Data science applied to environmental sciences

Received 26 April 2016; | Revised 6 June 2016; | Accepted 6 June 2016

DOI: xxx/xxxx

ARTICLE TYPE

Data Science Applied to Environmental Sciences[†]

Paulo Canas Rodrigues*[1,2] | Elisabetta Carfagna*[3,4]

[1]Department of Statistics, Federal University of Bahia, Salvador, BA, Brazil
[2]Chair of the ISI Special Interest Group on Data Science
[3]Department of Statistical Sciences, University of Bologna, Italy
[4]Vice-Chair of the ISI Special Interest Group on Data Science

Correspondence
*Paulo Canas Rodrigues: Email: paulocanas@gmail.com; Elisabetta Carfagna: Email: elisabetta.carfagna@unibo.it

Summary

In recent years, immense amounts of data have been generated, from sensors to purchase transaction records, mobile GPS signals, digital satellite images and social media. The raise of data collection has brought the need for quantitative minded professionals able to transform that data into information and decision making. In this opinion piece, we will share some of our views and experiences about the general role that data science plays nowadays, with a special interest in the field of environmetrics. We will include a limited number of examples that highlight the usefulness of data science in environmetrics, and a specific illustration of the behavior of the wildfires in Brazil between January and December of 2021.

KEYWORDS:
data science, environmetrics, statistics, Brazilian wildfires

1 | INTRODUCTION

Data science has emerged as a very strong, visible, and publicly recognized label for problem-solving using ever-growing, large data sets and new data sources, but the crucial questions are: what is “data science”? Should it be within the field of statistics or computer science? What is its relationship with “applied statistics”, “big data”, “machine learning”, “deep learning,” and “artificial intelligence”?

Many authors have discussed these concepts and their relationship in many areas of statistics and in many fields of application. Donoho (2017) reviewed some ingredients and points of view of the current “data science moment,” and discussed how/whether data science is really different from statistics, stating that the field of data science amounts to a superset of the fields of statistics and machine learning, which adds some technology for “scaling up” to “big data.”

Clearly data science does not completely coincide with statistics, but the overlap is wide. For example, as clarified in the book by Agresti and Kateri (2021), a data scientist should be familiar with statistical science, including probability distributions, descriptive and inferential statistical methods, and be able to use software to implement statistical methods and to conduct simulations to illustrate key concepts.

Computer Science and programming represents the language of the Data Science, whereas Statistics is the logic of the Data Science that allows for the modeling, analysis and interpretations that transform data into information. Besides these two fields, other knowledge such as of scientific method, mathematics, visualization, data engineering, and domain expertise, represent the toolbox of Data Science, that can be seen as a team sport with experts from all these subjects. In this point of view, there is not just a data science, but data sciences.

Data Science adopts and/or develops appropriate methodologies for purposes of knowledge discovery, forecasting, and decision-making in the face of an increasingly complex reality often characterized by large and non-structured amounts of data

[†]The authors would like to thank the Guest Editors of this Special Issue for the invitation to write this opinion piece, and to all our colleagues of the ISI Special Interest Group on Data Science for fruitful discussions on this topic.

("big data"), of various types (e.g. numeric, ordinal, nominal, symbolic, text, images, audio, video, click streams, networks, etc.), coming from disparate sources (sensors, social media, surveys, GPS signals, transactions, digital satellite images, etc.).

The analysis and interpretation of these kinds of data include the adaptation of existing methods, such as regression, calibration, and classification, and the development of novel statistical methods, such as deep learning methodologies, for specific data science applications.

The discussion on advantages, disadvantages, limitations, and requirements of the use of alternative methodologies and data sources in all areas of knowledge is setting the stage for the debate in national and international communities of statisticians, computer scientists, and other communities, all over the world. However, caution should be exercised. As stressed out by Utts (2021), as sources of data become more plentiful and massive data sets are easier to acquire, new ethical issues arise involving data quality and privacy, and the analysis, interpretation and dissemination of data-driven decisions. Statisticians and Data Scientists can help raise awareness and encourage implementation of ethical best practices. Moreover, they are in a privileged position for selecting, interpreting, and combining the most appropriate tools for data collection, data modeling and prediction.

There are indeed many challenges in the field of data science, from the development of proper models to high performance computational needs, and from ethical issues to analysts with quantitative knowledge (i.e. Data Scientists), among others. And these have been and will continue to be widely discussed.

In the next section we present a few examples of the usefulness of data science in environmetrics, and then present an illustration of a spatio-temporal virtualization of the wildfires in the Brazilian territory during 2021, before drawing some concluding remarks about the role of data science in statistics and environmetrics.

2 | THE USEFULNESS OF DATA SCIENCE IN ENVIRONMETRICS

Statistics, big data, machine learning techniques, and the integration of various data sources for addressing environmental problems have widespread over the years. We mention here just a few examples of the usefulness of data science in environmetrics, for illustration purposes.

One widely considered application of statistical and data science techniques in environmetrics is related to the modeling of particle pollution such as PM10, i.e. inhalable particles, with diameters that are generally 10 micrometers and smaller. For example, Encalada-Malca, Cochachi-Bustamante, Rodrigues, Salas, and López-Gonzales (2021) provided a spatio-temporal visualization approach of PM10 concentration data in metropolitan Lima, Peru, and Cordova et al. (2021) considered the same data for time series forecasting using artificial neural networks. On the other hand, Lee, Robertson, Ramsay, and Pyper (2020) integrated various information layers with different supports and the analysis of the impact of the modifiable areal unit problem when estimating the health effects of air pollution.

Other applications of data science in environmetrics are related, e.g. with environmental variables. For example, Ahmed, Maume-Deschamps, and Ribereau (2021) adopted a deep learning approach for recognizing a spatial extreme dependence structure with application to data on air temperature rainfall. Other developments and applications are widely available in the literature.

The field of environmetrics has much to gain, not only from the attention that data science gave to the community, but also from some of the models and computational power that it brought with. The environmental databases are becoming larger and from different sources, which makes them difficult to manipulate and analyze in standard personal computers, and more diverse, including, e.g., data from sensors and satellite images. Consequently, the analysis of this data requires computational power and specific expertise in different fields. This is where an interdisciplinary team, that can be called a data science team, mastering subjects such as statistics, advanced computing, data engineering, visualization, and domain expertise, is required. The combination of these expertise allows for a great advance in the field of environmetrics and helps solving meaningful problems in environmental sciences.

3 | SPATIO-TEMPORAL VISUALIZATION OF THE BRAZILIAN WILDFIRES

Wildfires are one of the most common natural disasters in many world regions that actively impact everyday life, which result mostly from human activities and climate change. In this section, we present a spatio-temporal visualization of all fire spots detected in the Brazilian territory by the reference satellite AQUA M-T, between January and December 2021. This

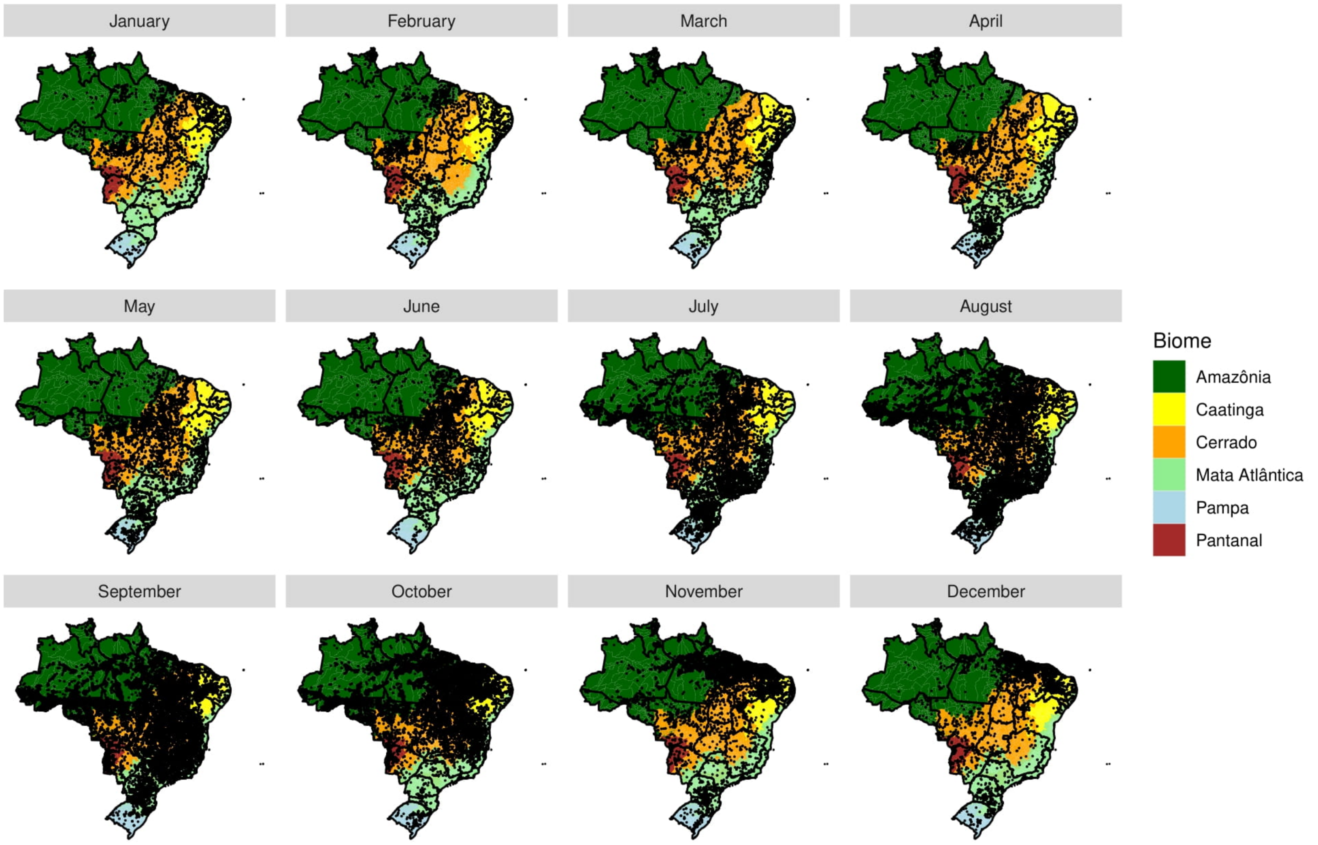

FIGURE 1 Locations of the fire spots per month, between January and December 2021, for each of the six biomes: Amazônia, Caatinga, Cerrado, Mata Atlântica, Pampas, and Pantanal.

data were obtained from the Brazilian National Institute for Space Research (INPE; Instituto Nacional de Pesquisas Espaciais; queimadas.dgi.inpe.br), and includes the time and location of the 184081 detected fire spots.

Figure 1 shows the locations of the fire spots per month, between January and December 2021, being the map colored by each of the six biomes (official ecosystem types): Amazônia, Caatinga, Cerrado, Mata Atlântica, Pampas, and Pantanal. The months between July and October are particularly intense in the number of fire spots, especially in the south Amazonia, Pantanal, and Cerrado.

After this original spatio-temporal data visualization, the next step would be to better understand the dynamics of the data and the possible explanatory variables in a spatio-temporal modeling approach. Pimentel, Bulhões, and Rodrigues (2022) considered the historical data with the geographical locations of all the fire spots detected by the reference satellites that cover the whole Brazilian territory between January and December 2021, comprising more than 1.8 million fire spots, and used a spatial econometric model with meteorological and human variables as covariates. They found out that the change in land use from forest and green areas to farming has a significant positive impact on the number of fire spots for all six Brazilian biomes.

4 | CONCLUDING REMARKS

Data science is not a passing trend, it has come to stay! And we should adapt to that. Statistics is its cornerstone and should embrace and lead the changes, both in terms of education and research. We need to think about how we can exploit this opportunity, and to act accordingly. We need to look beyond our backyards.

Statisticians, like computer scientists, should lead (and be part of) teams that aim at solving large and meaningful problems in data science for a wide spectrum of areas of application. This is particularly true and relevant for the field of environmetrics in an era where climate change, pollution, deforestation, and environmental sustainability are in the agenda. These collaborations and interdisciplinary teams will make a great impact in the field of environmetrics.

Indeed, we need data scientists, statistical minded people with programming skills that are curious and that can transform data into information. They will help in getting trustworthy analyses for proper decision making.

We wrote this opinion piece while being the current Chair and Vice-Chair of the recently established Special Interest Group (SIG) on Data Science of the International Statistical Institute (ISI). This SIG includes representatives from the ISI associations and other ISI SIG, and aims to: (i) strengthen the collaboration between statisticians, computer scientists and other communities in the field of data science; (ii) provide a "big tent" to all ISI members, and other colleagues interested in data science and data analytics, to sit under; (iii) emphasize the importance of data science in the activities promoted by the ISI and its associations; (iv) foster capacity building on data science, including computational skills among statisticians, and statistical skills among computational scientists; and (v) coordinate and promote, in collaboration with the ISI and its Associations, specific activities in the field of data science.

ACKNOWLEDGMENTS

The authors would like to acknowledge the Guest Editors of this Special Issue for the invitation to write this opinion piece and to all our colleagues of the ISI Special Interest Group on Data Science, especially Daniel Jeske and Jürgen Symanzik, for fruitful discussions on this topic. We also thank Jonatha Pimentel from the Federal University of Pernambuco, Brazil, for helping design the figure in this paper and to the associate editor and the anonymous reviewer for their generous time in providing detailed comments and suggestions that helped us improving the paper.

References

Agresti, A., & Kateri, M. (2021). *Foundations of statistics for data scientists: With r and python*. Chapman and Hall/CRC.

Ahmed, M., Maume-Deschamps, V., & Ribereau, P. (2021). Recognizing a spatial extreme dependence structure: A deep learning approach. *Environmetrics*, e2714.

Cordova, C. H., Portocarrero, M. N. L., Salas, R., Torres, R., Rodrigues, P. C., & López-Gonzales, J. L. (2021). Air quality assessment and pollution forecasting using artificial neural networks in metropolitan lima-peru. *Scientific Reports*, *11*(1), 1–19.

Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics*, *26*(4), 745–766.

Encalada-Malca, A. A., Cochachi-Bustamante, J. D., Rodrigues, P. C., Salas, R., & López-Gonzales, J. L. (2021). A spatio-temporal visualization approach of pm10 concentration data in metropolitan lima. *Atmosphere*, *12*(5), 609.

Lee, D., Robertson, C., Ramsay, C., & Pyper, K. (2020). Quantifying the impact of the modifiable areal unit problem when estimating the health effects of air pollution. *Environmetrics*, *31*(8), e2643.

Pimentel, J., Bulhões, R., & Rodrigues, P. (2022). Spatio-temporal modelling of the brazilian wildfires: The influence of human and meteorological variables. *submitted*.

Utts, J. (2021). Enhancing data science ethics through statistical education and practice. *International Statistical Review*, *89*(1), 1–17.

How to cite this article: Rodrigues, P.C., and E. Carfagna (2022), Data Science Applied to Environmental Sciences, *Environmetrics*, *2022;00:1–3*.