The Journal of Open Source Education

Spatial data analysis with R: wrangling, visualization and econometric models

Jaime A. Prudencio-Vázquez[1]

1 Economics Department, Universidad Autónoma Metropolitana, Unidad Azcapotzalco, CDMX, México.

DOI: 10.21105/jose.00173

Software

- Review
- Repository
- Archive

Submitted: 01 March 2022
Published: 08 June 2023

Summary

Spatial Analysis with R: wrangling, visualization and econometric models (Análisis Espacial con R: manejo, visualización y modelos econométricos)is a full open course in Spanish for the analysis of spatial information in R software. The course is intended for Spanish-speaking economics undergraduate students interested in acquiring technical skills for quantitative analysis required by the regional sciences (Fischer & Nijkamp, 2014) and the spatial approach to economics (Fujita et al., 2000), but it could be useful for other students of social sciences attracted by the topic.

The objective of the book is to guide the student, from a basic practical approach, in the knowledge and handling of spatial information exploration, analysis and modeling techniques through the use of R (R Core Team, 2021) and RStudio, which are, respectively, a program computer science and programming language focused on statistical analysis and information visualization, on the one hand, and an integrated development environment (IDE), on the other.

The course is presented in the form of an electronic book and it is structured in five chapters of gradual learning. In Chapter 1, the basics of R and RStudio are introduced using the exploratory data analysis approach (Croarkin & Tobias, 2014) not only with the basic R package, but also with the popular set of tools provided by tidyverse [@ tidyverse], that is, the student is completely introduced to the use of the software to propose and solve questions related to the information structure. Chapter 2 shows how to create various types of choropleth maps and the enormous styling customization flexibility for this purpose via the tmap (Tennekes, 2018) package. Chapter 3 presents how to carry out an exploratory analysis of spatial data where the student will find how to define the interrelationships that take place in space through the construction of spatial weight matrices and she will learn about spatial autocorrelation and its implications in the information analysis, through various packages such as spdep (Bivand, 2022) and rgeoda (Li & Anselin, 2022). Meanwhile, in Chapter 4, a very synthetic review of simple regression models is presented, emphasizing the autocorrelation problem that could occur when estimating a linear model with spatial data. Finally, in Chapter 5, two of the different spatial econometric modeling alternatives available in R it is shown, with spatialreg (Bivand et al., 2021).

The logic of each chapter integrates three elements: i) explanation of the fundamental concepts covered, ii) the use of real information in the software that serves to illustrate the highlighted concepts, iii) exercises proposed for the student to delve into the topics exposed.

The examples and exercises presented in the course are based on a database on the situation of the COVID19 pandemic between March and September 2020 in the Metropolitan

Zone of the Valley of Mexico (Zona Metropolitana del Valle de México), the largest metropolitan area in Mexico, made up of 76 local administrative units or municipalities with more than 21 million inhabitants. In addition, the database used provides information on the economic structure at the municipal level and a set of variables of sociodemographic characteristics from the population census.

Story of the project

The project arises and is fed by two academic experiences. The first, which dates back to 2014, corresponds to my work as a teacher of tools for spatial analysis through various introductory courses. The second is a course on the use of R software that I taught in the fall of 2020, from the Economics Department of the Autonomous Metropolitan University (Universidad Autónoma Metropolitana Unidad Azcapotzalco), based on Azcapotzalco, Mexico City.

As a result, multiple notes and materials were generated and gradually incorporated into my teaching activity in the Spatial Econometrics course, for which I have been responsible for more than 3 years. Spatial Econometrics is part of the academic offer of the specialization line of the Degree in Economics called "Economics of Innovation: firms, networks and territory," of the aforementioned university. I frequently found myself in need of generating materials for teaching the contents of Spatial Econometrics. Thus, instead of isolated and unstructured materials, I decided to compile and order with a coherent expository logic and gradual learning the set of materials worked up to now and that now make up this course.

Since January 2022, when the integrated version of this series of materials was put into circulation in the form of an electronic book, the Spatial Econometrics course has been taught continuously in quarterly promotions to almost 50 students.

Statement of Need

Regional sciences are characterized by their multidisciplinary character and their solid quantitative support (Fischer & Nijkamp, 2014). Although it is true that in all economics Majors we find a solid repertoire of quantitative instruments: mathematics, statistics and, of course, econometrics, the necessary tools for the analysis of reality from a spatial and regional perspective are still scarce within the economy. training at the undergraduate level. Analysis of spatial data with R intends to be a contribution, albeit minimal, to remedy this situation.

In addition, much of the literature on data management and analysis of spatial information is still in a language other than Spanish, mostly in English, which makes access to these tools difficult for those who do not yet master the language. This becomes a barrier to knowledge, notably in countries like Mexico where English proficiency is still very low (IMCO et al., 2015; Matt, 2020). Thus, this material in Spanish becomes a gateway and facilitates the learning process for the student interested in these topics.

The book, focused on undergraduate students who are not specialists in the subject, is intended to be an accessible material, because the used language is didactic and as simple as possible, without losing rigor.

As it is an introductory book, it recommends and invites the use of multiple materials, both in Spanish and English, so that the student deepens her own knowledge of spatial information management and the computer platform in which it is carried out.

Suggestions for following the course

Each chapter contains both theoretical and practical elements related to the treatment of spatial data in the context of economics. These are illustrated through segments of R code that are explained in their logic and structure so that the student can not only replicate the results, but also understand what each piece of code does. In addition, throughout each chapter, exercises are proposed to deepen, as a challenge, the knowledge about the tools that are exposed, for which the answers are not provided.

The suggested way to follow the course is through the sequence proposed by the structure of the book, from chapter 1 to 5, since the tools are exposed gradually, trying to ensure adequate assimilation. However, if the student feels already comfortable with handling one topic, she can move on to the next one without difficulty.

Experience in the use of this material indicates that it can be covered in 5 or 6 weeks, spending between 4.5 and 5 hours per week of study. Chapter 1 can be covered in approximately 4.5 hours, however if the student is not familiar with R and RStudio it could take a bit longer. The choropleth mapping section would be smoothly covered in about 3 hours. Meanwhile, Chapter 3 on the exploratory analysis of spatial data, one of the central parts of the book, may require at least 6 hours of study. Chapter 4, dedicated to the elementary review of linear regression, can be covered in 3 hours of study, since some previous knowledge on the subject is assumed, however, more could be required if the student needs to review these topics in greater depth. Finally, Chapter 5, also central to this material, would require around 6 hours of study.

Contributions

This book is, to some extent, an effort to compile and systematize multiple materials that the active R community, interested in spatial analysis in Mexico and the world, selflessly shares.

I believe that this is how knowledge should always be: free, open and collaborative, like the software that is used here. Thus, this project is a living one, in permanent construction and modification, so all comments and observations will be welcome, both from the students who have used it, and from the teachers who consider it appropriate to include it in their reference materials.

A guide on how to contribute to this project can be found in the GitHub repository, here, but the interested people could contact directly via e-mail (japv@azc.uam.mx) or even in social media on twitter. All comments are welcome.

Acknowledgements

The author thanks Montserrat Romero Martínez (vmrm@azc.uam.mx) and Alvaro Martínez Rodríguez (amr@azc.uam.mx), assistants in the Productive Relations Area, who were respectively in charge of reviewing Preliminary materials for this book and its edition for publication online with Bookdown on GitHub.

References

Bivand, R. (2022). R packages for analyzing spatial data: A comparative case study with areal data. *Geographical Analysis*, *54*(3), 488–518. https://doi.org/10.1111/

gean.12319

Bivand, R., Millo, G., & Piras, G. (2021). A review of software for spatial econometrics in R. *Mathematics*, *9*(11). https://doi.org/10.3390/math9111276

Croarkin, C., & Tobias, P. (2014). NIST/SEMATECH e-handbook of statistical methods. In *Retrieved January* (Vol. 1).

Fischer, M. M., & Nijkamp, P. (2014). Handbook of regional science. In *Handbook of Regional Science*. https://doi.org/10.1007/978-3-642-23430-9

Fujita, M., Krugman, P., & Venables, A. J. (2000). *Economia espacial : Las ciudades, las regiones y el comercio internacional*. Ariel.

IMCO, COMCE, & social, I. para la competitividad y la movilidad. (2015). *Inglés es posible propuesta de una agenda nacional*. Instituto Mexicano para la Competitividad, A.C. https://imco.org.mx/wp-content/uploads/2015/04/2015_Documento_completo_Ingles_es_posible.pdf

Li, X., & Anselin, L. (2022). *Rgeoda: R library for spatial data analysis*. https://CRAN.R-project.org/package=rgeoda

Matt. (2020). EF EPI 2019: El nivel de inglés en méxico sigue disminuyendo ‹ GO blog | EF blog mexico. In *EF Educación Internacional*. https://www.ef.com.mx/blog/language/nivel-de-ingles-en-mexico-sigue-disminuyendo/

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Tennekes, M. (2018). tmap: Thematic maps in R. *Journal of Statistical Software*, *84*(6), 1–39. https://doi.org/10.18637/jss.v084.i06