International Journal of Advances in Computer Science and Technology

A Review On Rise Of Big Data On Urban Studies

T. Kiranmayee [1], A. Harshitha[2], M. Neelima [3]

3rd Year B.Tech Computer Science & Engineering, Andhra Loyola Institute of Engineering and Technology, Vijayawada, Andhra Pradesh.

tkiranmayee1997@gmail.com, harshithaakula9@gmail.com, Neelima1525@gmail.com

Abstract

Big Data is identically modernistic and sizzling topic in today's scenario. Cities are growing very rapidly worldwide. This growth entails many challenges which cut across different city layers. In terms of demography, we are facing many issues to do with migration and aging of the population. In terms of land use, a big challenge involves how we deal with congestion in terms of high densities and sprawl in cities and also how we can tackle segregation so that we might decrease inequality and deprivation. The availability of resources is of concern in terms of how efficiently and sustainably we use energy. The transport sector faces big tests with respect to congestion in infrastructure across all travel modes, growing levels of pollution and noise, and accidents. To add to the complexity of the challenges just mentioned, they span different spatial and temporal scales as well. In this paper, we describe a number of new urban data sources and their problems and also illustrate how they can be used to improve the study and function of cities in coming days.

Keywords: demography, spatial and urban data.

1. INTRODUCTION

The big data revolution is going to change urban science and turns a cross section of space into living data of urban life ever

been available before. There are some issues that are to be addressed in urban areas. The urban planners need to somehow juggle these issues through the use of a variety of tools. One of these tools is the so called Land Use Transport interaction (LUTi) model. This is really a family of models that aim to estimate how cities will develop on a long term basis (typically over a period of 30 to 50 years) through the interaction of three main factors: population, land use and transport services. Among the many processes addressed by LUTi models, the main one is perhaps the interplay and feedback of information from the land use system to the transport system and vice versa[1]. This reflects the influence of land use patterns on mobility patterns and the evolution of transport infrastructure in one direction, and in the other direction, how transport systems have an impact on how urban form evolves and how people engage in various land use activities. Typical plans evaluated using this family of models include the estimation of the impacts around a change in transport infrastructure, e.g. a new railway line between two areas, or the building of a new development in the region, e.g. a new industrial estate. This would include economic impacts (regional and/or national), often disaggregated by industrial sectors; and the prediction of diverse data on households, population (by type) and the number of additional jobs for each of the modelled areas[8].

2. HISTORIAL BACKGROUND

There was a first surge of LUTi models during the 1960s in the US which lasted until the mid 1970s. The main difficulties around this first family of urban models were synthesised by Lee in his *Requiem for large-scale models* article in 1973 where he listed 'seven sins' of the models from this period: hypercomprehensiveness,

grossness, hungriness, wrongheadedness, complicatedness, mechanicalness and expensiveness. This was the rhetoric of those times and in essence, the models did not match up to the intellectual and policy needs for effective forecasting and they were difficult to implement due to limits on data and computer resources. It is therefore not surprising that these kinds of urban model were then more or less abandoned during the late 1970s and 1980s[2].

However, advances in computing, the birth of Geographic Information Systems (GIS) and better data led to a second surge of urban models during the 1990s. New models advanced from previous implementations by further disaggregating zonal components and also population and employment groups using socio-economic attributes. New modelling techniques were explored, such as discrete choice theory and agent-based modelling. Another important factor was visualisation, which made results more interpretable and thus clearer to wider audiences[3]. In particular, this includes better usability for planning practitioners through what have come to be called planning support systems (PSS). These improvements were further supported in the US and the EU by specific initiatives, which meant that many of these models ended up being applied in an increasing number of cities.

Perhaps the most important advance in recent years has been the adoption of new modelling approaches - amongst which activity-based modelling and microsimulation transport modelling are particularly exciting. At the same time, our evolving knowledge around model development continues to challenge the limits of model computation, which means we are also observing an increasing trend on the parallelisation of processes to run models. Simply put, the sheer amount of calculations within such models, with constant feedback loops from sub-models to other sub-models, means that multiple models are being integrated together in new ways[4].

3. BIG DATA AND URBAN MODELS

Understandably, the Big Data era, where information relevant to urban planning is starting to be available from

unconventional sources, has brought integrated urban models back into the spotlight. Data based on crowd-sourcing, remote sensing, online social networking, smart transit ticketing, mobile phone usage and credit card transactions have a common denominator: they all contain geo-located information. As a result, we are moving from structured, static, demographic and economic activity data (e.g. census data) to unstructured, dynamic data able to provide new insights about urban dynamics[5].
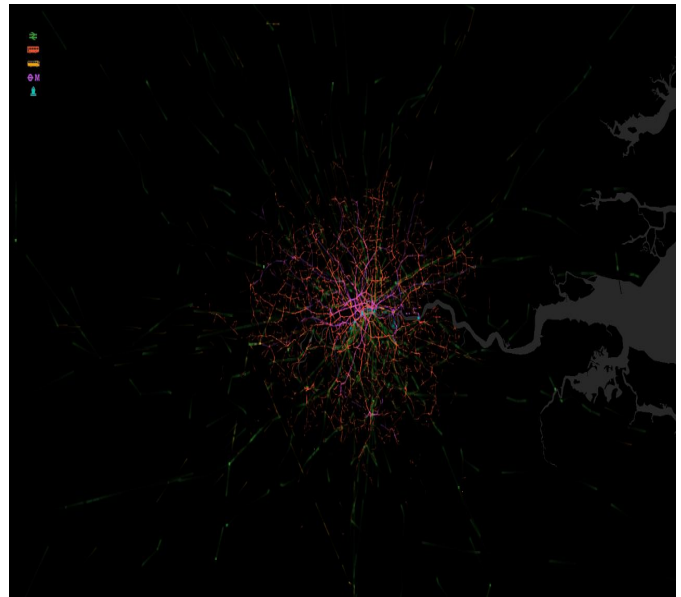

Figure 1: Public transport services operating in London

While the potential of the data is huge, it also comes with many hurdles. We have more data, but often with lower explanatory power about the underlying decisions and behaviours of city users. Another important issue around some of these data types which is very relevant to us as urban modellers is its 'representatively' in our case study areas. In this sense, we are just starting to learn how to cope with this huge paradigm shift. In the past, behavioural patterns in the population data used in urban models were deduced using 1% population sample surveys (or similar values). From many of these new datasets we have now a much higher coverage, which can reach 40% or 50% of the population - but this sample size often comes at the expense of low quality, noisy or biased data. Data mining and the ability to blend data

from multiple sources are becoming increasingly important for the identification of biases and inconsistencies in what can be quite gigantic datasets where billions of records are commonplace. Computational science enables us to process such data faster and new kinds of statistics based on data mining are essential for its analysis [6].

To make matters more complicated, the very same technologies which allow us to collect and use all this data have a much more important impact: research suggests that ICT is changing the way we live. Or, in other words, our daily activities in their various purposes (work, shopping, leisure or education to name a few) are being affected by the way we interact with our surroundings through ICT. Not just in the way we plan them but also in the way we get to interact with them and experience them. Urban models face a big challenge in capturing these behavioural changes.

EUNOIA is a European research project that investigates how data from multiple sources available in the context of the Big Data and the smart city movement (including data from smart cards, mobile phone traces, online social networks or credit cards, among others) can be integrated, analysed and visualised to understand mobility and location patterns in cities. New data sources can be used to replace or augment traditional data collection methods, but also to inform and enable the development of new modelling approaches. These in turn support researchers and practitioners with new insights about how city users live and move about in cities. Mobile phone data can be used to obtain origin-destination matrices at a much lower cost than from traditional household travel surveys, or can be combined with surveys to provide a richest pool of data. Data on credit card usage provides very rich information on expenditure flows across the city which can be used to formulate, calibrate and validate retail location models. Online social networks can be used to investigate the role of social interaction on mobility. The list is endless, and its exploration is still a largely unknown domain.
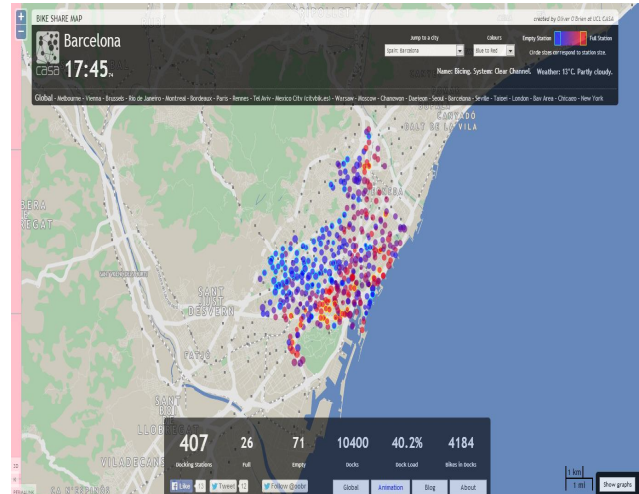

Figure 2: Occupancy in bike share stations in Barcelona

EUNOIA is looking at these and other questions with the aim of developing improved models and integrating them into large-scale, state-of-the art urban simulation tools, such as the agent-based transport simulation framework MATSim or the more aggregated LUTi framework SIMULACRA. The project also aims at developing user-friendly visual interfaces and data representations enabling analytical reasoning and interpretation of the simulation results. A number of case studies, defined in collaboration with planning authorities and mobility stakeholders from the three cities participating in the project (Barcelona, London, and Zurich) are aimed at evaluating the potential of the newly developed tools to address relevant policy questions, such as the planning and operation of the bike sharing systems in London or Barcelona.

4. CLOSING STATEMENTS

Urban models have become a useful tool for planners to tackle many of the problems around the growth of cities. These models are now over 40 years old which means they have gone through many re-evaluations to improve their accuracy. Having said that, urban models still face many challenges ahead. We will emphasise four of them. First, they require a lot of computer processing power, especially regarding transport modelling. Fast and yet realistic implementations need to be

sought to enable various model runs using shorter times. Second, the visual interface showing the results of LUTi models still has much room for improvement. More interactive and comprehensive tools to understand the results need to be implemented to help practitioners and other stakeholders. Third, there has been some discussion in the modelling community around the concept of dynamic versus static model implementations. As discussed by Ying and Wegener, this is a very challenging topic as it points to the core of the model design in order to better capture a world whose equilibrium is most likely dynamic. Finally, and perhaps most relevant in the context of our current research in EUNOIA, we need to find out whether more representative samples such as the ones from big data lead to potential advances in urban modelling. We believe breakthroughs in any of these areas will allow urban planners to be in a better position to tackle many of the challenges that cities are currently facing[7].
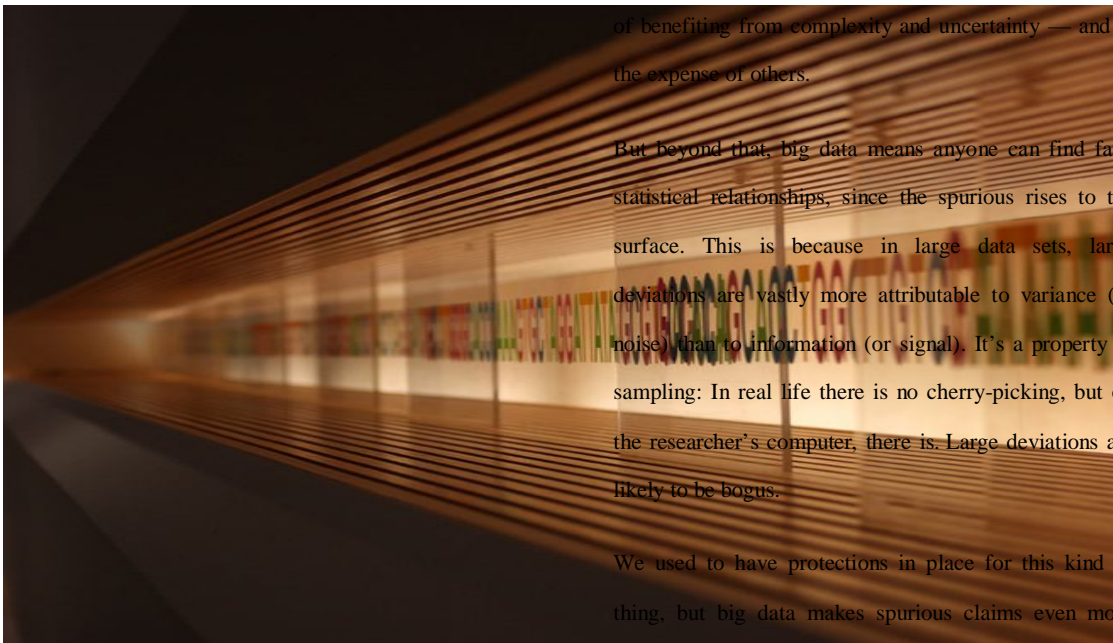
5. PROBLEMS WITH BIG DATA:

Figure 3: problems with big data

We're more fooled by noise than ever before, and it's because of a nasty phenomenon called "big data." With big data, researchers have brought cherry-picking to an industrial level.

Modernity provides too many variables, but too little data per variable. So the spurious relationships grow much, much faster than real information.

In other words: Big data may mean more information, but it also means more falseinformation.

Just like bankers who own a free option — where they make the profits and transfer losses to others – researchers have the ability to pick whatever statistics confirm their beliefs (or show good results) … and then ditch the rest.

Big-data researchers have the option to stop doing their research once they have the right result. In options language: The researcher gets the "upside" and truth gets the "downside." It makes him antifragile, that is, capable of benefiting from complexity and uncertainty — and at the expense of others.

But beyond that, big data means anyone can find fake statistical relationships, since the spurious rises to the surface. This is because in large data sets, large deviations are vastly more attributable to variance (or noise) than to information (or signal). It's a property of sampling: In real life there is no cherry-picking, but on the researcher's computer, there is. Large deviations are likely to be bogus.

We used to have protections in place for this kind of thing, but big data makes spurious claims even more tempting. And fewer and fewer papers today have results that replicate: Not only is it hard to get funding for repeat studies, but this kind of research doesn't make anyone a hero. Despite claims to advance knowledge, you can

hardly trust statistically oriented sciences or empirical studies these days.

This is not all bad news though: If such studies cannot be used to confirm, they can be effectively used to debunk — to tell us what's wrong with a theory, not whether a theory is right.

Another issue with big data is the distinction between real life and libraries. Because of excess data as compared to real signals, someone looking at history from the vantage point of a library will necessarily find many more spurious relationships than one who sees matters in the making; he will be duped by more epiphenomena. Even experiments can be marred with bias, especially when researchers hide failed attempts or formulate a hypothesis after the results — thus fitting the hypothesis to the experiment (though the bias is smaller there).
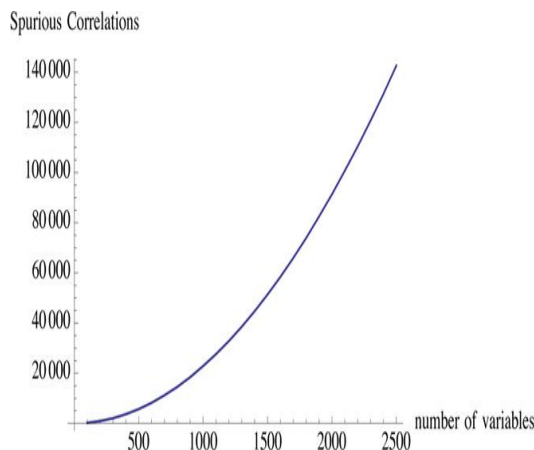

Figure 4: Tragedy of big data: The more variables, the more correlations that can show significance. Falsity also grows faster than information; it is nonlinear (convex) with respect to data (this convexity in fact resembles that of a financial option payoff). Noise is antifragile. *Source: N.N. Taleb*

The problem with big data, in fact, is not unlike the problem with observational studies in medical research.

In observational studies, statistical relationships are examined on the researcher's computer. In double-blind cohort experiments, however, information is extracted in a way that mimics real life. The former produces all manner of results that tend to be spurious (as last computed by John Ioannidis) *more than eight times out of 10.*

Yet these observational studies get reported in the media and in some scientific journals. (Thankfully, they're not accepted by the Food and Drug Administration). Stan Young, an activist against spurious statistics, and I found a genetics-basedstudy claiming significance from statistical data even in the reputable *New England Journal of Medicine* — where the results, according to us, were no better than random.

6. BIG DATA CAN TELL US WHAT'S WRONG, NOT WHAT'S RIGHT.

And speaking of genetics, why haven't we found much of significance in the dozen or so years since we've decoded the human genome?

Well, if I generate (by simulation) a set of 200 variables — completely random and totally unrelated to each other — with about 1,000 data points for each, then it would be near impossible *not* to find in it a certain number of "significant" correlations of sorts. But these correlations would be entirely spurious. And while there are techniques to control the cherry-picking (such as the Bonferroniadjustment), they don't catch the culprits — much as regulation didn't stop insiders from gaming the system. You can't really police researchers, particularly when they are free agents toying with the large data available on the web.

I am not saying here that there is no information in big data. There is plenty of information. The problem — the central issue — is that the needle comes in an increasingly larger haystack.
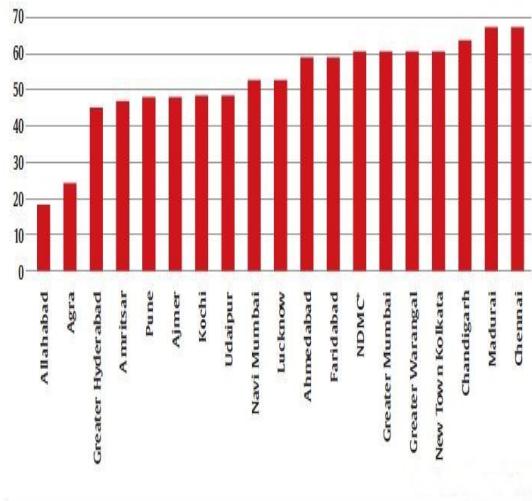

Figure 5: Evaluation of the preparedness of smart city contenders

7. **CONCLUSION**

In this paper we have broadly classified how urban planners have used different tools to address the problems on growth of cities. We have studied some of the problems around growth of the cities and showing efficient solutions to toggle the issues by using new tools and techniques. Due to increase in population and expansion of cities still some of the issues are needed to be addressed to store data in better way.

REFERENCES

[1] Al-Hader, M., Rodzi, A., Sharif, A.R., & Ahmad, N. (2009). Smart city components architecture. In Proceedings of the International Conference on Computational Intelligence, Modelling and Simulation, Brno, Czech Republic, September 7-9.

[2] Al-Hader, M., Rodzi, A., Sharif, A.R., & Ahmad, N. (2009). SOA of smart city geospatial management. In Proceedings of the 3rd UKSim European Symposium on Computer Modeling and Simulation, Athens, Greece, November 25-27.

[3] Anthopoulos, L., &Fitsilis, P. (2010). From digital to ubiquitous cities: Defining a common architecture for urban development. In Proceedings of the 6th International Conference on Intelligent Environments, Kuala Lumpur, Malaysia, July 19-21.

[4] Anthopoulos, L., &Fitsilis, P. (2010). From online to ubiquitous cities: The technical transformation of virtual communities. In Sideridis, A. B., &Patrikakis, C. Z. (Eds.), Next Generation Society: Technological and Legal Issues (Vol. 26, pp. 360-372). Proceedings of the Third International Conference, e-Democracy 2009 (Athens, Greece, September 23-25), Berlin, Germany: Springer. Available from http://www.springerlink.com/content/g644776482968k 36/fulltext.pdf.

[5] Barzilai-Nahon, K. (2006). Gaps and bits: Conceptualizing measurements for digital divide/s. The Information Society, 22(5), 269-278.

[6] Barzilai-Nahon, K. (2009). Gatekeeping: A critical review. Annual Review of Information Science and Technology, 43(1), 1-79.

[7] Belissent, J. (2011). The Core of a Smart City Must Be Smart Governance. Cambridge, MA: Forrester Research, Inc.

[8] Borja, J. (2007). Counterpoint: Intelligent cities and innovative cities. UniversitatOberta de Catalunya (UOC) Papers: E-Journal on the Knowledge Society, 5. Available from http://www.uoc.edu/uocpapers/5/dt/eng/mitchell.pdf.