International Journal of
Geo-Information

Editorial

Spatial Data Science

Fernando Bacao [1,*], Maribel Yasmina Santos [2] and Martin Behnisch [3]

[1] NOVA Information Management School (NOVA IMS), Universidade Nova de Lisboa, Campus de Campolide, 1070-312 Lisboa, Portugal
[2] Department of Information Systems, Campus de Azurém, University of Minho, 4800-058 Guimarães, Portugal; maribel@dsi.uminho.pt
[3] Leibniz Institute of Ecological Urban and Regional Development, 01217 Dresden, Saxony, Germany; m.behnisch@ioer.de
* Correspondence: bacao@novaims.unl.pt

Received: 3 July 2020; Accepted: 6 July 2020; Published: 8 July 2020

The field of data science has had a significant impact in both academia and industry, and with good reason. The ability to make use of large amounts of data to find solutions for pressing problems in society, the environment, and business, constitutes both an opportunity and a challenge. The concept of data is our best prospect to improve our understanding of the world significantly, ease the attrition in human/environment interaction, optimize resource allocation, and mitigate human suffering and deprivation.

Recently, there have been many examples of the "unreasonable effectiveness of data" (Haley et al. 2009 [1]), where sizable high-quality datasets unlock the solution to difficult and perennial problems. The ImageNet LargeScale Visual Recognition Challenge (ILSVRC) (Russakovsky et al. 2015 [2]) is probably one of the most spectacular examples of how data can have a pivotal role in advancing a whole field of research. The competition, that ran from 2010 to 2017, completely transformed the landscape of image recognition in a mere seven years. In this period, the winning accuracy in the classification of objects in the dataset rose from 71.8% to 97.3%, and the difference in the performance of the different teams drastically reduced. In the last year of the competition, 29 of the 38 teams achieved an accuracy rate above 95% (for an interesting account of the competition origins and development, see Gershgorn 2017 [3]). The spectacular results of this competition promoted a paradigm shift, where data take center stage, and its impact on improving the performance of old models and the development of new and improved ones becomes evident. After all, the idea is to let the data do the heavy lifting (Domingos 2012 [4]), and large-high-quality datasets proved to be up to par with the task.

The ImageNet competition example is far from being unique, as far as the relevance of data goes, but it is especially appropriate when talking about spatial data science. In fact, the ImageNet story also includes an interesting lesson for spatial data science, which is related to the pivotal role of convolution neural networks in the results of the ImageNet competition. The year 2012 was a turning point with the results achieved by AlexNet (Krizhevsky et al. 2012 [5]), which beat the competition by a massive 10.8% margin. This feat was probably one of the most critical events in the establishment of the deep learning phenomenon and the (re)boost of interest in machine learning and artificial intelligence. From 2012 onwards, convolution neural networks (CNN) dominated the competition and bled into many other areas of application. However, the compelling aspect of CNN for this Special Issue, and spatial data science in general, is the smart way in which they take into account the spatial structure of data, effectively encoding the first law of geography ("everything is related to everything else, but near things are more related than distant things." (Tobler 1970 [6])) into the algorithm.

The data deluge and the consequent digital transformation processes in the economy and society [7] also created new opportunities and challenges in the study of geographical phenomena. Due to the plethora of georeferenced data collected today by sensors and people, the transition from theory-driven research to data-driven research has been discussed in the literature (Miller and Goodchild 2015 [8]; "geographic research has shifted from a data-scarce to a data-rich environment"). This view is exaggerated by the emergence of the so-called fourth paradigm of science, i.e., after experimental science, theoretical science and computational science (simulating of complex phenomena) comes data science (data-intensive) (Hey et al. 2009 [9]; Kitchin 2014 [10]).

While in the 1980s and 1990s, the geographic information science community debated if there was something special in spatial data (Gahegan 2003 [11], Anselin 1990 [12] and Bação et al. 2005 [13]), today, the question does not seem to be so relevant, as data science is forced to deal with a myriad of data types, most of them suffering from similar pathologies as spatial data. Let us take the example of spatial dependence, which can be seen as a particular form of dependency between observations. The problem is not one of violating the independence assumption, as most data science methods are essentially assumption-free. The problem is that, if we do not account for spatial dependency in the model, the results will probably never be either very good or relevant. This is assuming that every phenomenon is defined by a process and expressed in a context, where the process represents the factors underlying the phenomena, and the context represents the frame in which the phenomena are observed (e.g., space and time). Spatial dependency indicates that the context has a meaningful impact in the process, in other words, the phenomenon in a particular location is a function of the underlying factors, but also of the intensity of that same phenomenon in neighboring locations. This factor adds complexity to the analysis, for it would be much simpler to concentrate our attention on the underlying factors and assume a neutral context. This facet is the reason why spatial data science needs to produce spatially explicit models.

The question now is what do we mean by spatially explicit models, according to (Goodchild 2001 [14]); these are not invariant under relocation, include spatial representations in their implementations, include spatial concepts in their formulations, and the spatial structures of inputs and outcomes are different. The important thing about spatially explicit models is that they harness the geographic frame to produce better results, whenever space is the relevant context of expression of the phenomenon. Therefore, building spatially explicit models in spatial data science is not so much a philosophical question; instead, it is a utilitarian approach.

Several authors (Miller and Goodchild 2015 [8]; Li et al. 2015 [15]; Jiang and Shekhar 2017 [16]) have already highlighted that spatial data science must support decision making in a meaningful way and not aim to replace human decisions, which are usually made by intelligence and skepticism (see Miller and Goodchild 2015 [8]; 'data dictatorship'). Thus, knowledge and theories of the disciplines should not be ignored in the course of spatial analyses, because otherwise, results (e.g., patterns and correlations in data) discovered by (big data) algorithms quickly tend to be uninteresting and less useful (Jiang and Shekhar 2017) [17]: "Ignoring domain knowledge and theories, patterns discovered by spatial big data science algorithms may be spurious."

The collection of papers accepted for this Special Issue is broad and eclectic and deals with topics that range from motion activity and trajectories to epidemic spreading. Some papers are more focused on developing theoretical aspects, and others on real-world applications, although all of them have reported experimental results. We are sure that the *International Journal of Geo-Information* reader will find some exciting and thought-provoking ideas in this Special Issue.

The paper "Spatio-Temporal Analysis of Intense Convective Storms Tracks in a Densely Urbanized Italian Basin" (Sangiorgio and Barindelli 2020) [17] combines both the spatial and temporal dimensions to identify the most favorable conditions for the formation of convective events. Intense convective storms usually produce large rainfall volumes in short time periods, leading to an increase in floods and corresponding damages. The use of visualization solutions allows for an improved understanding of the phenomenon and identifies the geographic areas where these convective thunderstorms are more frequent.

The paper "Analyzing Road Coverage of Public Vehicles According to Number and Time Period for Installation of Road Inspection Systems" (Sangiorgio et al. 2020) [18] deals with the problem of using sensors to address the monitoring of aging road infrastructure efficiently. They focus on a methodology to automate road inspection based on the use of a smartphone-based system and analyze the data collected from public vehicles with a long-term global positioning system (GPS), in two Japanese cities. The authors conclude that, with only a fraction of the public vehicles, the entire road inspection area can be achieved efficiently.

Living in the current pandemic situation, we are all too aware of the relevance of having appropriate spatial-temporal tools to identify, understand, and promptly react to the spread of pathogens. Hamer et al. 2020 [19] propose papros, an R package for spatial-temporal prediction based on local data, using various deterministic, geostatistical regionalization, and machine learning methods. To showcase the package, the authors present a use case—based on the prediction of powdery mildew infestation events.

Moreover, "Quantitative Identification of Urban Functions with Fishers' Exact Test and POI Data Applied in Classifying Urban Districts: A Case Study within the Sixth Ring Road in Beijing" (Yi et al. 2019) [20] puts forward a quantitative methodology to identify urban functions. The authors use Fisher's test and point of interest (POI) data, and apply the methodology to determine the urban districts, based on their urban functions within the Sixth Ring Road in Beijing. After the application of a k-modes clustering algorithm, the authors identify four main groups of districts based on their urban functions.

Dealing with trajectory data continues to be a challenge; there are still many problems to tackle in order to be able to extract relevant and accurate knowledge from trajectory data. Pulshashi et al. [21] propose an application to simplify trajectory data, for both batch and streaming environments, in their paper "Simplification and Detection of Outlying Trajectories from Batch and Streaming Data Recorded in Harsh Environments." The application seeks to reduce noise, and especially outlying point-locations that can mislead the analysis and alter the statistical properties of trajectories. They conclude with an experimental evaluation of the proposed method and compare it with other outlier detection algorithms.

Finally, the last paper of this Special Issue [22] (Crivellari and Beinat 2019) uses motion traces to build a behavioral portrait of places based on how people move between them. In their proposal, they ignore geographical coordinates and spatial proximity, and based on the word2vec concept, create a motion-to-vector (Mot2vec). They start by transforming the original trajectories into sequences of locations, and then they use the skip-gram word2vec model to build the location embedding. According to the authors, these embeddings constitute a meaningful representation of locations, "allowing a direct way of comparing locations' connections and providing analogous similarity distributions for places of the same type."

With this Special Issue of the *ISPRS International Journal of Geo-Information*, based on spatial data science, we hope to contribute to promoting the discussion and interest around the role of spatial in data science. More importantly, we hope that this volume can be seen as a contribution to encourage the geographic information science community to become (even more) involved, and contribute to the advance of this exciting and thriving field.

Author Contributions: The three authors contributed equally to this work. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially funded by the Portuguese Foundation for Science and Technology (FCT), under projects IPSTERS (DSAIPA/AI/0100/2018), and foRESTER (PCIF/SSI/0102/2017).

Acknowledgments: The guest editors would like to thank the whole editorial team and all reviewers who provided constructive and helpful comments on the articles in our special issue.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analysis, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Halevy, A.; Norvig, P.; Pereira, F. The Unreasonable Effectiveness of Data. *IEEE Intell. Syst.* **2009**, *24*, 8–12. [CrossRef]
2. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]
3. Gershgorn, D. The Data that Transformed AI Research—And Possibly the World. *Quartz*, 26 July 2017.
4. Domingos, P. A few useful things to know about machine learning. *Commun. ACM* **2012**, *55*, 78–87. [CrossRef]
5. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Pdf ImageNet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
6. Tobler, W.R. A Computer Movie Simulating Urban Growth in the Detroit Region. *Econ. Geogr.* **1970**, *46*, 234. [CrossRef]
7. Mayer-Schonberger, V.; Cukier, K. *Big Data: A Revolution That Will Change How We Live, Work and Think*; Eamon Dolan: Boston, MA, USA, 2013.
8. Miller, H.; Goodchild, M.F. Data-driven geography. *GeoJournal* **2014**, *80*, 449–461. [CrossRef]
9. Hey, T.; Tansley, S.; Tolle, K. (Eds.) *The Fourth Paradigm—Data-Intensive Scientific Discovery*; Microsoft Research: New York, NY, USA, 2009.
10. Kitchin, R. Big data and human geography. *Dialog-Hum. Geogr.* **2013**, *3*, 262–267. [CrossRef]
11. Gahegan, M. Is inductive machine learning just another wild goose (or might it lay the golden egg)? *Int. J. Geogr. Inf. Sci.* **2003**, *17*, 69–92. [CrossRef]
12. Anselin, L. What is Special About Spatial Data? In *Alternative Perspectives on Spatial Data Analysis, in Spatial Statistics, Past, Present and Future*; Griffith, D.A., Ed.; Institute of Mathematical Geography: Ann Arbor, ML, USA, 1990; pp. 63–77.
13. Bação, F.; Lobo, V.; Painho, M. On the particular characteristics of spatial data and its similarities to secondary data used in data mining. In Proceedings of the GIS PLANET 2005, II International Conference and Exhibition on Geographic Information, Estoril, Portugal, 30 May–2 June 2005.
14. Goodchild, M. Issues in spatially explicit modeling. In Proceedings of the Agent-Based Models of Land-Use and Land-Cover Change Report and Review of An International Workshop, Irvine, CA, USA, 4–7 October 2001; Parker, C.D., Berger, T., Manso, S.M., Eds.; LUCC Focus 1 Office: Bloomington, IN, USA, 2001; pp. 12–15.
15. Li, S.; Dragićević, S.; Castro, F.A.; Sester, M.; Winter, S.; Çötekin, A.; Pettit, C.J.; Jiang, B.; Haworth, J.; Stein, A.; et al. Geospatial big data handling theory and methods: A review and research challenges. *ISPRS. J. Photogramm. Remote Sens.* **2016**, *115*, 119–133. [CrossRef]
16. Jiang, Z.; Shekhar, S. *Spatial Big Data Science-Classification Techniques for Earth Observation Imagery*; Springer: Cham, Switzerland, 2017.
17. Sangiorgio, M.; Barindelli, S. Spatio-Temporal Analysis of Intense Convective Storms Tracks in a Densely Urbanized Italian Basin. *ISPRS Int. J. Geo-Inf.* **2020**, *9*, 183. [CrossRef]
18. Kashiyama, T.; Sekimoto, Y.; Seto, T.; Lwin, K.K. Analyzing Road Coverage of Public Vehicles According to Number and Time Period for Installation of Road Inspection Systems. *ISPRS. Int. J. Geo-Inf.* **2020**, *9*, 161. [CrossRef]
19. Hamer, W.B.; Birr, T.; Verreet, J.-A.; Duttmann, R.; Klink, H. Spatio-Temporal Prediction of the Epidemic Spread of Dangerous Pathogens Using Machine Learning Methods. *ISPRS. Int. J. Geo-Inf.* **2020**, *9*, 44. [CrossRef]
20. Yi, D.; Yang, J.; Liu, J.; Liu, Y.; Zhang, A.J. Liu Quantitative Identification of Urban Functions with Fishers' Exact Test and POI Data Applied in Classifying Urban Districts: A Case Study within the Sixth Ring Road in Beijing. *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 555. [CrossRef]

21. Pulshashi, I.R.; Bae, H.; Choi, H.; Mun, S.; Sutrisnowati, R.A. Simplification and Detection of Outlying Trajectories from Batch and Streaming Data Recorded in Harsh Environments. *ISPRS. Int. J. Geo-Inf.* **2019**, *8*, 272. [CrossRef]
22. Crivellari, A.; Beinat, E. From Motion Activity to Geo-Embeddings: Generating and Exploring Vector Representations of Locations, Traces and Visitors through Large-Scale Mobility Data. *ISPRS. Int. J. Geo-Inf.* **2019**, *8*, 134. [CrossRef]