Statistics and Probability Letters xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

Quality of life, big data and the power of statistics

Shivam Gupta [a,*], Jorge Mateu [b], Auriol Degbelo [a], Edzer Pebesma [a]

[a] Westfälische Wilhelms-Universität, Münster, Germany
[b] Universitat Jaume I, Castellon, Spain

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Air quality
Big data
Land use regression
Optimal location
Smart city

ABSTRACT

The digital era has opened up new possibilities for data-driven research. This paper discusses big data challenges in environmental monitoring and reflects on the use of statistical methods in tackling these challenges for improving the quality of life in cities.

1. Introduction

Quality of life (QoL) is tied to the perception of 'meaning'. The quest for meaning is central to the human condition, and we are brought in touch with a sense of meaning when we reflect on what we have created, loved, believed in or left as a legacy (Barcaccia, 2013). QoL is associated with multi-dimensional issues and features such as environmental pressure, total water management, total waste management, noise and level of air pollution (Eusuf et al., 2014). A significant amount of data is needed to understand all these dimensions. Such knowledge is necessary to realize the vision of a smart city, which involves the use of data-driven approaches to improve the quality of life of the inhabitants and city infrastructures (Degbelo et al., 2016).

Technologies such as Radio-Frequency Identification (RFID) or the Internet of Things (IoT) are producing a large volume of data. Koh et al. (2015) pointed out that approximately 2.5 quintillion bytes of data are generated every day, and 90 percent of the data in the world has been created in the past two years alone. Managing this large amount of data, and analyzing it efficiently can help making more informed decisions while solving many of the societal challenges (e.g., exposure analysis, disaster preparedness, climate change). As discussed in Goodchild (2016), the attractiveness of big data can be summarized in one word, namely *spatial prediction* - the prediction of both the *where and when*.

This article focuses on the 5Vs of big data (volume, velocity, variety, value, veracity). The challenges associated with big data in the context of environmental monitoring at a city level are briefly presented in Section 2. Section 3 discusses the use of statistical methods like Land Use Regression (LUR) and Spatial Simulated Annealing (SSA) as two promising ways of addressing the challenges of big data.

2. Environmental monitoring and big data challenges

With an increasing number of people moving in (and to) urban areas, there is an urgent need of examining what this rising number means for the environment and QoL in cities. Air quality has an effect on the population's QoL (Darçın, 2014), which is also the major environmental risk factor for health. In 2012, one in eight deaths could be attributed to exposure to air pollution according to the World Health Organization.[1] Air quality has high fluctuation at a fine scale due to its very complex

[*] Corresponding author.
E-mail address: shivam.gupta@uni-muenster.de (S. Gupta).

[1] See http://www.who.int/phe/health_topics/outdoorair/global_platform/en/ (last accessed: December 4, 2017).

https://doi.org/10.1016/j.spl.2018.02.030
0167-7152/© 2018 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

distribution, the structure of the city, and dispersion processes. Institutions such as the European Environmental Agency have produced maps of air quality across Europe. Nonetheless, these maps have two drawbacks: first, their spatial resolution is coarse (i.e., they are usually available for the member state level), and second, they do not give a real-time account of the situation. Projects such as the World Air Quality Index provide real-time air quality maps (see http://aqicn.org/), but again they have a relatively coarse spatial resolution.

Data for environmental and meteorological analysis are not only of a significant volume but are also complex in space and time. Formats and types of data are also very diverse (e.g., netCDF, GDB, CSV, GeoTIFF, shapefile, JSON, etc.), and many interconnections prevail within data, which make it complicated for traditional data analysis procedures. Fusing official monitoring stations data with methods like IoT based crowd-sourced data sources can increase redundancy and make data management a serious challenge. Using this example, challenges associated with big data can be illustrated as:

Volume: The large data volume is induced by fusing data from monitoring stations, with crowd-sourcing sensors which can further be integrated with significant environmental data, city dynamics data and other parameters like city land use information. The data size for some variables varies from MBs to TBs (e.g., a single data file for atmospheric data is around 2GB's for a single point of interest). Handling this amount of data needs proper planning; otherwise, the analysis may take longer time because of the mixture of redundant or less relevant data.

Velocity: The speed at which the data from monitoring stations, added sensors and other data sources are created, captured, extracted, processed and stored also needs to be dealt with appropriately. Statistical issues arise from fusing together different data source streams at different spatiotemporal scales. Delay in data fetching from remote storage devices or geographical constraints may also impact the process. Velocity is one crucial characteristic that defines the kind of outcomes we can develop from the data sources.

Variety: Environmental data are in various formats (e.g., NetCDF files for environmental variables, GeoTIFF files for land use, shapefiles of the city for road networks and traffic congestion), which represents heterogeneity challenges, entity resolution issues arising by merging data from different data sources and interaction challenges between big data and data applications.

Veracity: With the variety of data pouring in the analysis, the level of uncertainty also increases. Outcomes expected from the analysis may be affected by some offsets and origin errors of data sources. To maintain data veracity, it is sometimes advised to discard noisy sources and include only reliable sources. However, ignoring some data points may lead to missing some air quality pattern in the city.

Value: A large amount of data is of no use until it is converted into value. For air quality, the value can be considered as the extraction of intelligence to improve QoL in the city through the development of applications which help city dwellers become aware of their air quality exposure. However, issues such as inefficient handling of large amounts of data, inability to provide quality results on a timely basis, the bottleneck in sharing processed data, high computational cost of big data processing hinder the provision of efficient, easy outcomes for public use.

3. Statistics and environmental monitoring

As Scott (2017) said, statistics remains highly relevant irrespective of 'bigness' of data. It provides the basis to make data speak while taking into account the inherent uncertainties. Statistical analysis involves developing data collection procedures to further handle different data sources and to propose formal models for analysis and predictions. There are a number of statistical methods varying from sophisticated data requirement (e.g., dispersion models) to simple inference models (e.g., proximity-based models) for air quality prediction. Each of the methods has their specific data and computational requirements. Some methods cannot always be implemented due to the cost, time and resources involved. Notable air quality modeling methods, such as dispersion models, are very sophisticated and require deep insight into the chemical and physical assumptions of the pollutant along with pollutant monitoring sites in the city at a very fine spatiotemporal resolution. The downfall of these methods also includes the cost of the data needed for the study with disputable assumptions about the dispersion pattern (i.e., Gaussian dispersion) and extensive cross-validation with monitoring station data (Jerrett et al., 2005). The next subsections highlight the potential of land use regression and spatial simulating annealing in addressing both big data challenges, and shortcomings of previous work.

3.1. Land use regression (LUR)

Land use regression requires simple geographical variables for predicting environmental factors such as air pollution or sound pollution in the city. It is one of the standard methods used by epidemiologists and health care researchers for exposure analysis. LUR helps in breaking the limitations in developing the models while offering the flexibility to use already available data sources. Regarding performance, LUR-models have been outperforming geostatistical methods and may perform equally, or sometimes better, than dispersion models (Gulliver et al., 2011). With LUR, researchers can estimate individual exposures from statistical models that combine the predictive power of several surrogates based on their relationship with measured concentrations.

Advantages. The advantage of the LUR approach is the flexibility of incorporating more theoretical knowledge about the process governing the spatial and spatiotemporal variation. This way the challenges due to the addition of new data (e.g., IoT data) can be handled with the context-based variable selection. This restricts the amount of input in the analysis and hence