RUI ZHU

STKO Lab
Department of Geography
University of California, Santa Barbara
Email: ruizhu@geog.ucsb.edu

Rui Zhu is a PhD candidate at the Department of Geography, University of California, Santa Barbara. His major is Geographic Information Science (GIS) with emphases on spatial statistics and geospatial semantics. His dissertation focuses on revisiting fundamentals in spatiotemporal analysis (e.g., high-order spatial interactions and direction effect) and adopting spatiotemporal thinking into deep learning in order to explore complex spatial structures of geographic information (e.g., remotely sensed images, points of interest in urban cities, and the simulation of cellular automata). The semantics of extracted patterns are further investigated to uncover meaningful knowledge. He is also working on applying spatial statistics to understand the semantics of geospatial ontologies in a data-driven approach; one example is to answer the classic geographic question *what is a mountain and what is a hill*? Furthermore, Zhu has experiences researching household travel behaviors, spatiotemporally enriched knowledge graph, geospatial question answering, urban dynamics and so on.

Before joining UCSB, Zhu received his master degree in Information Sciences from the University of Pittsburgh, where he was involved in multiple research projects that are related to spatial cognition, high performance computing and mobile GIS.

http://www.geog.ucsb.edu/~zhu/

Challenges in Spatial Data Science

The science of data is not new to us. However, it is not until these past several years that we witness a significant emergence of data science in both academia and industry, thanks to the ubiquitous sensors and powerful computational capacities. It is, therefore, unsurprising that geographers start to join the data science troop with goals of applying newly developed technologies and methods to address geospatial problems, as well as contributing our spatiotemporal thinking and knowledge to the community. Artificial intelligence (AI), which is the most popular topic in data science, is a good example. We see a dramatic increase of papers, workshops and proposals discussing geographic artificial intelligence (GeoAI) in the past five years. However, as Dr. Goodchild and many other geographers concerned at AAG 2019, we still lack of "compelling examples" and a "deeper understanding" of how we should appropriately interact with this new trend of science. *Spatial is special*[1], but questions such as what types of spatial information are useful in the big data

era, how to intellectually leverage them into new models, and how to interpret the results thus making appropriate decisions are still open to be answered in spatial data science.

This statement attempts to emphasize some of these challenges and propose my insights and solutions to address them. Before I delve into these challenges, I briefly discuss about my understanding of the history of spatial data science. History is essential as it helps us inspect which stage we are currently at, and reflect on where we should go next.

Spatial Data Science: When GIScience meets Big Data

Around the 1960s, geographers started to extensively store, retrieve, model and visualize geospatial phenomena using computers. Many computational models and statistical methods were developed since then, which significantly contributed to the growth of geographic information science (GIScience). With about 50 years' development, GIScience has become a multidisciplinary field in science, and since the 2010s, research on traditional computational models and spatial statistics tended to retard; instead, we saw a trend of analyzing geospatial phenomena *in a big data fashion*[3], in which numerous new techniques, such as convolutional neural networks, word2vec, and generative adversarial network, were greatly emphasized. This is probably the reason why we are inclined to use "spatial data science" more frequently nowadays compared with traditional "geographic information science." In such a big data era, the problems are not only "geographic" but also "spatial" in a broader sense. Meanwhile, we are now able to collect richer but more noisy and heterogeneous geospatial "data," while traditional spatial statistics and models were mainly designed for sampled but homogeneous "information." This fundamental change affords us opportunities to address geospatial problems using new methods, but also brings new challenges.

Challenges

It is admitted that many existing data science models could be directly applied in a geospatial context, such as using convolutional neural network to classify remotely sensed images. However, geospatial problems are far more complex than just classification, which is a lower-order scientific task. Spatial prediction, for instance, is in significant demands in spatial data science but receives less attention in the new models. On the other hand, geospatial data is fundamentally more complicated than nonspatial data, but most data science models are designed without explicit consideration of spatial heterogeneity and spatial dependency. In addition to the few efforts of simply incorporating distance into existing models, I think it is time for us to revisit the fundamental principles of spatiotemporal data in a new context (i.e., the big data era and AI). For instances, what information make neural networks spatially and temporarily explicit? Is it necessary, and how, to model spatial interactions beyond pairs[5]? What is the role of directional information (anisotropicity) in spatial analysis[4]?

Furthermore, uncertainty analysis is a tradition in GIScience. In fact, most of the developed spatial statistical models are capable of providing uncertainty analysis, as they are all based on classic probabilistic distributions and/or Bayesian theory. In modern data science, however, many popular models (e.g., deep neural network) are designed as *point-based estimation and prediction* (i.e., one

set of weights for all). This works fine for simple cat-and-dog classification, but becomes problematic when studying geographic phenomena that are far more complicated and sensitive to outliers (e.g., weather prediction). Therefore, as spatial data scientists, we have to leverage spatial uncertainty into these new computational models in order to accommodate our unique research questions.

Last but not least, I highly value the effort of promoting interpretability and reproducibility in spatial data science[2]. Since data are becoming larger and models are becoming more complicated, it is often difficult to interpret and reproduce experiments, which impedes the development of spatial data science. Consequently, to have widely agreed geospatial problems and an open source data platform are necessary to advance spatial data science, which is a challenge though mainly due to ethical issues.

References

[1] Anselin, Luc. What is special about spatial data? Alternative perspectives on spatial data analysis. *Technical Report 89-4 (NCGIA)*, 1989.

[2] Kedron, Peter, Amy E. Frazier, Andrew B. Trgovac, Trisalyn Nelson, and A. Stewart Fotheringham. Reproducibility and replicability in geographical analysis. *Geographical Analysis*, 2019.

[3] Miller, Harvey J. and Michael F. Goodchild. Data-driven geography. *GeoJournal* 80(4): 449–461, 2015.

[4] Zhu, Rui, Krzysztof Janowicz, and Gengchen Mai. Making direction a first-class citizen of Tobler's first law of geography. *Transactions in GIS* 23(3): 398–416, 2019.

[5] Zhu, Rui, Phaedon C. Kyriakidis, and Krzysztof Janowicz. Beyond pairs: generalizing the geo-dipole for quantifying spatial patterns in geographic fields. In *The Annual International Conference on Geographic Information Science*, pages 331–348. Springer, 2017.