Challenges and Prospects of Uncertainties in Spatial Big Data Analytics

Wenzhong Shi,* Anshu Zhang,* Xiaolin Zhou,* and Min Zhang[†]

*Department of Land Surveying and Geo-informatics, The Hong Kong Polytechnic University, Hong Kong

[†] School of Remote Sensing and Information Engineering, Wuhan University, Wuhan

Abstract: Knowledge extraction from spatial big data (SBD) with advanced analytics has become a major trend in research and industry. Meanwhile, the increasingly complex SBD and its analytics face proliferating challenges posed by uncertainties in them. Linked to various characteristics of SBD, the uncertainties emerge and propagate in each stage of SBD analytics. To avoid unreliable knowledge and losses resulted from the uncertainties, and to ensure the value of authentic knowledge, this article proposes uncertainty-based SBD analytics. Uncertainty-based SBD analytics strive to understand, control, and alleviate uncertainties and their propagation in each stage if geographic knowledge extraction. Key topics involved in uncertainty-based SBD analytics include, for example, place-based heuristics for learning urban structure and place-based analytics on broader knowledge extraction tasks; dealing with the biases and inferencing the semantics in cellphone tracking data; quality assessment of unstructured spatial user-generated contents, and the rectification of location shifts and time elapses between humans' activities and corresponding online contents they generate; and uncertainty handling in sophisticated black-box analytics with SBD such as deep learning. Challenges and latest advances in each of these topics are presented, and further research for addressing these challenges are suggested in this article. *Keywords: big data, geographical knowledge discovery, social networks, time geography, uncertainties*

Introduction

Big data poses serious challenges to existing data analytics due to its vast volume, velocity, and variety (Gerhardt et al. 2012). The rise of *spatial big data (SBD)* that comes with references to geographic location or place is a prominent trend in the big data realm. SBD may be divided into two categories: (a) *Earth observations* through airborne and ground sensors, for example, satellites, unmanned aerial vehicles, LiDAR systems, and weather and pollution monitoring instruments; and (b) *Human activity observations*, such as GPS tracks from cellphones, smartcard tap-ins/outs, surveillance videos, and online user activities, posts and image uploads that are geographically referenced.

SBD analytics, sometimes called SBD analyses, is widely expected to generate great value in research and applications. In this article, analytics refers to various forms of knowledge extraction from data, including tasks commonly termed data mining and predictive learning of data. On the other hand, people tend to be confident in advanced SBD analytics and use the extracted knowledge to resolve essential scientific and practical issues. Incorrect, spurious or biased resultant knowledge can thus mislead people to wrong actions with severer losses.

This article aims at investigating the framework, challenges, and prospects regarding uncertainties in SBD analytics. We advocate that SBD analytics need to be *uncertainty-based*, that is, to understand, control and alleviate the ubiquitous uncertainties in the real world and each stage of knowledge extraction from SBD, thereby assuring and improving the reliability and the value of resultant knowledge. Under this proposed framework, we select a few key research topics, discuss the challenges and latest advances in these topics, and suggest their future directions.

The next section of this article describes the key uncertainties in SBD. The "Uncertainty-Based SBD Analytics" section proposes the uncertainty-based framework of SBD analytics. The

"Advances and Challenges in Selected Uncertainty Issues of SBD Analytics" section presents the challenges, latest solutions, and prospects on some key topics under the proposed framework. The last section gives concluding remarks.

Uncertainties in SBD and Implication to SBD Analytics

SBD inherits and further develops the major characteristics of big data, such as variety, velocity, volume, veracity, and value (International Business Machines 2016). Veracity, or the quality and trustworthiness, directly concerns and attracts intensive research efforts on uncertainties of SBD. However, most of the other characteristics of SBD are also linked to uncertainties in SBD (Li et al. 2016) and eventually its analytical results.

The wide *variety* of SBD, as exemplified in the introduction, can either be structured and fit into standard databases, or be semi-structured and unstructured, such as texts, images, videos, and complex human behaviors. Despite richer information and knowledge it provides, multisource data can also generate uncertainties. First, multisource datasets possess heterogeneous acquisition methods, units, quality, and presentation of the same geographic subjects, which can result in different and even contradictory knowledge. It is nontrivial to decide which datasets are suitable for specific research problems and applications. The situation becomes more complicated when considering that semi-structured and unstructured datasets themselves are not yet clearly represented. Second, if people decide to fuse and jointly use multisource datasets or their individual analysis results, heterogeneous units in the datasets normally need to be unified. Aggregating and segmenting data into different areas (Openshaw 1983) and time intervals (Cheng and Adepeju 2014) can significantly impact the result of statistical data analyses. Inappropriate unit unification in the data or result fusion can thus incur unreliable or biased knowledge. Third, fusing datasets into

higher dimensional ones can increase the potential to learn useful data correlations for constituting richer knowledge. However, the chances of getting accumulated noises from multiple datasets, spurious data correlations, and finally dubious knowledge are also increased (Fan et al. 2014).

SBD streams generated in real time at high *velocity* enable more realistic capturing and analysis of spatiotemporal dynamics (Miller and Goodchild 2015). Most existing methods for evaluating or enhancing the reliability of spatial data and its analytics are designed for processing all data in one run or in batches. These methods often involve mechanisms that are incompatible with the processing of ongoing data streams, or statistical assumptions that no longer hold for continuously changing attributes to be learned from data streams. Reliable SBD stream analytics call for further development of these existing techniques and new techniques.

Veracity is undoubtedly the most relevant to uncertainty issues. Veracity is especially concerned for the upsurging *user-generated SBD*, including cellphone tracking data, user check-ins at points of interest (POIs) in location-based social networks (LBSNs), and spatial *user-generated content (UGC)* like crowdsourced maps and geotagged online user-created texts and images. Compared with data from professional vendors, spatial UGC faces severer authenticity issues, such as mistaken data due to carelessness or inadequate expertise, and intentional creation of data noises (e.g., vandalistic "Graffiti" on OpenStreetMap; OpenStreetMap Wiki 2017). User-generated SBD has also been criticized for being self-selective and thus generating high risks to produce biased knowledge. The biases are particularly problematic in LBSNs, where only small fractions of data, for example, less than 1% tweets in Greater London during 2013 (Longley and Adnan 2016), are geotagged. It is then highly questionable whether the knowledge extracted from geotagged tweets applies to general Twitter users, not to say the general public. Studies have also reported how LBSN data could be biased by various factors, such as ethnics and income of users (Fekete 2017), incentives to users, and data policy of LBSN operators (Thatcher 2014).

The *volume* of the data means more than the need for powerful computing techniques. The massive data enables the exploitation of subtle data correlations that are unsupported by smaller sized data. The newly discovered correlations can have high chances to be spurious, as is the case of mining high-dimensional data. Besides, the exploding data size is often due to the finer resolution of data acquired by modern techniques. Yet data in the finest resolution might not yield the best analytical results: people sometimes need to aggregate the data to seek for higher-level regularities that are more meaningful for the investigated problems. Aggregating data into different spatiotemporal units in turn influences the analytical results.

Ultimately, the *value* of SBD relies on uncertainty handling and the reliability of extracted knowledge, as emphasized in the introduction. The application of unreliable knowledge may rather lead to big losses.

Uncertainty-Based SBD Analytics

Uncertainties, including but not limited to those described in the last section, arise from each stage before and during the process of SBD analytics, as listed as follows: (a) Uncertainties in real-world geographic entities and relations, in that they often cannot be precisely described, and their characteristics may vary with environment, time, and scale; (b) Uncertainties in human cognition and abstraction, or modeling, of the real world into data. It can be obscure to select an appropriate data model from multiple ones, and improper data modelling can lead to highly deficient results of the analytics; (c) Uncertainties in data capture using various techniques, including ground and airborne earth observation methods, latest urban sensing, crowdsourcing, and other techniques for acquiring the wide variety of SBD; and (d) Uncertainties due to

methodological limitations in the knowledge extraction process, such as omissions of important variables to investigated problems, and enforcement to information loss in data processing. Once generated, the uncertainties will propagate in all subsequent stages and finally into the unreliability of resultant knowledge (Shi 2010).

In response to the above acute uncertainty issues, we propose *uncertainty-based SBD analytics*, the process of geographic knowledge extraction that incorporates understanding, control, and alleviation of uncertainties and their propagation in the real world and each stage of the analytical process. There are two primary tasks in uncertainty-based SBD analytics: (a) *Reliability assurance*: to quantify previously unclear uncertainties and reliability of the resultant knowledge. This task is the ground of identifying the scopes to which the knowledge is generalizable, and the risks of accepting any untrue knowledge; and (b) *Reliability enhancement*: to reduce the uncertainties and obtain more reliable knowledge through novel analytics or improvements of existing analytics. This task expands the scopes and benefits while reducing the risks of applying the knowledge (Figure 1). The ultimate objective of uncertainty-based SBD analytics, to be achieved through the reliability assurance and enhancement, is to steer SBD and its analytics to real "big value."

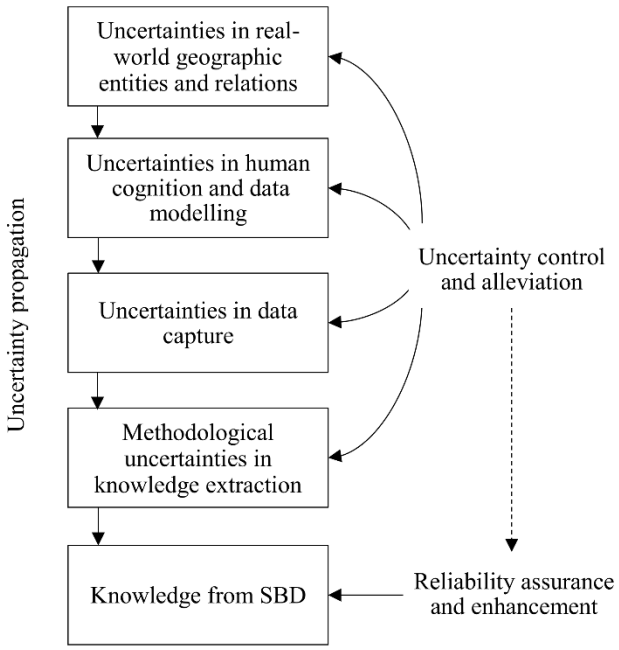

Figure 1. Framework of uncertainty-based SBD analytics.

The core of reliability assurance is to effectively assess the quality of the data and analytical results and identify those deficient ones. The characteristics of SBD make its quality assessment more challenging than that of conventional spatial data. Reliability assurance also requires relevant quality standards, as well as feedback mechanisms to isolate deficient data or analytical results and ask for quality improvement. Techniques for reliability enhancement highly depend on specific SBD types and analytical tasks. In the next section, parts of the latter two subsections discuss the latest techniques for assessing the reliability of unstructured spatial UGC and complicated black-box SBD analytics. The other parts are dedicated to reliability issues and enhancement in specific SBD analytical fields.

Advances and Challenges in Selected Uncertainty Issues of SBD Analytics

This section discusses the challenges, latest advances, and prospects of one representative uncertainty issues on each of the four stages of SBD analytics listed in the last section. Each issue can arise on multiple analytical stages, especially due to the propagation of the uncertainty through these stages. Here we match each issue to the earliest stage at which it becomes significant. These four issues are only a small subset of the important uncertainty concerns in SBD analytics. For example, the aforementioned bias of spatial UGC is of key concern but is not further discussed below due to space limitation.

On Uncertain Real World: Place-Based Heuristics and Analytics

Natural language geotags in UGC and human activity areas are expressed in terms of vague places instead of precise locations. Geometries and hierarchies of places vary with human activities and time. A national park may be seen by tourists as their accessible area, but by ecologists as the region of protected ecosystem. Adjustments of administrative divisions often lead to place hierarchy changes. Improper perception and modeling of places can severely distort the results of SBD analytics. This issue lately motivates the place-based GIScience.

A recent focus in place-based research is to extract urban functional regions. Urban functions, such as transport, tourism, and education, represent humans' common usage of the urban space (Tu et al. 2017). Such usage information is often inferred from LBSN user check-ins or cell phone tracking records (Rösler and Liebig 2013, Zhou and Zhang 2016, Zhou et al. 2017). In most studies, functions of the extracted regions are predefined or interpreted manually by researchers. Gao, Janowicz, and Couclelis (2017) developed the first bottom-up approach to construct urban function ontologies using the data itself. They applied latent Dirichlet allocation (LDA), a

technique traditionally for mining latent topics from texts, to user check-ins at Foursquare POIs. Each Foursquare POI was associated with a detailed type like French Restaurant and Tennis Court. LDA was applied by taking all the POIs surrounding a location as a document, each POI type as a word, and the number of user check-ins at POIs of that types as the word frequency. The resultant topics were probability vectors of POI types and could infer certain functions. Then urban functional areas were extracted by K-means and Delaunay triangulation spatial constraint clustering of the spatially distributed topics. This work inspires future studies, such as how to incorporate other online UGC to infer user activities and construct the functional regions more accurately. Bottom-up place-based research like their work is promising with rich, high-dimensional SBD and machine learning innovations.

Most current place-based studies end up with cognized place models like the functional regions, which by themselves are very useful in understanding urban systems. Yet a step forward is to use place-based models in further SBD analytics and machine learning tasks, for example, traffic prediction and POI recommendation. These tasks, when conducted over appropriate place models instead of precise locations or regular grids, can generate more reliable human activity patterns and final knowledge. Chen et al. (2017) conducted one of the first such studies: they presented four place-based measures of accessibility to urban services that consider stochastic variations in travel time. They demonstrated that conventional accessibility measures without considering travel time uncertainty overestimate accessibility, especially for suburbs when compared with downtown.

On Cognition and Data Modelling: Biases and Inferred Semantics of Cellphone Tracking Data

Cellphone tracking data is among the most popular data sources for human mobility studies, because of its very broad coverage in urban population, reasonable spatiotemporal resolution, and cost-effective access for researchers. However, the study results can be considerably influenced by different practices of modeling human movements using the data, for example, when inferring users' movement trajectories using their recorded locations. Kwan (2016) made a comprehensive discussion on this issue. In particular, the data used by most existing studies are cellphones' call detail records (CDRs) which are made when users make/receive calls and send/receive messages. The uneven spatiotemporal distribution of these phone communication activities can incur biases of research results. In the empirical study of Zhao et al. (2017), CDRs underestimated users' travel distance by 35%-85% and movement entropy by 12%-67%, compared with complete cellphone 'signaling' data which includes other regular updates of user locations. The underestimations became severer as the ratio of CDRs in the signaling data decreased.

While the signaling data enables much less biased study results, it still suffers from the absence of semantic information. With cell-tower-level spatial resolution, the data cannot locate users to specific POIs. Thus it is difficult to obtain knowledge about the users' activities. Tu et al. (2017) developed a method to infer the activities of cellphone users by integrating LBSN data. In the method, cellphone users' social activities such as shopping, schooling, and entertainment were labeled using a hidden Markov model, based on spatiotemporal distributions and variations of LBSN user check-ins at different types of POIs as well as cellphone user movements. Ratios between the overall frequency inferred activities generally showed good matches with those in governmental household travel survey. Then the inferred activities were used to analyze hourly urban functions like residence, commerce, and education across the study area. While this method

is a notable advance, further investigations are needed regarding how accurate the inference of user activities is in different localities of a study area, and how much the remaining inaccuracy affects the results of relevant human mobility studies.

On Data Capture: Quality Assessment of Unstructured Spatial UGC and Misregistered Spatial Footprints

Quality assessments of collaborative maps have been extensively researched, and relevant studies are comprehensively reviewed by Senaratne et al. (2017). The quality of unstructured spatial UGC, for example, geotagged user-generated images, has been much less assessed but can be harder to quantify and more problematic. Existing quality assessments for geotagged user-generated photos focused on the positional accuracy of the geotags (Zielstra and Hochmair 2013, Hollenstein and Purves 2014). The semantic accuracy, or that if photos are mislabeled to other places than where they were taken, is more difficult to judge whether manually or automatically.

Latest advances in image mining techniques, especially those adopting deep learning, could be very helpful in assessing and enhancing the semantic reliability of geotagged photos. Crandall et al. (2015) developed an approach to automatically recognize popular places like Venice and the Louvre from LBSN photos, by using a deep convolutional neural network (CNN) to classify the photos, where each class was a place label. The training dataset was built without human intervention using over 30 million geotagged Flickr photos, including numerous misplaced ones and ambiguous ones like tourists' selfies. The geotags were taken as 'true' places of the training photos. Despite the noisy training data, the CNN classifier achieved over 75% accuracy in recognizing each top-10 venues, outperforming traditional classifiers such as the support vector machine, and sometimes even well-traveled human observers. Weyand et al. (2016) developed another CNN-based approach to infer geolocations of photos for both famous spots and ordinary

local places. They obtained over 60% accuracy in registering photos into 200Km grid cells on the Earth and nearly 80% accuracy for 750Km cells. These classifiers can also be adapted to geotagged photos to assess their semantic accuracy. Photos with geotags contradicting the CNN classification result can be highlighted for further quality improvement. With additional clues from users' information and photo streams, the image quality assessment could be more accurate than using CNNs to locate non-geotagged photos.

Another challenge more specifically for LBSN data lies in 'misregistered spatial footprints' of users: a geotagged and timestamped user message is unnecessarily related to that place and time. Instead, the message may be about the user's current feeling that is irrelevant to any place, or the user's previously visited location, but now her location has shifted, and the message has a time lag.

The abovementioned deep learning of images may also be a tool for detecting misregistered photo posts in a similar way to detecting misplaced ones. As for textual posts, researchers have also used non-georeferenced online texts to infer geolocations these texts refer to, through topic modeling and co-occurrence of spatial entities in the texts. This approach can register most texts to within 500 Km of their true locations (Adams and Janowicz 2012). A similar method may thus identify texts irrelevant to or with large shifts from current user locations.

Latest image and text mining solutions do not yet work well in differentiating places inside the same city. Therefore, rectifying misplaced and misregistered spatial UGC, especially at in-city scale, remains a big challenge. The solution to this challenge calls for either more powerful image and text mining techniques or effective integrations of state-of-art techniques and extra online information.

On Knowledge Extraction Methods: Deep Learning and Other Sophisticated Black-box Techniques

Deep learning uses multiple layers of processing units to automatically extract features from complex data for classification or pattern recognition. Compared with conventional analytics which requires substantial human intervention, deep learning is more efficient, and sometimes has significantly higher accuracy and better performance due to its utilization of empirically unknown features to improve the learning process. In current geography research, deep learning is mainly used for more reliable data production and preprocessing, for example, classifying land uses with satellite images, extracting human movement trajectories from videos, and possibly detecting deficient user-generated photos and texts as proposed in last subsection. The use of deep learning as the major analytical tool in geographical studies is scarce, but might be a future trend.

While seeing the promise of deep learning and other sophisticated black-box machine learning techniques for enhancing the reliability of SBD and SBD analytical results, geographers also need to be aware of specific uncertainties of these techniques. First, higher accuracy of deep learning classifiers relies on much larger training datasets with labeled classes than conventional classifiers. Obtaining sufficient training data can take prohibitive cost, and becomes infeasible when the classifiers face evolving real-world situations and require continuous relabeling of the training data. A recent solution of this issue is the weakly supervised learning that works on incomplete, multivalued or other kinds of low-quality labels that can be quickly obtained. Weakly supervised deep learning has made achievements in object detection from satellite images (Han et al. 2015). Another possible solution is crowdsourcing the training data. For instance, Chen and Zipf (2017) used images with buildings and roads labeled by volunteers from OpenStreetMap and the MapSwipe mobile application to train a deep learning model for satellite image classification.

Another key issue that it is very difficult to assess the reliability, such as accuracy and residual, of deep learning and other complex black-box machine learning models. The mechanisms of these models are extremely complicated and hardly known, which makes it almost impossible to derive analytical results on their model uncertainty. Meanwhile, estimating the reliability of these techniques is crucial for users to control the risk and loss incurred by possible false resultant knowledge. Several pioneering works have been done on estimating the uncertainty of deep learning. For example, Gal and Ghahramani (2016) utilized dropout training to conduct approximate Bayesian computation on the uncertainty of deep neural networks. Dropout training is a technique in training the networks for avoiding overfitting. In any case, geographers need to note the uncertainties brought by new computational analytics, and keep up with latest solutions to these uncertainties.

Conclusion

Uncertainties are inherent in the real geographical world, SBD and each stage of SBD analytics. Linked to various characteristics of SBD, these uncertainties can lead to unreliable results of the analytics and losses of research and practical value of SBD analytics. Motivated by this issue, this article proposes uncertainty-based SBD analytics, the knowledge extraction from SBD with stage-by-stage uncertainty handling to assure and enhance the reliability of resultant knowledge. Challenges, advances and prospects for selected topics in uncertainty-based SBD analytics, including place-based heuristics and analytics; biases and semantic information inferences of cellphone tracking data; quality assessment of user-generated images and texts, and rectification of the misregistered ones; and uncertainties in deep learning and other sophisticated black-box analytics for SBD, are also discussed in this article.

Tremendous efforts on addressing uncertainties and pursuing more reliable knowledge are demanded to keep pace with the exploding capacity and complexity of SBD analytics. GIScience, like other disciplines, has no means to bypass the data-driven trend and accompanying uncertainties. On the gigantic liner of SBD analytics sailing in either right or wrong directions, that we head ourselves towards benefits and value lies in the reliability of the chart.

Acknowledgement

We sincerely thank the editor Prof. Mei-Po Kwan and the anonymous reviewers for their insightful comments and help on improving this article. Thanks to Pengfei Chen, Lipeng Gao, Zhewei Liu, Pan Shao, Yiling Wan and Xiaokang Zhang for helping provide opinions and collect references for this work.

Funding

This work was supported by the National Natural Science Foundation of China (41331175)，Minister of Science and Technology of P.R. China (2017YFB0503604), and the Hong Kong Polytechnic University (1-ZE24, 4-ZZFZ).

References

Adams, B., and Janowicz, K. 2012. On the geo-indicativeness of non-georeferenced text. In *Proceedings of the Sixth International Conference on Web and Social Media*, Dublin, Ireland, 375–78.

Chen, B., Yuan, H., Li, Q., Wang, D., Shaw, S.-L., Chen, H., and Lam, W.H.K. 2017. Measuring place-based accessibility under travel time uncertainty. *International Journal of Geographical Information Science* 31(4): 783–804.

Chen, J., and Zipf, A. 2017. DeepVGI: Deep learning with volunteered geographic information. In *the 26th International World Wide Web Conference (WWW'17) Companion*, April 3–7, 2017, Perth, Australia, 771–72.

Cheng, T., and Adepeju, M. 2014. Modifiable Temporal Unit Problem (MTUP) and its effect on space-time cluster detection. PLoS ONE 9(6): e100465.

Crandall, D.J., Li, Y., Lee, S., and Huttenlocher, D.P. 2015. Recognizing Landmarks in Large-Scale Social Image Collections. In: Zamir, A.R., et al. (eds.) *Large-Scale Visual Geo-Localization*. Switzerland: Springer, 121–144.

Fan, J., Han, F., and Liu, H. 2014. Challenges of big data analysis. *National Science Review* 1(2): 293–314.

Fekete, E. 2017. Foursquare in the city of fountains: Using Kansas city as a case study for combining demographic and social media data. In Thatcher, J. et al. (eds.) *Thinking Big Data in Geography: New Regimes, New Research*. Lincoln, NE: University of Nebraska Press, 165–88.

Gal, Y. and Ghahramani, Z. 2016. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *Proceedings of The 33rd International Conference on Machine Learning*, 48: 1050–59, 2016.

Gao, S., Janowicz, K., and Couclelis, H. 2017. Extracting urban functional regions from points of interest and human activities on location-based social networks. *Transactions in GIS* 21(3): 446–67.

Gerhardt, B., Griffin, K., and Klemann, R. 2012. Unlocking value in the fragmented world of big data analytics. Cisco Internet Business Solutions Group. http://www.cisco.com/web/about/ac79/docs/sp/Information-Infomediaries.pdf [last accessed 4 August 2017].

Han, J., Zhang, D., Cheng, G., Guo L., and Ren, J. 2015. Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning. *IEEE Transactions on Geoscience and Remote Sensing* 53(6): 3325–37.

Hollenstein, L., and Purves, R. 2014. Exploring place through user-generated content: Using Flickr tags to describe city cores. *Journal of Spatial Information Science* 1: 21–48.

International Business Machines, 2016. *Extracting business value from the 4 V's of big data*. http://www.ibmbigdatahub.com/infographic/extracting-business-value-4-vs-big-data [last accessed 4 August 2017].

Kwan, M.P. 2016. Algorithmic geographies: Big data, algorithmic uncertainty, and the production of geographic knowledge. *Annals of the American Association of Geographers* 106(2): 274–82.

Li, S., Dragicevic, S., Castro, F.A., et al. 2016. Geospatial big data handling theory and methods: A review and research challenges. *ISPRS Journal of Photogrammetry and Remote Sensing* 115(2016): 119–33.

Longley, P.A., and Adnan, M. 2016. Geo-temporal Twitter demographics. *International Journal of Geographical Information Science* 30(2): 369–89.

Miller, H.J., and Goodchild, M.F. 2015. Data-driven geography. *GeoJournal* 80: 449–61.

Openshaw, S. 1983. *The Modifiable Areal Unit Problem*. Norwick: Geo Books.

OpenStreetMap Wiki. 2017. Vandalism. http://wiki.openstreetmap.org/wiki/Vandalism (last accessed 7 August 2017).

Rösler, R., and Liebig, T., 2013. Using data from location based social networks for urban activity clustering. In: Vandenbroucke, D. et al. (eds.) *Geographic Information Science at the Heart of Europe*, Lecture Notes in Geoinformation and Cartography: 55–72.

Senaratne, H., Mobasheri, A., Ali, A.L., Capineri, C., and Haklay, M. 2017. A review of volunteered geographic information quality assessment methods. *International Journal of Geographical Information Science* 31(1): 139–67.

Shi, W., 2010. *Principle of modeling uncertainties in spatial data and spatial analyses*. New York: Taylor & Francis Group/CRC Press.

Thatcher, J. 2014. Living on fumes: Digital footprints, data fumes, and the limitations of spatial big data. *International Journal of Communication* 8: 1765–83.

Tu, W., Cao, J., Yue, Y., Shaw, S.L., Zhou, M., Wang, Z., Chang, X., Xu, Y., and Li, Q. 2017. Coupling mobile phone and social media data: A new approach to understanding urban functions and diurnal patterns. *International Journal of Geographical Information Science* 31(12): 2331–58.

Weyand, T., Kostrikov, I., and Philbin, J. 2016. PlaNet - Photo Geolocation with Convolutional Neural Networks. In: Leibe B. et al. (eds.) *Computer Vision – ECCV 2016, Lecture Notes in Computer Science*, 9912. Cham: Springer, 37–55.

Zhao, Z., Shaw, S.L., Xu, Y., Lu, F., Chen, J., and Yin, L. 2016. Understanding the bias of call detail records in human mobility research. *International Journal of Geographical Information Science* 30(9): 1738–62.

Zhou, M., Yue, Y., Li, Q., and Wang, D. 2017. Portraying temporal dynamics of urban spatial divisions with mobile phone positioning data: A complex network approach. ISPRS International Journal of Geo-Information 5(240).

Zhou, X., and Zhang, L. 2016. Crowdsourcing functions of the living city from Twitter and Foursquare data. *Cartography and Geographic Information Science* 43(5): 393–404.

Zielstra, D., and Hochmair, H.H. 2013. Positional accuracy analysis of Flickr and Panoramio images for selected world regions. *Journal of Spatial Science* 58(2): 251–273.

Author Biographies

WENZHONG SHI is the Head and Chair Professor of Geographical Information Science and Remote Sensing in the Department of Land Surveying and Geo-informatics at the Hong Kong Polytechnic University, Hung Hom, Hong Kong, P.R. China. Email: lswzshi@polyu.edu.hk. His research interests include GIScience and remote sensing, with focusing on uncertainties and quality control of spatial data, satellite images and LiDAR data, 3D modelling, and human dynamics.

ANSHU ZHANG is a Postdoctoral Fellow in the Department of Land Surveying and Geo-informatics at the Hong Kong Polytechnic University, Hung Hom, Hong Kong, P.R. China. E-mail: anshu.zhang@connect.polyu.hk. Her research interests include spatial data mining, human dynamics, and machine learning for human geography.

XIAOLIN ZHOU is a PhD student in the Department of Land Surveying and Geo-informatics at the Hong Kong Polytechnic University, Hung Hom, Hong Kong, China. E-mail: xiaolin.zhou@connect.polyu.hk. Her research interests include GIScience, location-based social networks, and commercial site selection.

MIN ZHANG is a PhD student in the School of Remote Sensing and Information Engineering at Wuhan University, Wuhan, Hubei, P.R. China. E-mail: zincmin@gmail.com. His research interests include GIScience, spatial data quality, change detection with satellite images, and deep learning for remote sensing.