Environmental Modelling & Software 106 (2018) 1–3

Contents lists available at ScienceDirect

Environmental Modelling & Software

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

Preface

Preface to the thematic issue on Environmental Data Science. Applications to air quality and water cycle

With new and cheaper methods for remote sensing, sensor deployment, and other methods of observation, the volume of environmental data continues to grow at an unprecedented rate – so much so that there is now often a gap and disconnect between data and decision making. Data Science combines data analysis with data processing methods and domain expertise to transform data into understandable and actionable knowledge relevant for decision making. With deep roots in computer science, artificial intelligence, and statistics, and with many existing applications in the business domain within the last decade, Data Science methods have also emerged in the environmental sciences to the extent that the corpus of data, methods, algorithms, and techniques now available make up a multidisciplinary field we refer to here as Environmental Data Science.

The Data Mining Techniques for Environmental Sciences (DMTES) workshop series started in 2006 as a special session and associated workshop of the biennial meeting of the International Environmental Modelling and Software Society (iEMSs). DMTES was as a pioneer initiative to provide a forum for discussion around technological and methodological advances related to data mining and its applications within the environmental sciences. For more than a decade, DMTES has been providing a valuable opportunity for scientific exchange between data scientists and the environmental community. The DMTES workshops and sessions originally set out to disseminate research in Data Mining (also known as knowledge discovery and data mining (KDD) but evolving in more recent years toward the term Data Science) to the environmental science community, aiming to inspire novel and useful applications of Data Mining techniques for real environmental problem solving and decision-making. DMTES also emerged as a valuable opportunity for the KDD/Data Science community to have closer contact with real environmental problems, managers, and decision makers for a better understanding of their perception of data-intensive procedures and, most importantly, their needs and demands. Thus, data scientists have received relevant inputs to drive further research on data-driven methods. Over the years of the DMTES workshop series, the research of both environmental scientists and data scientists has been enriched by multidisciplinary contacts and fruitful discussions held during the meetings.

After ten uninterrupted years of biennial meetings, in 2016 a stable community of data miners/data scientists and environmental scientists had grown around DMTES, with participants from all over the world. The papers included in this thematic issue reflect this world-wide dimension, and, show only partially the results of the long interactions over the past decade. The first call for this thematic issue was launched at the end of the 2016 iEMSs meeting in Toulouse, France, when its usual session and workshop on DMTEs were held.

This thematic issue provides a view on how Data Science can contribute to better understanding, better forecasting, and better managing environmental systems with special focus on two critical environmental domains: air and water, among the most important pieces of environment, addressing challenges linked to pollution and the quality of air and water, and the provisioning and managing of water resources.

Air quality is a major environmental issue in many areas of the world that fundamentally affects human health and can adversely affect ecosystems and climate. Air pollution is one of the top environmental cause of mortality worldwide, with millions of people losing their lives each year as a result of exposure to household and ambient air pollution. Flora and fauna within the natural environment are also seriously affected by air pollution and related consequences driven by human activity, including climate change.

The importance of sustainable water management is well known and is critical to guarantee healthy and resilient ecosystems able to provide the services needed to sustain human well-being and society's development by minimizing the impact on the natural environment. According to European Environment Agency, although water is one of the world's most abundant substances, only about 1% of the water on Earth is suitable for human usage, and in 2008 humans were only able to exploit about 0.08% of all the world's water. Despite efforts made in the past two decades and the improvements achieved, large numbers of the world's population still do not have access to clean, safe water. This remains as a critical challenge in many areas of the world, and will get even worse with the population rise.

Challenges related to water and air, as well as those associated with other environmental media (e.g., soils and biota) inherently span multiple scientific domains and environmental media, with links to climate variability, land use change, population growth, and economic development. The technological development requisite for collecting the volume and resolution of data required to study these phenomena has matured, but the diversity and volume of data required to holistically address big problems renders many classical data analysis methods insufficient to cope with the volume, velocity, and diversity of information sources providing evidence under the variety of formats that must be assimilated and analyzed to extract knowledge to support higher level decision making. Many investigators, including those who submitted papers to this thematic issue, are already investigating how Data Science

https://doi.org/10.1016/j.envsoft.2018.03.020

can address this deficiency.

The issue begins with a paper from the guest co-editors providing an overview of the Environmental Data Science field itself. Then, two papers on air pollution and six papers related to the water cycle were selected for this issue, providing a perspective on how different issues in air pollution and water cycle-related problems can benefit from a Data Science approach with different data structures and analysis methodologies.

The paper "Environmental Data Science," written by the guest co-editors, describes the origins and history of Data Science and discusses the new profile of a data scientist. It describes the relevance of promoting multidisciplinary teams including both data scientists and environmental scientists and the importance of Data Science to extract decisional knowledge from data, providing added value to bridge the gap between data processing and decision-making layers in a world concerned with decisions of increasing complexity. Furthermore, it reviews recent applications of Data Science in the Environmental Sciences and articulates current challenges, potential solutions, trends, and a vision for how Environmental Data Science may advance over the next several years. The guest co-editors hope that this work will spark a discussion within the Environmental Modelling & Software community around how and where Data Science fits within the broad spectrum of methods and techniques employed for solving environmental challenges. It is also expected that this thematic issue promotes further development from both research and applications in Environmental Data Science field.

The paper "Modelling background air pollution exposure in urban environments: implications for epidemiological research," written in collaboration among the European Commission, Portugal, and Spain, shows how Data Science approaches can provide a better understanding of the baseline levels of exposure to air pollutants in multiple urban areas in Spain. In this paper, the original essence of clustering methods as categorization tools is used to identify the threshold levels of baseline air pollution that citizens are permanently exposed to.

The paper "Environmental data stream mining through a case-based stochastic learning approach" is also a result of an international collaboration between Spain and Mexico. It focuses on a data streaming approach for air quality indicators in a single city in Mexico. Case-based reasoning methods were used to provide dynamic assessment of air pollution under a non-supervised approach.

The following six papers highlight different Data Science methodologies for providing insights into environmental issues related to the water cycle, from rainfall to drinking water distribution networks to river ecosystems. The paper "Data-driven rainfall/runoff modelling based on a neuro-fuzzy inference system," contributed by an Italian research team, provides an interesting hybrid methodology, combining artificial intelligence and statistical methods to build predictive models of rainfall/runoff in minor catchments in central Italy. Using sensor data, the paper encompasses principal components analysis with fuzzy artificial neural networks to better cope with changing characteristics of catchments.

The paper "Imbalanced classification techniques for monsoon forecasting based on a new climatic time series," developed by a Spanish research team, focuses on forecasting the Western North Pacific Summer Monsoon based on climate index time series. The authors tackled the frequent issue of modelling rare events, i.e., dealing with imbalanced data sets in classification problems. The paper is centered on different classifier methods including models based on trees, models based on rules, black box models, and ensemble techniques.

The paper "Model-based analysis of the relationship between macroinvertebrate traits and environmental river conditions," developed by an international team from Belgium, Philippines, and Australia, assessed water quality in rivers and the capacity to predict it through macroinvertebrate traits. Special preprocessing of qualitative, ordinal variables is tackled, and multivalued qualitative variables are treated through fuzzy encoding. Negative binomial regression and generalized linear models considering linear and nonlinear scenarios are discussed.

The paper "Quantitatively scoring behavior from video-recorded, long-lasting fish trajectories," developed by a research team in Catalonia, analyzes fish behavior by using one of the alternative data sources exploited today in data science: videos. Thus, important efforts on preprocessing the videos are invested through innovative feature extraction methods to synthesize videos into relevant metrics for scoring fish behavior. The work provides a methodology to analyze dynamics of fish populations through video-tracking in scenarios far from optimal as an effective alternative to traditional and costly experimental trials. These results may be interesting for living resources managers and for ecologists.

The paper "Hybrid SOM + k-means clustering to improve planning, operation and management in water distribution systems," developed by a collaboration between investigators in Mexico, Brazil and Spain, provides a Data Science methodology to support management of water distribution systems as the last step of the water cycle. Three particular case-studies in the field are presented to show the versatility of the methods: (1) a classification of Brazilian cities in terms of their water utilities; (2) district metered area creation to improve pressure control; and (3) transient pressure signal analysis to identify burst pipes. The work presents a hybrid Data Science methodology encompassing self-organizing maps with k-means algorithms to discover clusters in water distribution system (WDS) oriented to decision-making support.

The last paper of this issue "Transforming data into knowledge for improved wastewater treatment operation: A critical review of techniques," is an international collaboration between investigators in Spain, USA, Switzerland and Sweden, and provides an overview of computer-based techniques for data analysis to improve operation of wastewater treatment plants. A comprehensive review of peer-reviewed papers shows that European researchers have led academic computer-based method development during the last two decades and highlights that there is still hard work needed to achieve integral decision support systems in water management that smoothly integrate several types of knowledge and different methods of reasoning.

The selection of papers for this issue was done with a rigorous blind peer-review process and high rate of rejection. The issue originally received 29 submissions, and the 8 papers most favorably evaluated by reviewers were selected. The reviewing process involved more than 170 reviewers. The selected papers provide a nice overview of research conducted in multiple countries, including international collaborations and some collaboration between academia and corporations. We believe that the set of selected papers provides a useful perspective of how Data Science tackles nonlinear, spatio-temporal environmental phenomena using a variety of data sources (numeric variables, qualitative, ordinal, time series, videos, smart sensor data, etc.) for both descriptive and predictive processes. The included papers also use techniques ranging from classical statistical methods to innovative preprocessing methods or hybrid machine learning methods to extract knowledge from data.

The Guest Editors wish to thank the authors for contributing their original papers and for their patience throughout the peer review and publication process. We would also like to express gratitude to the numerous reviewers who contributed in constructively reviewing and improving the quality of the contributions within this issue.

Karina Gibert, fiEMSs, Guest Editor*
Dep. Statistics and Operations Research; Knowledge Engineering and Machine Learning Group at Intelligent Data Science and Artificial Intelligence Research Center; Research Institute on Science and Technology for Sustainability; Universitat Politècnica de Catalunya-BarcelonaTech, Spain

Jeffery S. Horsburgh, Guest Editor
Department of Civil and Environmental Engineering and Utah Water Research Laboratory, Utah State University, Logan, UT, 84322-8200 USA

Ioannis N. Athanasiadis, fiEMSs, Guest Editor
Information Technology Group, Wageningen University, The Netherlands

Geoff Holmes, Guest Editor
Department of Computer Science, University of Waikato, New Zealand

* Corresponding author.
E-mail address: karina.gibert@upc.edu (K. Gibert).

22 March 2018
Available online 12 April 2018