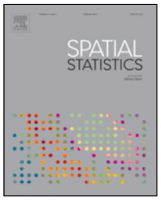
Spatial Statistics 42 (2021) 100466

Contents lists available at ScienceDirect

Spatial Statistics

journal homepage: www.elsevier.com/locate/spasta



Introduction to the special issue *Towards Spatial Data Science*

Jorge Mateu [a,*], Alfred Stein [b]

[a] *Department of Mathematics, University Jaume I, Campus Riu Sec, E-12071 Castellón, Spain*
[b] *Faculty of Geo-information Science and Earth Observation (ITC), University of Twente, The Netherlands*

ARTICLE INFO

Article history:
Received 9 July 2020
Accepted 10 July 2020
Available online 13 August 2020

Keywords:
Data Science
Disease mapping
Environmental applications
Geostatistics
Machine learning
Spatial and spatio temporal point patterns

ABSTRACT

This is the editorial letter for the Special Issue dedicated to the conference Spatial Statistics 2019 *Towards Spatial Data Science* held in Sitges (Spain) from July 10 to 13, 2019. This fifth international conference on Spatial Statistics was run under the theme *Towards Spatial Data Science* with the aim to honour the emerging field of Data Science with a focus on spatial and spatio-temporal methods and real-data problems. The conference was home for more than 250 delegates from about 50 countries worldwide, and the conference included four workshops, more than 100 oral presentations, and nine widely recognised keynote speakers.

This special issue summarises a selection of the main contributions presented at this workshop, related to spatial and spatio-temporal point processes, geostatistics, machine learning and neural networks, disease mapping, extremes and information and complexity. The methodology is embedded in a large number of applications from the environment, traffic and crime, health and socio-economical problems that are the true motivation and justification of this new field.

1. Introduction

In the past two decades we have seen how the field of spatial statistics has found a central and innovative place in the area of Statistics, Mathematics and Computer Science. Indeed, the appropriate combination and complement of theoretical and methodological developments, together with

* Corresponding editor.
E-mail addresses: mateu@uji.es (J. Mateu), a.stein@utwente.nl (A. Stein).

https://doi.org/10.1016/j.spasta.2020.100466
2211-6753/

applications and computational algorithms are at the very core of statistics in general, and spatial statistics in particular. During the last years the inclusion of time has become expanded into the field as well. Spatial and spatio-temporal modelling has become a most interesting and challenging research area in the natural and social sciences. This has been largely fuelled by the increased availability of inexpensive, high-speed computing. Such availability has enabled the collection of large spatial and spatio-temporal datasets, it has facilitated the widespread usage of sophisticated geographic information systems (*GIS*) software to create attractive displays, and it has endowed the ability to investigate challenging, evermore appropriate and realistic models. The relevant literature is growing fast and along directions that range from theoretical contributions through methodological developments to real world applications. Spatio-temporal systems modelling involves the synthesis of a rich interdisciplinary body of knowledge for which it is necessary to establish a solid theoretical foundation and a science-based methodology with both researchers and practitioners in mind. In the end the impact of this research matters most.

In this context, the fifth edition of the biannual conference Spatial Statistics, held in Sitges (Spain) from July 10 to 13, 2019, was run under the theme 'Towards Spatial Data Science'. The conference brought together leading scientists in the field of spatial statistics that presented, debated and discussed how to make statistically sound decisions and judgements in the domain of spatial data science. The conference was home for more than 250 delegates from about 50 countries worldwide, and there were more than 100 oral presentations, and 9 widely recognised keynote speakers.

The selected theme aimed to honour the now topical field of Data Science with a focus on spatial and spatio-temporal techniques and real-data problems. In this context, the conference welcomed methods related to space–time statistics, new spatial data sources (e.g. social media, Google, citizen science, crowd source maps), stochastic geometry (tessellations, point processes, random sets), causal statistical modelling, trajectory/movement modelling, predictive modelling, spatial data quality and uncertainty, and statistical learning. The application area covered many of the domains one can encounter under the umbrella of Data Science. To name a few, at the conference we had the opportunity to address issues in image analyses, traffic and transport, global change, ecology, epidemiology, hazards, disasters and risks, risk mapping, crime and poverty mapping, geohealth and global health, and spatial econometrics. Methods and applications were issued in the new context of spatial data science.

Despite the widely used title of spatial data science (*SDS*), it may sometimes be unclear what we evoke under this title. *SDS* is the practice of distilling insight from spatial data using algorithms and analytical techniques. *SDS* can be seen as a sub-field of Data Science that focuses on the unique characteristics of spatial data, moving beyond simply looking at where events occur to understand why they occur where they occur. *SDS* treats location, distance and spatial interactions as core aspects of the data. It uses specialised methods and *GIS* software facilities to handle, analyse and visualise spatial and spatio-temporal data. Such software allows one to make the most out of the data that describe natural or human features and actions. Most importantly, it also allows one to include the relevant context, e.g. given by topography, political boundaries, urban systems, weather and climate patterns, road networks, shipping and logistics systems and further demographics. In this context the processes occur that are addressed in *SDS* and that were central in our conference.

Within a *GIS* environment, spatial and spatio-temporal data typically are either raster or vector based. Both ways allow one to describe the space, or the space–time, and represent features. Despite recent developments, however, they work rather differently. Vector features often have descriptive metadata — the name of a road, say, or the population of a state. These extra, non-spatial metadata of a feature are usually called "attributes", and are often represented in an "attribute table". In *SDS*, scientists will combine the spatial dimensions (coordinates — for points, or coordinate arrays — for lines and polygons) with non-spatial dimensions in their analysis. GeoJSON and .shp files commonly contain vector data.

Spatial data scientists try to extract relevant information from these datasets, e.g. to better understand the system or phenomenon that they are studying. Some incredible (and often free) software tools make this possible. Most programming languages like Python, R and Javascript have amazing spatial analysis libraries like geopandas and turf.js, and desktop programs like QGIS make visualising and analysing spatial data accessible to less technical people. There are also powerful

online tools like Mapbox, Carto and Google BigQuery to help with these analysis and visualisation challenges. JavaScript libraries like Leaflet and Mapbox GL JS enable web developers to create interactive maps in the browser. By analysing spatial data with this specialised statistical toolkit, spatial data scientists are able to better understand spatial relationships, and, possibly, work out why things happen where, and predict where things will happen next. So far, methods and software were typically addressing medium sized datasets. A major challenge in the years to come will be to make the software ready for the methods to allow statistically sound decisions given the availability of very large databases.

In light of the above considerations, the articles of this special issue have been carefully selected to present a variety of conceptual frameworks, powerful methods and comprehensive techniques that address a number of interesting problems in environmental sciences (such as earthquakes or forest fires), traffic accidents in cities, health problems (mainly disease mapping), or other socio-economical issues. In particular, the selected papers present contributions related to spatial and spatio-temporal point patterns on the Euclidean space and on networks. We also have contributions having a more geostatistical flavour. The field of machine learning, neural networks and data science in general is also touched in several contributions. There is a contribution on spatial extremes, and several interesting ones in the field of disease mapping. We highlight two more contributions, one on spatial data imputation and measurement error, and a review in information and complexity analysis for spatial data, a topic which will be certainly crucial in the next coming years.

At this place, we like to express our gratitude to all reviewers that collaborated in the edition of this special issue. We are specially grateful to Elsevier for supporting this conference and this special issue of Spatial Statistics. We are confident that these contributions will further enhance the current interest in statistical methods in a spatial and a spatio-temporal framework within the emerging context of Data Science.