Developing smart statistics for urban mobility: challenges and opportunities

Dr Andrew McHugh
Senior Data Science Manager, Urban Big Data Centre, University of Glasgow

Prof. Vonu Thakuriah
Founding Director, Urban Big Data Centre, University of Glasgow

Abstract. *The Urban Big Data Centre (UBDC), funded primarily by Research Councils of UK's Economic and Social Research Council, aims to promote innovative research methods and the use of big data to improve social, economic, and environmental well-being in cities. UBDC operates a research-led data service. Its diverse portfolio of research across no less than ten academic disciplines informs and in turn benefits from a developing data collection and community of users. This presentation will focus on UBDC's urban mobility research in the context of ICT and data trends including smart and connected transport and buildings, integrated smart systems and personal and wearable technology. It will cover sources of urban big data, and showcase methodological and infrastructural approaches to facilitate data management and the production of information insights. It will touch too on organisational aspects around big data including team composition and skills, quality assurance and information governance. The presentation will align UBDC's approaches with a wider smart statistics agenda, and outline through a series of examples the ways in which data capture, processing and analysis can be systematically embedded to enable the production of comparable, reproducible, scalable and robust data for policy.*

1. The Urban Big Data Centre

The Urban Big Data Centre (UBDC) based at the University of Glasgow and funded by the Economic and Social Research Council exists to promote innovative research methods and the use of big data to improve social, economic, and environmental well-being in cities[1]. In support of this, UBDC operates a research-led data service, accessible to academic, policy and business user communities. The Centre provides essential data infrastructure (including data collections for research use and derived data products) and knowledge derived from these unique data sources, in support of several priority areas, including transport, education, automation of work, smart statistics, the UK Industrial Strategy, and smart cities. UBDC is unique in combining

[1] The Urban Big Data Centre at the University of Glasgow is supported by the Economic and Social Research Council (grant number ES/L011921/1) – http://www.ubdc.ac.uk.

social sciences research with data analytics and computing science. The Centre's programme of research and knowledge exchange across multiple disciplines informs and benefits from the data service. The Centre's aim is to deliver innovative urban analyses using novel data that have substantial and demonstrable impacts on public policy, to evaluate critically the value and role of big data and urban analytics for understanding cities and influencing planning and policy-making and to enhance quality of urban big data and methods for undertaking urban analysis. Key research themes include transport and infrastructure; neighbourhood, housing and environment; educational attainment, skills and productivity; and big data and urban governance. This short paper introduces the Centre's research around urban mobility and the associated challenges and opportunities, and explores their applicability to trusted smart statistics agenda.

2. Urban data context for delivering smart statistics

2.1. The challenge for urban analytics

21st Century cities offer tremendous concentrations of innovation and economic growth, but within these settings there remain extreme human and social, environment and physical and economic challenges. Collectively these add significant complexities to urban governance. Within this context, the widespread and rapid emergence of ICT solutions and data sources such as personal and wearable tech, smart transport and buildings, sensor network and connected infrastructure poses as many questions as it provides answers. How to deliver urban management that is efficient and effective, that supports sustainable economic growth, delivers fairness to its citizens and ensures resilience to both man-made and natural threats? In seeking to explore such issues, UBDC is building approaches that have a great deal in common with the challenges and processes of delivering smart statistics. The centre's big data strategy builds upon foundations of infrastructure, analytics capacity and knowledge discovery, leading to implementation and ultimately impact. Our aspirations are towards smart systems that integrate data capture, processing and analysis transparently, where contextual risks are managed while delivering actionable insights.

2.2. Data sources

Within the urban environment, sources of data are many and varied [12]. Sensor devices produce vast quantities of data spanning a range of urban themes, including transportation, health, energy, water, waste, weather, buildings and environments. These offer the potential for information on a range of aspects of demand, usage and other behaviours. They present

challenges too – often they are operated or owned by private organisations or are bound by otherwise restrictive access conditions. “Social” or “human” sensors produce additional, user generated content (e.g., social media, GPS, web analytics), either directly echoing urban events, problems or disruptions, or surmised retrospectively through analysis of online data [13]. Sensor networks, and the Internet of Things [2] promise ever-increasing and ever-transforming opportunities for data accumulation and variety.

Transaction data provides further opportunities for big data collection. Governments collect data on citizens, their activities and behaviour, ranging from health and social care to taxes and revenues to property and vehicle sales and registration. Open datasets at non-disclosive levels of aggregation are frequently available. Processes for obtaining controlled access to even individual level data are increasingly mature. Private organisations administer swathes of data that document customer interactions and transactions that can be hugely variable in terms of their accessibility, reliability and coverage.

Complementing these sources are more traditional datasets, from arts and humanities text, image and sound recordings datasets to film and other time based media.

In combination, these sources can offer insights in excess of those achievable by studying each in isolation. UBDC’s *Integrated Multimedia City Data* (iMCD) project combines a wide-ranging social survey with GPS, personal sensor, earth observation and online social media data linked spatially and at the individual level. For urban mobility this enables exploration of wide ranging issues including evaluation of commuter behaviours and Internet usage [6, 7], indoor and outdoor walking and social isolation and worker wellbeing.

In all cases, data offer little evidential value in and of themselves. They demand broader disciplinary insights, appropriate legal, ethical and cultural governance and methodological interventions in order to distil information and actionable insights.

2.3. Defining the context

At least four aspects define the context that drives and surrounds our work, and within which a smart statistics agenda must function. The first is technological – the management of information from a myriad of potential sources. The generation and capture of data, its management and processing, archiving, curation and storage and its ultimate dissemination.

Second is methodological – the preparation of data and distillation of information that can be deployed as evidence or as actionable insights. Primarily these methods will be focused on preparing data (information retrieval or extraction, on linkage and integration or in cleaning, anonymization and quality assessment) and its analysis (often domain specific, and designed to identify uncertainties, biases and error).

Third is theoretical and epistemological, demanding domain specific understanding of metrics, definitions and ideologies, broader knowledge of limitations of a data-driven approach, and information paradoxes (Jevons paradox), of user equilibrium versus system equilibrium.

Finally, the fourth is associated with the political economy – issues around data entrepreneurship, innovation networks and power structures, value propositions and economic implications, barriers to data availability and sharing, privacy, security and trust management and responsible innovation and emergent ethics.

2.4. Urban Mobility – Developing a big-data driven approach

UBDC's urban mobility agenda is wide ranging and reflects the changing nature of transport. Emergence of autonomous vehicles, connectedness, sharing, integration with other services and questions relating to infrastructure investment characterise our approach.

The work has explored areas such as urban metabolism – using a real time analytics approach to support city management, deriving from social media (Twitter, FourSquare) and GPS data spatio-temporal activity clusters that in turn offer insights into functional usage of place and stay duration within the city [11]. Where would new infrastructure or transporation service investment deliver greatest benefits? Where is there evidence of dissatisfaction with existing services and resources?

Complementing this work have been efforts to improve geolocalisation of tweets, using content and metadata cues, prompted by the fact that only 1-2% of Twitter content is explicitly geotagged [5]. This has formed the basis of work to identify road traffic incidents, a response to a perception that in certain urban neighbourhoods incidences of crashes are significantly under-reported [4].

In complementary work, image and wearable multi-sensor data has been combined to demonstrate other patterns and behaviours, such as indoor walking, travel modality and social isolation [8]. Elsewhere, GPS movement data has been cleaned and semantically annotated to

connect land use, points of interest and transport networks, facilitating activities such as urban planning, people trajectory, traffic analysis and fleet tracking [15].

Reflecting an increasing decentralisation of jobs, the modern 24 hour economy and high cost of private and public transport research has also explored links between transport and labour markets. Where are the most "transport poor" areas? How do these vary geographically and what are the implications for new policies aimed at local growth? Transit feed data provides the basis for small area transit availability indices, juxtaposed with traditional census, labour market and job listings data to reveal areas of high transport poverty, and low access to work. UBDC's Spatial Urban Data System collates these and other data sources within a multifaceted system to supports research and analysis of urban areas, policy and business decision-making, private sector innovation and public engagement [1].

A further strand of urban mobility research is associated with active travel. Researchers have used GPS-derived data from activity tracking app Strava (origin destination flows; link and junction level counts and other information) to validate and inform infrastructural investment and to explore relationships with air pollution [9].

3. Key challenges and opportunities

3.1. Skills, disciplinary knowledge and team composition

Urban transport analytics is not a discipline in its own right, instead pulling from a variety of traditional academic fields, urban management specialisms and scientific approaches. To be successful demands substantive knowledge of core fields, most notably urban studies, transport planning and engineering. To these one must add technical and methodological expertise, from spatial and statistical analysis to traditional computing science disciplines such as information and data management, information retrieval and human computer interaction. Given the often uncertain legal, governance and political contexts, these must be complemented by understanding of economics, law and information science.

Skills requirements for this work are broad and team structures must reflect this. Needs relate partially to workflows for data acquisition, management and analysis. Gathering data demands knowledge of the new data sources. The science of sensors – cooperative or connected vehicle systems, vehicle-to-grid systems, smart grid systems or assistive technologies for those with mobility support needs – and remote sensing technologies captured from satellites, aircraft or drones are vital, but also more traditional methods of survey design. The management and

curation of data requires contributions from systems, database, programming and information science specialists. Consistent with data science disciplines more generally, relevant analytics skillsets include machine learning, advanced statistical analysis, urban and transport modelling and simulations, GIS, spatial analysis and visualisation. Data driven work also imposes significant non-technical demands. Information governance, characterised by legal and economic aspects of data management, privacy and security; and business management, including project management, business case development, and contractual management are as vital as any technical or disciplinary contribution.

New academic programmes promise the emergence of a new breed of urban analytics professionals, but for now at least, few if any individuals will offer a full complement of skills. In most cases, several individuals and roles will collectively meet these skill demands. Successful teams for delivering urban insights or producing smart statistics will be heterogeneous in nature. Domain actors, information scientists, statisticians and analysts, legal and ethical experts, consumer needs specialists, communications and outreach specialists and business modellers will each perform critical roles.

3.2. Technology and structure

Among the foremost technical barriers are capacity constraints and concerns regarding the appropriateness of technological choices. Programmes of training and capacity building, partnerships with academics, industry and local governments and the adoption of standardised approaches to data management, technology (including algorithms) and methodologies can mitigate these challenges. There is scope and value in the establishment of local and national champions to highlight the value of smart statistics and data-derived insights, and to drive a policy and wider public benefit impact agenda. Community-based models of endorsement and validation are also effective, demonstrably so within the academic community. Peer-to-peer networks are effective means of establishing a basis for collaboration and community based learning. Systems and infrastructures for querying and mining data on an exploratory basis can reveal initial insights and offer reassurances of the legitimacy of particular approaches. Demonstrating public good, and engaging directly with communities further emphasises benefits and offers a means to mitigate and address perceived risks, such as those associated with personal freedoms and privacy.

3.3. Data sharing

As a national data service, UBDC has established considerable experience and expertise in the negotiation of access to data, not only in support of the Centre's own research activities, but also to facilitate efforts elsewhere. There are many associated challenges, which may be addressed using top-down or bottom up approaches. For smart statistics, there is scope to achieve the former through regulation, subject to legislative context. As an academic research centre, UBDC has primarily relied on the latter. Demonstrating value and engaging closely with data owners is a means of establishing shared ownership in the outcomes of any data-driven work. Offering appropriate reassurances regarding data governance is vital, irrespective of the nature of implicit sensitivities, be they related to concerns around personal privacy, commercial disadvantage or risks to reputation. Data sharing agreements and licensing are critical foundations and must reflect proposed and anticipated future usage. Data quality assessments should not focus solely on information content, formats and methodological basis of a given dataset. They must reflect too on the available licensing and data sharing conditions. How robust is a given agreement? What opportunities or reassurances are available to ensure continued supply over time? Will the data continue to conform to an initially agreed specification? Does the data owner exhibit signs of organisational instability? Changes within senior management, new business models or corporate mergers can significantly impact the continuing viability of an agreement to share data.

One response to these challenges would be to rely only upon *open* data sources. In comparison with open data, shared data demands additional negotiation, controls and governance mechanisms but offers opportunities to achieve broader, higher resolution insights. The value of open data, and its desirability as a public good are beyond question, but factors, often related to privacy or commercial sensitivities, limit its viability in many cases. The benefits of data sharing for doing research work have been widely discussed [3], albeit offset with concerns – that wealthier stakeholders are best positioned to benefit, at the cost of poorer communities, or that data subjects' privacy may be at risk because of the practice [14]. UBDC aims to minimise barriers to the use of data in the resolution of urban challenges. Broadening access means providing a service that is free at the point of use, and negotiating with data owners to agree terms for data sharing that are as unrestrictive as possible, while protecting the interests of individuals and organisations affected. UBDC partly achieves this by offering data owners reassurances through its policies for managing data access.

UBDC approaches the accessioning of a given dataset with its safe accessibility of foremost importance, reflected by infrastructure and governance controls in place. Data accessioning comprises of seven stages. These are *negotiation of dataset licensing*, where data sharing agreements and end user licensing arrangements are agreed and formalised; *physical acquisition of data*, where data and associated metadata are physically transferred and received; *dataset assessment*, where datasets are evaluated and additional processing requirements identified; *dataset processing*, where applicable processing is undertaken; *data documentation*, where accompanying documentation is created, validated and standardised; *dataset definition*, where one or more agreed data packages are defined and their manifests recorded; and *dataset publication*, where data is published to one or more delivery platforms. The counterpoint, UBDC's process for end user access also defines several stages. End users' purposes are defined and compared with relevant data sharing policy(ies); sub-licensing documentation is exchanged, completed and stored; and data is securely transferred or made accessible to authorised, authenticated users through a secure platform. For the most sensitive controlled data (e.g. individual-level health data) additional governance processes require prospective users to satisfy an independent committee of the scientific and public benefit impacts of their proposed work, and of the appropriate mitigation of associated risks. Predictably, negotiating these policies and processes is much simpler for acquiring and sharing open data, than, for example, commercially sensitive business data.

3.4. Trust and validation

A further consideration relates to the difference between a dataset viewed in isolation, and its deployment as the basis of trusted smart statistics. The establishment of recognised, accredited and independent authorities to review and provide oversight functions is worth pursuing. To deliver on the promise of trusted smart statistics we must aspire to produce agreed ground truth data. We must have means for weighting results that come from different methods, datasets or analysts. Finally, we require novel methods to assess and capture uncertainty and to understand the behaviour of statistics and indicators from 'black box' algorithms at *every stage* of a data-to-output lifecycle.

4. Summary

This short paper presents some of the Urban Big Data Centre's experiences in undertaking urban mobility analytics work, outlining not only the methodological challenges and opportunities, but also the contextual factors likely to be influential to the success of these

efforts. There are several facets of this work. Many are not particularly domain-specific and appear relevant for efforts to produce trusted smart statistics. Methodological, data and skills gaps continue to be evident, presenting notable challenges, at least commensurate with the value that may be achieved.

5. References

[1] Anejionu, O., Thakuriah, P., McHugh, A., Sun, Y., McArthur, D., Mason, P., Walpole, R. (2018). Spatial Urban Data System: A Cloud-enabled Big Data Infrastructure for Social and Economic Urban Analytics. Forthcoming in Future Generation Computer Systems. (under review)

[2] Ashton, K. (2009). That "Internet of Things" Thing: In the Real World Things Matter More than Ideas. RFID Journal. http://www.rfidjournal.com/articles/view?4986

[3] Chatham House Data Sharing Advisory Group. Public health surveillance: a call to share data. International Association of National Public Health Institutes, 2016

[4] Gonzalez Paule, J. D., Sun, Y. and Moshfeghi, Y. (2018) On fine-grained geolocalisation of tweets and real-time traffic incident detection. Information Processing and Management, (doi:10.1016/j.ipm.2018.03.011)(Early Online Publication)

[5] Graham, M., Hale, S. A., & Gaffney, D. (2014). Where in the world are you? geolocation and language identification in twitter. The Professional Geographer, 66(4), 568–578.

[6] Hong, J. and Mcarthur, D. P. (2017) How does Internet usage influence young travellers' choices? Journal of Planning Education and Research, (doi:10.1177/0739456X17736811) (Early Online Publication)

[7] Hong, J. and Thakuriah, P. (2018) Examining the relationship between different urbanization settings, smartphone use to access the Internet and trip frequencies. Journal of Transport Geography, 69, pp. 11-18.(doi:10.1016/j.jtrangeo.2018.04.006)

[8] Nowicka, K-S, P. Thakuriah (2016). The trade-off between privacy and geographic data resolution. a case of GPS trajectories combined with the social survey results. Proc. ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences

[9] Sun, Y. , Moshfeghi, Y. and Liu, Z. (2017) Exploiting crowdsourced geographic information and GIS for assessment of air pollution exposure during active travel. Journal of Transport and Health, 6, pp. 93-104.(doi:10.1016/j.jth.2017.06.004)

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

[11] Thakuriah, P., K. Sila-Nowicka and J. Gonzalez-Paule (2016). Sensing Spatiotemporal Patterns in Urban Areas with a Multi-Modal Urban Data. Forthcoming in Big Data and the City, a special issue of Built Environment

[12] Thakuriah, P., N. Tilahun and M. Zellner (2016). Big Data and Urban Informatics: Innovations and Challenges to Urban Planning and Knowledge Discovery. In Seeing Cities through Big Data: Research, Methods and Applications in Urban Informatics, Springer, NY, pp. 11- 48.

[13] Thakuriah P, Geers G (2013) Transportation and information: trends in technology and policy. Springer, New York

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

[15] Wang, Y. and McArthur, D. (2018) Enhancing data privacy with semantic trajectories: a raster-based framework for GPS stop/move management. Transactions in GIS, (doi:10.1111/tgis.12334) (Early Online Publication)