Applied Spatial Data Analysis and Statistics

Antonio Paez

3/11/2021

> "Patterns cannot be weighed or measured. Patterns must be mapped."
>
> — Fritjof Capra, The Web of Life: A New Scientific Understanding of Living Systems

Preface

Spatial Analysis and Spatial Statistics

The field of spatial statistics has experienced phenomenal growth in the past 20 years.

From being a niche subdiscipline in quantitative geography, statistics, regional science, and ecology at the beginning of the 1990s, it is now a mainstay in applications in a multitude of fields, including medical imaging, remote sensing, civil engineering, geology, statistics and probability, spatial epidemiology, end ecology, to name just a few disciplines.

The growth in research and applications in spatial statistics has been in good measure fueled by the explosive growth in geotechnologies: technologies for sensing and describing the natural, social, and built environments on Earth. An outcome of this is that spatial data are, to an unprecedented level, within the reach of multitudes. Hardware and software have become cheaper and increasingly powerful, and we have transitioned from a data poor environment (in all respects, but particularly in terms of spatial data) to a data rich environment. Twenty years ago, for instance, technical skills in spatial analysis included tasks such as digitizing. In the mid-1990s, as a Masters student, I spent many boring hours digitizing paper maps before I could do any analysis on the single-seat (and relatively expensive) Geographic Information System (GIS) available in my laboratory. In that place at that time I was more or less a geographical freak: although there was an institutional push to adopt GIS, relatively few in my academic environment saw the value of spending hours digitizing maps, something that nowadays would be considered relatively low-level technical work. Surely, the time of a Masters student, let alone a professional researcher or business analyst, is more valuable than that. Indeed, very little time is spent anymore in such low-level tasks, as data increasingly are collected and disseminated in native digital formats. Instead, there is a huge appetite for what could be called the *brainware* of spatial analysis, the intelligence counterpart of the hardware, software, and data provided by geotechnologies.

The contribution of brainware to spatial analysis is to make sense of vast amounts of data, in effect transforming them into information. This information in turn can be useful to understand basic scientific questions (e.g., changes in land cover), to support public policy (e.g., what is the value capture of public infrastructure), and to inform business decisions (e.g., what levels of demand can be expected given the distribution of retail outlets). There are numerous forms of spatial analysis, including normative techniques [such as spatial optimization; see @Tong2011] and geometric and cartographic analysis [for instance, map algebra; @Tomlin1990map]. Among these, spatial statistics is one of the key elements in the family of toolboxes for spatial analysis.

So what is spatial statistics?

Very quickly, I will define spatial statistics as the application of statistical techniques to data that have geographical references - in other words, to the statistical analysis of maps.

Like statistics more generally, spatial statistics is interested in hypothesis testing and inference. What distinguishes it as a branch of the broader field of statistics is its explicit interest in situations where data are not independent from each other (like throws of fair dice) but rather display systemic associations. These associations, when seen through the lens of cartography, can manifest themselves as patterns of similarities (e.g., birds of a feather flock together) or dissimilarities (e.g., repulsion due spatial competition among firms) - as two common examples of spatial patterns.

Spatial statistics covers a broad array of techniques for the analysis of spatial patterns, including tools for testing whether patterns are random or not, and a wide variety of modeling approaches as well. These tools enhance the brainware of analysts by allowing them to identify and possibly model patterns for inferring processes and/or for making spatial predictions.

Why this Text?

The objective of this book is to introduce selected topics in applied spatial statistics.

The foundations for the book are the notes that I have developed over several years of teaching applied spatial statistics at McMaster University. This course is a specialist course for senior-level undergraduate geographers and students in other disciplines who are often working towards specializations in GIS.

Over the course of the years, my colleagues at McMaster and I have used at least three different textbooks for teaching spatial statistics. I have personally used the book by McGrew and Monroe [-@Mcgrew2009] to introduce fundamental statistical concepts to geographers. McGrew and Monroe (currently on a third edition with Lembo) do a fine job of introducing statistics as a tool for decision making, and therefore offer a very valuable resource to learn matters of inference, for instance. Many of the examples in the book are geographical in nature; however, the book is relatively limited in its coverage of *spatial statistics* (particularly models for spatial processes), which is a limitation for teaching a specialist course on this topic.

My text of choice early on (approximately between 2003 and 2010) was the excellent book *Interactive Spatial Data Analysis* by Bailey and Gatrell [-@Bailey1995]. A notable aspect of Bailey and Gatrell was that the book was accompanied by a software application to implement the techniques it discussed. I started using this book as a graduate student around 1998, but even then the limitations of the software that accompanied the book were apparent - in particular the absence of updates or a central repository for code (the book had a sleeve to store a $3\frac{1}{2}$ floppy disk to install the software). Despite the regrettable obsolescence of the software, the book provided then, and still does, a very accessible yet rigorous treatment of many topics of interest in spatial statistics. Bailey and Gatrell's book was, I believe, the first attempt to bridge, on the one hand, the need to teach mid- and upper-level university courses in spatial statistics, and on the other, the challenges of doing so with the very specialized texts on this topic that existed at the time, including the excellent but demanding *Spatial Econometrics* [@Anselin1988], *Advanced Spatial Statistics* [@Griffith1988], *Spatial Data Analysis in the Social and Environmental Sciences* [@Haining1990], not to mention *Statistics for Spatial Data* [@Cressie1993].

More recently, as Bailey and Gatrell aged, my book of choice for teaching spatial statistics became O'Sullivan and Unwin's *Geographical Information Analysis* [@Osullivan2010]. This book updated a number of topics that were not covered by Bailey and Gatrell. To give one example, much work happened in the mid- to late-1990s with the development of *local forms* of spatial analysis, including pioneering research by Getis and Ord on concentration statistics [@Getis1992], Anselin's Local Indicators of Spatial Association [@Anselin1995], and Brunsdon, Fotheringham, and Charlton's research on geographically weighted regression [@Brunsdon1996]. These and related local forms of spatial analysis have become hugely influential in the intervening years, and are duly covered by O'Sullivan and Unwin in a way that merges well with a course focusing on spatial statistics - although other specialist texts also exist that delve in much more depth into some of these topics [e.g., @Fotheringham1999; and @Lloyd2010local].

These resources, and many more, have proved invaluable for my teaching for the past few years, and I am sure that their influence will be evident in the present book. Other excellent sources are also available, including *Applied Spatial Data Analysis in R* [@Bivand2008], *Spatial Data Analysis in Ecology and Agriculture Using R* [@Plant2012], *An Introduction to R for Spatial Analysis & Mapping* [@Brunsdon2015R], *Spatial Point*

Patterns: Methodology and Applications with R [@Baddeley2015], and *Geocomputation with R* [@Lovelace2019]. This is in addition to other resources available online, such as M. Gismond's Intro to GIS and Spatial Analysis and R. Hijmans's Spatial Data Analysis and Modeling with R.

So, if there are some excellent resources for teaching and learning spatial statistics, why am I moved to unleash on the world yet another text on this topic?

I am convinced that there is richness in variety.

As demand for training in spatial statistics grows, there is potential for different sources to satisfy different needs. Some books are geared towards specialized topics [e.g., point pattern analysis; @Baddeley2015] and cover their subject matter in much more depth than I could in an undergraduate course. For this reason, they are more useful as a reference or a tool for learning for researchers and graduate students. Other books focus more heavily on mapping in `R` than a course on spatial statistics can comfortably accommodate [e.g., @Brunsdon2015R; @Lovelace2019]. And yet other books are geared towards specific disciplines [e.g., ecology and agriculture; @Plant2012]. Bivand et al. [-@Bivand2008] is an excellent reference. At the time of their writing, much work was devoted to issues of spatial data representation. As a consequence, a good portion of their book is concerned with the critical issue of handling spatial data, including data classes and import/export operations which, while essential, happen for most practitioners at a baser level.

My approach can be seen as complementary to some of the texts above.

I have tried to write a text that introduces key concepts of data handling and mapping in `R` as they are needed to learn and practice spatial statistical analysis. This I have tried to do as intuitively as I could. Readers will see that the computational part of the book - everything that usually lives "under the hood", so to speak - is all bare in the open. The code is extensively documented as it is introduced (with extensive repetition for pedagogical purposes). Once that a reader has seen and used some commands, we proceed to introduce more sophisticated computational approaches, which are in turn documented extensively when they first appear. I like to think of this approach as introducing coding by stealth, with a gentle ramp for those students who may not have extensive experience in computer-speak. These computational aspects constitute the "how to" of the book. How to calculate a summary statistic. How to create a plot. How to map a variable. How to estimate a model.

The how to is an essential foundation for then exercising the brainware. By introducing the tools needed to accomplish data analysis tasks in a relatively gentle way, I have been able to concentrate in introducing (again, in what I hope is an intuitive way!) key concepts in spatial statistics. The text is not meant to be used as a reference, although some lectors may find that it works in that way in particular with respect to the implementation of techniques. Rather, the text is more suitable to be read linearly - indeed as a course on the topic of spatial statistics. Readers who have familiarized themselves with the text can possibly find it useful as a reference, but I do not recommend using it as a reference in the first place.

Lastly, the focus of the text is on *applied* spatial statistics. There is, inevitably, a component of math, but I have tried, to the extent of my ability, to make the underlying math as intuitive and accessible as possible. As noted above, there is also an important computational component - in particular, as per the title, using the `R` statistical language. As McElreath [-@Mcelreath2016] notes, in addition to the pedagogical value of teaching statistics using a coding approach, much of statistics has in fact become so computational that coding skills are increasingly indispensable. I tend to agree with this, and there are reasons to believe that one of the strengths of this approach as well is to make statistical work as open, clear, and reproducible as possible [see @Rey2009open].

Plan

My aim with this book is to introduce key concepts and techniques in the statistical analysis of spatial data in an intuitive way. While there are other resources that offer more advanced treatments of every single one of these topics, this book should be appealing to undergraduate students or others who are approaching the topic for the first time.

The book is organized thematically following the canonical approach seen, for instance, in Bailey and Gatrell

[-@Bailey1995], Bivand et al. [-@Bivand2008], and O'Sullivan and Unwin [-@Osullivan2010]. This approach is to conceptualize data by their unit of support. Accordingly, data are seen as being represented by:

1. Discrete processes in space (e.g., points and events).
2. Aggregations into zones for statistical purposes (e.g. demographic variables into census areas).
3. As discrete measurements in space of an underlying continuous process (e.g. weather stations monitoring temperature)

The book is organized in such a way that each chapter covers a topic that builds on previous material. All chapters, starting with Chapter 3, are followed by an activity.

I have used the materials presented in this texts (in a variety of incarnations) for teaching spatial data analysis in different settings. Primarily, these notes have been used in the course **GEOG 4GA3** *Applied Spatial Statistics* at McMaster University. This course is a full (Canadian) academic term, which typically means 13 weeks of classes. The course is organized as a 2-hour-per-week class, with a GIS-lab component which uses a complementary set of notes. For this reason, each chapter is designed to cover very approximately the material that I am used to cover in a 50 minutes lecture in a traditional classroom-lecturing setting. In this case, the activities that accompany each chapter could be assigned as homework, optional materials, or as lab materials. For instructors who do not have a lab component, the activities could easily be adapted as lab exercises.

More recently, I have experimented with delivery of contents in a flipped classroom format (see here for a discussion of flipped classrooms).

Briefly, a flipped classroom changes the role of the instructor and the delivery of contents. In a flipped classroom, the instructor minimizes lecturing time, opting instead for offering study materials in advance (often the materials are online and may have an interactive component). This frees the instructor from the tyranny of lecturing, so that in-class time can be dedicated instead to hands-on activities. The instructor is no longer a magical source of wisdom, but rather a partner in the learning process. Under this scenario, students are responsible for reading the chapter or chapters required in advance to a class. The class then is dedicated to the activity that follows the chapter, with students working individually or in small groups in the activity. I have broken a 50-minutes session of this type as follows: 10 minutes for a short mini-lecture and to discuss any questions about the preceding reading/study materials, followed by 30 minutes to complete the activity; during this time I engage individually or in small groups with the students as they work; and before the end of the 50-minutes session a 10 minute recap, where I summarize the key aspects of the lesson, clearly identify the threshold concepts covered, and indicate how this relates to the next lesson. Increasingly I see this format as a form of apprenticeship, where the students learn by doing, and see links (which I have yet to explore) to experiential learning.

In addition to the two formats above (traditional classroom-lecture and flipped classroom), I have also used portions of these notes to teach short courses in different places, including at Universidade de Sao Paulo in Brazil, the University of Western Australia, at the Gran Sasso Scientific Institute in Italy, and Universidad Politecnica de Madrid, in Spain, among other places. The materials can, with only relatively minor modifications, be used in this way.

As I continue to work on these notes, I hope to be able to add optional (or bonus) chapters, that could be used 1) to extend a course on spatial statistics beyond the 13 week horizon of the Canadian term, and/or 2) to offer more advanced material to interested readers see here for an example on spatial filtering.

Audience

The notes were designed for a course in geography, but in fact, could be easily adjusted for an audience of earth scientists, environmental scientists, econometricians, planners, or students in other disciplines who have an interest in and work with georeferenced datasets. The prerequisites are an introductory college/university level course on multivariate statistics, ideally covering the fundamentals of probability, hypothesis testing, and multivariate linear regression analysis.

Requisites

To fully benefit from this text, up-to-date copies of R and RStudio are highly recommended. Many examples in the text use datasets that have been packaged for convenience as an `R` package. To install the package (`geog4ga3`) use the following command(which requires **devtools**):

```
library(devtools)
devtools::install_github("paezha/Spatial-Statistics-Course", subdir = "geog4ga3")
```

The source files for the chapters and activities can be obtained from the following GitHub repository:

https://github.com/paezha/paezha.github.io/tree/master/applied_spatial_statistics

Words of Appreciation

I would like to express my gratitude to the Paul R. MacPherson Institute for Leadership, Innovation and Excellence in Teaching. The Institute supported, through its Student Partners program, my work with some amazing student partners. As part of this program, I worked with Mr. Rajveer Ubhi in the Fall of 2018 and Winter of 2019 organizing all the materials for the text, documenting the code, and ensuring that it satisfied student needs. I also had the opportunity to work with Ms. Megan Coad and Ms. Alexis Polidoro in the Fall of 2019 and Winter of 2020. As former students of the course, Ms. Coad and Polidoro helped to develop a set of mini-lectures to accompany the materials, continued to document the code, and tested the activities. In the Winter 2020 they also accompanied me in the classroom to work directly with new students. Dr. Anastasios Dardas helped develop illustrative applications that helped us understand the value of interactivity in delivering many of the contents.

Working with these wonderful individuals has been a pleasure, and I am grateful for their contributions to this effort.

Versioning

These notes were developed using the following version of `R`:

```
##                _
## platform       x86_64-w64-mingw32
## arch           x86_64
## os             mingw32
## system         x86_64, mingw32
## status
## major          4
## minor          0.3
## year           2020
## month          10
## day            10
## svn rev        79318
## language       R
## version.string R version 4.0.3 (2020-10-10)
## nickname       Bunny-Wunnies Freak Out
```