Spatiotemporal data science: theoretical advances and applications

Federico Amato[1] · Luigi Lombardo[2] · Marj Tonini[3] · Antonino Marvuglia[4] · Daniela Castro-Camilo[5] ·
Fabian Guignard[6]

1 Introduction

In 2009, a book dedicated to the Turing Award winner Jim Gray introduced the idea of a *fourth paradigm of science* (Hey et al. 2009). The concept behind this new paradigm was based on a provocation: the existing paradigms, i.e. those of experimental science, theoretical science, and computational science, were, on the one hand, unsuitable for analyzing the increasing amounts of available data, while on the other hand, they were contributing to the generation of such masses of data. In this context, data is measured by instruments or generated by simulations before being processed by software and the resulting

✉ Federico Amato
federico.amato@epfl.ch

Luigi Lombardo
l.lombardo@utwente.nl

Marj Tonini
marj.tonini@unil.ch

Antonino Marvuglia
antonino.marvuglia@list.lu

Daniela Castro-Camilo
daniela.castrocamilo@glasgow.ac.uk

Fabian Guignard
fabian.guignard@stat.unibe.ch

1   Swiss Data Science Centre, École Polytechnique Fédérale de Lausanne (EPFL) and Eidgenössische Technische Hochschule Zurich (ETH), Zurich, Switzerland

2   Faculty of Geo-Information Science and Earth Observation (ITC), University of Twente, Enschede, The Netherlands

3   Institute of Earth Surface Dynamics, University of Lausanne, Ecublens, Switzerland

4   Luxembourg Institute of Science and Technology (LIST), Esch-sur-Alzette, Luxembourg

5   School of Mathematics and Statistics, University of Glasgow, Glasgow, UK

6   Institute of Mathematical Statistics and Actuarial Science, University of Bern, Bern, Switzerland

information or knowledge being saved in computers. As a result, scientists do not have access to their data until much later in the process. This data-intensive science demands the definition of new techniques and technologies that differ from those utilized in earlier scientific paradigms. Therefore, a new paradigm is required: the one of data science (Kotu and Deshpande 2019). In a broader sense, Data Science is a discipline that combines domain expertise, computer science skills, mathematical and statistical algorithms to transform data into actionable knowledge allowing to support and validate decisions as well as performing predictions.

While this new paradigm has found widespread application in a variety of scientific domains, progresses in geo-environmental sciences have been sluggish and have had a limited impact on our understanding of environmental, climatic, and social processes. This situation is exacerbated given the massive amount of data generated by earth-observing systems, in-situ observations (including *crowd-sourced data*), and climate-related models. To turn these data into valuable knowledge, scientists will have to cope with the so-called five V's of big data: velocity, volume, value, variety and veracity. They will, however, also confront challenges due to the unique characteristics of geo-environmental data, the most evident one being linked to their spatial nature. In 1992, Carl Franklin estimated that about 80% of data included a spatial component (Franklin and Hane 1992). It can be imagined that this proportion has increased over time, as more data have been collected by satellites, mobile devices, and internet of things (IoT) tools. Furthermore, most of these data include a temporal dimension, often inextricably linked to the spatial one by the underlying physical process that the data represent. The analysis of these spatiotemporal data is rendered more complex by other compelling issues involving, among others: (i) data heterogeneity, due to different sources or multiple spatiotemporal scales; (ii) collection biases, due to clustered observation networks, the presence of under-sampled/oversampled regions or short observational records; (iii) complex spatiotemporal dependencies,

including lagged and long-distance relationships between variables. In addition, most of the common data analysis approaches are not designed for autocorrelated, heterogeneous, and non-linear socio-environmental data. For all of these reasons, mining geo-environmental data sets raises difficulties that are rarely addressed in other domains.

We believe that spatiotemporal data science will have a crucial role in addressing many of the great environmental and social challenges of the XXI century and of the other centuries to come. In this context, this special issue contributes to the open discussion on the methodological advances needed to analyze complex spatiotemporal processes. At the same time, several applied case studies highlight the potential of spatiotemporal data science in giving reliable solutions to real-world problems. Papers in this issue include case studies from hydrology and hydromorphology (Budiman et al. 2021; Wang et al. 2021; Rolim et al. 2021; Tang et al. 2021; Tian et al. 2021; Sottile et al. 2021; Han and Morrison 2021), to geology (Giaccone et al. 2021), geomechanics (Li et al. 2021; Luo et al. 2021; Lombardo and Tanyas 2021; Aguilera et al. 2022; D'Angelo et al. 2022; Bryce et al. 2022; Grimm et al. 2022), atmospheric phenomena and renewable energy (La Fata et al. 2022; Amato et al. 2022; Kajbaf et al. 2022), pathogenic viruses and associated diseases (Niraula et al. 2022; Temple et al. 2022).

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Aguilera Q, Lombardo L, Tanyas H, Lipani A (2022) On the prediction of landslide occurrences and sizesvia hierarchical neural networks. Stoch Environ Res Risk Assess. https://doi.org/10.1007/s00477-022-02215-0

Al Kajbaf A, Bensi M, Brubaker KL (2022) Temporal downscaling of precipitation from climate model projections using machine learning. Stoch Environ Res Risk Assess. https://doi.org/10.1007/s00477-022-02259-2

Amato F, Guignard F, Walch A, Mohajeri N, Scartezzini J-L, Kanevski M (2022) Spatio-temporal estimation of wind speed and wind power using extreme learning machines: predictions, uncertainty and technical potential. Stoch Environ Res Risk Assess. https://doi.org/10.1007/s00477-022-02219-w

Bryce E, Lombardo L, van Westen C, Tanyas H, Castro-Camilo D (2022) Unified landslide hazard assessment using hurdle models: a case study in the island of dominica. Stoch Environ Res Risk Assess. https://doi.org/10.1007/s00477-022-02239-6

Budiman JS, Al-Amri NS, Chaabani A, Elfeki AM (2021) Geostatistical based framework for spatial modeling of groundwater level during dry and wet seasons in an arid region: a case study at hadat ash-sham experimental station, saudi arabia. Stoch Environ Res Risk Assess. https://doi.org/10.1007/s00477-021-01971-9

D'Angelo N, Di Benedetto A, Adelfio G, D'Alessandro A, Chiodi M (2022) A new picking algorithm based on the variance piecewise constant models. Stoch Environ Res Risk Assess. https://doi.org/10.1007/s00477-022-02218-x

Franklin Carl, Hane Paula (1992) An introduction to geographic information systems: linking maps to databases [and] maps for the rest of us: Affordable and fun. Database 15(2):12–15

Giaccone E, Oriani F, Tonini M, Lambiel C, Mariéthoz G (2021) Using data-driven algorithms for semi-automated geomorphological mapping. Stoch Environ Res Risk Assess. https://doi.org/10.1007/s00477-021-02062-5

Grimm C, Hainzl S, Käser M, Küchenhoff H (2022) Solving three major biases of the etas model to improve forecasts of the 2019 ridgecrest sequence. Stoch Environ Res Risk Assess. https://doi.org/10.1007/s00477-022-02221-2

Han H, Morrison RR (2021) Data-driven approaches for runoff prediction using distributed data. Stoch Environ Res Risk Assess. https://doi.org/10.1007/s00477-021-01993-3

Hey AJG, Tansley S, Tolle KM et al (2009) The fourth paradigm: data-intensive scientific discovery, vol 1. Microsoft research Redmond, WA

Kotu V, Deshpande B (2019) Data Science. Concepts and Practice, 2nd edn. Morgan Kaufmann, US

Fata AL, Amato F, Bernardi M, D'Andrea M, Procopio R, Fiori E (2022) Horizontal grid spacing comparison among random forest algorithms to nowcast cloud-to-ground lightning occurrence. Stoch Environ Res Risk Assess. https://doi.org/10.1007/s00477-022-02222-1

Li W, Fang Z, Wang Y (2021) Stacking ensemble of deep learning methods for landslide susceptibility mapping in the three gorges reservoir area, china. Stoch Environ Res Risk Assess. https://doi.org/10.1007/s00477-021-02032-x

Lombardo L, Tanyas H (2021) From scenario-based seismic hazard to scenario-based landslide hazard: fast-forwarding to the future via statistical simulations. Stoch Environ Res Risk Assess. https://doi.org/10.1007/s00477-021-02020-1

Luo L, Lombardo L, van Westen C, Pei X, Huang R (2021) From scenario-based seismic hazard to scenario-based landslide hazard: rewinding to the past via statistical simulations. Stoch Environ Res Risk Assess. https://doi.org/10.1007/s00477-020-01959-x

Niraula P, Mateu J, Chaudhuri S (2022) A bayesian machine learning approach for spatio-temporal prediction of covid-19 cases. Stoch Environ Res Risk Assess. https://doi.org/10.1007/s00477-021-02168-w

Rolim LZR, da Silva SMO, de Assis de Souza Filho F (2021) Analysis of precipitation dynamics at different timescales based on entropy theory: an application to the state of ceará, brazil. Stoch Environ Res Risk Assess. https://doi.org/10.1007/s00477-021-02112-y

Sottile G, Francipane A, Adelfio G, Noto LV (2021) A pca-based clustering algorithm for the identification of stratiform and convective precipitation at the event scale: an application to the sub-hourly precipitation of sicily, italy. Stoch Environ Res Risk Assess. https://doi.org/10.1007/s00477-021-02028-7

Tang X, Li J, Liu W, Huafei Y, Wang F (2021) A method to increase the number of positive samples for machine learning-based urban waterlogging susceptibility assessments. Stoch Environ Res Risk Assess. https://doi.org/10.1007/s00477-021-02035-8

Temple SD, Manore CA, Kaufeld KA (2022) Bayesian time-varying occupancy model for west Nile virus in Ontario, Canada. Stoch Environ Res Risk Assess. https://doi.org/10.1007/s00477-022-02257-4

Tian D, He X, Srivastava P, Kalin L (2021) A hybrid framework for forecasting monthly reservoir inflow based on machine learning techniques with dynamic climate forecasts, satellite-based data, and climate phenomenon information. Stoch Environ Res Risk Assess. https://doi.org/10.1007/s00477-021-02023-y

Wang N, Cheng W, Lombardo L, Xiong J, Guo L (2021) Statistical spatiotemporal analysis of hydro-morphological processes in china during 1950–2015. Stoch Environ Res Risk Assess. https://doi.org/10.1007/s00477-021-02007-y