
Public transport GPS location	Good	Fair	Poor re. drivers, good re. public	Fair	Fair	Good	Good	Fair	Fair	Good	Good	Good
Public transport at stops (bus, tram and train)–Real-time passenger information	Fair	Not ready	Good	Fair	Good	Good	Fair	Fair	Good	Fair	Not ready	Good
Travel time/ Roadway traversal time	Fair	Good	Good	Fair	Good	Not Ready	Fair	Fair	Good	Fair	Fair	Fair
Car park spaces	Fair	Good	Good	Good	Good	Not ready	Fair	Fair	Fair	Good	Fair	Not ready
Inductive loop counters	Good	Good	Good	Good	Good	Not ready	Good	Not Ready	Good	Good	Good	Good
Flight locations	Good	Good	Good	Good	Good	Fair	Good	Good	Good	Good	Good	Good
Flight arrivals/ departures	Good	Good	Good	Good	Good	Good	Fair	Good	Good	Good	Good	Good
Marine vessel location	Good	Good	Good	Good	Good	Fair	Good	Good	Good	Good	Good	Fair

Table 4. Continued

Hyperdimension	Source			Data				Metadata				
Data	Sustainability/ Reliability Status	Transparency and interpretability	Privacy	Fidelity	Cleanliness	Completeness	Coherence	Metadata availability	Changes over time	Standardisation	Methodological transparency	Relevance
Bikeshare	Fair	Good	Good	Good	Good	Good	Good	Fair	Good	Fair	Not ready	Fair
CCTV cameras	Fair	Fair	Fair	Good	Good	Not ready	Poor	Fair	Good	Fair	Good	Poor
Sound levels	Good	Good	Good	Good	Good	Not ready	Good	Fair	Good	Fair	Good	Good
Air quality	Good	Good	Good	Fair	Good	Fair	Not ready	Fair	Good	Good	Fair	Fair
Pollution levels	Good	Not ready	Good	Good	Not ready	Not ready	Good	Good	Not ready	Good	Good	Good
Tide level	Good	Good	Good	Good	Good	Fair	Good	Good	Fair	Good	Fair	Good
River level	Good	Good	Good	Good	Fair	Fair	Good	Fair	Fair	Good	Good	Good
Weather	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good

of an open government agenda and are not necessarily open across jurisdictions. Many are tracked and released by local authorities or governmental institutions, but others are generated and made available by private owners. In our experience, private data generating entities are responsive to format requests, but access can be hindered by providers changing their access permissions and methods. Some restrictions to open access exist for several of our data sets. Bus location data is provided only through a two-month sample. For one data set – pollution data – we do not have access directly to the data themselves, but rather a visualization of the data for each measurement site: a graph that is updated in real-time and shows change over a variable number of days. Gaining access to the flight and marine tracking data requires some technical investment, purchasing an antenna and sensor, setting up an API, and hosting a feed to the data network. It is also possible to buy access to the data from a third-party app such FlightRadar24. The data in Table 3 are not universally accessible across cities, with some data unavailable to us in our second study city. In addition, some of the data are only partially available. For example, car park data is only available for state-owned car parks and not commercially-owned car parks.

2.1.2. Sustainability of the Entity-Data Provider and Reliability Status

Sustainability of the entity-data provider refers to whether their data host will be stable and steady, and *reliability status* is the trustworthiness of the data source to provide it. Official statistics require measurement systems to have longevity and reliably deliver data. When an agency is the producer they have charge of on-going data production. However, big data is usually controlled by other public and private entities and there is no guarantee that data will be reliably generated in perpetuity. Some of our data sets have been produced for a long time, for example weather measurements. Others are more recent, such as road speeds, real-time passenger information, car park spaces, bikeshare usage, and sound levels. It is difficult to know if bikeshare will still be operational in a decade, or if the scheme will be operated in the same way, producing the same kinds of data (our city has deployed a third generation scheme with fixed bike stations, fourth generation schemes in other cities have a different configuration) (Bradshaw 2019). Two additional data sets for which we originally had access were subsequently halted: emergency service response times (mainly because the data showed that the emergency vehicles were not meeting target response times) and footfall (because the private provider decided the data was too commercially sensitive to continue sharing publicly). For other data sets we have had intermittent access, for example, CCTV footage. In part, this is to do with local capacity to maintain data infrastructures and keep data stores and APIs up-to-date. Judging if a data source is still going to be in operation in a decade's time, and that it will continue to be measured in the same way, is difficult.

2.1.3. Transparency and Interpretability

Transparency and interpretability refers to confidence in and reputation of the data source and the portal where the data can be accessed, specifically through the availability of information about data generating processes. The OECD (2011) also discusses transparency through the concept of the source's credibility and is particularly concerned about political interference in the timing and content of data release, which is not a factor with streaming

data. Perceptions of confidence in the producers of real-time data do factor into the credibility of data that we have access to. Because of the frequency of public use and consequences for inaccurate bus arrival and location data, the perception of transparency and interpretability of Real-Time Passenger Information (RTPI) data is reduced, despite its actual or relative fidelity. Some data sources, such as roadway traversal times, have minor issues, but users expect inexactness and rely on knowledge of normal conditions to make use of imperfect data. Data portals share in the transparency and interpretability of the data they host. The provision of metadata is a responsibility to both data provider and host.

2.1.4. Privacy

Privacy concerns the selective revealing of aspects of oneself to others and protections regarding the accessing and disclosing of personal and sensitive information (Solove 2006). At an individual level, privacy consists of a number of different facets relating to personal and confidential information, and the protection of personal space, location, movement, and communication and transaction behaviour. In the big data age, individual level statistics are coveted and the need to maintain privacy and confidentiality present a dilemma for official statistics (MacFeely 2018). Privacy is considered a basic human right in most countries and its protection is enshrined in national and supra-national laws. Smart city technologies are known to pose a number of threats to privacy because they generate data at the individual scale and track and trace movement (e.g., ANPR cameras, MAC address tracking of smartphones, smartcard tap-in and outs) (Kitchin 2016). Rightly, privacy concerns are a key factor limiting access to these data and the 16 data sets that we have access to do not contain personal identifiable information (PII). For example, we know how many bikes are being used, but not who is riding them; or how many car parking spaces are free, but not whose car is in the car park; or what the noise level is, but not what or who is causing it. There are just two data sets that raise privacy concerns: the location of public transit vehicles, where drivers perceive the system as spying on their work performance; and the CCTV footage where the potential for an individual or vehicle to be individually recognised (the footage released in our case is too low resolution and at too far a distance for this to be the case). Some systems might also use PII to produce data, but the PII data themselves are not shared: for example, some travel time data is produced by tracking mobile network devices or number plates.

2.2. Data

In practical application of the UNECE framework it became clear that the 'accuracy and selectivity of data' measure needed to be unpacked further to more strongly differentiate different components of data quality: fidelity (accuracy, precision, bias), cleanliness (gaps, errors), coverage (spatial, temporal, attribute), and granularity (resolution). This was important in order to assess in detail the extent to which the data are fit-for-purpose for official statistics.

2.2.1. Fidelity

Fidelity refers to a collection of data issues that influence the ability to trust the data. Accuracy, precision, and bias are aspects of fidelity that are knowable and sharable to data users. Any data set using GPS can be prone to accuracy errors of several metres or more,

especially where there are channels, tunnels, tall buildings, or other obstructions to satellite coverage, and they can be misplaced onto the street network. Sensor instruments often record 'noise' or residual errors in their readings. With extensive observation of this data, regular noise can be modelled and negated for long-term use, but that does not help in the case of irregular instances of noise or when what appears to be noise is actually an unmodelled signal. There are general fidelity issues in most data sets, but what we are most interested in is whether they are made known by the data providers, with sufficient detail on how they are propagated through the process of public release so that any subsequent effects on official statistics can be planned for and presented. One example involves the presence of "ghost buses" which appear as arriving vehicles in the transport stops data, but not in the real world. These are known to exist, but their frequency is not presented or knowable without being present to observe the arrival of buses at stops. Most data sources are good and fair-to-good, but our impression from using the data sources and talking to data providers is that actual levels are unknown and are not reported in metadata.

2.2.2. Cleanliness

Cleanliness is a measure of the degree of effort required to make the data usable for display and analysis and standardisable for database storage. The cleaner the data, the less work is necessary to wrangle it into serviceable data. Cleaning and wrangling data is usually necessary if there are data formatting issues, missing values and errors, database structure issues. In a study of the veracity of open and real-time urban data, McArdle and Kitchin (2016) noted that smart city data often possess cleanliness issues, which are not addressed at source but need to be tackled by data users. This is also the case in our study, although in some cases it is difficult to achieve in practice because errors and miscodings are not always obvious. For example, the sound monitoring data occasionally reports date-times in an inconsistent format.

2.2.3. Completeness

Completeness refers to the attribute coverage and exhaustivity of the data, as well as any gaps in data production that would create a commensurate time-series gap in the database (which is a function of consistency of production). It can be the case that a system measures a number of attributes, but for reasons such as commercial sensitivity or privacy, only makes a limited set available, thus providing an incomplete set of data, which might limit analysis and modelling. While smart city technologies would claim to be exhaustive within a system (e.g., an Automatic Name Place Recognition (ANPR) system captures every single license plate passing a camera), this is nonetheless a sample of the entire population (e.g., not every car drives past cameras and not all roads are monitored for traffic). With respect to car parks, in our case we only have access to those owned by the local authority, not private enterprises (which far outnumber public ones).

The UNECE quality dimension of accuracy and selectivity refers to the representativeness of the data sample to the population as a whole. We further specify this dimension into granularity – resolution and level of disaggregation in the data set – and coverage.

2.2.4. Spatial Granularity

Spatial granularity concerns the spatial scale at which data are generated and published. Many official statistics do not currently have local scale application, though some are published at regional and sub-regional scales (in the EU context, NUTS scales 1–3), and big data sources have the potential to describe more localised variation. In the majority of our cases, the data can be tied to a specific known location (see Table 3). In one case, the data correspond to road segments, another to the site of the airport. Only one data set has a very coarse spatial scale, with the air quality index being reported for the city as a whole. While the recording of data at specific locations provides strong granularity, unless there is a dense network of recording sites it also introduces spatial sampling issues and potentially ecological fallacies. For example, vehicle counters on a handful of roads that record high traffic volume does not necessarily mean that all roads in the city are busy. A dense network of a couple of hundred sensors spread across the network would provide a better overall picture. In contrast, only a handful of weather sensors are required as weather has relatively low spatial variability. Ideally then, we would want the data to be from a network of sites with a sufficient density to minimise ecological fallacy issues.

2.2.5. Temporal Granularity

Temporal granularity refers to the velocity of data recording and publication. Ideally, we would want a relatively fine temporal granularity to be able to see patterns and trends over a micro-temporal scale, such as the course of a day or week, and to help identify noise/residuals in the data set. In our case, most of the data are updated every few seconds to five minutes (see Table 3). There are a couple that report every 30 minutes and a couple whose intervals are unknown. It should be noted that while the data are recorded in a timely fashion, in some cases their publication is delayed for various reasons. These delays are generally not for very long and should have little effect on their use in official statistics.

2.2.6. Spatial Coverage

Spatial coverage concerns the spatial selectivity in what is measured and recorded and the geographic extent to which the data refers. As Table 3 reveals, while in some cases the spatial coverage of data sets is national, with data being recorded at specific sites across a country-wide network (and in a couple of cases at the global scale), in other cases the spatial coverage is the city, or just part of the city (e.g., the city centre). Due to their cost, sensors are placed at particular locations of interest and have limited spatial coverage. Although this limited network can then be used to estimate values elsewhere in the city through extrapolation, the number of sensors must be sufficiently large to approximate a significant range of urban conditions. Limited spatial coverage makes comparison of phenomena across the state, and between states, unobtainable. Ideally, for use in official statistics, we would want data sets that have national coverage.

2.2.7. Coherence

The *coherence* between processes and methods and observed data values provides a sense of measurement validity; that is, how well the data signifies the phenomenon that it measures. In general, our sixteen data sets are considered to have good coherence, being

direct measures of phenomena (e.g., actual location, sound level, temperature) using scientific instruments and established procedures rather than composites, or proxies or approximations. Three of the measures are calculated data. The real-time passenger information and airport flight arrivals are projected measures based on the present location of vehicles. The road segment times is a projected travel time between two points on the road network given present traffic conditions. Car park spaces are calculated by comparing the number of cars entering and leaving through the entrance/exit barriers (in other cities individual spaces can be monitored by sensors in the ground or cameras). These estimates are considered by the agencies who produce them as robust. While CCTV photos do present a picture of a street, it is not clear what the data are representative of – that is, it gives an impression of the street, but is not directly measuring a particular phenomena.

2.3. Metadata Issues

2.3.1. Metadata

Metadata – data about data – is critical for being able to understand and assess a data set (Riley 2017). Metadata typically details information about data formats, standards, spatial and temporal coverage, methods, instruments, intellectual property rights, access, quality issues such as fidelity, calibration and errors, as well as provenance and lineage information with respect to who generated the data, for what purpose, history, ownership, and contact information. Metadata tells a data user important information about how the data should be used and provides a basis for assessing data quality and establishing trust. Metadata for the smart city data we assessed is generally weak, with little provided with the data other than very basic metadata such as attribute labels and units. Sometimes more information can be ascertained from searching documents/webpages on the data source website or through direct communication with the data provider. Adequate metadata may partially resolve many of the issues discussed simply by providing documentation that provides specific details and clarity. Essentially, though, a user is asked to accept the data as being valid and trustworthy without any supporting detail or documentation. This is a fundamental weakness with respect to using the data for official statistics as without strong metadata the integrity and accountability of the data set cannot be established. This is a significant issue for NSIs, which highly value public trust.

2.3.2. Changes Through Time

Changes through time refers to the stability of procedures, instruments, and modes of sharing data being used. There can be changes over time to a system, such as sensor being added to the network or removed from specific locations, as well as variations between sensors, such as inconsistent timing and sensor calibration. There can also be periodic changes in the API format used to access the data that affect data flow and the time-series archive. In general, smart city data seems to be quite consistent in production within a jurisdiction, though access can be patchy, and it is difficult to know if there have been instrument or procedural changes unless documented. Frequently, temporary monitoring will focus on a specific geographic area, then shift to another, creating spatio-temporal inconsistency.

2.3.3. Standardisation

Standardisation is a common issue when data crosses between jurisdictions, which is why there is so much effort being invested in standardisation initiatives by national and supranational bodies (ANSI 2016; White 2019). For example, road speed is already calculated in different ways within and between jurisdictions (by mobile devices, by ANPR, by inductive loops); will this become standardised, or will it be measured in an entirely different way in years to come as technology develops? The geopolitical structure of Dublin requires consistency in data sharing between four local authorities that comprise the city, and sometimes competing management strategies at the local and county levels. This is a serious hindrance to using such data for official statistics and largely prevents global- and some national-scale statistics. Standardisation is a high priority for improving the comparability of new statistics across spatial scales.

2.3.4. Methodological Transparency

Methodological transparency concerns the degree to which the data user has full access to information on how the data were generated and are interpretable. This includes information on: the instruments used; selection criteria for measurement sites, sampling frameworks and parameters; any techniques of data preparation/processing/validation (including calibration, error checking/fixing, cleaning or wrangling); and any methodological analysis applied prior to data being circulated. Methodological transparency is the big data parallel to paradata in survey methods, similarly used to better understand the data collection process. Such information provides the user with important contextual insight for any downstream processing, analysis and interpretation. Discovering such information for smart city data is often difficult. In one example, the lack of methodological transparency in bikeshare administration impacts the coherence and standardisation of bikeshare data. Bike-station sensors measure the number of available bikes and empty stations available to customers, but includes no indication of when or how often the bikeshare operator moves bikes wholesale by truck across the city to redistribute bikes and restock popular stands. Although reflected in the data, this is not a pattern resulting from normal operation, so would be important to account for in creating usage ridership statistics. Another important consideration of methodological transparency is knowing how much, if any, quality control is performed on the data before it is released. Intervention by the data manager, usually a state agency concerned about privacy or misrepresentation, could reduce timeliness and introduce bias. Sometimes methodological information can be found by hunting around on websites, or can be sourced through communication with the provider. If some aspects are included in the metadata it is often rudimentary in nature. Methodological transparency is highly valued by NSIs because it enables trust and accountability, and its general absence will be an issue for them if the data are to be used as official statistics.

2.3.5. Relevance

Relevance is difficult to evaluate because it is a measure of the data's ability to serve the purpose of users (OECD 2011). We add relevance to our framework to explicitly consider the users and uses of data as they help understand its practicality. Knowing who uses the

data and for what purposes are difficult questions for data providers to answer (Reboot 2017). Relevance is high where data is a direct measurement of what users want to know about the city – locations of features, measurements of sound, weather observations, and so on. The signal of good data relevance might be surveys of users of open data through portals or other data providers, ensuring high response rates for a broad user spectrum, and frequent sample rates to capture changes in user objectives and satisfaction. Relevance is a necessary measure in the context of official statistics because it is the only measure concerned directly with how the data is used, rather than evaluating its quality.

3. Evaluating Our 16 Data Sources for Official Statistics

Using our assessment measures we consider possible official statistics that each of our 16 data sources can feasibly generate. Ideally, for a data set to be considered suitable for official statistics it should be accessible, have national coverage (or at least coverage in every city), be timely (see Table 3), and be rated as 'good' for all the categories in Table 5. As Tables 3 and 4 make clear, of our 16 data sets, only weather data fits these criteria precisely. Several others – flight locations, flight arrivals/departures, marine vessel locations, tide level, and river level – are rated as 'good' with only a few 'fair' rankings for our criteria. Most of these are long-standing official data sets produced by state agencies, with records often going back many decades. The other data sets while having operational utility for managing systems and infrastructure, and public utility in informing and aiding citizens' decision-making, do not have sufficient robustness to be used to construct official statistics at present.

However, presuming that the issues preventing these data sources from being used reliably at present can be fixed, we consider the official statistics that they might help to generate. Although fixing the deficiencies in these data is beyond the scope of this article, Table 5 describes two sets of statistical product that can be derived from each of our data sets. Key Performance Indicators (KPI) are operational measures for monitoring how a system is performing over time, often with respect to targets. They are usually direct counts of a phenomena, sometimes calculated against a norm, that are of practical importance. They can be produced across organisations producing data about their own systems and are common in smart city contexts for tracking how urban systems are performing. They are a staple component of city dashboards. Official statistics – produced by a variety of agencies – extend beyond the tracking of operational performance to collate and convert that information into higher level, comparable statistics that span systems to enable the answering of questions of strategic interest. While official statistics can be used as KPIs by organisations, not all KPIs are suitable to be used as official statistics given their operational rather than strategic utility. We divide our discussion into environment and transport domains.

3.1. Environmental Statistics

Weather presents the most optimistic view of big urban data as a source for official statistics on environmental conditions. They are already considered official state data and are managed by dedicated national statistical organisations to produce trustworthy and timely reports. Weather information is collected at fine spatial and temporal scales with

Table 5. Smart city KPIs and potential official statistics.

Data	KPIs (want to know in the here and now)	Stat (higher level system comparison)
Public transport GPS locations and at stops (bus, tram and train) – Real-time passenger information	# vehicles on the system % of late buses at each individual stop across course of day	% deviation from timetable of all buses/light rail/ trains on the system # of vehicles which arrive within 5 minutes of schedule
Travel time/Roadway traversal time	Congestion against normal (free flow) speed	Is congestion getting better or worse over time? % time per day where congestion above free flow.
Car park spaces	% number of spaces in use against capacity % number of spaces against norm at that time of day	Is parking getting better or worse over time? Average % spaces in use over course of day Number of commuters into the city
Inductive loop counters	Number of vehicles passing through a junction at a given time	Is the volume of traffic growing/decreasing over time? Total volume cars passing through measured intersections over course of day Intensity of rush hour What is the hour of peak congestion?
Flight locations and arrivals/departures	# flights arriving/departing	% of flights arriving more than 15 minutes late % of flights departing more than 15 minutes late % of flights from/to different geographic locations
Maritime boat locations	# of active cargo/passenger ships # of incoming and outgoing ships	International commercial activity How many travellers are visiting the city via passenger ships?
Bikeshare	Stand occupancy at time/date Bikes available at time/date Bikes in transit at time/date Empty stations at time/date Full stations at time/date	Is the use of bikeshare increasing/decreasing over time Average % of total stand occupancy Amount of non-motor-vehicle travel % occupancy above normal for time/date
CCTV cameras (with image processing)	Traffic concentrations Concentrations of certain types of vehicles Pedestrian volume	

Table 5. Continued

Data	KPIs (want to know in the here and now)	Stat (higher level system comparison)
Sound levels	Decibel level per station	Is sound decibel levels increasing/decreasing?
	Number of time maximum decibel level breached.	Average decibel readings across day, across all stations
		How frequently are EU sound limits breached over a timeframe?
		% above normal for time/date
Air quality	AQIH code for a region	How high is pollution in a city or region?
Pollution levels	Concentration of: particulate matter at multiple granularities, ozone, nitrogen oxide, sulfur dioxide	Are pollution levels above or below agency-defined limits to exposure?
		How frequently are exposure limits passed in a given timeframe
Tide level	Tide level at a given time/date	Is average tidal level increasing/decreasing over time?
	Wave height	
River level	Water level at a given time/date	Number of times flood level reached per year
		Is average water level increasing/decreasing over time?
		Extent to which water level above norm for time/date
Weather	Temperature	Average temperature for city
	Rainfall	Average wind direction/speed
	Wind speed/direction	Road temperature
		Temperature/rain anomalies

near exhaustive geographic coverage, and is consistent in the features that it measures. Big data weather sources include crowdsourced and distributed sensors across a city-region providing real-time conditions at several sites, rather than regional and national forecasts. Together, these make weather ideal as a comparative statistic when real-time observations are tallied over meaningful time periods. Statistics easily compiled from big data include average high and low temperature, hours of sunlight, amount of precipitation, wind direction and speed, and other measurements as collected by weather sensors. Although extensive histories of weather statistics exist already, big data can continue to contribute to the comprehensive representation of weather statistics with crowdsourced observations increasing the spatial coverage and representativeness of its measurements.

Two other environmental sensors – river level and tide level – are nearly suitable in their current state to be used in the creation of official statistics. In the case of river level

data, there has been a recent expansion in the network of sensors, but data access has been restricted for fear of causing panic with respect to flood events. The KPI of measuring water level at points along urban waterways is potentially misleading without information regarding normal water levels or historical flood levels or contextual events such as high tides. Smart cities integrate live water levels with predictive models and social media to raise situational awareness in the event of flooding, yielding greater coverage of waterways with active sensors. Thus, the official statistic we designate for river level sensing data considers the frequency at which the water level surpasses the flood threshold at any point along the monitored waterways. Measuring river levels with high temporal frequency increases preparedness in the event of rapidly rising waters, but defining and comparing historical occurrences of high water with respect to a defined threshold allows for suitable comparison of the tendency of an area to experience objectively high waters. With sufficient sustainability to establish long-term, high resolution records, such sensors of water level rise will provide valuable quantitative documentation of climate change.

In addition to the data that are mostly rated as 'good,' there are a couple of data sets that could be used for official statistics with some minor modifications to their constitution. The most obvious data set is pollution sensor observations. The automatically-generated images we have access to are not useful for creating and comparing statistics, but if the raw data were made openly available (and it could be made available to an official state agency, given it is created by such a body), this data set would help create critical environmental indicators for urban areas. Most sensors measure similar pollution conditions, including particulate matter and concentration of contaminants including ozone, nitrogen oxide, sulphur dioxide, and others, enabling comparison of specific concentrations over time and between locations. International and national standards are set for these pollutants, so an ideal official statistic would compare the frequency at which areas pass the specific limits established by the European Union and the European Environmental Agency (European Union 2008), and the World Health Organization (WHO 2006). Statistics reporting how frequently each of these agency-specific limits are breached each year are stated in reports by the Environmental Protection Agency already. The addition of real-time monitoring information can indicate the current status of urban pollution monitoring with respect to established limits, and also indicate rolling year-to-date measures and other timeframe-specific frequencies for comparing places.

Air quality data does have national coverage, but its spatial granularity is regional, its reporting done in a limited number of categories, and it lacks methodological transparency. The Air Quality Index of Health (AQIH) is a numerical code between one and ten that summarizes the presence of the same particulate matter and pollutants as measured by the previously discussed pollution sensors. The AQIH codes are defined by quantitative intervals of each of the pollutant concentrations and should already be considered an official statistic, as it is validated and published by the responsible state agency. However, with its reduced methodological transparency – the derivations of the specific numerical limits between categories is unknown – and aggregated spatial and temporal scales, this statistic still violates some of our measures. Similar to the specificity of pollution sensors, the AQIH could be a comparative tool as a summary of pollution sources at high temporal resolution, albeit with a spatial resolution which lacks usable detail.

The sound sensors are perhaps the data set with the most promise, given they are a robust, clear, consistent set of data, but are presently hampered by relatively weak spatial coverage, availability issues in other jurisdictions, standardisation issues across jurisdictions (there are 40 sound sensor across four local authorities that make up our city, only 14 of which are publicly available, all from a local authority), and the long-term insecurity associated with a private sector monitoring company. Although the EU requires that member nations report noise maps to their publics, it does not specify how or if sound monitoring should be carried out. Often, such reports on noise pollution are created with computational models involving no sensor measurements at all (Khan et al. 2018), creating inconsistency in reporting and an opportunity to leverage active sensing. The EU also specifies several value limits that define certain levels of sound above which constitute undesirable noise. These value limits provide a resource for creating official statistics, measuring the frequency at which measurements pass the given limits and creating a comparable measure of noise pollution. Any composite statistic on sound at the scale of the city depends on the proximity of the sensors to sources of sound such as vehicle traffic and industry. Ensuring comparability between jurisdictions requires placement of sensors in locations with similar sound sources in range. Sound statistics are valuable tools of environmental health and with some improvement in spatial coverage of monitoring stations, this data source could be a valuable statistical resource.

3.2. Transportation Statistics

Open data pertaining to the current operation of public transportation is a service to users of busses, trains, and light rail systems, but producers less frequently release historical records of such data. Data on bus location and arrival times can be used to indicate the efficiency of service and deviations from the expected schedule. One way to do this might represent the frequency at which busses arrive within five minutes of their scheduled time in a given time period and provides a way to compare public services across different cities. However, despite the difference between data problems and transportation systems problems reflected in data, transparency issues might hinder the creation of trusted official statistics.

Inductive loop counters do have national coverage, but their local spatial coverage is usually restricted to busy intersections on main roads and within cities. That said, its other characteristics are strong and it could provide a useful data set for measuring traffic volume and congestion in major traffic-prone areas. As with weather, local and national transportation institutions manage some statistics on vehicle movement already. At the local level, our ability to access this data in sufficient detail to enumerate directional travel is limited by inadequate metadata to explain the numerical codes used to designate direction and lane of travel. The national roadway sensors do not have this limitation. However, local roadways still produce volume of vehicles on the road at any given moment of observation so it is possible to monitor critical traffic patterns associated with the working day and commuting. KPIs pertaining to traffic sensing data will typically correspond to current traffic volumes at specific locations, but an official statistic might consider the patterns formed throughout the day by measuring the difference between free flow volume and rush hour volume, the peak hours of high congestion, the differences

between days of the week in traffic volumes, and many others. Traffic patterns associated with specific areas are useful indicators of local activity, so classifying time-series patterns of traffic volumes is an important comparative process (Celikoglu 2013).

Car parks frequently track their current occupancy in order to communicate the number of available spaces for parking vehicles. The opening of that data to public interfaces is increasingly common but has issues that prevent it from being an ideal source of statistics, specifically its inconsistency associated with privatisation of the data and car parks themselves, lack of indication as to how occupancy is calculated, rapid rate of change, and missing metadata. Key performance indicators from this data reflect the occupancy of specific car parks at a given point in time, and can also reflect the deviation in occupancy from what might be considered normal at the given time. Since data from car parks is not a comprehensive representation of all parking in an urban area, this data is only representative of the vehicles at a specific place, and official statistics cannot represent the full presence of parked vehicles in the city. But assuming that additional data can be obtained, car park use would be a valuable part of a composite statistic on the use of transportation networks. Changes in the vehicles parked at specific locations can indicate commuting patterns, but in order to represent such a comprehensive set of spatial and temporal patterns, this data must also incorporate other sources of information about where vehicles are, including street parking, private car parks, carpooling, public transportation, and more.

Bikeshare data is increasingly ubiquitous in smart cities, but presents some barriers to creating official statistics. Although some bikeshare schemes release data about specific trips including origin and destination stations and time spent en-route, our bikeshare scheme data only produces the number of currently available bikes and open stands at each station every two minutes. This presents some methodological issues, where the exact number of incoming and outgoing bikes during that two-minute interval is unknown. Because of this, as well as the private nature of bikeshare operators, standardisation between jurisdictions and over time are critical issues limiting the consistency and longevity of the data. Although bicycles represent only a fraction of the means by which people move through a city, their use contains patterns indicative of commuting and non-vehicle transportation. KPIs thus measure bike use at the level of the station. The number of bikes in use, number and location of empty and full stations, and the deviation from normal for a given station can reveal spatial and temporal patterns. Official statistics for bike use data, like many sources of transportation information, are more useful when combined with other data that enumerates how people move throughout the city. Combined with car park data, traffic sensors, footfall, and others, bike share data could approximate the temporal patterns of commuting, and a subset of these available transportation data suggest "green" commuting modes. Alone however, this data is only able to represent the use of this particular bikeshare program, its average use over a given time period, and whether bike ridership is changing over time. Such information could still provide useful comparisons across jurisdictions, but improvement of the source for consistent use is necessary.

Flight arrivals and departures is a valuable data set despite the appearance of its limited spatial scope. Flight information is limited to the point of airports, but also contain information about the movement of people and goods in and out of a city. Since flight arrivals and departures are strictly scheduled, this data not only explains the occurrence of

air travel, but how current flights compare to their intended times of arrival and departure. We enumerate this data in a KPI which explains the proportion of current arrivals and departures which are on-time with their schedules. This measures the current state of air travel, with more specific and long-term official statistics that measure the tendency of lateness in arrivals and departures as a function of time and the amount of lateness. No official airline industry standard is used to quantify lateness, but the Official Aviation Guide has measured airport and airline performance using 15 minutes from scheduled departure and arrival times to signify late aircraft (OAG 2019). OAG is considering a more dynamic measurement of lateness based on flight duration, time of day, cultural expectations, and other factors, but in the absence of an industry standard, 15 minutes is considered an acceptable standard. Although this is a reflection of the airport more than the geographic area, the lateness tendencies of flights in and out of the city are important metrics for visitors, especially when multiple airports carry passengers, or it is a connecting hub. Finally, we suggest a statistic aimed at measuring the international networks facilitated by air travel via the origins and destinations of scheduled flights. The proportion of incoming and outgoing flights to each continent highlights the political, social, and economic connections in the movement of people and goods between global entities. Locations of arriving and departing flights do not change frequently, but NSIs and others may collect origin and destination statistics on a monthly or quarterly basis, revealing geographic networks of travel and their changes over time.

Similarly, marine vessel locations might measure the transport of goods and people to urban areas with major port systems. In contrast to flight tracking, only some maritime transportation is comparable with known schedules. Automatic Identification System (AIS) marine traffic is used as a trusted data source by international statistics institutes already (Eurostat 2018). As the quantity and type of marine vessels change, official statistics can also estimate the movement of goods, whether or not the data actually contains reference to the nature of a ship's cargo. With origin and destination information of tracked maritime vehicles, it is possible to quantify current, past, and future loadings and off-loadings of cargo ships. Thus, a useful official statistic pertaining to city commercial activity might be derived, at least in part, by data representing maritime trips associated with a given location, as the UN Conference on Trade and Development currently does (Zein 2020). Better inter-jurisdictional standardisation would allow for combining data from ports to acquire a global picture of maritime trade at a fine temporal scale.

It is not clear what official statistic the CCTV photos might be used for since they are uniquely image data and have not been designed to measure a particular phenomenon. This undirected monitoring, such as exists as well in social media as a data source, requires methodological transparency to convert such data into specific official statistics (Severo et al. 2016). CCTV cameras at traffic intersections have been used previously in license plate recognition, speed detection and enforcement, and others, so automatic image processing could be utilised to count vehicles, identify certain styles of vehicles, measure pedestrian volume, detect crime incidents, observe weather events, and more. The value of these statistics have been discussed in previous sections, where they were detected directly with specific sensors without the violations to privacy that ANPR makes possible. Despite recent advances in image processing for these purposes, CCTV is an inefficient and inaccurate method of deriving official statistics about many urban features. We have given

this data source a rating of "poor" in multiple categories, reflecting the lack of coherence of any specific urban feature and undefinable relevance as it is unclear what uses public users get from this imagery. In our view, CCTV has little to offer official statistics that cannot be satisfied through other sensing means.

4. Conclusion

Through our analysis we have considered the potential utility and limitations of smart city big data for producing official statistics. Our adaptation and application of the UNECE (2014) framework for evaluating big data indicates that while the data might be useful for producing operational KPIs, where the data are sufficient quality for the intended purpose, they generally lack sufficient qualities for creating official statistics. Our limited access to open data sources that are environment and transport related, which lack personal identifiable information limit us from making a generalised assessment of privacy. This is not the case for many big data sets being considered for producing official statistics and in this sense, the data we have evaluated are outliers with respect to privacy concerns. Our positive assessment of privacy then is a function of data type and our strict use of open data sources. Agencies compiling official statistics and data protection offices should maintain proactive legal measures to prioritise effective management of PII in potential data sources.

Access is a key issue. Our focus has been limited to environment and transport data because data produced by most of the systems and infrastructure set out in Table 2 has not been available to us. Similarly, access is a major issue with other big data sets being considered for use in official statistics. Many piloted studies have used samples of otherwise closed data, such as mobile network data and consumer transaction data, which can make assessment difficult. Even in cases where we have gained access, our ability to fully evaluate the data against our criteria was sometimes limited, but may be solvable with effective collaboration. Since such closed data is rarely made available for public consumption, even with adequate anonymisation and individual protections it raises concerns with respect to interpretability, privacy, and methodological transparency. For closed data to be trusted – meaning it has fidelity and transparency and maintains privacy – official statistics providers must have strong relationships with data holders. Good communication remains a necessary means of resolving some data inadequacies, especially metadata, but when internal communication is prioritised over openness, methodological and source transparency as well as relevance could suffer for non-institutional users.

We need to distinguish between interesting and useful data, and data that are also suitable for official statistics. There is no doubt that much smart city real-time data are high value data sets that provide useful data for operational purposes, for understanding how a city is functioning, and for citizens making decisions. The data can be used to build apps and for undertaking prediction, simulation, optimisation tasks, and building analytic models that combine various data. This does not mean though that they are necessarily available or suitable for producing official statistics. Indeed, our analysis indicates that only a limited set of smart city real-time data is presently suitable for creating new or augmenting established official statistics. Others could be made suitable with some changes to data management.

We can also consider the complexity of the changes necessary to fix some of the issues we have observed here. Some issues – such as access, standardisation, and sustainability – are further out of our influence as data users than others, such as NSIs. Although institutional limits impede much of our efforts to improve the issues of access, cross-jurisdictional standardisation, metadata, and sustainability, we do anticipate that some less comprehensive changes can reduce some of the problems we have observed. First, built-in procedures for ensuring anonymisation would immediately reduce privacy concerns, but are also necessary to satisfy some of the concerns that are held by, for example, public transit drivers, over the public release of their locations and other information. Spatial coverage and temporal granularity can be improved upon by investment in the monitoring data management infrastructure. Additional sensors improve spatial coverage, although the issue still persists without complete coverage, depending on sensor range and the physical landscape. Temporal granularity can be improved by increasing the sampling rate of real-time sensors, which places a larger demand on data storage volume and transfer speeds. Finally, many of the issues we are faced with are not data issues, but factors stemming from inadequate documentation. Methodological transparency, metadata, and to a lesser extent, relevance, depend on adequate documentation of the processes of data generation, data dictionaries, provenance, and data descriptions. If a real-time data set is created to address a specific need or as a by-product of a new smart city technology, then documentation may be a low priority upon releasing the data openly, causing usability issues. Stricter requirements on documentation of open data would demonstrate a commitment from data managers to producing high quality usable urban data.

In response to acknowledgements stated by the Scheveningen and Bucharest Memoranda to produce official statistics from real-time data, NSIs and academics have been piloting studies of various big data sources. The pilot studies which have emerged to fulfil these objectives has been limited in number and scope, in part due to the issues we have identified. With lots of interesting but inconsistent data available to drive these efforts currently, it becomes necessary for NSIs to engage with data producers to address issues of data access, data quality, metadata and standards to create consistent, long-term longitudinal data sets and facilitate the demand for real-time and granular official statistics. This will have to be accompanied by governance arrangements that set out who is responsible for the various parts of the data life cycle and statistical production. Prior to this, further systematic testing of the data is required to validate the suitability and utility of the data, including the pilot generation of potential new official statistics as proposed in Table 5. Finally, the advantages that smart city technologies provide for statistics over survey-based methods should be emphasised, increasing the demand for quality data and motivation to increase its usability. Given this current state of play, despite the potential opportunities afforded by smart city data for official statistics, it is likely to be some time before such sources become a trusted part of the statistical system of NSIs.

5. References

ANSI 2016. *Directory of Smart and Sustainable Cities Standardization Initiatives and Related Activities.* American National Standards Institute Network on Smart and

Sustainable Cities (ANSSC). Available at: https://www.ansi.org/standards_activities/-standards_boards_panels/anssc/overview#Standards (accessed February 2021).

Bradshaw, R. 2019. "Instrumentalization in the Public Smart Bikeshare Sector." PhD Doctoral thesis, Maynooth University. Available at: http://mural.maynoothuniversity.ie/10509/ (accessed February 2021).

Celikoglu, H.B. 2013. "An approach to dynamic classification of traffic flow patterns." *Computer-Aided Civil and Infrastructure Engineering*; 28, 4: 273–288. DOI: doi.org/10.1111/j.1467-8667.2012.00792.x.

ECOSOC. 2015. *Report of the Global Working Group on Big data for official statistics. United National Economic and Social Council.* 46th Statistical Commission. Available at: http://unstats.un.org/unsd/statcom/doc15/2015-4-BigData.pdf (accessed February 2021).

ESSC. 2013. *The Scheveningen Memorandum on Big Data and Official Statistics.* Scheveningen: Directors General of the National Statistical Institutes (DGINS). Available at: https://ec.europa.eu/eurostat/cros/system/files/SCHEVENINGEN_MEMORANDUM%20Final%20version.pdf. (accessed February 2021).

ESSC. 2014. "ESS Big Data Action Plan and Roadmap 1.0." Riga: European Statistical System Committee. 26th September 2014, Riga. Latvia. Available at: https://ec.europa.eu/eurostat/cros/content/ess-big-data-action-plan-and-roadmap-10_en (accessed February 2021)

ESSC. 2018. "The Bucharest Memorandum on Official Statistics in a Datafied Society (Trusted Smart Statistics)." Bucharest: Directors General of the National Statistics Institutes (DGINS), 12th October 2018, Bucharest, Bulgaria. Available at: https://ec.europa.eu/eurostat/documents/7330775/7339482/The + Bucharest + Memorandum + on + Trusted + Smart + Statistics + FINAL.pdf (accessed February 2021).

European Union. 2008. "Directive 2008/50/EC of the European Parliament and of the Council of 21 May 2008 on ambient air quality and cleaner air for Europe." *Official Journal of the European Union* 51: 1–44. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32008L0050&from=EN (accessed February 2021).

Eurostat. 2014a. *Accreditation procedure for statistical data from non-official sources*. European Commission. Available at: https://ec.europa.eu/eurostat/cros/system/files/-D5_Accreditation%20procedure%20for%20statistical%20data%20from%20non-official%20sources_20140206_0.pdf (accessed February 2021).

Eurostat. 2014b. "Big data – an opportunity or a threat to official statistics?" Paper presented at the Conference of European Statisticians, 62nd plenary session, Paris, 9–11 April, 2014. Available at: http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/2014/32-Eurostat-Big_Data.pdf (accessed February 2021).

Eurostat. 2018. *Report describing the quality aspects of Big Data for Official Statistics*. ESSnet Big Data, Work Package 8, Deliverable 8.2. Available at: https://ec.europa.eu/eurostat/cros/sites/crosportal/files/WP8_Deliverable_8.2_Quality_aspects.pdf (accessed February 2021).

Florescu, D., M. Karlberg, F. Reis, P.R. Del Castillo, M. Skaliotis, and A. Wirthmann. 2014. "Will 'big data' transform official statistics?" European Conference on the

Quality of Official Statistics. Available at: http://www.q2014.at/fileadmin/user_upload/ESTAT-Q2014-BigDataOS-v1a.pdf (accessed February 2021).

Karvonen, A., F. Cugurullo, and F. Caprotti. 2018. *Inside Smart Cities: Place, Politics and Urban Innovation*. London: Routledge.

Kahn, J., M. Ketzel, K. Kakosimos M. Sorensen, and S.S. Jensen. 2018. "Road traffic air and noise pollution exposure assessment – A review of tools and techniques." *Science of The Total Environment* 634 no. 1: 661–676. DOI: https://doi.org/10.1016/j.scitotenv.2018.03.374.

Kitchin, R. 2015. "The Opportunities, Challenges and Risks of Big Data for Official Statistics." *Statistical Journal of the International Association of Official Statistics* 31. 3: 471–481. DOI: https://doi.org/10.3233/SJI-150906.

Kitchin, R. 2016. *Getting smarter about smart cities: Improving data privacy and data security*. Data Protection Unit, Department of the Taoiseach, Dublin, Ireland. Available at: http://mural.maynoothuniversity.ie/7242/1/Smart (accessed February 2021).

Kitchin, R. 2017. "Data-Driven Urbanism". In: *Data and the City*. Kitchin R., G. McArdle, T. Lauriault, eds.: 44–56. Routledge: London.

Kitchin, R, T. Lauriault, and G. McArdle. 2015. "Knowing and governing cities through urban indicators, city benchmarking and real-time dashboards." *Regional Studies, Regional Science* 2: 1–28. DOI: https://doi.org/10.1080/21681376.2014.983149.

Kitchin, R., and G. McArdle. 2016. "What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets." *Big Data and Society* 3: 1–10. DOI: https://doi. org/10.1177/2053951716631130.

Landefeld, S. 2014. "Uses of Big Data for Official Statistics: Privacy, Incentives, Statistical Challenges, and Other Issues." Paper presented at International Conference on Big Data for Official Statistics, 28–30 October 2014, Beijing, China. Available at: https://unstats.un.org/unsd/trade/events/2014/beijing/Steve%20Landefeld%20-%20Uses%20of%20Big%20Data%20for%20official%20statistics.pdf (accessed February 2021).

MacFeely, S. 2018. "The privacy dilemma for official statistics in a big data world." *Open Access Biostatistics and Bioinformatics* 2. 1: 1:3. DOI: https://doi.org/10.31031/OABB.2018.02.000526.

McArdle, G, and R. Kitchin. 2016. "Improving the Veracity of Open and Real-Time Urban Data." *Built Environment* 42, 3: 446–462. DOI: https://doi.org/10.2148/benv.42.3.457.

OAG. 2019. *Defining Late: is 15 minutes the right measure?* Official Aviation Guide, Worldwide Aviation Limited. Available at: https://www.oag.com/hubfs/Defining_Late/Defining-Late-Report.pdf?hsLang = en-gb (accessed February 2021).

OECD. 2011. *Quality Framework and Guidelines for OECD Statistical Activities*. Organisation for Economic Co-operation and Development, Statistics Directorate. Available at: https://www.oecd.org/sdd/qualityframeworkforoecdstatisticalactivities.htm (accessed February 2021).

Reboot. 2017. *Understanding the Users of Open Data*. Reboot Inc. for NYC Open Data. Available at: https://opendata.cityofnewyork.us/wp-content/uploads/2017/07/Understanding-the-Users-of-Open-Data_Reboot.pdf (accessed February 2021).

Riley, J. 2017. *Understanding Metadata: What is Metadata, and What is it For?: A Primer*. Baltimore: National Information Standards Organization. Available at: http://

www.niso.org/publications/press/UnderstandingMetadata.pdf (accessed February 2021).

Severo, M, A. Feredj, and A. Romele. 2016. "Soft Data and Public Policy: Can Social Media Offer Alternatives to Official Statistics in Urban Policymaking." *Policy and Internet* 8, 3: 354–372. DOI: https://doi.org/10.1002/poi3.127.

Solove, D.J. 2006. "A Taxonomy of Privacy." *University of Pennsylvania Law Review* 154. 3: 477–560. Available at: https://scholarship.law.upenn.edu/penn_law_review/-vol154/iss3/1 (accessed February 2021).

Struijs, P., B. Braaksma, and P.J.H. Daas. 2014. "Official statistics and Big Data." *Big Data and Society* 1, 1: 1–6. DOI: https://doi.org/10.1177/2053951714538417.

Townsend, A. 2013. *Smart Cities: Big data, Civic Hackers, and the Quest for a New Utopia*. New York: W.W. Norton & Co.

UN Big Data Working Group. 2017. *Bogota Declaration*. United Nations 4th International Conference on Big Data for Official Statistics. 8–10 November 2017, Bogotá, Colombia. Available at: https://unstats.un.org/unsd/bigdata/conferences/2017/Bogo-ta%20declaration%20-%20Final%20version.pdf (accessed February 2021).

UNECE. 1992. *Fundamental Principles of Official Statistics*. United Nations Economic Commission for Europe. Available at: https://unece.org/statistics/fundamental-princi-ples-official-statistics (accessed February 2021).

UNECE. 2014. *A Suggested Framework for the Quality of Big Data*. United Nations Economic Commission for Europe. Available at: https://statswiki.unece.org/downloa-d/attachments/108102944/Big%20Data%20Quality%20Framework%20-%20final-%20Jan08-2015.pdf (accessed February 2021).

White, J.M. 2019. "Standardising the city: A material-discursive genealogy of CPA-I_001, ISO 37120 and BSI PAS 181." Doctoral thesis, Maynooth University. Available at: http://mural.maynoothuniversity.ie/10848/ (accessed February 2021).

WHO. 2006. *WHO Air quality guidelines Global update 2005: particulate matter, ozone, nitrogen dioxide and sulfur dioxide*. World Health Organization Regional Office for Europe. Copenhagen. Available at: https://apps.who.int/iris/handle/10665/107823 (accessed February 2021).

Zein, A. 2020. "Short-term effects of the coronoavirus outbreak: what does the shipping data say?" *UNCTAD Transport and Trade Facilitation Newsletter N°85 – First Quarter 2020*. 48.

Received October 2019
Revised June 2020
Accepted December 2020