Advanced Geospatial Data Analysis in R: Environmental Applications

Marj Tonini, Haokun Liu

2024-06-04

Contents

Preface

Earth surface environmental processes exhibit distinctive characteristics, encompassing both spatial and temporal dimensions, along with various attributes and predictive variables. Furthermore, in the era of Data Science, the wealth of available data and the rapid development of analytical models have emerged as distinctive aspects in the realm of **Geospatial Data Analysis (GDA)**. Coupled with uncertainty and complexity issues, all this contribute to making this field of research highly challenging. The domain of GDA encompasses data exploration, manipulation, and modelling, from the acquisition phase up to the visualization and interpretation of the results. Mapping where events are located and how they relate to each other provides a better understanding of the process being studied. Finally, as an increasing volume of geo-environmental data becomes more and more accessible, the demand for experts in this domain is growing rapidly, both in public research institutions as well as in private companies.

Defined for the first time by Naur as *"The science of dealing with data"*, the term **Data Science** evolved over time around the original concept of *"…converting data into information and knowledge"* (IASC, 1977). In the disciplines like Environmental and Earth Sciences, Physical Geography, Humanities and Social Sciences, the use of Data Science procedures is emerging only recently, proving to be extremely efficient to deal with the complexity of the investigated process and the heterogeneity of the underlying data sources. This leads to a cultural shift, moving scientists away from individual working within their own research domain. Indeed, disciplinary boundaries are more and more permeable, pushing scientists to be more open to collaborate among them and with decision makers on the investigation and understanding of real-world problems. Modern earth and environmental scientists need to interact with other disciplines, apparently far from their domain. This openness is increasingly important as society struggles to respond to the implication of anthropogenic pressures on different issues, such as natural hazards and climate change, or the harmful impacts of human activities on biodiversity, water and air quality, human health.

The target audience of this eBook are master and graduate (PhD) students in Earth/Environmental Sciences, Biology and Spatial Ecology, Physical Geography, and equivalent disciplines. We strive to empower students by guiding

them through both theoretical knowledge and hands-on practical applications, enabling them to cultivate effective problem-solving skills. The primary focus here delves into applying Data Science methodologies to understand and analyze Earth's surface environmental processes. The scientific approaches related to this emerging discipline, ranging from statistics, mathematics, geomatics and computer science, are often hard to be acquired. While maintaining a strong emphasis on rigorous mathematical and statistical formalism in the methods presented, this eBook primarily accentuates their practical applications within the realm of Geosciences

The book is addressed to intermediate to advanced R users with some experience on geospatial data and a great interest in geocomputing. If you have some very basic knowledge in these fields, we encourage you to explore links provided in each chapter, which redirect you to further useful documentation.

This eBook seeks to provide the audience with:

- A good understanding of main practical concepts and applied aspects in GDA;
- Advanced tools designed to proficiently navigate various techniques for analysing spatial datasets in geosciences.

Specifically, the methodologies outlined will equip readers with expertise in various algorithms tailored for analyzing complex geo-environmental datasets. This knowledge will empower them to extract valuable insights and translate them into actionable decisions.

Explored algorithms include:

- *Geographically Weighted Summary Statistics* for exploratory data analyses and visualization of geographical variations in the statistical data distribution;
- *Ripley's K-function, Kernel Density Estimator, DBSCAN* for cluster detection and mapping;
- *Self-Organizing Maps* as an example of unsupervised machine learning approach to data clustering and visualization;
- *Random Forest* as an example of a supervised machine learning algorithm, applied here for classification.

Author Information

Marj Tonini is a spatial data scientist with a strong interest on geospatial modelling for risk assessment, mainly related to wildfires and landslides. She gained her PhD in 2002 at Sant'Anna School of advanced studies (Pisa, Italy), defending a thesis on agro-environmental modelling. She started working at the

University of Lausanne in 2004, as post-doc and, in 2008, she was appointed senior research manager at the Institute of Earth Surface Dynamics (current position).

Her research focuses on the development of innovative approaches allowing to enable efficient learning from complex environmental datasets. She seeks to set up a methodological framework to understand the spatio-temporal dynamic of environmental processes and to assess the influence of predictor variables. Her current research targets are on land use/land cover change analyses and on the elaboration of predictive scenario and susceptibility / risk assessment for natural hazard.

Haokun Liu is a Ph.D. student at the Group of Cities and dynamics of networks (CITADYNE Group), University of Lausanne. Meanwhile, he is also a student assistant at Swiss Geocomputing Center (SGC). Benefiting from the strict and effective training in China and Switzerland, his research interests and experience include, but are not limited to: 1) Urban analytics, 2) Health geography, 3) Spatial data science. With the goal of bridging the gap between theories and practical applications, he am focusing on the intersections among urban health, complex perspective, activity space, and the Spatial-social-semantic framework.

Prerequisites

- Knowledge of basic statistics: methods of descriptive statistics (measures of central tendency and dispersion); how to assess relationships between variables; concepts of correlation and regression.
- Basic knowledge in geomatics (GIS): basic operations with raster and vector datasets.
- R programming basics and RStudio.

Acknowledgements

The case study presented in each chapter came from different projects carried out in collaboration with several colleagues, including master and PhD students. All the produced scientific papers are duly cited in the bibliography.

We would like to thank the following collaborators:

- Mário Gonzalez Pereira and Joana Parente, for the extensive and fruitful collaboration we have had in investigating the spatio-temporal distribution of wildfires in Portugal; one of our studies seeking to investigate the evolution of forest fires, from spatio-temporal point events to smoothed density maps, is an integral part of Chapters 2 and 3.

- Stuart Lane and Natan Micheletti, for introducing me to the fascinating world of rock glacier research; notably, the 3D-points clouds dataset analysed in Chapter 4 was acquired and processed by Natan during his PhD studies.

- Axelle Bersier, for the meticulous work she did in acquiring and preprocessing the Swiss national population census dataset, which have been used for the exercise about unsupervised learning in Chapter 5.

- Julien Riese, for producing the input dataset and collaborating with me in developing the code allowing to assess the landslides susceptibility in canton Vaud, that is the main topic of Chapter 6.

- Both Julien and Axelle serve as a shining example of a highly successful master's students whom I had the pleasure of supervising.