Synthesis of Statistics, Data Mining and Environmental Sciences in Pursuit of Knowledge Discovery

Yulia R. Gel (University of Texas at Dallas),
L. Leticia Ramirez Ramirez (Centro de Investigación en Matemáticas),
Vyacheslav Lyubchich (University of Maryland Center for Environmental Science)

29 Oct – 3 Nov, 2017

1 Overview of the Field and Recent Developments

The technology advances have revolutionized our capacity for collecting data, altering the way in which scientists sense and analyze the available information, in virtually all research areas. Due to this, the problems and opportunities originated from the "Big Data" paradigm, attract the attention of almost all scientists, from those working in finance and business analytics, to biology-omics studies to earth and climate research.

While Big Data holds the promises for discovering new patters in the underlying phenomena, there are several challenges associated to handling and analyzing massive data. Some of these problems, like scalability and storage bottleneck, are addressed by computer tools and architectures. Some others, are associated to the use and analysis of this massive information in applied scientific areas. This latter scenario calls for interdisciplinary discussion to take advantage of previous experiences, developed and developing methods and theoretic results. The primary goal of this workshop was to bridge together disciplines and methodologies that typically never interact, but that are intrinsically close. This workshop, held at Casa Matematica Oaxaca, highlighted four themes: climate, epidemiology, social media, climate change and risk modeling, analyzed from the perspective of machine learning, nonparametric and Bayesian statistics.

The speakers presented a variety of cutting-edge methods in statistics and machine learning methods to tackle big data in a temporal and spatio-temporal context, including:

- Multi-model ensembles;
- Bayesian statistics;
- Nonparametric statistics;
- Machine learning;
- Information theory;
- Dimension reduction.

2 Presentation Highlights

Monday, October 30:

The morning presentations focused on the topic of *urban analytics and climate sustainability*. **Katherine Ensor** presented one of her most important projects. She and her team, are developing a data platform (Urban Data Platform) for the area around Houston, Texas with the aim to collect large amount of data across many activities, environmental and health variables. The overall goal is to have a system that helps to understand the activities of the inhabitants and how this, along with some other variables, affects or improves their wellbeing. One of the potential use of this system is quantifying the environmental risk associated to some illness. On this matter, Ensor uses part of the information collected for the platform and presents a new space-time model based on a Gaussian process (that generalized the presented in [4]) to understand and measure the relation of air quality and health problems, such as asthma attacks and cardiac arrests.

Lizzy Warner showed several different projects developed in the Sustainability and Data Sciences Laboratory using mathematical, engineering, and computational methods and tools. Some of these projects study transportation networks, their power grid demand, resilience and recovery. Some other projects are dedicated to evaluate the extreme sensitivity to initial conditions in complex systems, and identification and understanding of extreme events.

The following two talks were part of the session *climate models 1*. **Kyo Lee** introduced the problem of weighting selection in a multi-model ensemble, and the need to propose a method that systematically selects such weights in a computer efficient manner. This last characteristic is key, when handling high resolution climate models that uses massive data and require intensive numerical computations. The proposed model to select the weights is base on an multi-objective optimization problem and it is combined with the Virtual Information Fabric Infrastructure architecture, in order to address the Big Data in climate models.

Singdhansu Chatterjee presented a new statistical method that uses multivariate quantiles, extremes and nonparametric statistics for studying the geometry of data. By representing the information as objects in a p-dimension some concepts can be applied, such as data-depth functions. This, together with resampling methods, are used to score the parameters under candidate models. This methodology has the advantage of remain simple when p grows. This method was illustrated in the problem of monsoons prediction.

The afternoon talks were aggregated by the topic of *health and climate*. **Geoffrey Fairchild** introduced the advantages and opportunities of using real-time social Internet data for modeling and forecasting diseases. He presented three undergoing projects at Los Alamos National Laboratory.

The talk of **Georgiy Bobashev** rotated around synthetic populations, how they are constructed at the RTI, and the huge potential they have to develop Agent-Based models in diverse applications. Using multiple source information, the synthetic populations are recreated to individual level. Using a forecasted synthetic population, then epidemic forecast can also consider future geographical information.

Leticia Ramirez Ramirez showed a proposal for a multi-source model that considers both, official epidemic surveillance reports and searches in Google, in order to predict emerging infectious diseases. She presented a particular model for the case of chikungunya in the Americas. This model incorporates the steps of uncertainty quantification for a compartmental epidemic model, and TRUST algorithm [2], for time-space clustering.

Matthew Dixon addressed the challenge of spatio-temporal models (such as epidemics and percolation) and presented deep learning (DL) as an alternative to predict complex systems. His proposed method incorporates the Bayesian paradigm for uncertainty quantification of spatio-temporal flows. The method was illustrated in the settings of traffic flow.

Tuesday, October 31:

The two morning talks were part of the session *climate models 2*. **Alan Gelfand** discussed studying climate change from another perspective; by considering its velocity. This velocity has been measured, or approximated, by the change over time, of some environmental variables (temperature, plants, animals, precipitation) in terms of its value, area, volume etc. The main contribution of Gelfand is extending the previous definition of velocity to a continuous function, that allows capturing spatial structure in velocities and assessing the velocity change over time.

Robert Lund pointed out the need to detect change points when estimating trends in time series data. He reviewed some of the most important statistical methods and genetic algorithms for change point estimation and presented a new method based on information theory. The selection model associated to this method generalizes the AIC and BIC.

All afternoon talks comprised the session *Space-time climate processes 1* **Vadim Sokolov** discussed the deep learning method from a statistical perspective. This allows to optimize some inference procedures such as parameter tuning. The potential use of this methodology was illustrated with several applications.

Based on machine learning and Markov random fields methods, **Sloan Coats** presented a paloeclimate spatio-temporal model for droughts. These allow to reconstruct, simulate and predict droughts in large areas of the Northern Hemisphere.

In order to describe some ocean phenomenons, such as currents and turbulence, **Adam Sykulski** discussed a new stochastic spatio-temporal model for the movement of surface drifters that are freely moving in the world oceans. The proposed model generalizes [6]. The used information correspond to over 70 million data points.

Wednesday, November 1:

The morning session was named *Space-time climate processes 2* and **Alexander Brenning** addressed some challenges for the analysis of high-dimensional data in remote sensing for environmental research. He also presented statistical analysis approaches developed in Jena for three case studies: 1) landslide modeling for prediction purposes, 2) detection of rock glacier, and 3) crop classification.

Murali Haran presented a proposed approach for modeling non-Gaussian spatial data using spatial generalized linear mixed models [3]. These models are very flexible but face two strong challenges when considering high-dimensional data [1]. One is on the computational demand side and the other is in relation with the results interpretation. Haran presented a method using projections that addresses both of the problems.

Juan Martin Barrios Vargas discussed two different distribution models that are implemented at the Mexican National Commission for the Knowledge and Use of the Biodiversity (CONABIO). The goal of these models is to facilitate the studies and protection of different species and biodiversity, under different climate change scenarios.

The last four talks were part of the session *climate and insurance*. **Robert Beach** addressed the risk management for agricultural production. This research has the goal to update the risk models to consider climate change and draw some alternative climate and policy scenarios for different crop yields.

Alicia Mastretta-Yanes discussed the outcomes and challenges for characterizing, modeling, conserving, and using plant genetic diversity. This diversity represents a protection against the changing environment. She also addressed the data that CONABIO is capturing for different species.

Vyacheslav Lyubchich presented a proposal for modeling agricultural insurance risks using deep learning algorithms based on a weather-based index insurance. The flexible deep learning tools allow to consider complex nonlinear relationship among the variables and improve the quality of yield forecasts compared with alternative techniques. As a result, this approach can improve the risk management in agricultural insurance.

Ola Haug was the last presenter of the session and he addressed the problem of trend analysis for data in varying spatial grid scales. The proposed model estimates the parameters via R-INLA [5] and is implemented for analyzing trends in seasonal means for Europe.

Thursday, November 2:

During morning we had a poster session with the works:

Igor Barahona “Sensory Evaluation of Coffee Beverages through Statistical Methods”.

Eduardo Alvarez Rodríguez “Measures of Causal Influence between Environmental Signals in the Frequency Domain (FD)”.

Irving Gómez “EWMA Control Charts for Monitoring Pollution”.

Sadoth, Sandoval Torres “Modeling Groundwater Vulnerability to Climate Change in Agricultural Areas of Oaxaca”.

The afternoon was filled with social activities and motivating informal discussions.

Friday, November 3:

This day was dedicated to informal discussions.

3 Outcome of the Meeting

The main objective of this workshop was bringing together researchers, who are facing the big data problem with novel proposals. Most of the presentations were on ongoing research and it was an exceptional opportunity to gain better understanding of the modern problems, exchange methodological advancements and ideas, during the presentations and informal discussions.

References

[1] J. Besag, J. York and A. Mollié, Bayesian Image Restoration, with Two Applications in Spatial Statistics.*The Annals of the Institute of Statistical Mathematics*, 43(1) (1991), 1–20.

[2] A. Ciampi; A. Appice and D. Malerba, Discovering trend- based clusters in spatially distributed data streams, *in ECML/PKDD Workshop on Mining Ubiquitous and Social Environments* (2010).

[3] Guan and Haran, A Computationally Efficient Projection-Based Approach for Spatial Generalized Linear Mixed Models, arxiv.org (2017).

[4] K.B. Ensor and L. RaunHeart, Attack Incidence and Air Quality Levels for Houston. *Technical Report*(2010).

[5] H. Rue, S. Martino, and N. Chopin. Approximate Bayesian inference for latent Gaussian models using inte- grated nested Laplace approximations (with discussion). *Journal of the Royal Statistical Society*, Series B, 71(2) (2009):319–392.

[6] A.M. Sykulski, S.C. Olhede, J.M. Lilly and J.J Early (2017). Frequency-Domain Stochastic Modeling of Stationary Bivariate or Complex-Valued Signals. *IEEE Transactions on Signal Processing*. (2017): 1–1.