Big Data Tsunami at the Interface of Statistics, Environmental Sciences and Beyond

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

March 11–13, 2016

1 Overview of the Field and Recent Developments

The rampant growth of digital technologies and information storage have revolutionized the volume, velocity and variety of collected information, leading to the so-called "Big Data" paradigm. In turn, this alters the way in which scientists sense and analyze the available information, and ignites the interest in Big Data phenomenon virtually everywhere, from climate research to omics studies to business analytics.

One of the greatest challenges of massive data (and the one that typically constitutes the primary focus of all Big Data workshops) is an increasing demand for developing innovative computational methodology and more powerful computational tools. What is typically less discussed at many computational workshops on Big Data is the whelm of ground-breaking new interdisciplinary links that emerge from these massive information volumes. The primary goal and the key difference of this workshop from other Big Data events is that it aimed to bridge together disciplines and methodologies that typically never meet and interact at other conferences and scientific gatherings but which are in fact intrinsically close. In particular, the workshop highlighted three tightly woven themes: climate, infectious epidemiology and social media; weather, climate and complex socio-ecological networks, and climate change vulnerability, risk mitigation and adaptation.

The speakers presented a variety of modern statistical and machine learning methods to tackle big data in a spatio-temporal context, including such approaches as:

- causal discovery;
- Bayesian networks;
- dynamic networks and graphs;
- statistical compression;
- statistical downscaling and data assimilation;
- distributional calibration;
- penalized likelihood.

2 Presentation Highlights

Saturday, March 12: The presentations in the morning session focused on the topic of *methods for big data in climate science*. The conference was opened by talks of **Imme Ebert-Uphoff** and **Dorit Hammerling**. The first speaker introduced a new framework for constructing climate networks [9] based on causal relationship, competing with the existing methods based on correlations, mutual information, and phase synchronization. One of the main advantages of the new approach is the ability to discover patterns of climate interactions [2]. The second speaker presented and illustrated a new method for high-performance computing in climate sciences, the multi-resolution approximation (MRA).

Presentations by **Joe Guinness** and **Daniel Griffith** continued the session. Guinness targeted the emerging problem of increasing memory requirements for the data storage, by suggesting a framework to compress the data in a form of statistical model with all the parameters obtained from original data. The decompression would result in a surrogate data set with the same statistical qualities, or a model-generated sample. Griffith closed the session with a talk on spatial correlations and uncertainties associated with remotely sensed data [3], which can be seen as an extension of spatial statistics to the remote sensing work of [1].

The first afternoon session incorporated three presentations on forecasting of infectious diseases, highlighting different transmission pathways of diseases and analysis outcomes. **Teresa Yamana** presented extensions of their work with J. Shaman on probabilistic prediction of seasonal influenza outbreaks that is rooted in ensemble methodology for weather and climate forecasting [10]. **Lilia Leticia Ramirez Ramirez** discussed new results on probabilistic forecasting of influenza, based on initializing epidemic models on networks of contacts [7] with online social media. **Chris McMahan** developed Bayesian hierarchical space-time methods for long-term prediction of zootonic and vector-borne diseases using vaccination data in the conterminous United States [6].

The conference continued with presentations by **Ola Haug** and **Andrew Finley**. Haug discussed the application of climate model outputs in forecasting of insurance claims [8], and stressed the demand for reliable procedures to calibrate the climate model outputs, for each spatial location. In a case study of forest biomass prediction across Alaska, Finley showcased a highly scalable Nearest Neighbor Gaussian Process (NNGP) to provide model-based spatial inference within a hierarchical modeling framework.

The last session of the day featured **Elizabeth Martinez-Gomez** and **Robert Lund** who focused on the big data problems in particular applications. Martinez-Gomez discussed the data volumes produced by modern telescopes and statistical challenges [5] associated with timely analysis of the large amounts of data, and utilization of the information for comprehensive investigations of the Universe. Lund described a novel approach of identifying regime shifts, coupled with a genetic algorithm to more efficiently analyze long climate time series [4].

Sunday, March 13: The final morning was filled with informal discussions.

3 Outcome of the Meeting

The main goal of the workshop was to bring researchers tackling the big data problems at the interface of statistics and the broad field of environmental science. The meeting presented a unique opportunity to exchange the ideas and methodological advancements across the countries (Canada, Mexico, Norway, and USA) and the areas of applications (climate science, astronomy, disease surveillance, remote sensing, and insurance).

References

[1] X.-L. Chen, H-M. Zhao, P.-X. Li, Z.-Y. Yin, Remote sensing image-based analysis of the relationship between urban heat island and land use/cover changes. *Remote Sensing of Environment* **104** (2006), 133–146.

[2] I. Ebert-Uphoff and Y. Deng, Identifying physical interactions from climate data: Challenges and opportunities. *Computing in Science & Engineering* **17**(6) (2015), 27–34.

[3] D. A. Griffith and Y. Chun, Spatial autocorrelation in spatial interactions models: geographic scale and resolution implications for network resilience and vulnerability. *Networks and Spatial Economics* **15**(2) (2015), 337–365.

[4] S. Li, and R. Lund, Multiple changepoint detection via genetic algorithms. *Journal of Climate*, **25**(2) (2012), 674–686.

[5] E. Martnez-Gómez, V. M. Guerrero and F. Estrada, A Multivariate Time Series Analysis for Climate Modeling based on the Solar-Terrestrial Relationship. *Proceedings of the JSM 2009 Section on Physical and Engineering Sciences* (2009).

[6] C. S. McMahan, D. Wang, M. J. Beall, D. D. Bowman, S. E. Little, P. O. Pithua, J. L. Sharp, R. W. Stich, M. J. Yabsley and R. B. Lund, Factors associated with *Anaplasma spp.* seroprevalence among dogs in the United States. *Parasites & Vectors* **9** (2016), 169.

[7] L. L. Ramirez Ramirez, Y. R. Gel, M. E. Thompson, E. de Villa and M. McPherson, A new surveillance and spatio-temporal visualization tool SIMID: SIMulation of Infectious Diseases using random networks and GIS. *Computer Models and Programs in Biomedicine*, **110**(3) (2013), 455–470.

[8] I. Scheel, E. Ferkingstad, A. Frigessi, O. Haug O, M. Hinnerichsen, E. Meze-Hausken, A Bayesian hierarchical model with spatial variable selection: the effect of weather on insurance claims. *Journal of the Royal Statistical Society Series A* **62** (2013) 85–100.

[9] A. A. Tsonis and P. J. Roebber, The architecture of the climate network. *Physica A: Statistical Mechanics and its Applications* **333** (2004), 497–504.

[10] W. Yang, B. J. Cowling, E. H. Y. Lau and J. Shaman, Forecasting influenza epidemics in Hong Kong. *PLOS Computational Biology*, **11**(7) (2015), e1004383, doi:10.1371/journal.pcbi.1004383