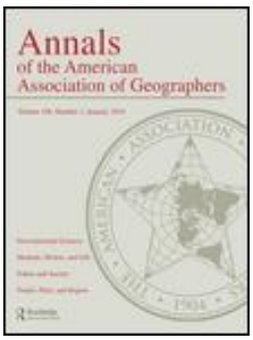

Annals of the American Association of Geographers

ISSN: 2469-4452 (Print) 2469-4460 (Online) Journal homepage: http://www.tandfonline.com/loi/raag21

Algorithmic Geographies: Big Data, Algorithmic Uncertainty, and the Production of Geographic Knowledge

Mei-Po Kwan

To cite this article: Mei-Po Kwan (2016) Algorithmic Geographies: Big Data, Algorithmic Uncertainty, and the Production of Geographic Knowledge, Annals of the American Association of Geographers, 106:2, 274-282

To link to this article: http://dx.doi.org/10.1080/00045608.2015.1117937

Published online: 09 Feb 2016.

Algorithmic Geographies: Big Data, Algorithmic Uncertainty, and the Production of Geographic Knowledge

Mei-Po Kwan

Department of Geography and Geographic Information Science, University of Illinois at Urbana–Champaign

Drawing on examples from human mobility research, I argue in this article that the advent of big data has significantly increased the role of algorithms in mediating the geographic knowledge production process. This increased centrality of algorithmic mediation introduces much more uncertainty to the geographic knowledge generated when compared to traditional modes of geographic inquiry. This article reflects on important changes in the geographic knowledge production process associated with the shift from using traditional "small data" to using big data and explores how computerized algorithms could considerably influence research results. I call into question the much touted notion of data-driven geography, which ignores the potentially significant influence of algorithms on research results, and the fact that knowledge about the world generated with big data might be more an artifact of the algorithms used than the data itself. As the production of geographic knowledge is now far more dependent on computerized algorithms than before, this article asserts that it is more appropriate to refer to this new kind of geographic inquiry as algorithm-driven geographies (or algorithmic geographies) rather than data-driven geography. The notion of algorithmic geographies also foregrounds the need to pay attention to the effects of algorithms on the content, reliability, and social implications of the geographic knowledge these algorithms help generate. The article highlights the need for geographers to remain attentive to the omissions, exclusions, and marginalizing power of big data. It stresses the importance of practicing critical reflexivity with respect to both the knowledge production process and the data and algorithms used in the process. *Key Words: algorithms, algorithmic geographies, big data, geographic knowledge, human mobility.*

我运用人类能动性研究的案例, 于本文中主张, 大数据的出现, 已显着地增加了演算法在中介地理知识生产过程中的角色。相较于传统的地理探问模式而言, 演算法中介的中心性之强化, 为地理知识生产带来了更多的不确定性。本文反思从运用传统的 "小数据" 转而运用大数据的地理知识生产过程中的重要改变, 并探讨电脑化的演算法如何能够大幅影响研究结果。我质问 "数据驱动的地理" 此一备受吹捧之概念, 该概念忽略了演算法对于研究结果所具有的潜在显着影响, 以及透过大数据生产的世界知识, 或许较数据本身而言更像演算法的人工产物之事实。随着当今地理知识的生产较以往更为依赖电脑化的演算法, 本文主张, 将此般新形式的地理探问指称为 "演算法驱动的地理" (或演算地理) , 相较于 "数据驱动的地理" 之指称更为合适。演算地理的概念, 同时凸显出必须关注演算法对于其所促成的地理知识的内容、可信度与社会意涵的效应。本文强调地理学者必须持续留意大数据的遗漏、排除与边缘化的力量, 并着重对于知识生产过程以及该过程中使用的数据和演算法, 进行批判性反思的重要性。 *关键词: 演算法, 演算地理, 大数据, 地理知识, 人类能动性。*

Con base en ejemplos de investigación sobre movilidad humana, sostengo en este artículo que el advenimiento de los *big data* ha incrementado significativamente el papel de los algoritmos para mediar el proceso de producción de conocimiento geográfico. Esta creciente centralidad de la mediación algorítmica introduce un mayor grado de incertidumbre en el conocimiento geográfico generado cuando se la compara con los modos tradicionales de pesquisa geográfica. Este artículo reflexiona sobre las transformaciones importantes del proceso de producción de conocimiento geográfico asociados con el cambio de usar "datos pequeños" tradicionales por el uso de datos mayores, y explora la manera como los algoritmos computarizados podrían influenciar considerablemente los resultados de la investigación. Cuestiono la noción muy publicitada de la geografía orientada por datos, que ignora la influencia potencialmente significativa de los algoritmos en los resultados de la investigación, y el hecho de que el conocimiento acerca del mundo generado con *big data* podría ser más un artefacto de los algoritmos usados que los propios datos. Como la producción de conocimiento geográfico es ahora mucho más dependiente de algoritmos computarizados que antes, este artículo afirma que es mucho más apropiado referirnos a este nuevo tipo de pesquisa geográfica como geografía de orientación algorítmica (o geografías algorítmicas) que geografía orientada por datos. La noción de geografías algorítmicas también pone en primer plano la necesidad de dar atención a los efectos de los algoritmos en el contenido, confiabilidad e implicaciones del

Annals of the American Association of Geographers, 106(2) 2016, pp. 274–282 © 2016 by American Association of Geographers
Initial submission, February 2015; revised submissions, August and October 2015; final acceptance, October 2015
Published by Taylor & Francis, LLC.

conocimiento geográfico que estos algoritmos ayudan a generar. El artículo resalta la necesidad que tienen los geógrafos de permanecer atentos a las omisiones, exclusiones y poder marginador de los *big data*. Enfatiza la importancia de practicar la reflexividad crítica con respecto del proceso de producción de conocimiento y de los datos y algoritmos utilizados en ese proceso. *Palabras clave: algoritmos, geografías algorítmicas, big data, conocimiento geográfico, movilidad humana.*

Although much has been written about the advent of big data, the implications of using big data for the generation of geographic knowledge are still far from clear. One of the most important elements missing in this discussion to date, with a few exceptions (e.g., Kitchin 2014b), is the role of algorithms in generating, processing, and analyzing big data in the process of geographic knowledge production. Although algorithms have been used to handle and analyze geographic data for decades, there are indications that the process of geographic knowledge production is increasingly mediated by computerized algorithms with the emergence of big data. Such an increase in algorithmic mediation introduces much more uncertainty to the geographic knowledge generated when compared to traditional modes of geographic inquiry. Drawing on examples from research on human mobility and activity-travel patterns and with a focus on how geoprocessing algorithms could influence research results, this article discusses various sources of algorithmic uncertainty. It reflects on important changes in the geographic knowledge production process associated with the shift from using traditional "small data" to using big data. I call into question the much touted notion of data-driven geography, which neglects the potentially significant influence of algorithms on research results and the fact that knowledge about the world generated with big data might be more an artifact of the algorithms used than the data itself. As the production of geographic knowledge is now far more dependent on computerized algorithms than before, this article asserts that it is more appropriate to refer to this new kind of geographic inquiry as algorithm-driven geographies (or algorithmic geographies) rather than data-driven geography. Through the notion of algorithmic geographies, the article also foregrounds the need to pay attention to the effects of algorithms on the content, reliability, and social implications of the geographic knowledge these algorithms help generate. I highlight the need for geographers to remain attentive to the omissions, exclusions, and marginalizing power of big data. I stress the importance of practicing critical reflexivity with respect to both the knowledge production process and the data and algorithms used in the process.

The Algorithmic Mediation of Geographic Knowledge Production

In the long chains of events that happen before results in geographic research are obtained, many sets of procedures need to be implemented to collect, process, and analyze relevant data, which could be qualitative or quantitative data. Some of these procedures are or can be performed manually (e.g., transferring data from survey instruments to digital data files), whereas many are implemented as computerized procedures because it would be extremely tedious and time consuming to perform them manually on most empirical geographic data sets. This article refers to these sets of procedures for collecting, processing, and analyzing data in geographic research as algorithms, which are well-defined sequences of steps for solving problems or performing specific tasks with or without computerized implementations.

It should be noted that implementation as computerized procedures (in the form of computer code or software) is not a necessary condition for a set of procedures to be considered an algorithm, contrary to the definition used in computer science or in some recent works by geographers (e.g., Graham and Shelton 2013; Kitchin 2014b). For instance, the shortest path between a source node and a destination node in a small network can be found using Dijkstra's algorithm without implementing any computerized procedures. It should also be noted that although many analytical methods might be considered conceptually separate from and can be described without referring to the algorithms that implement them (e.g., the Moran's *I* or accessibility measures can be expressed as mathematical equations), some other methods can only be expressed in the form of specific algorithms (e.g., Dijkstra's algorithm and evolutionary algorithms; Kwan, Xiao, and Ding 2014). It is thus often impossible to maintain a clear conceptual distinction among methods, procedures, techniques, and algorithms.

Further, because geographic data concern a variety of geographic entities and relations, algorithms are used to perform not only mathematical computation but also complex geoprocessing operations on georeferenced data (e.g., line simplification, spatial search,

spatial interpolation, surface modeling, or identifying the topological relationships between two geographic objects; Shi 2010; Xiao 2016). In addition, geoprocessing algorithms are often necessary for representing geographic data at suitable spatial and temporal scales using appropriate data models before any analysis can be performed (e.g., addressing geocoding or digital elevation models). The data used in much of geographic research are thus the product of prior processing using a wide variety of algorithms. It is therefore important to recognize that algorithms are an essential and integral element of geographic data.

Because no results from geographic research involving data can be generated without using algorithms, the production of geographic knowledge derived from data is necessarily mediated by algorithms even before the widespread use of computerized procedures. This is a fundamental reality of the geographic knowledge production process. When algorithms or procedures are applied to generate, process, and analyze geographic data, some uncertainty or error might be introduced and research findings might differ slightly, or even considerably, depending on the specific algorithms used. For example, an algorithm identified the wrong street segments for about 20 percent of the trips in a Global Positioning System (GPS) data set (Gong et al. 2012). Further, different algorithms that implement the same analysis or different implementations of the same algorithm could lead to different results, and the differences in research findings might vary considerably. As cogently demonstrated by Fisher's (1993) study, there can be more than 50 percent difference in the results obtained with different viewshed analysis algorithms. Even the computer languages, compilers, and computational platforms used might introduce some differences to the results generated with the same algorithm due to different processor precision and methods of handling interim values. Algorithmic uncertainty is thus an essential element in the geographic knowledge production process due to the use of different sets of procedures, implementations, data environments (data model and data structure), or computational platforms. These uncertainties are often magnified in big data research.

Algorithms and Traditional Data in Human Mobility Research

For decades, human mobility studies conducted by geographers and transport researchers collected the needed data largely through activity-travel diary surveys. These traditional data sets were obtained with custom-designed survey instruments. They were created to answer specific questions about human activity-travel patterns based on some prior theoretical understanding of these patterns. These data sets tended to be small to moderate in size (with several hundred to several thousand participants) and were often collected with specific sample designs that seek to obtain representative samples of a population. Although activity-travel diary data sets are costly and time consuming to collect, they contain highly detailed information about participants' sociodemographic attributes and activity-travel behavior (e.g., activity purpose and location, start and end time of activities, travel mode, and travel route) that enables rich description and analysis of their mobility patterns.

Because these data sets are not huge, computerized procedures were typically not necessary and were often not used in the data generation and preparation phases. For instance, data tend to be transferred from survey instruments to digital data files manually. Little algorithmic uncertainty is introduced by computerized data transformation operations because definite mathematical relationships govern how new variables are generated. In addition, it is still possible to examine particular data records or items to check and clean the data as well as to address specific data quality issues (e.g., missing data or anomalies such as outliers) via researchers' experience and expertise. For instance, in past research I have contacted research participants to rectify missing or inconsistent data in their surveys and corrected errors in data spreadsheets with hundreds of records. Further, even when computerized algorithms are implemented to analyze traditional "small data," it is still quite feasible to examine the effects of different algorithms on research results. This is because it is not prohibitively costly or time consuming to rerun statistical tests or analyses using different algorithms and to use researchers' experience to identify and correct errors in analytical procedures or results.

Thus the algorithmic mediation of the geographic knowledge production process is limited when using traditional "small" data sets (especially in the data generation and preparation phases, as most of the tasks during these phases can be performed manually). Meanwhile, computerized procedures or algorithms are mainly used in the data analysis and modeling phase when using these traditional data sets (except when analyzing most qualitative data). It is important to note that when data are or can be handled manually, researchers and data interact more directly, and both the data and

procedures are more tangible and visible in the research process. This is especially true for procedures that handle or analyze qualitative data, as researchers or their associates often need to perform these tasks themselves (e.g., coding or interpreting interview transcripts) instead of relegating them to computerized procedures (note that this is true even when computer-aided qualitative data analysis software is used).

The Advent of Big Data

In recent years, the widespread use of location-aware technologies, mobile devices, and social media has made it possible to assemble huge amounts of data about people's location and movement from various sources (e.g., cell phone companies, public transit and taxi companies, real-time GPS/geographic information system (GIS) functionalities in mobile devices, Internet search engines, and social media providers). The rapid increase in the volume, diversity, and intensity of data from these sources has led to the emergence of big data and stimulated new developments in human mobility research. Big data are not just massive in volume (e.g., about 10 million geotagged tweets are generated every day); it is generated continuously at high velocity and high space–time resolution (Richardson et al. 2013).

Although big data seem promising for advancing human mobility research, how algorithms might influence the data that researchers obtain and the research findings they report have received very little attention to date. A fundamental fact about big data should be noted, however: No big data can be generated without using some computerized procedures or algorithms (e.g., searching, selecting, or ranking using specific parameters, such as the PageRank algorithm used by Google to generate search results). Thus, no big data or research results obtained using big data are unmediated by algorithms. An instructive example of how big data are the result of prior processing before reaching the public or research community is provided by Fischer (2014). In the process of developing a software tool for mapping geotagged tweets from Twitter, he observed a banding phenomenon: The original tweet locations tend to align with the closest latitude or longitude. This suggests that tweet locations might have been fuzzed by Twitter through snapping them to the closest latitude or longitude to prevent people's exact locations being disclosed. Further, the study observed a strange phenomenon of missing data at the Prime Meridian when zooming in on London and suggested that Twitter might have filtered out the tweets on the Prime Meridian for some reason. Two algorithms were then implemented to address these issues and to make the tweet maps look more natural. As a result, every map generated by the mapping tool is mediated by Fischer's own algorithms, Twitter's privacy-protection algorithms, or both.

An important change in the geographic knowledge production process associated with the increasing use of big data is the greatly expanded use of computerized algorithms, which have become necessary even in the data collection and generation phase. As algorithms are increasingly implemented as computerized procedures, many of which are now relegated to computer programmers, they become increasingly detached from and less visible to researchers who use them. One important reason for this trend is the limited informational content of big data. For instance, many variables needed for addressing specific questions about human mobility (e.g., home and workplace location, travel route, travel mode, gender, income, and race) are often not available in popular big data sets. Algorithms are thus needed to infer their values indirectly from the big data used (e.g., the filtering and clustering algorithms used in Widhalm et al. [2015] to infer trip sequences from cell phone data). In addition, algorithmic uncertainty has increased because checking raw data with researchers' experience and expertise, rerunning analyses, or comparing the effects of different algorithms on research results are often infeasible or prohibitively costly, given the huge volume, complexity, and dynamic nature of big data.

Another important factor contributing to increased algorithmic uncertainty when using big data is that search algorithms used to generate data could change over time (algorithmic dynamics). Lazer et al. (2014) presented a highly instructive example that illustrates the unstable and changing nature of the algorithms used in the generation of big data as well as the significant influence of algorithms on the knowledge produced. In early 2013, Google Flu Trends (GFT), the flu tracking system created by Google, made significant prediction error about influenza-related doctor visits in the United States. Lazer et al. (2014) explained that Google search algorithms are constantly being modified by its engineers to improve its service and that "GFT was an unstable reflection of the prevalence of the flu because of algorithm dynamics affecting Google's search algorithm" (1204). The authors further highlighted the fact that because the algorithms underlying Google, Twitter, and Facebook are always being modified and constantly changing, it is far from

clear "whether studies conducted even a year ago on data collected from these platforms can be replicated in later or earlier periods" (Lazer et al. 2014, 1204). Thus, the production of geographic knowledge in the big data era is far more uncertain and affected by algorithms than before. The section that follows examines issues concerning the use of computerized algorithms in human mobility research in greater detail.

Algorithms and Human Mobility Research Using Big Mobile Phone Data Sets

Popular sources of big data for human mobility research include GPS tracks of vehicles, public transit smart card data, GPS data from bicycle sharing programs, and social media data (e.g., Ma et al. 2013; Corcoran and Li 2014; Hawelka et al. 2014). Among them, passive cell phone data, which are recorded automatically by cell phone companies without implementing additional data collection procedures such as GPS tracking, is perhaps the most popular because of its advantages over conventional surveys (e.g., González, Hidalgo, and Barabasi 2008; Ahas et al. 2010; Bayir, Demirbas, and Eagle 2010; Silm and Ahas 2014; Widhalm et al. 2015). For instance, when phone companies are willing to provide these data at reasonable prices, researchers can have cost-effective access to very large numbers of communication records (e.g., over 1 billion) from a large number of users (e.g., several million users) that cover a high proportion of the population over large areas and long periods of time (e.g., six months or one year; Song et al. 2010; Wang, Chen, and Ma 2014). Researchers also do not need to worry about the problem of sample attrition over time because most data provided by phone companies are kept for a long period of time. In light of these advantages, big cell phone data sets have become a major data source in recent human mobility research.

There has been little discussion to date, however, about how algorithms might affect the data or research results when big cell phone data sets are used. Algorithms for processing and analyzing big mobile phone data sets are necessary largely because of their limited informational content. For instance, the actual location of users is not recorded and thus is unknown in passive cell phone data. Instead, the recorded location is the geographic location of the serving cell towers that handled users' communication activities (e.g., a call or a text message). Further, unless complemented by data obtained directly from individuals through surveys (e.g., Licoppe et al. 2008), people's activity location, activity duration, activity purpose, travel route, travel mode, and other trip characteristics (e.g., trip distance) are unknown and can only be inferred using algorithms. For instance, whether an individual is staying at a location instead of moving around needs to be inferred from the data based on spatial and temporal constraints that identify a sequence of consecutive cell phone records as an activity location. A spatial constraint sets the roaming distance within which a user is considered staying at a location. A temporal constraint sets the minimum duration a user needs to spend at a location for it to be considered an activity location (instead of moving around). For instance, Jiang et al. (2013) used a 300-m roaming distance as the spatial constraint and a temporal constraint of ten minutes to identify certain stay points as activity locations. Because different constraints and parameters can be used in these inferential algorithms, slightly or even considerably different patterns of activity location could be observed, depending on the exact algorithms and constraints used.

Similarly, both trip distance and activity need to be inferred from the data using algorithms and specific parameters. When trip distance is estimated using inferred activity locations, considerable positional error can occur. For instance, for one respondent in Ahas et al. (2010), the home location inferred from cell phone data is 830 m from the real home location, and the inferred workplace location is 300 m from the real workplace location. Although this does not seem to be a lot of error, the home–work distance derived from these two inferred locations is about 1 km longer than the real home–work distance. Given that the original home–work distance is about 2 km (estimated based on a visual examination of Figure 4 in Ahas et al. 2010), this amounts to a 50 percent error. The potential uncertainty introduced by algorithms used to process big cell phone data sets could thus be considerable. The difficulty is, unlike with traditional data sets, estimating and correcting this kind of positional error are prohibitively costly and often infeasible when there are millions of records in the data set.

Information about the activity being performed when data are recorded is available only for certain types of data sets (e.g., people are riding taxis in taxi-tracking data sets). For big cell phone data sets, we know almost nothing about what people are actually doing when they performed various communication events. Some studies have used algorithms to infer people's most likely activities at different locations based

on the space–time characteristics of the activity (e.g., a long stay at a location from the evening to early morning as staying home and a long stay at a location during the day as work). Some studies used algorithms based on land use data to infer the activity being performed when people's communication events took place. Even if a person is located in a particular type of land use (e.g., commercial or recreational land use), though, we still do not know and cannot verify what the person is actually doing because many different activities can be performed in a given type of land use (e.g., shopping, dining, and running errands can happen in commercial land use). This is particularly true as people can now perform a wide range of activities in the same type of land use or at the same place with the greatly increased use of information and communication technologies (Kwan 2007; Couclelis 2009).

Further, although human mobility research using big cell phone data sets often represents people's movement trajectories as if they can be directly observed from the data, these data sets do not record people's actual location or movement over space and time. The location associated with a particular communication event in passive mobile phone records is actually the location of the cell tower that handled that event. The movement trajectories derived from big cell phone data sets are thus not the actual movement trajectories of phone users but the paths connecting consecutive cell towers that handled users' communication activities. To reconstruct these paths, algorithms are used to infer whether a phone user was moving or not and the likely routes traversed. Because the routes of movement are inferred and are just arbitrary lines connecting consecutive serving cell towers, trip distance cannot be accurately estimated and such use of movement inference algorithms introduces an unknown amount of uncertainty into analytic results.

Several issues arise when using algorithms to infer the movement trajectories of cell phone users in big data mobility research. For example, movement is detected only when the current serving cell tower of a phone shifts to another one. Short trips that do not lead to such a cell tower shift (i.e., when the actual movement is not long enough to shift the serving cell tower to another one), however, will not be recorded in the data set. Because most human trips are over a short distance, considerable measurement error can occur when using big cell phone data sets. In fact, recent studies have observed that mobility measures derived from big cell phone data sets using cell tower for location tend to underestimate people's daily travel distance when compared to those obtained from GPS data (Wang, Chen, and Ma 2014).

Other issues arise when cell phone users make short trips across cell tower service areas or when they are located in the overlapping service areas of two or more cell towers. In the former case, a short trip might be recorded as a much longer one because the recorded movement distance is the distance between the serving cell towers instead of the actual trip distance. In the latter case, when a phone is located in the overlapping service areas of two or more cell towers, a cell phone might switch or oscillate between neighboring cell towers for load balancing or signal strength optimization (Ahas et al. 2010; Wang, Chen, and Ma 2014). Referred to as the *ping-pong effect*, this kind of oscillation was also observed in Wi-Fi networks (Bayir, Demirbas, and Eagle 2010). When this happens, a phone user, although not moving, could appear to have traveled back and forth for several kilometers in just a few seconds. Algorithms are thus necessary to assign a phone user to the most likely cell tower, but it remains unclear which cell tower better approximates the actual location of the phone user.

To highlight the algorithmic uncertainty involved due to the need to use algorithms to infer unavailable information, I provide two examples in what follows to illustrate how the precise movement trajectories obtained from passive big cell phone data can be affected by the inferential algorithms and their interactions with particular data environments. In Figures 1 and 2, the actual movement trajectory of an individual is represented by the solid black arrow that runs from

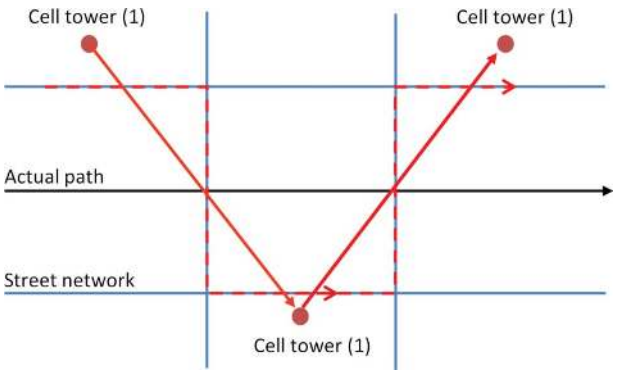

Figure 1. The effects of different trajectory inference algorithms on the inferred distance and movement trajectory with cell tower Set 1 (red dots). The actual movement trajectory is represented by the solid black arrow that runs from left to right across the center of the figure. This movement took place along one of the streets of the local street network, which is represented by a grid of light blue lines. (Color figure available online.)

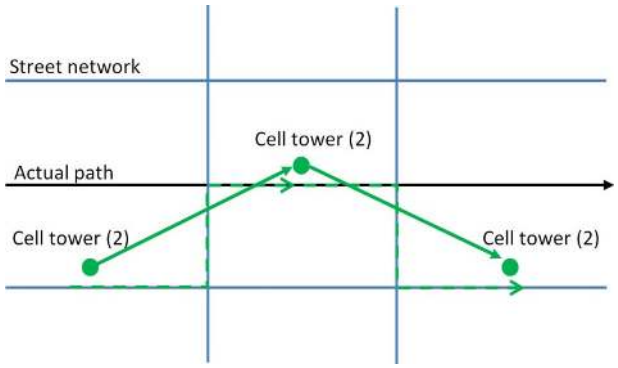

Figure 2. The effects of different trajectory inference algorithms on the inferred distance and movement trajectory with cell tower Set 2 (green dots). The actual movement trajectory is represented by the solid black arrow that runs from left to right across the center of the figure. This movement took place along one of the streets of the local street network, which is represented by a grid of light blue lines. (Color figure available online.)

left to right across the center of the figures. This movement took place along one of the streets of the local street network, which is represented by a grid of light blue lines. Two possible geographic distributions of cell towers are represented: cell tower Set 1 in Figure 1 is represented as red dots, and cell tower Set 2 in Figure 2 is represented as green dots.

Let us first consider Figure 1. If the trajectory inference algorithm uses cell towers as a proxy of the phone user's actual location, then the inferred movement trajectory obtained will be the solid red line. If the algorithm attempts to take into account actual movement possibilities in the area and snaps the movement path to the street network, then the inferred movement trajectory obtained will be the red dashed line. Note that both of these lines are different from and longer than the actual movement trajectory represented by the solid black arrow that runs from left to right across the center of Figure 1. This means that the inferred travel distance is longer than the actual travel distance, no matter which trajectory inference algorithm is used. Note that even when the algorithm uses the real street network to better approximate people's actual movement paths, this might be helpful only under specific conditions—for example, when the street network is not dense or when there is only one possible transport route in the study area, as illustrated by Gong et al. (2012), where wrong street segments were identified for about 20 percent of trips due to the dense street network of the study area.

A comparison of Figures 1 and 2 highlights how the location of cell towers in the study area could affect both the inferred movement trajectory and inferred travel distance even when using the same algorithm. Although the actual movement is the same (depicted by the solid black arrow that runs from left to right across the center of both Figures 1 and 2), either the red or green solid lines are inferred as the movement trajectories, depending on which set of cell towers is present in the study area. Note that the inferred travel distance based on the red cell towers is longer than that obtained with the green cell towers, and both of these inferred distances are longer that the actual movement distance represented by the solid black arrow, independent of whether the trajectory inference algorithm takes the street network into account or not. These examples show how algorithms and their interactions with particular data environments (cell tower location) could introduce a considerable amount of uncertainty to the inferred movement trajectories and inferred travel distance when using big cell phone data sets.

Toward Reflexive Algorithmic Geographies

No geographic knowledge derived from data can be created without using algorithms, which are sets of procedures for collecting, generating, processing, and analyzing data. All knowledge created with any data is thus necessarily mediated by algorithms, which might or might not be implemented as computerized procedures and are thus more than the computer code or software that implements them. Understanding algorithms this way renders them amenable even to research with qualitative methods and to the practice of critical reflexivity discussed later.

Using computerized algorithms in geographic research is not new. They have been used in studies that collected and used traditional data sets of small to moderate size. The advent of big data, however, significantly increases the role of algorithms in the geographic knowledge production process, especially in the data generation and preparation phases. Drawing from examples from human mobility research with big cell phone data sets, this article shows that many more procedures in the geographic knowledge production process are now performed by computerized algorithms, frequently due to the need to infer variables that are not directly available in big data sets. An important consequence of this trend is a considerable potential increase in the algorithmic uncertainty in the knowledge created, as many algorithms introduce uncertainty

that will propagate in the process, leading to slightly or even significantly different findings. Further, as the algorithmic mediation of the knowledge production process increases, the precise ways in which data are generated, processed, and analyzed tend to become increasingly invisible to and detached from researchers.

Many have argued that the advent of big data is leading to the rise of a new paradigm of scientific inquiry (the fourth paradigm) and a new mode of data-driven knowledge discovery, which entails a shift from deductive forms of inquiry (based on the sequence of hypothesis formulation, data collection, and analysis) toward more inductive and emergent forms of analysis that allow the data to speak for itself (Kitchin 2014a). Following this line of thinking, some geographers have begun to advocate the data-driven generation of geographic knowledge based on the latest advances in big data geographic information science, spatial data mining, and visual analytics. Yet, the notion of data-driven geography is misleading and untenable. It ignores the potentially significant influence of algorithms on research results and the fact that knowledge about the world generated with big data might be more an artifact of the algorithms used than the data itself (Lazer et al. 2014). As the examples of this article illuminate, the existence of pristine and pure big data is largely a myth because the generation of big data itself necessitates the use of computerized algorithms, not to mention further processing and analysis. Thus, no big data can reach researchers or the public untainted by some algorithmic uncertainty. Further, as the production of geographic knowledge is now far more dependent on and affected by the algorithms used in the process, it seems appropriate to refer to this new kind of geographic inquiry as algorithm-driven geographies (or algorithmic geographies) rather than data-driven geography.

The notion of algorithmic geographies foregrounds the need for geographers to pay attention to the effects of algorithms on the content, reliability, and social implications of the geographic knowledge these algorithms help generate. It alerts us to the perils in elevating the promise of big data at the expense of ignoring critical issues concerning the scientific and social consequences of the knowledge generated with such data. It also highlights the need for geographers to mitigate the tendency and consequences of becoming increasingly detached from both algorithms and data and to remain attentive to the omissions, exclusions, and marginalizing power they entail. It stresses the imperative for us to practice critical reflexivity with respect to both the knowledge production process and the data used in the process. Geographers need to recognize that algorithms might have played an important role in determining their research results. We also need to be aware of how using big data could lead us to address questions that are less central to pressing societal concerns when compared to using traditional data. For instance, past studies using traditional data are often interested in questions concerning how social difference such as gender or race affects people's mobility experience, but these kinds of questions cannot be addressed by big data due to their lack of detailed sociodemographic information.

To address the scientific and social consequences of algorithmic uncertainty when using big data in geographic research, several steps can be undertaken to practice critical reflexivity: (1) evaluate how different algorithms and their interactions with data might lead to different results, omissions, and exclusions; (2) assess the amount of algorithmic uncertainty involved, how much confidence about the research findings is warranted, and whether it is acceptable with respect to the research questions and geographic scale of the study (e.g., error of 1 km or less may be tolerated for studies on long-distance intercity travel but might not be acceptable for research on intraurban travel where people make a lot of short trips); (3) examine or validate the algorithms using smaller subsets of the data that have been enriched with additional information; (4) complement big data by traditional data, especially with regard to information that is not available in big data sets but is often obtained directly from research participants; (5) evaluate whether the algorithms are capable of revealing the effects of interpersonal and social difference such as gender, race, and class (e.g., some accessibility measures just mimic the spatial patterns of urban opportunities and are not capable of capturing interpersonal differences in individual accessibility; see Kwan 1998); and (6) assess the stability of the algorithms used to generate the data and its implications for the replicability of the findings.

In the final analysis, geographers need to proceed with great caution when using big data in their research. It is important to bear in mind that "big data sets do not, by virtue of their size, inherently possess answers to interesting questions" (Reich 2015, 34). Using data sets of enormous size "does not mean that one can ignore foundational issues of measurement. . . . The core challenge is that most big data that have received popular attention are not the output of instruments designed to produce valid and reliable data amenable for scientific analysis" (Lazer et al. 2014, 1203).

Acknowledgments

I thank Tim Schwanen for handling the blind review process and making editorial decisions for this article (including the abstract). His helpful suggestions and the thoughtful comments of three anonymous reviewers have helped improve the article considerably.

Funding

This article was written while I was supported by grants NSF IIS-1354329, NSF BCS-1244691, and NSFC-41529101.

References

Ahas, R., S. Silm, O. Järv, E. Saluveer, and M. Tiru. 2010. Using mobile positioning data to model locations meaningful to users of mobile phones. *Journal of Urban Technology* 17 (1): 3–27.

Bayir, M. A., M. Demirbas, and N. Eagle. 2010. Mobility profiler: A framework for discovering mobility profiles of cell phone users. *Pervasive and Mobile Computing* 6 (4): 534–54.

Corcoran, J., and T. Li. 2014. Spatial analytical approaches in public bicycle sharing programs. *Journal of Transport Geography* 41:268–71.

Couclelis, H. 2009. Rethinking time geography in the information age. *Environment and Planning A* 41 (7): 1556–75.

Fischer, E. 2014. Making the most detailed tweet map ever. https://www.mapbox.com/blog/twitter-map-every-tweet/ (last accessed 8 October 2015).

Fisher, P. F. 1993. Algorithm and implementation uncertainty in viewshed analysis. *International Journal of Geographical Information Science* 7 (4): 331–47.

Gong, H., C. Chen, E. Bialostozky, and C. T. Lawson. 2012. A GPS/GIS method for travel mode detection in New York City. *Computers, Environment and Urban Systems* 36 (2): 131–39.

González, M. C., C. A. Hidalgo, and A. L. Barabasi. 2008. Understanding individual human mobility patterns. *Nature* 453 (7196): 779–82.

Graham, M., and T. Shelton. 2013. Geography and the future of big data, big data and the future of geography. *Dialogues in Human Geography* 3 (3): 255–61.

Hawelka, B., I. Sitko, E. Beinat, S. Sobolevsky, P. Kazakopoulos, and C. Ratti. 2014. Geo-located Twitter as a proxy for global mobility patterns. *Cartography and Geographic Information Science* 41 (3): 260–71.

Jiang, S., G. A. Fiore, Y. Yang, J. Ferreira, E. Frazzoli, and M. C. González. 2013. A review of urban computing for mobile phone traces: Current methods, challenges and opportunities. Paper presented at the ACM SIGKDD International Workshop on Urban Computing, Chicago.

Kitchin, R. 2014a. Big data, new epistemologies and paradigm shifts. *Big Data and Society* 1 (1): 1–12.

———. 2014b. Thinking critically about and researching algorithms. Article presented at the Programmable City Working, National University of Ireland, Maynooth, Ireland. http://articles.ssrn.com/sol3/articles.cfm?abstract_id=2515786 (last accessed 8 October 2015).

Kwan, M.-P. 1998. Space–time and integral measures of individual accessibility: A comparative analysis using a point-based framework. *Geographical Analysis* 30 (3): 191–216.

———. 2007. Mobile communications, social networks, and urban travel: Hypertext as a new metaphor for conceptualizing spatial interaction. *The Professional Geographer* 59 (4): 434–46.

Kwan, M.-P., N. Xiao, and G. Ding. 2014. Assessing activity pattern similarity with multidimensional sequence alignment based on a multiobjective optimization evolutionary algorithm. *Geographical Analysis* 46 (3): 297–320.

Lazer, D., R. Kennedy, G. King, and A. Vespignani. 2014. The parable of Google Flu: Traps in big data analysis. *Science* 343:1203–05.

Licoppe, C., D. Diminescu, Z. Smoreda, and C. Ziemlicki. 2008. Using mobile phone geolocalisation for "socio-geographic" analysis of co-ordination, urban mobilities, and social integration patterns. *Tijdschrift voor Economische en Sociale Geografie* 99 (5): 584–601.

Ma, X., Y.-J. Wu, Y. Wang, F. Chen, and J. Liu. 2013. Mining smart card data for transit riders' travel patterns. *Transportation Research Part C* 36:1–12.

Reich, J. 2015. Rebooting MOOC research. *Science* 347 (6217): 34–35.

Richardson, D. B., N. D. Volkow, M.-P. Kwan, R. M. Kaplan, M. F. Goodchild, and R. T. Croyle. 2013. Spatial turn in health research. *Science* 339 (6126): 1390–92.

Shi, J. 2010. *Principles of modeling uncertainties in spatial data and spatial analyses*. Boca Raton, FL: CRC.

Silm, S., and R. Ahas. 2014. Ethnic differences in activity spaces: A study of out-of-home nonemployment activities with mobile phone data. *Annals of the Association of American Geographers* 104 (3): 542–59.

Song, C., Z. Qu, N. Blumm, and A.-L. Barabási. 2010. Limits of predictability in human mobility. *Science* 327 (5968): 1018–21.

Wang, T., C. Chen, and J. Ma. 2014. Mobile phone data as an alternative data source for travel behavior studies. Paper presented at the Transportation Research Board 93rd annual meeting, Washington, DC.

Widhalm, P., Y. Yang, M. Ulm, S. Athavale, and M.C. González. 2015. Discovering urban activity patterns in cell phone data. *Transportation* 42:597–623.

Xiao, N. 2016. *GIS algorithms*. London: Sage.

MEI-PO KWAN is a Professor in the Department of Geography and Geographic Information Science at the University of Illinois at Urbana–Champaign, Champaign, IL 61820. E-mail: mpk654@gmail.com. Her research interests include GIScience, environmental health, human mobility, sustainable transport and cities, and GIS methods in geographic research.