Statistics and Probability Letters 136 (2018) 101–104

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 27 February 2018

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

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 for 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

Table 1
Challenges of big data, and potential of the combined use of the proposed methods.

Dimension	Challenge	Solution
Volume	Data reduction techniques (Koh et al., 2015)	LUR selects variables and SSA selects the optimal locations, hence reducing data for analysis
Velocity	Quick and constant access of data (Namiot and Sneps-Sneppe, 2012)	By identifying optimal locations we can decrease the data to process and accelerate data access
Variety	Creating a knowledge base from different data formats (Koh et al., 2015)	LUR reduces the number of variables to process, possibly reducing the variety of data sets for analysis
Veracity	Avoid inessential data (Koh et al., 2015)	LUR variable selection and SSA optimal location can avoid the inessential data
Value	Complexity restricts timely processing (Villanueva et al., 2014)	LUR selected variables provide context and SSA puts cost function to achieve context-aware outcomes timely

can help in tackling the volume, variety, veracity and velocity data challenges. The perk of LUR also is its ability to run models within raster spatial environments, which allows rapid computation. Hence, it can help with challenges related to the value aspect of big data analysis. Another major advantage of the LUR-model over dispersion and interpolation models is to gain the spatial scale desired at the city level. LUR-models are better at describing hot-spots in cities, unlike aforementioned methods which provide smoother concentration maps (Marshall et al., 2008).

Drawbacks. Compared to dispersion models, the LUR method requires less detailed input data at the expense of the need to obtain monitoring data for a sufficiently large number of sites. Moreover, LUR-models have limited capacity to separate the impact of some pollutants because they are collinear to each other, which is the same case of other exposure study methods. LUR methods can benefit from a more systematic selection and description of space–time attributes of monitoring locations.

Building reliable models for big data requires strengthening the sampling process towards locations and time points which can improve predictability. The reliability of methods always depends on the quality of input data. The selection of monitoring sites to develop the air quality models has been identified as one of the factors affecting the quality of models' outcomes. We still lack rigorous methods to determine the number and distribution of monitoring sites (Hoek et al., 2008). Using a large number of monitoring sites to build a model, improves its ability to estimate the pollutants. However, improvement in models' predictive power can be achieved by a certain number and specific distribution of monitoring stations. Selecting optimal locations may assist in minimizing data redundancy and can enhance the computational time. Various statistical methods exist for optimizing the sampling process. Here, we discuss the method called "Spatial Simulated Annealing (SSA)" for optimization of an air quality monitoring network.

3.2. Spatial simulated annealing for optimizing monitoring network

Placing sensors at certain locations often should fulfill several purposes, and these can be achieved by combining the respective cost functions. By defining the cost function, we try moving each sensor in their neighboring cells/locations and find the best places where a cost function can be achieved so that the purpose of placing the sensor is worth. SSA takes into account the spatial neighborhood to optimize spatial sampling schemes based on a defined cost function. During the process, both the size of the movement of sensors around the specified area of interest, and the probability to agree to the worst results decrease with a decreasing annealing temperature. By using this approach, we can decrease the amount of data needed to perform the analysis with optimal results.

Recent works regarding spatial sample configuration optimization using SSA, emphasized on the following aims: (a) conditioned latin hyper cube sampling (Roudier et al., 2012); (b) variogram identification and estimation using constraints like pairs contributing to each lag-distance class (Truong et al., 2013); (c) spatial interpolation using constraints like minimization of the kriging variance in a space–time setting (Delmelle, 2014).

Advantages. SSA enables the specification of various types of optimization goals during the spatial analysis. Once the goal is decided, we can limit the area of interest along with the geostatistical criterion, i.e., we magnify our research goals at certain spatial vicinity for a proposed outcome. This method can take into account the weight of the area we are more keen on collecting knowledge about. By limiting the area and incorporating context-aware goals in location selection processes, we also curb the creation of a large amount of nonessential data, thus helps in overcoming aforementioned big data challenges.

Drawbacks. Based on convergence analysis, different forms of temperature updating functions are followed concerning different kinds of probability density functions employed. The convergence of the objective using SSA depends on the input of appropriate conditions for both the probability density function and the temperature updating function. Calculating these inputs for SSA can be a time-consuming process and needs practical experience. Depending on the objective and size of an area, the processing of the algorithm is intermittently time-consuming too. However, time-consuming processes will pay off at a later phase as they help in selecting the possible best locations for data collection and hence improve the overall flow of the analysis. Table 1 summarizes the key points of the combined use of both methods discussed earlier to tackle challenges of big data for environmental monitoring.

4. Conclusions

In this paper, we have focused on the role of statistics in handling the five Vs of big data, and the challenges posed. Big data analytics demands different methodologies from traditional statistical approaches which can enable efficient computer processing and timely outcomes for the efficient use of data. We propose to combine two well-established statistical methods to optimize the selection of variables and locations for spatial and temporal analysis of environmental data sources. The combined use of both methods will help in designing data acquisition processes so that the maximum information can be extracted given a specific number of possible measurement sites. Limiting the data sources can increase the speed of the analysis. The key highlight of integrating LUR and SSA is to make processes like air quality monitoring flexible because LUR can consider limited but accessible data sources. However, we need to consider some crucial aspects. First, variables should be selected carefully and used correctly in the models. Second, the design of SSA-based optimization relies on the quality of input from the LUR for putting weight on those areas for the cost function we want to achieve. It is also helpful in reflecting on the temporal dependency of air quality at a location and the spatial correlation among other locations. Third, SSA needs inputs about probability distributions and temperature change functions which is a critical aspect of optimal location selection. Hence, by using such statistical tools, big data analysis can be effective regardless of the "bigness". Statistics has been a major component of data analysis for centuries and will be crucial in the era of big data.

Acknowledgments

The authors gratefully acknowledge funding from the European Commission through the GEO-C project (H2020-MSCA-ITN-2014, Grant Agreement Number 642332, http://www.geo-c.eu/). Jorge Mateu has also been partially funded by the grant MTM2016-78917-R from the Spanish Government of Science and Competitiveness.

References

Barcaccia, B., 2013. Definitions and domains of health-related quality of life. In: Outcomes Assessment in End-Stage Kidney Disease-Measurements and Applications in Clinical Practice. pp. 12–24.
Darçın, M., 2014. Association between air quality and quality of life. Environ. Sci. Pollut. Res. 21 (3), 1954–1959.
Degbelo, A., Granell, C., Trilles, S., Bhattacharya, D., Casteleyn, S., Kray, C., 2016. Opening up smart cities: citizen-centric challenges and opportunities from GIScience. ISPRS Int. J. Geo-Inf. 5 (2), 16.
Delmelle, E.M., 2014. Spatial sampling. In: Handbook of Regional Science. Springer, pp. 1385–1399.
Eusuf, M.A., Mohit, M.A., Eusuf, M.S., Ibrahim, M., 2014. Impact of outdoor environment to the quality of life. Proc. Soc. Behav. Sci. 153, 639–654.
Goodchild, M.F., 2016. GIS in the era of big data. Cybergeo: Eur. J. Geogr.
Gulliver, J., de Hoogh, K., Fecht, D., Vienneau, D., Briggs, D., 2011. Comparative assessment of GIS-based methods and metrics for estimating long-term exposures to air pollution. Atmos. Environ. 45 (39), 7072–7080.
Hoek, G., Beelen, R., De Hoogh, K., Vienneau, D., Gulliver, J., Fischer, P., Briggs, D., 2008. A review of land-use regression models to assess spatial variation of outdoor air pollution. Atmos. Environ. 42 (33), 7561–7578.
Jerrett, M., Arain, A., Kanaroglou, P., Beckerman, B., Potoglou, D., Sahsuvaroglu, T., Morrison, J., Giovis, C., 2005. A review and evaluation of intraurban air pollution exposure models. J. Expo. Sci. Environ. Epidemiol. 15 (2), 185–204.
Koh, J.M., Sak, M., Tan, H.-X., Liang, H., Folianto, F., Quek, T., 2015. Efficient data retrieval for large-scale smart city applications through applied bayesian inference. In: Intelligent Sensors, Sensor Networks and Information Processing, ISSNIP, 2015 IEEE Tenth International Conference on. IEEE, pp. 1–6.
Marshall, J.D., Nethery, E., Brauer, M., 2008. Within-urban variability in ambient air pollution: comparison of estimation methods. Atmos. Environ. 42 (6), 1359–1369.
Namiot, D., Sneps-Sneppe, M., 2012. Context-aware data discovery. In: Intelligence in Next Generation Networks, ICIN, 2012 16th International Conference on. IEEE, pp. 134–141.
Roudier, P., Beaudette, D., Hewitt, A., 2012. A conditioned latin hypercube sampling algorithm incorporating operational constraints. In: Digital Soil Assessments and Beyond. CRC Press, Sydney, NSW, Australia, pp. 227–231.
Scott, E.M., 2018. The role of Statistics in the era of big data: Crucial, critical and under-valued. Statist. Probab. Lett. 136, 20–24. Special Issue on "The role of Statistics in the era of Big Data".
Truong, P.N., Heuvelink, G.B., Gosling, J.P., 2013. Web-based tool for expert elicitation of the variogram. Comput. Geosci. 51, 390–399.
Villanueva, F.J., Aguirre, C., Villa, D., Santofimia, M.J., López, J.C., 2014. Smart City data stream visualization using Glyphs. In: Innovative Mobile and Internet Services in Ubiquitous Computing, IMIS, 2014 Eighth International Conference on. IEEE, pp. 399–403.