Chapter 17
GeoAI and the Future of Spatial Analytics

Wenwen Li and Samantha T. Arundel

Abstract This chapter discusses the challenges of traditional spatial analytical methods in their limited capacity to handle big and messy data, as well as mining unknown or latent patterns. It then introduces a new form of spatial analytics—geospatial artificial intelligence (GeoAI)—and describes the advantages of this new strategy in big data analytics and data-driven discovery. Finally, a convergent spatial analytical framework is suggested as a potential future pathway for spatial analysis.

Keywords Spatial analysis · GeoAI · Artificial intelligence · Deep learning · Data-driven discovery

17.1 Challenges in Spatial Analytics

As a set of quantitative and computational approaches for analyzing geospatial data, spatial analytics is the core of Geographic Information Science (GIScience) for exploration, knowledge discovery, and decision making in the spatial realm. Identified by Golledge (2009) as the unique contribution by geographers to the scientific community, spatial analysis is defined as the methods developed exclusively for analyzing location-based information. Location-based data need specialized analytics to handle spatial dependence, scale dependence, and ecological fallacy, which are not sufficiently accounted for using conventional statistical methods. In the past decades, as spatial theory and computing technology advanced, spatial analysis expanded considerably to cover spatial statistics (for example, exploratory spatial data analysis and spatial regression), spatial simulation (such as agent-based modeling and

W. Li (✉)
School of Geographical Sciences and Urban Planning, Arizona State University, Tempe, AZ 85287-5302, USA
e-mail: wenwen@asu.edu

S. T. Arundel
U.S. Geological Survey, Center of Excellence for Geospatial Information Science (CEGIS), Rolla, MO 65401, USA
e-mail: sarundel@usgs.gov

B. Li et al. (eds.), *New Thinking in GIScience*,
https://doi.org/10.1007/978-981-19-3816-0_17

microsimulation), spatial optimization (Murray, 2021), and data-driven techniques, such as data mining and artificial intelligence (Li, 2020).

Despite covering remarkable breadth, spatial analytics still faces substantial challenges. Goodchild (2009) identified notable issues that spatial analysis is facing. From the perspective of technology, the trend towards the migration of spatial analytical functions to the Web necessitates new business models. New models would ideally handle server-client communication and interoperability and manage data innovatively for online parallel processing services that require use of server-client communication. They also would ideally promote transparency in spatial analysis modules available online. From the science perspective, a (re)formulation of GIScience based on how spatial analytics are being used in scientific and practical problem solving would be beneficial. Over a decade later, we ask "how has the research landscape of spatial analysis changed, how well were Goodchild's challenges addressed, and what new challenges are emerging?".

The last 10 years have witnessed revolutionary advances in technology. Although the term 'cloud computing' was new a decade ago, it has become prevalent today to support the storage, computing, and analysis of geospatial data and its applications (Li et al., 2016). Instead of maintaining a dedicated server, geographic information system (GIS) users and developers have increasingly used cloud infrastructure based on highly reliable virtualized cloud machines capable of elastic computing to meet the different needs of end users. For example, Google Earth Engine, Google's cloud platform that hosts multi-decades of remote sensing images, offers the public rapid access to massive geospatial data and planetary-scale spatial analytics (Gorelick et al., 2017; Yang et al., 2018). The emergence of cyberinfrastructure and CyberGIS has also revolutionized the landscape of spatial analysis to allow collaborative data sharing, analytics, and decision-making (Anselin & Rey, 2012; Li et al., 2016, 2019a, 2019b; Wang, 2010; Yang et al., 2017).

Despite these advances, spatial analytics still have existing and new challenges. Here we present a few examples of these challenges from the computational and data science perspectives.

17.1.1 The Size Challenge of Big Data

Big data have changed nearly every aspect of our lives and the way we conduct science. Datasets, such as earth observation and remote sensing images, images from unmanned aerial vehicles (UAVs), and georeferenced data from social media platforms and sensors for the Internet of Things (IoT) have yielded the production and availability of geospatial data at unprecedented spatial and temporal coverage, resolution, and collection frequency (Li et al., 2020). Handling these data at high throughput and in real-time has presented considerable challenges for traditional analytical methods designed for processing small, clean datasets (Li et al., 2022). Spatial statistical methods, for instance, often require an abstraction of raw data to point data in tabular forms to identify clustering patterns or the associations between

certain numerical attributes through linear regression. These methods have reached limitations when it comes to analyzing big data, which are, by definition, large, noisy, diverse, and complex. Although redesigning existing statistical methods to handle big data has been attempted (Laura et al., 2015; Li et al., 2019a, 2019b), many widely used spatial statistical software, such as PySAL (Python Spatial Analysis Library) (Rey et al., 2015) and Geographically Weighted Regression (GWR) (Oshan et al., 2019), continue deployment in desktop computing environments and lack the utilization of advanced computing devices, such as Graphics Processing Units (GPUs). This is likely because the focus of innovation remains on methodology rather than computational performance. In addition, to handle big data, sampling approaches are often introduced. However, in a large dataset with an unknown distribution, it is difficult to guarantee that conventional sampling does not introduce bias into the data, for example in sub-setting training and test sets.

17.1.2 Navigating Through the Messiness of Big Data

Conventionally, big data equals messy data. At the rates data are generated today, the diversity in data collection methods makes (timely) quality control difficult. For example, very fast sampling of some phenomena, such as an event of interest that occurs sporadically, can lead to many empty records. Data reduction can introduce problems, such as when stacking large numbers of raster images over time and then computing a mean or median response in co-located pixels, one can end up with a median image that is too dark in areas of dense cloud cover. Resampling issues result in less accurate results when images are not registered uniformly, and their pixels are aligned. Such issues are easier to detect in small datasets than in large ones. Hence, the ability to navigate through big, complex data becomes a new challenge that calls for innovative techniques designed for big data analytics. Census data for the 2020 Census alone cost the U.S. Census Bureau over $14 billion for compilation and delivery (GAO, 2021). This is one example of high quality, official data managed by governments. However, many other big datasets are created from social media and crowd sourcing platforms, such as Twitter, which have been increasingly used for research because of their broad spatial coverage, richness of content, and low collection cost. However, data from these platforms inevitably contain a substantial amount of noise due partially to their openness, which allows anyone to say anything at any time. In Bayesian statistics, where random variables are introduced, determining the proper prior distribution is often needed to make the estimated posterior distribution match with reality. In such cases, data noise will impede the accurate estimation of a prior distribution. The resulting errors will propagate to later stages of the inference process and lead to imprecise results.

17.1.3 Hypothesis Test Versus Knowledge Mining

Besides relying on well processed data, the traditional spatial analytical approach also requires an accurate understanding and prior knowledge of the underlying process. For instance, in agent-based modeling, heuristic rules need to be defined to guide how an agent moves in space and interacts with the environment and other agents (Li et al., 2020). When applying regression analysis, one needs to specifically define both the independent (X) and dependent variables (y) when building the model, which means we should have knowledge about how X are affecting y. The goal of the analytics is to explain whether and how these independent variables (for example, income or climate) affect the dependent variable (such as housing price) in a geographical region. To incorporate geographically varying effects resulting from spatial heterogeneity, local modeling, such as GWR, is introduced to determine the variation of effects across space. These analyses belong in general to the testing of a hypothesis or identifying the degree of effect between X and y in a predefined model. Whereas such methods are known to be effective in identifying patterns that are expected, their ability to discover or learn unknown relations is weak.

Confronting these challenges requires new spatial analytical methods capable of mining new knowledge from large datasets containing unanticipated or previously unknown patterns, as well as being tolerant to noise. The methods also would ideally be able to learn to model the process itself rather than relying on definitions drawn from prior knowledge. GeoAI has emerged as a new arena for attacking these challenges.

17.2 GeoAI: A New Form of Spatial Analytics

GeoAI, or geospatial artificial intelligence, is a transdisciplinary research area integrating cutting edge AI to solve geospatial problems (Li, 2020). In the past decade, amazing progress has been made in the field of AI, particularly in machine learning and deep learning. The convolutional neural network (CNN) framework is a milestone development (Reichstein et al., 2019). The CNN framework adopts the novel concept of artificial neural network (ANN) in building a computer model mimicking the biological neural network of the human brain even as it brings transformative changes through the introduction of the convolution modules (Fukushima, 2007; Li, 2021; Li et al., 2012; Zhang, 1988). Such modules can conduct information extraction (also known as feature extraction, with each feature treated being the independent variable X in a regression process) from the raw data. CNN-based techniques, therefore, can directly act on the raw data and uncover hidden patterns through deep mining and iterative learning. This kind of data-driven analysis relaxes the constraint in traditional spatial analytics for assuming any predefined rules or relationships between the

data (input) and the objective (output), thus supporting discovery and pattern recognition directly from data. This is also known as data-driven discovery (Miller & Goodchild, 2015; Yuan et al., 2004).

Another breakthrough in the design of CNNs is that each convolution layer (Albawi et al., 2017) performs local operations on the data, making parallel computation possible. This design lifts the computation constraint in traditional ANN that has high dependency among the artificial neurons across the fully connected layers. The recent development of high-speed GPUs that contain a few hundred to several thousand micro-processing units allows the high-performance training of CNNs, even with complex structures, on its computing units running in parallel. This also empowers a deep learning model to process big data, furthering its ability to detect new patterns, extract useful information, and create high-quality foundational datasets to aid the elucidation of important scientific questions (Arundel et al. 2020).

Moreover, deep learning models are arguably better at handling noise in training labels than traditional statistical methods (Rolnick et al., 2017). Because many such models are designed to learn complex relations, they tend to overfit the training data. Overfitting occurs when a model fits the training data exactly. When this happens, the model's performance on unseen data will be inferior. One solution is to add noise to the training data such that the model will fit less perfectly, reducing the likelihood of overfitting, and increasing predictive accuracy. In addition, strategies, such as increasing the batch size and thus exposing the model to more samples for updating its parameters during the iterative learning process, lowering learning rates, allowing a more thorough search for the optimal solutions, and providing enough correctly labeled samples, will enable a deep learning model to tackle even extremely noisy data (Rolnick et al., 2017). Although noise in big data is inevitable, the way deep learning is designed and how it handles the data makes deep learning more robust towards dealing with noise than traditional spatial analytical approaches. On the other hand, deep learning requires thousands to billions of training examples to develop abstractions that the human brain can easily intuit through explicit, verbal definition (Marcus, 2018). Interpretability of the results and extension beyond the scope of the training data are also limitations to deep learning systems (Reichstein et al., 2019) that must be overcome.

17.3 Concluding Remarks

As a new form of spatial analytics, GeoAI is exciting because of its outstanding performance in big data analytics, especially in classification, prediction, and pattern recognition. However, the GeoAI domain is still in its infancy and more research is needed for it to become a well-established scientific field. The role of GeoAI in (re)formulating GIScience also needs to be more clearly defined. This need echoes insights shared by Goodchild (2009) in terms of the challenges of spatial analytics in general. We know that the complexity and black-box nature of GeoAI models render the model's reasoning process more difficult to explain than that of traditional spatial

analytical approaches (Goodchild & Li, 2021). But this also offers an opportunity to create an even more powerful analytical framework by combining GeoAI and traditional methods. GeoAI can serve as a data pre-processing module that directly interacts with raw big data to achieve high-yield analysis and data filtering (Li et al., 2022).

For instance, a GeoAI-based analytical framework can achieve near real-time processing of satellite remote sensing imagery to create a national to global scale database characterizing natural and human-made features on Earth (Li & Hsu, 2020). This dataset, for which scientists and researchers have waited decades, can be integrated into subsequently processed statistical models to understand crucial environmental and climate change problems (Reichstein et al., 2019). The data and models may jointly contribute to a convergent research agenda for spatial analytics.

Clearly, the development and refinement of existing and future spatial analytics (GeoAI and beyond) should consider fundamental geospatial principles, such as location, scale, spatial autocorrelation, spatial heterogeneity, and geographic similarity. As data and systems become more open, they are less likely to follow fundamental principles and best practices. This concern is like that expressed by scholars during the early years of the development of GIS. Concerns included whether users would utilize the correct projection for the variable studied, correct their statistical analyses for bias in location, or analyze error by combining the variables of the spatial themes.

Whereas some elements of these potential problems are now controlled inherently by software systems, other problems persist or may not be envisioned in the present. Like GIS, GeoAI and subsequent technologies would ideally balance the accessibility of the approach with its applicability, the enforcement of the principles with the flexibility of application. This is the grand challenge of the spatial science community: to not only create and disseminate new tools towards the goal of empowering more vast and ethical utilization, but more importantly to leverage these tools to improve analysis of spatial information to address critical global, regional, and local problems.

Acknowledgements This work is supported in part by National Science Foundation (NSF) under grant BCS-1853864. Li acknowledges additional funding support from NSF (BCS-1455349, GCR-2021147, PLR-2120943, and OIA-2033521). Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U. S. Government.

References

Albawi, S., Mohammed, T. A., & Al-Zawi, S. 2017. Understanding of a convolutional neural network. In *2017 International Conference on Engineering and Technology (ICET)* (pp. 1–6).

Anselin, L., & Rey, S. (2012). Spatial econometrics in an age of CyberGIScience. *International Journal of Geographical Information Science, 26*(12), 2211–2226. https://doi.org/10.1080/13658816.2012.664276.S2CID942116

Arundel, S., T., Li, W., & Wang, S. (2020). GeoNat v1.0: A dataset for natural feature mapping with artificial intelligence and supervised learning. *Transactions in GIS, 24*(3), 556–572.

Fukushima, K. (2007). Neocognitron. *Scholarpedia, 2*(1), 1717.
GAO (2021). *2020 Census: Innovations helped with implementation, but Bureau can do more to realize future benefits*. United States Government Accountability Office (GAO). https://www.gao.gov/assets/gao-21-478.pdf
Goodchild, M. F. (2009). Challenges in spatial analysis. In A. S. Fotheringham, & P. A. Rogerson (Eds.), *The SAGE handbook of spatial analysis* (pp. 465–480). SAGE Publishing.
Goodchild, M. F., & Li, W. (2021). Replication across space and time must be weak in the social and environmental sciences. *Proceedings of the National Academy of Sciences*, *118*(35).
Golledge, R. G. (2009). The future for spatial analysis. In A. S. Fotheringham, & P. A. Rogerson (Eds.), *The SAGE handbook of spatial analysis* (pp. 465–480). SAGE.
Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., & Moore, R. (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment, 202*, 18–27.
Laura, J., Li, W., Rey, S. J., & Anselin, L. (2015). Parallelization of a regionalization heuristic in distributed computing platforms—A case study of parallel-p-compact-regions problem. *International Journal of Geographical Information Science, 29*(4), 536–555.
Li, W. (2020). GeoAI: Where machine learning and big data converge in GIScience. *Journal of Spatial Information Science, 20*, 71–77.
Li, W. (2021). GeoAI and deep learning. *International Encyclopedia of Geography: People, the Earth, Environment and Technology*, 1–6.https://doi.org/10.1002/9781118786352.wbieg2083
Li, W., Batty, M., & Goodchild, M. F. (2020). Real-time GIS for smart cities. *International Journal of Geographical Information Science, 34*(2), 311–324.
Li, Z., Fotheringham, A. S., Li, W., & Oshan, T. (2019a). Fast Geographically Weighted Regression (FastGWR): A scalable algorithm to investigate spatial process heterogeneity in millions of observations. *International Journal of Geographical Information Science, 33*(1), 155–175.
Li, W., Goodchild, M. F., Anselin, L., & Weber, K. T. (2019b). A smart service-oriented CyberGIS framework for solving data-intensive geospatial problems. In *CyberGIS for geospatial discovery and innovation* (pp. 189–211). Springer.
Li, W., & Hsu, C. Y. (2020). Automated terrain feature identification from remote sensing imagery: A deep learning approach. *International Journal of Geographical Information Science, 34*(4), 637–660.
Li, W., Liu, Y., & Wang, S. (2022). Real-time GIS and geocomputation. In J. P. Wilson (Ed.), *The geographic information science & technology body of knowledge* (3rd Quarter 2021 Edition) (in press).
Li, W., Raskin, R., & Goodchild, M. F. (2012). Semantic similarity measurement based on knowledge mining: An artificial neural net approach. *International Journal of Geographical Information Science, 26*(8), 1415–1435.
Li, W., Shao, H., Wang, S., Zhou, X., & Wu, S. (2016). A2CI: A cloud-based, service-oriented geospatial cyberinfrastructure to support atmospheric research. In *Cloud Computing in Ocean and Atmospheric Sciences* (pp. 137–161). Academic Press.
Marcus, G. (2018). Deep learning: A critical appraisal (pp. 1–27). arXiv preprint arXiv:1801.00631.
Miller, H. J., & Goodchild, M. F. (2015). Data-driven geography. *GeoJournal, 80*, 449–461. https://doi.org/10.1007/s10708-014-9602-6
Murray, A. T. (2021). Significance assessment in the application of spatial analytics. *Annals of the American Association of Geographers, 111*(6), 1740–1755.
Oshan, T. M., Li, Z., Kang, W., Wolf, L. J., & Fotheringham, A. S. (2019). mgwr: A Python implementation of multiscale geographically weighted regression for investigating process spatial heterogeneity and scale. *ISPRS International Journal of Geo-Information, 8*(6), 269.
Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., & Carvalhais, N. (2019). Deep learning and process understanding for data-driven Earth system science. *Nature, 566*(7743), 195–204.
Rey, S. J., Anselin, L., Li, X., Pahle, R., Laura, J., Li, W., & Koschinsky, J. (2015). Open geospatial analytics with PySAL. *ISPRS International Journal of Geo-Information, 4*(2), 815–836.

Rolnick, D., Veit, A., Belongie, S., & Shavit, N. (2017). Deep learning is robust to massive label noise. arXiv preprint arXiv:1705.10694.

Wang, S. (2010). A cyberGIS framework for the synthesis of cyberinfrastructure, GIS, and spatial analysis. *Annals of the Association of American Geographers, 100*(3), 535–557.

Yang, C., Huang, Q., Li, Z., Liu, K., & Hu, F. (2017). Big Data and cloud computing: Innovation opportunities and challenges. *International Journal of Digital Earth, 10*, 13–53. https://doi.org/10.1080/17538947.2016.1239771

Yang, Z., Li, W., Chen, Q., Wu, S., Liu, S., & Gong, J. (2018). A scalable cyberinfrastructure and cloud computing platform for forest aboveground biomass estimation based on the Google Earth Engine. *International Journal of Digital Earth, 12*(9), 995–1012.

Yuan, M., Buttenfield, B. P., Gahegan, M. N., & Miller, H. (2004). Geospatial data mining and knowledge discovery. In *A research agenda for geographic information science* (p. 24). CRC Press.

Zhang, W. (1988). Shift-invariant pattern recognition neural network and its optical architecture. In *Proceedings of Annual Conference of the Japan Society of Applied Physics.*