Title
Study on the Effects of Human Intention on Spatial Data Quality

Journal
International Conference on GIScience Short Paper Proceedings, 1(1)

Author
Zhao, Bo

Publication Date
2016

DOI
10.21433/B31180p357ks

Peer reviewed

Study on the Effects of Human Intention on Spatial Data Quality

Bo Zhao

College of Earth, Ocean, and Atmospheric Sciences, Oregon State University, 101 SW 26th St, Corvallis, OR 97331
Email: zhao2@oregonstate.edu

Abstract

The GIScience community has devoted considerable research to locational uncertainty rather than other types of spatial data quality issues. Today, the flood of spatial big data has brought about new concerns, such as location spoofing, GPS jamming, or AIS (Automatic Identification System) hacking. Yet, the current data quality assessment framework falls short in defining, interpreting and analyzing these critical issues. By examining the reasons for measuring the geographic world, I suggest a modification of the hypothesis of the rational geographer in this paper, and further to analyze the distinctions among mistakes, spoofing and uncertainties, with the goal of placing the identified types of locational inconsistencies into a more holistic theoretical framework for spatial data quality. I call on GIScientists to pay more attention to the role of human intentions and advocate for a human-centric assessment of spatial data generation. Only then can we more effectively handle the emerging quality issues in the era of big data.

1. Introduction

The GIScience community frequently focuses on uncertainty with regard to spatial data quality (Devillers et al., 2010). In the era of big data, the advent of mobile, social and geospatial technologies has created a considerable degree of heterogeneous, real-time and geo-referenced data. A large percentage of such data, especially VGI (Volunteered Geographic Information), geo-tagged social media, or location based service feeds, may be generated by ones' mistakes and/or created deliberately rather than being merely affected by inherent uncertainties. Although geographers have recognized the significance of human intention, the motivations of the data generator are seldom examined, and the recent popularity of spoofings, especially those in the form of location spoofing (Zhao, 2015), GPS jamming (Grant et al., 2009) and AIS hacking (McCauley et al., 2016) are often dismissed by GIScientists. In order to more effectively address this critical issue, I will investigate the role of human intentions in the process of spatial data generation and clearly distinguish among uncertainty, mistakes and spoofing.

In the remainder of this paper, I will review the concepts of error, accuracy, precision, uncertainty, mistakes, and spoofing in the context of geography. Then, I will examine the role of human motivation with regard to spatial data quality, and end with a brief concluding remark.

2. Uncertainty, mistakes, and spoofing

Since human beings are forced to view the world through a fuzzy and distorting lens, the measured data are inevitably generalized, approximated, and subject to uncertainties (Zhang & Goodchild, 2002). In other words, the way we observe the world has invoked an inevitable locational (or positional) inconsistency between any observed object in the geographic world and the data that it produces (value of the object being measured). When referring to such locational inconsistencies, geographers usually consider them to be underlying uncertainties. One older and simpler term to describe uncertainty is *error*. By definition, error is the difference between the measured value and the "true" value of the object being measured. It is represented as an estimation of the range of values within which the true value is likely to be found. There are two types of errors: *systematic* and *random error*. While the former affects the *accuracy* of a measurement and can be reduced only by refining the method of measurement or technique, the latter affects the *precision* of a measurement and can be improved by repeating those measurements. It should be noted that systematic and random errors are inherent in the measuring process and cannot be totally eliminated. In the past three decades, the GIScience community began to avoid the term "error" by equating uncertainty as how confident about the measured value. For example, if I say the shortest distance between downtown Portland and Corvallis is 85.3 miles, +/- 0.5 miles at a 95% confidence level. This indicates I am 95% sure that the shortest distance is between 84.8 to 85.8 miles.

Notably, the underlying locational uncertainty is measured by one crucial assumption, that the measuring process was conducted by a *rational geographer* who has a basic geographic knowledge and is acting within a certain set of circumstances with ordinary prudence to control spatial data quality. However, in today's data-intensive society, rational geographers are not the only ones who can contribute to the massive amount of spatial data. Most human beings are irrational, imperfect and have a less than professional knowledge of measuring methods. Even the most careful geographer in the finest laboratory is likely to make mistakes and miscalculations. In such circumstances, the measured locational inconsistency can be regarded as a *locational mistake*. Notably, the term "mistake" is different from error. An error, inherent within spatial data, is locationally inconsistent at the uncertainty level; while a mistake, externally generated contrary to the data generator's original intention, is locationally inconsistent beyond the uncertainty level. Additionally, people may intentionally exaggerate locational inconsistencies even beyond the uncertainty level. For example, a social media user at a local night bar purposely put a geo-tag to his place of employment via a location spoofing app (Andev, 2013), with the goal of deceiving the user's followers as if he was working late rather than out drinking. Obviously, the distance between these two locations greatly surpasses the locational uncertainty level. This type of inconsistency is termed *locational spoofing*.

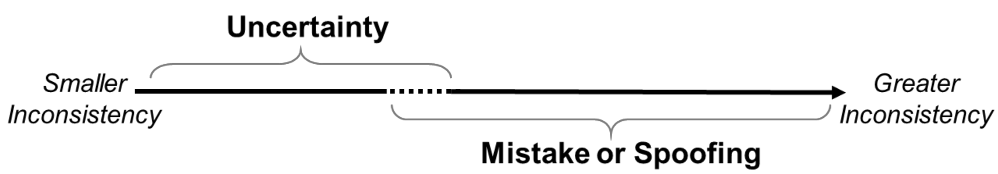

Figure 1. Comparing uncertainty to a mistake or spoofing

Notably, mistakes and spoofing usually bring about more inconsistencies than uncertainties (see figure 1). However, with respect to the degree of locational inconsistency, there is no clear demarcation between a locational mistake and spoofing. For any value that falls in the overlapping region (indicated by the dashed line), the locational inconsistency could be caused by either uncertainty, a mistake or spoofing. Regarding to the values greater than the overlapping region, it is very difficult to distinguish between a mistake and spoofing only by the degree of locational inconsistency *per se*. At this point, it is crucial to examine the intentions of the data generator.

3. The role of human intentions with regard to spatial data

In order to build a holistic framework for spatial data quality assessment, it is imperative to modify the longstanding hypothesis of rational geographer. Indeed, people may have an infinite number of reasons for wanting to generate spatial data. For example, Coleman, Georgiadou, and Labonte

(2009) have discovered eight positive intentions, including, altruism, professional or personal interest, intellectual stimulation, protection or enhancement of a personal investment, social reward, enhanced personal reputation, creative or independent self-expression, and pride of place; as well as three negative aspects: mischief, agenda setting and malicious and/or criminal intent. Furthermore, a legalistic interpretation of human intentions helps to identify hidden motivations in these spoofing cases. In Latin, the standard common law test of criminal liability states, "*Actus reus non facit reum nisi mens sit rea* (it means that the act is not culpable unless the mind is guilty)". In other words, to be guilty of committing a crime, the defendant must have knowingly committed the act (*actus reus*), accompanied by some level of guilt (*mens rea*). According to Model Penal Code[i] (Wechsler, Schwartz, Ploscowe, & Tappan, 1962), there are four levels of guilt for which a suspect is potentially culpable: purposely, knowingly, recklessly, and negligently (in descending order of severity). In addition, if a criminal act is committed out of ignorance (no explicit purposefulness is detected) or by mistake (the mind negates the guilty act), such circumstances can serve as defenses[ii] to withdraw the charge of a crime.

Table 1. Locational inconsistencies categorized by intention

	Intentional+	**Mistaken+**	**Ignorant**	**Mistaken-**	**Intentional-**
Intention	controlling locational inconsistency		no explicit intention	amplifying locational inconsistency	
Measurement	in accord with the intention	contrary to the intention	not specified	contrary to the intention	in accord with the intention
Locational Inconsistency	*uncertainty*	*a mistake*	*a mistake or uncertainty*	*uncertainty*	*spoofing*
Example	Accurate earthquake locations were monitored and then published by USGS geoscientists.	The route to India was prudently measured by Christopher Columbus. However, it led his ship to America.	Locational data was measured through a total station; the operator had little knowledge of how to measure with that equipment.	Accurate AIS data of a fishing ship was sent to the data hub. However, the captain originally wanted to hack/jam the AIS data in order to illegally fish in a prohibited fishing zone.	False locational information of a truck was generated because the driver had jammed all surrounding GPS signals to deceive his boss about the truck's whereabouts.

By the same token, I modify the hypothesis of rational geographer and argue that most measurements (*actus reus*) are the result of five major levels of intention (*mens rea*), including intentional+, mistaken+, ignorant, mistaken-, and intentional- (refer to Table 1). The symbol "+" and "-" denotes two opposite intentions of the data generators: "+" denotes that the intention is to control spatial data quality, or in other words, to control locational inconsistency at the uncertainty level; while "-" denotes the intention to amplify the locational inconsistency of an object. These five levels of intention can be qualitatively determined through participatory observation, questionnaire or online interviews. Specifically, an individual or organization measures an object of the geographic world with an intention to control locational inconsistency, if the measurement aligns with such intention, the measured locational inconsistency will be represented as uncertainty; if contrary, the measured locational inconsistency will be represented as a mistake. If the measurement is not driven by any explicit intention, the measured locational inconsistency may be viewed as either a mistake or uncertainty. Or with an intention to amplify the locational inconsistency, if the measurement aligns with such intention, the measured locational inconsistency will be represented as spoofing, if contrary, the measured locational inconsistency will be represented as uncertainty (the data generator consciously wants to amplify locational inconsistency in the measurement but fails). Thus, based upon underlying motivations, the proposed framework enables us to distinguish locational uncertainty, mistakes and spoofing. In practice, it also allows researchers to detect different types of locational inconsistencies from a dataset, thereby generating a more accurate subset for reuse. For example, geo-tagged Twitter feed enables researchers to track Oregon residents' opinions on specific topics (i.e., terrorist attacks, presidential campaign). Before drawing any conclusions from a set of geo-tagged tweets sent out from Oregon, researchers must delete those tweets that are falsely attributed to Oregon residents because of spoofing or mistakes, and then estimate the collective opinion using the rest of the tweets. More importantly, this framework transcends a purely technical interpretation by providing a holistic perspective from which to examine spatial data quality. In this sense, this framework encourages GIScientists to devote more effort to discovering the generative mechanism of various locational inconsistencies at the human intention level.

4. Concluding remark

In sum, I investigated the effects of human intention on spatial data. Through a detailed examination of various motivations associated with measurement, I suggested to modify the hypothesis of rational geographer and put spoofing and mistakes align with uncertainty under a greater umbrella topic - spatial data quality. I call on GIScientists to establish a human-centric assessment strategy to better understand this topic. Only then can we more effectively handle the emerging quality issues in the era of big data.

Acknowledgements

For their invaluable assistance with this article, I wish to thank Daniel Sui, Li He, Xining Yang, Peter Chen, and three anonymous reviewers.

References

Andev. (2013). Fake gps - fake location. Retrieved on May 4, 2016 from https://play.google.com/store/apps/details?id=com.fakegps.mock

Coleman, D. J., Georgiadou, Y., & Labonte, J. (2009). Volunteered geographic information: The nature and motivation of produsers. *International Journal of Spatial Data Infrastructures Research, 4*(1), 332-358.

Devillers, R., Stein, A., Bédard, Y., Chrisman, N., Fisher, P., & Shi, W. (2010). Thirty years of research on spatial data quality: Achievements, failures, and opportunities. *Transactions in GIS, 14*(4), 387-400.

Grant, A., Williams, P., Ward, N., & Basker, S. (2009). GPS jamming and the impact on maritime navigation. *Journal of Navigation, 62*(02), 173-187.

McCauley, D. J., Woods, P., Sullivan, B., Bergman, B., Jablonicky, C., Roan, A., . . . Worm, B. (2016). Ending hide and seek at sea. *science, 351*(6278), 1148-1150.

Wechsler, H., Schwartz, L. B., Ploscowe, M., & Tappan, P. W. (1962). *Model Penal Code*. Philadelphia, PA: ALI (American Law Institute).

Zhang, J., & Goodchild, M. F. (2002). *Uncertainty in geographical information*: CRC press.

Zhao, B. (2015). Detecting location spoofing in social media: Initial investigations of an emerging issue in geospatial big data. Ph.D. Disseration. The Ohio State University.

[i] MPC is highly influential across the United States for clarifying these various levels of culpability.

[ii] A defense can also be a reasonable excuse (e.g., infancy, insanity, involuntary intoxication, etc.) or justification (e.g., duress, necessity, self-defense, consent of the victim, entrapment, etc.). Since these intentions are far off the intention for a locational inconsistency, I did not discuss in details.