Inference and analysis across spatial supports in the big data era: Uncertain point observations and geographic contexts

Colin Robertson1 | Rob Feick2

1 Department of Geography & Environmental Studies, Wilfrid Laurier University, Waterloo, Ontario, Canada
2 School of Planning, Faculty of Environmental Studies, University of Waterloo, Waterloo, Ontario, Canada

Correspondence
Colin Robertson, Department of Geography & Environmental Studies, Wilfrid Laurier University, 75 University Ave West, Waterloo, ON N2L 3C5, Canada.
Email: crobertson@wlu.ca

Abstract
The ways in which geographic information are produced have expanded rapidly over recent decades. These advances have provided new opportunities for geographical information science and spatial analysis—allowing the tools and theories to be expanded to new domain areas and providing the impetus for theory and methodological development. In this light, old problems of inference and analysis are rediscovered and need to be reinterpreted, and new ones are made apparent. This article describes a new typology of geographical analysis problems that relates to uncertainties in the relationship between individual-level data, represented as point features, and the geographic context(s) that they are associated with. We describe how uncertainty in context linkage (uncertain geographic context problem) is also related to, but distinct from, uncertainty in point-event locations (uncertain point observation problem) and how these issues can impact spatial analysis. A case study analysis of a geosocial dataset demonstrates how alternative conclusions can result from failure to account for these sources of uncertainty. Sources of point observation uncertainties common in many forms of user-generated and big spatial data are outlined and methods for dealing with them are reviewed and discussed.

1 | INTRODUCTION

Geographical data have become increasingly pervasive in the social and physical sciences due to a greater number and variety of data sources, widespread use of geographic information system (GIS) software, and training in spatial analysis and spatial data handling techniques. In light of these changes, it is important to revisit and recast old problems, and articulate new ones made evident by the current "data-rich" environment (Warf & Arias, 2008).
in the modeling and analysis of geographical data are often contingent on how observations, spatial processes, and spatial relationships are represented in the analysis. Miller and Wentz (2003) note in their review of spatial representation issues in GIS that "A sterile geometry is associated with a simplified GIS that fails to fully represent some segments of society or complex geographic processes." While much progress has been made in the spatial analysis of multi-scale processes (Jones, 1991), time and space-varying relationships (Brunsdon, Fotheringham, & Charlton, 1998; Gelfand, Kim, Sirmans, & Banerjee, 2003; Smith, Lucey, Waller, Childs, & Real, 2002), and “spatializing” modes of analysis in new domain areas, problems of crisp spatial representations, static rather than dynamic data models, uncertainty in data linkages, and issues of statistical inference and generalizability persist (Boruff, Nathan, & Nijenstien, 2012; Puig & Ginebra, 2015; Wan, Lin Kan, & Wilson, 2017).

Linking data from different levels of support has become integral for building new spatial variables and generating inferences that link individuals to their environmental context. For example, advances in multi-level modeling methods allow differentials in individual outcomes to be apportioned more accurately between individual factors and contextual influences that span a range of spatial and temporal scales. This modeling framework provides a more nuanced treatment of individual–contextual relationships than was possible with earlier approaches, which conceived of person–environment relationships through a simple ecological representation (Chaix, Merlo, & Chauvin, 2005; Kestens, Wasfi, Naud, & Chaix, 2017). Similar trends are evident in other fields that rely on spatial processing and representation, such as wildlife research, where concepts of habitat and home range have been recast as multi-scale, hierarchical, spatially dependent, and uneven in spatial and temporal usage compared with earlier static models (Kie et al., 2010; McGarigal, Wan, Zeller, Timm, & Cushman, 2016), calling into question the utility of core conceptual constructs as new data and associated modeling tools become more widely used (Kie et al., 2010).

Recently, Kwan (2012) has described a new problem, termed the uncertain geographic context problem (UGCoP), which recounts how issues of spatial uncertainty can manifest in multi-level modeling designs, where the objective of analysis is to make inferences about individuals (e.g., health outcomes) based on area-level contextual variables (e.g., neighborhood income inequality). The uncertainty in the UGCoP is due to the arbitrariness of the areal unit boundaries used to represent the contextual variables, and their relationship to the unknown true contextual influences of these factors on individuals. The areal unit boundaries used for analysis (e.g., census tracts) may not match the spatial or temporal bounds of the true contextual influences and the degree of misalignment may vary across the study, along lines of gender, occupation, and/or socioeconomic status (Diez Roux & Mair, 2010). UGCoP is related to, but distinct from, the more well-known modifiable areal unit problem (MAUP), which geographers have been dealing with for decades. The MAUP pertains to the variation in outcomes of analysis that arise from changes in the configuration and/or scaling of areal unit boundaries (Flowerdew, Manley, & Sabel, 2008; Jelinski & Wu, 1996), whereas the UGCoP pertains to errors in individual-level inferences resulting from the spatial mismatch between the boundaries of the units used to measure contextual variables and the true (unknown) contextual influences of those variables. Both the MAUP and the UGCoP affect research designs using areal-aggregated data to measure causal relationships.

In this article, we introduce a new typology of geographical analysis problems that are derived from well-known issues of individual and group-level spatial analysis. The problems we describe may contaminate inferences made from spatial data. In addition to UGCoP, a related problem arises when findings about group-level differences or relationships are affected by the measurement of individual-level variables via their spatial location(s). We term this problem the uncertain point observation problem (UPOP), which occurs because of the spatial uncertainty in the individual–group-level (i.e., point–area) linkages. This uncertainty can be due to location error, such as geocoding error, as well as a host of other new and old sources of spatial error and uncertainty. We believe UPOP is of growing concern for two key reasons. First, increasing sources of individual-level data are available from GPS, transactional data, volunteered geographic information (VGI) and citizen science, and a variety of sensors. Second, who is conducting spatial analysis differs notably from even a decade ago. In addition to experts trained in spatial data handling and analysis methods, software coders without the same expertise can embed methods in software that use individual data to make improper inferences about
larger areas (Unwin, 2005). We argue that both UGCoP and UPOP can be situated within a typology of geographical analysis problems that depend on the form of spatial support (point or area) involved in the study and/or application.

2 | CONTEXT: BROADENING SPATIAL DATA USE, PRODUCTION, AND ANALYSIS

The growing interest in more nuanced methods of explaining person–context relationships in health parallels and is fueled by two related, but distinct, movements toward broad-based authorship of spatial data by persons ranging widely in interests and expertise, and data-centric analysis approaches that capitalize particularly on big data resources (Crampton et al., 2013). First, the shift in spatial data authorship from the sole purview of experts to also include broad swaths of society engaged in citizen science or VGI projects has been documented well by Goodchild (2007), Sui, Elwood, and Goodchild (2012), and Buytaert et al. (2014), among others. Irrespective of whether these citizen-sourced data are contributed deliberately by individuals according to their interests and concerns (e.g., bird watching, water quality monitoring, etc.) or are generated passively without a person’s conscious effort (e.g., connecting to public WiFi), the implications of this development for spatial analysis are widespread. Most pertinent are: (a) growing, heterogeneous, and often poorly documented sources of georeferenced data; and (b) data collection and quality control processes that are more varied and social in nature than industrial and statistical (Haklay, Singleton, & Parker, 2008; Regalia, McKenzie, Gao, & Janowicz, 2016; Song & Sun, 2010; Yang, Fan, & Jing, 2016).

Second, the impacts of big data, which are often characterized in terms of unprecedented volumes, velocity, and variety of data, on geographic inquiry have been discussed widely with respect to topics as diverse as personal privacy, civic participation, mobility and movement, resource use, and spatial cognition (Kitchin, 2014; Miller & Goodchild, 2015). The growth of data-driven approaches is transformative for geographic inference and analysis. Many of these data pertain to, or are created by, individuals and document their daily activities, movements, and interactions as point-event observational data. These datasets are highly disaggregated across space and time, and offer new, highly granular windows into the routine dynamics of human and natural processes (Batty et al., 2012; Fritz, Schuurman, Robertson, & Lear, 2013). Sensors, for example, can track when and where people board transit, point of sale records detail the times and locations of electronic purchases, and cellular phone call and social media metadata permit exploration of social connectivity, movements, and momentary expressions of perceptions and emotions across space (Ahas et al., 2015; Calabrese, Diao, Di Lorenzo, Ferreira, & Ratti, 2013; Shaughnessy et al., 2018; Shen & Cheng, 2016).

Further, unlike traditional scientific approaches where data collection follows the development of theory-based research questions and controlled sampling protocols, big data approaches often invert this sequence by first exploring what patterns are evident and what questions a specific dataset may be able to answer (Miller & Goodchild, 2015; Thatcher, 2014). As Kitchin (2014) and others note, this reflects the fact that big data resources are defined less by need and more by what is convenient or technologically feasible to monitor or repurpose. Finally, there is often uncertainty concerning the veracity or validity of big data resources. In part, these concerns relate to how representative a data source (e.g., Twitter) is of a community (e.g., a city) or a variable of interest (e.g., mobility). Interest has grown in furthering our understanding of the biases in big data sets, which are often constrained to narrowly defined sub-populations (e.g., only transit riders who use smart cards), describe very limited aspects of the human experience (e.g., only when and where transit rides begin and end, rather than why), and have biased spatial and temporal coverage (e.g., along transit line, predominantly during work hours) of people’s activities (Kwan, 2016; Robertson & Feick, 2016; Shelton, Poorthuis, & Zook, 2015). Given the low information content and uncertainties associated with most big data sources, several researchers have highlighted the need to examine the patterns of individual point-event data in light of other spatial-contextual data (Crampton et al., 2013; Graham & Shelton, 2013; Li et al., 2016). As spatial data become more widely used across disciplines and more distributed in terms of authorship and access (e.g., via data portals and open application programming interfaces), issues of spatial uncertainty and linkage become more pernicious and important to diagnose and characterize.
There is growing interest in exploiting new sources of individual-level spatial data in geographic research (Fritz et al., 2013; Ghosh & Guha, 2013) as well as in practical applications. For example, a recent patent for estimating creditworthiness of applicants includes provisions to use “historical instantaneous geographic data obtained from the digital device comprising the GPS-equipped smartphone” as well as social network data related to family and personal health status to determine credit scores for individuals at the point of a transactional credit approval (Hochstatter, Leonard, & McKinzie, 2016). Such individual and transactional data records are also shared among data exchanges and may be joined to or correlated with other geographic datasets. While such automated procedures for joining spatial and transactional data represent risks to individual privacy and autonomy, similar procedures may be used to characterize areas, such as in insurance rate estimation (e.g., Baecke & Bocca, 2017) or crime rate prediction and forecasting tools (Wang, Kifer, Graif, & Li, 2016). Embedded within these big data and algorithmic systems that combine geographically referenced data streams into real-time decision support applications are classical issues of geographic inference: representativeness, spatial uncertainty, and sampling error. What makes this new environment for automated spatial analyses potentially more problematic is that many of the algorithms that are shaping data and analyses are opaque and undocumented (Graham, 2005; Kitchin, 2014; Roche, 2017). Renewed interest among geographers in examining individual–context relationships posed by the big data environment raises important methodological issues for moving geographic inferences away from associative statistics and toward causal analysis and knowledge-based approaches.

In geographical research, individual-level data are typically employed in an inferential framework that aims to generalize findings to a broader population. This framework is derived from associational inference, where one aims to understand the association between one variable measured over a population of units and another variable measured over those same units (Holland, Glymour, & Granger, 1985). The representativeness of the units in the sample is crucial if sample associations are to be inferred to the wider population. With individual-level tracking data there are two representativeness problems: (a) how representative the sample points are of an individual’s true spatial context; and (b) how representative an individual’s true context is of the population which the study is aiming to characterize. When the objective of analysis is to determine the effect of some aspect of the environment (e.g., exposure) on an individual, full tracking and exposure sensing data for randomly sampled individuals from a population would theoretically support this for associational inference (addressing issue a). However, since we cannot know whether the effect would have occurred without the exposures, even in this case we cannot make causal inferences between spatial context and individual effects. In reality, most studies in geography employing individual-level data are partial in nature and only crudely approximate the true context. For example, exposure to air pollution, even with the best sensor technology, has a sampling interval, requires recharging, may not work in dense forests or indoor environments, etc. Factors such as ethical and legal restrictions (e.g., privacy laws), technological limitations (e.g., battery capacity on GPS wildlife collars), and the activity-specific nature of many newer sources of individual-level data (e.g., transit smart cards, credit cards, social media, etc.) cause these datasets to be incomplete approximations of spatial context. Critically, when the frequency of spatial sampling is reduced, this is typically done in a non-random way, thereby generating a non-representative spatial sample. When the units about which we wish to make inferences are geographical areas, linking these biased spatial point samples to areal containers can cause erroneous inferences about environmental–individual relationships which can accumulate into biased group-level spatial patterns.

3 | TYPOLOGY OF GEOGRAPHICAL ANALYSIS PROBLEMS RELATED TO INFERENCE WITH SPATIAL DATA

An increasing variety of technologies, research frameworks, and methods are available now for tracking individuals in space and time. There are many advantages to individual-level spatial data, such as greater spatial and temporal granularity and greater precision in estimates of variables of interest (health status, stress level, activity space, perceptions of neighborhood safety, etc.). Also, recent research has demonstrated the need for mobile methods that capture geographic context dynamically over various spatial and temporal scales to truly understand environment–individual relationships (Ahas et al., 2015; Shaughnessy et al., 2018; Sheller & Urry, 2006).
Diez Roux and Mair (2010) discuss spatial context definition in neighborhood health effects research, noting that measured spatial context variables will likely differ from the "true causally relevant spatial context." This difference and its impact on multi-level spatial analyses was described by Kwan (2012), introducing UGCoP, which describes how uncertainties in the measurement of the true causally relevant spatial context can contaminate inferences at the individual level.

We might consider how measurement of spatial context relates to more well-known inferential issues pertaining to group and individual data (see Figure 1). The ecological fallacy is a problem whereby erroneous inferences about causal relationships are made about individuals based on relationships estimated at the group level (Robinson, 1950). The fact that area-level findings frequently do not correspond to the same analysis done at the individual level has led to the huge increase in multi-level modeling, which aims to parse effects into aggregate/endogenous and contextual effects (Diez Roux & Mair, 2010; Subramanian, Jones, Kaddour, & Krieger, 2009). The inverse to the ecological fallacy is termed the atomistic or individualistic fallacy, which describes erroneous inferences made about causal relationships in aggregate units based on data measured at individual levels (Diez Roux, 2002). One of the sources of atomistic fallacy is sometimes called the "biological fallacy" when the errors in inference arise because contextual effects were not incorporated into the individual-level analysis. While the descriptions of inferential issues here relate to individual and area-level data, the fallacies actually occur for analysis between any lower-level and higher-level aggregations, such as census tracts and municipal boundaries (Diez Roux, 2002).

Ecological and atomistic fallacies have special interest to geographers as the sources of uncertainty that lead to discrepancies are often spatial in nature (Openshaw, 1984). UGCoP specifically concerns the use of arbitrary areal units for measuring context, as the precise contours of the true context in space and time are unknown. Interestingly, UGCoP arises out of the specification of multi-level analysis, which has grown in popularity for specifying models that aim to avoid ecological and/or atomistic issues.

We will distinguish between the measured and experienced (i.e., true) spatial contexts and discuss these in relation to broader issues in aggregate and individual-level analyses (Robinson, 1950). The true contextual unit (TCU) of an

**FIGURE 1** Classic inferential errors related to group and individual-level analysis

- *Ecological Fallacy*
  - Individuals
  - Groups

- *Atomistic Fallacy*

**FIGURE 2** Typology of geographical analysis problems related to inferential errors with spatial data

- *Ecological Fallacy*
  - Contextual Points
  - MAUP
  - MCI ≠ TCU
  - Spatial Relationship
  - Uncertain Point Observation

- *Atomistic Fallacy*
  - Polygons
  - Areas
individual is the set of locations that encapsulate the effect or exposure of some environmental variable on that person, and the measured contextual unit (MCU) is the measured representation of the TCU in a given study (see Figure 2). Typically, the TCU is unknown; a complex assemblage of daily, weekly, annual, spatial activities and exposures, latencies, and behaviors that comprise an individual's activity space, while the MCU is often represented by administrative units where data are collected, summarized, and distributed, such as health districts or census tracts. The TCU represents the spatial bounds for areas of influence on the individual, and will contain internal variability with respect to the causal relations under investigation. For example, when estimating associations between census data and individual outcomes linked by home address, we are inferring (i.e., the embedded spatial semantics) that the conditions measured over the MCU are causally related to the outcomes of individuals within that unit, since the home address is a signifier of where that person spends their time (i.e., a centroid exposome representation as per Jacquez, Sabel, & Shi, 2015). Such a study design is reducing the spatial tracking sample to a single location.

There are several problems with this representation. Firstly, the home address may be a poor approximation of the TCU, which will be impacted by individual mobility and environmental heterogeneity. This critique has been widely acknowledged in neighborhood effects research, as research has shifted to individual health-tracking studies (Su, Jerrett, Meng, Pickett, & Ritz, 2015), coupled indoor–outdoor exposure estimation (Quackenboss, Lebowitz, & Crutchfield, 1989; Steinle et al., 2015), and concepts such as spatial polygamy that recognize that individuals belong simultaneously to many proximate and distal physical, social, and digital contexts (Matthews & Yang, 2013). Secondly, even for individuals that spend all or most of their time at home, the home address location may differ substantively from the aggregate measure of the variable within the MCU, which may not represent any specific location (i.e., ecological fallacy). Finally, the “causally relevant” aspect of the TCU makes this construct application-specific, and therefore any evaluation of the quality of MCU–TCU alignment is dependent on the specific application being investigated (i.e., fitness for use) (Devillers et al., 2007).

Figure 2 illustrates the inter-relationships between fallacies and problems for within-level and multi-level spatial analyses. As noted by Subramanian et al. (2009), there are four potential study designs with respect to outcome and exposure variables measured over individuals (at point support) and groups (at area support). Given variable x as an exposure variable and variable y as an outcome variable, both measured with spatial point support, the estimated relationship \( \hat{Y} = f(x) \) can be obtained. Applying \( f(x) \) to \( X \) measured over areal-unit support to estimate \( \hat{Y} \) supposes that individual-level relationships hold at the group level. This may not be the case (i.e., atomistic fallacy) since, for example, constructs may have a different meaning when measured at individual or aggregate level (Klein & Kozlowski, 2000). Conversely, we may wish to aggregate x to obtain a new variable such as \( \hat{X} = g(x) \), where \( g \) is some aggregation function, and then estimate \( \hat{Y} = f(\hat{X}) \) at the area-unit level. Spatial uncertainty in the relationship between \( x \) and \( X \) can cause \( g(x) \) to be a biased estimate of \( X \). We term this issue the uncertain point observation problem (UPOP) and describe its sources and impacts below. Similarly, spatial uncertainty in the relationship between \( X \) and \( x \) when \( g(x) \) is used to construct a variable at point support from a variable at area support creates what we have described as UGCoP. Implementations of \( g(x) \) when used to generate point-level covariates might include buffering, zonal statistics, or GPS tracking. Conversely, \( g(x) \) may take the form of an arithmetic mean, geometric mean, sum, or rate calculation variable when used to generate a covariate at the area level. To summarize Figure 2, point-to-area inference raises the risks of UPOP and atomistic fallacy, whereas area-to-point inferences risk UGCoP and ecological fallacy.

UGCoP can be seen as a form of measurement error at the individual level whereby the contextual variable measured for an individual is a poor approximation of their causally relevant context. Analogously, UPOP can be seen as a measurement error at the areal-unit level whereby an aggregate variable is formed from a function (e.g., arithmetic mean) of individual samples which are not causally relevant to the areal unit (see Table 2 later). While issues of spatial aggregation have been well known for some time (Clark & Avery, 1976; Heuvelink & Pebesma, 1999), the big data environment has made this an issue of increasing significance—one that has been implicitly recognized—and methods to counter it have been developed in recent years.

The notion of MCU–TCU misalignment can be used to describe a typology of geographical analysis problems in geography that pertain to use of areal data (Figure 2). Reading from the left, individuals are represented as points, and
these data are joined through a spatial relationship with MCU data, creating point data enriched with spatial context information (i.e., contextual points). In some cases, individual data will already be available as aggregated to MCUs, such as with the case with census data. The key issue arises when the MCU does not match the TCU, creating the UGCoP (upper feedback arrow in Figure 2). Making inferences about individuals represented in these data is a source of ecological fallacy. Conversely, areas are represented as polygons, which are enriched through spatial relationships with point data. When UPOP exists, inference about areas may be contaminated, leading to the atomistic fallacy (lower feedback arrow). Each problem can be conceptualized in terms of MCUs and TCUs, and the problems that arise depend on the type of analysis undertaken. While typically MCUs have no inherent meaning as an object of study in and of themselves, available or derived context partitions themselves (e.g., pixels, census units, neighborhoods) often become the object of analysis when investigating spatial patterns. Census and household survey data avoid this issue by ensuring that respondents live at the address for which the responses are summarized (yet these may not be “causally relevant” units over which to sample for any given individual). With many forms of new observational data, the linkage between point observations and the underlying contextual units within which they are summarized through an aggregation function is unknown. The order in which relationship estimation and aggregation functions are carried out dictates the potential problems in any given analysis (Figure 3).

The UPOP and UGCoP are distinct from the modifiable areal unit problem (MAUP), which pertains to variable conclusions resulting from either reconfiguring areal units or scaling/aggregating areal units. Firstly, MAUP can be considered a type of data-analytical artefact resulting from aggregation/rezoning, and therefore can be solved or at least minimized by careful multi-scale analysis (Jelinski & Wu, 1996) or more explicit statistical estimation methods as described in Gelfand (2010). UPOP and UGCoP, however, both describe conceptual-level problems whereby even without any spatial error in locations, incorrect inferences can result from uncertainties in the veracity of TCU–MCU spatial relationships between individual point locations and the areal units these points are linked to.

There are two sub-problems that compose both UPOP and UGCoP. First is the problem of context definition and TCU measurement. In the UGCoP case, context is derived from areal units, in the UPOP case, it is measured over individuals. Each case can be conceptualized as a measurement error which misrepresents the true causally relevant process acting on the outcome under investigation. Definition of appropriate spatial context is a problem common to virtually all forms of spatial analysis. Table 1 identifies four broad classes of spatial analysis studies in terms of how they relate to the way spatial context is defined and analyzed, including spatial autocorrelation analysis and hotspot mapping (Aldstadt & Getis, 2006; Nelson & Boots, 2008; Nelson & Robertson, 2012), cluster detection (Kulldorff & Nagarwalla, 1995), spatial modeling (Stakhovych & Bijmolt, 2009; Wall, 2004), and kernel methods (Jones, Marron, & Sheather, 1996), among others. The sensitivity to UPOP and UGCoP varies based on how context is defined and analyzed. The second component, which is unique to UPOP and UGCoP, is the application of spatial context across spatial

FIGURE 3  Diagram of geographical analysis problems related to analysis across spatial supports (adapted from Heuvelink & Pebesma, 1999)
support levels. Using individual context variables to infer characteristics at the areal-unit level (i.e., through multi-level modeling or aggregation) is analogous to a spatially oriented atomistic fallacy, whereas using area-based context variables to infer relationships at the individual level is analogous to a spatially oriented ecological fallacy.

4 | THE UNCERTAIN POINT OBSERVATION PROBLEM

Many areas of geography are concerned with characterizing areal differentiation in order to identify spatial patterns (e.g., are there clusters or a trend?), evaluate spatial hypotheses (e.g., are disease rates higher near the coast?), or assess relationships between spatial variables (e.g., is there a relationship, is it constant across the study area?). A precursor to these forms of analysis is measuring the required variables across appropriate units of geography (i.e., contextual units). When individual-level data are erroneously linked to a contextual unit, leading to an incorrect characterization of that area, we have an instance of what we are calling UPOP. Such instances can accumulate, we believe, into erroneous conclusions about group-level spatial patterns.

It is important to note that UPOP is fundamentally an issue of data linkage and analysis and is not inherent to the data themselves. The problem of erroneously linking point data to polygon data has been explored in relation to locational uncertainty in the past. Krieger, Waterman, Lemieux, Zierler, and Hogan (2001) demonstrated geocoding error as a source of uncertainty at the block and census tract level in public health studies. Kravets and Hadden (2007) found that geocoding errors were more common in poorer and rural areas at the block group level. Malizia (2013) demonstrated how tests of space-time interaction were impacted by even slight errors in the location data of point-event data.

The source of UPOP can often be attributed to bias in sampling of a spatial process. If repeated sampling of a spatial process is not random, then the observations resulting from that sampling will not be a representative spatial sample. Further, while MAUP is concerned with uncertainty resulting from aggregation/rezoning of a given sampling of a spatial process, UPOP is concerned with uncertainty resulting from the nature of the sampling process itself, which increasingly may be related to behaviors and activity (e.g., tweeting at lunchtime at work), technology (e.g., temporal sampling interval for GPS), or more subjective personal characteristics that relate to the data authoring process.

Given data obtained over point and areal-unit support for the same study area, the objective of analysis is important for distinguishing between UGCoP and UPOP. When trying to learn about the characteristics of individuals, by examining relationships of attributes via their spatial relations (e.g., points contained within polygonal geographies), we are subject to both MAUP—Inferences change when boundaries change—and UGCoP—inferences are incorrect because the boundaries of the MCU do not reflect uncertain or unknown TCU boundaries (TCU ≠ MCU). When approaching the analysis from the opposite perspective, that is, trying to identify areal differentiation in geographical units by examining the individuals contained within their boundaries, errors in inference can occur if the locations of the points are not truly characteristic of the areas (TCU ≠ MCU). In short, even though the semantics of the spatial relationship of containment can be highly variable (e.g., a person passing through a neighborhood vs. a long-term resident), the relationship is treated as a binary relation (Egenhofer & Franzosa, 1991). While these place versus space

| TABLE 1 | Levels of analysis and spatial context definition in common types of studies in quantitative spatial research |
|-----------|--------------------------------------------------|--------------------------------------------------|
| **Context analyzed at:** | **Areal aggregate** | **Individual** |
| **Context defined at:** | **Individual** | **Social surveys, census analysis, GPS tracking/buffering, social media/citizen sensing, spatial “event data”** | **GPS tracking/buffering, spatial risk factor analysis** |
| | **Areal aggregate** | **Spatial weights matrices (local spatial analysis, GWR, spatial autoregressive modeling)** | **Neighborhood health effects studies, multi-level modeling** |
issues have been known in GIScience for many years, they are of increasing concern. Widespread use of GIS in data preparation for statistical analysis makes this issue particularly more prevalent than might otherwise be the case, and speaks to a deeper issue related to the distinction between computed spatial relationships, their semantics in the real world, and the objective of the analysis.

The notion of "platial" GIS operations has been proposed as a way to frame a recasting of classical spatial overlays with fuzzier place-based referents that can capture nuances and variations in individual spatial cognition (Goodchild & Li, 2011). Gao, Janowicz, and McKenzie (2013) give the example of a tornado point observation occurring near a state boundary. Naïve spatial overlay analysis would associate the point with one or another states, even when the attributes associated with that event (e.g., number of injured people) pertain to both sides of the state boundary.

There are several examples of UPOP potentially impacting inferences made in recent studies. For example, a study investigating the relationships between activity variables derived from geolocated tweets and unemployment in Spain found significant associations and high explanatory power in predicting unemployment (Llorente, Garcia-Herranz, Cebrian, & Moro, 2015). While the correlations identified in this article are not claimed to be causal, spatial associations are often interpreted as such: areas with rates of high unemployment are also areas with a high proportion of "mis-spellers"—individual Twitter users who misspelled a set of 617 commonly misspelled words. The problem here is atomistic. Population characteristics in each municipality (i.e., MCU) are inferred from the properties of individual users via the spatial overlap of Twitter messages represented as points and the municipalities for which unemployment data are available. Mitchell, Frank, Harris, Dodds, and Danforth (2013) link Twitter-derived sentiment measures to demographic and socioeconomic variables at state and city scales. Analysis here can also be quite granular thematically, for example finding significant positive correlation with keyword frequencies such as "wings" and "mcdonalds" with obesity rates, and negative correlation with keywords "cafe" and "sushi." Though purely correlational in nature, the findings and interpretations often hint at causal mechanisms. Many other studies have taken similar approaches, mostly identifying correlations at some aggregate level of geography with variables derived from individual point data that fall within areal units for which other demographic or thematic data are available. Yet the question of how granular this type of analysis can become is rarely explored or mentioned explicitly. As geographical units become smaller and/or individuals more mobile, the likelihood of UPOP errors increases.

Spatial analysis of individual-level data has become widespread in GPS tracking studies in health and environmental research. Many critiques of environmental epidemiological studies (Chaix, 2009; Matthews, 2011; Rainham, McDowell, Krewski, & Sawada, 2010) have recognized that data that sample individuals’ home locations are not sufficient to estimate relationships between environmental exposures and individual-level outcomes. The context variable measured over MCUs (e.g., vehicle air emissions) may not reflect the true value or even a reasonable estimate of conditions for any specific location within that area (Hystad et al., 2012). The use of tracking therefore provides a much greater spatial expression of an individual’s use of space and hence environmental exposures than a single home address location. However, Chaix et al. (2013) point out how GPS tracking studies of exposures often contain selective bias, which also contaminates our understanding of causal effects. In physical activity studies, for example, activity space estimation from GPS tracking has linked “access” to parks and green space with higher levels of physical activity, but this association may be an artefact of individuals purposely seeking those locations. However, often the research goal is to understand if parks and green space actually promote physical activity (i.e., change behaviors), which cannot be learned without spatial counterfactuals (i.e., matched individuals tracked in areas with lower access). It is increasingly apparent that individual-level data afforded by new geolocation technologies, while alleviating some of the issues of areal data analysis, bring new concerns and analytical issues.

5 SOURCES OF POINT OBSERVATION UNCERTAINTY

UPOP results from a range of social, conceptual, and technological factors that create uncertainty about the validity of making inferences about areas from individuals’ data. Here, individual-level data refers to disaggregated data that either
describe characteristics of individuals in space and time (e.g., location, age, sex, income) or spatial-temporal data that individuals actively or passively create (e.g., GPS traces, geotagged social media) as they move about their surroundings and record observations. Depending on the data and methods used in a study, the results may be affected by one or more of the sources of UPOP uncertainty listed in Table 2.

UPOP occurs most commonly as a form of sampling error or bias. Researchers rarely, if ever, have access to an entire population’s data at the individual level and instead strive to obtain a representative sample of individuals to base inferences on pertaining to the population as a whole. When systematic bias affects the likelihood that sampled individuals are representative of an area’s population, the potential to make erroneous inferences about the area increases due to heightened uncertainty about the degree of MCU–TCU match. This sampling bias results in either UGCoP or UPOP effects, depending on the direction of analysis. For UPOP, new forms of geographic data recorded at the individual level are particularly sensitive to sampling-UPOP when used for analysis with areal data. For example, it is widely acknowledged that user-generated content, and VGI specifically, are unrepresentative of wider populations (Elwood & Leszczynski, 2013; Haklay, 2013), yet researchers have drawn associations by linking these data to areal units (e.g., see Mullen et al., 2015). This unrepresentativeness is twofold and follows from classic ecological fallacies where population-level processes are incorrectly seen as simply the sum of individual processes (Schwartz, 1994). First, although the number of people who author data has increased markedly as personal and mobile technologies have become more pervasive, there is often bias in who participates in VGI data production directly (e.g., citizen science projects) and indirectly (e.g., social media platforms) (Brabham, 2012; Kelley, 2014; Preece, 2016). Second, there is also spatio-temporal bias associated with where and when these people engage with social media and, in a broader sense, author VGI, as seen in OpenStreetMap contribution patterns (Haklay, 2010; Li, Goodchild, & Xu, 2013; Mooney & Corcoran, 2012). VGI authoring in general tends to over-represent some locations (e.g., major roads, transit stops, shopping districts, popular landmarks), under-represent others (e.g., residential neighborhoods, industrial areas) (Li et al., 2013), and can also reflect broader socioeconomic gradients (Rabari & Storper, 2014), racial (Crutcher & Zook, 2009), and linguistic (Graham & Zook, 2013) in society. While the former aspect of sample bias is broadly acknowledged in studies using social media data, the impacts of spatio-temporal biases in UGC have only recently been acknowledged (Shelton, Poorthuis, Graham, & Zook, 2014). Consider, for example, teenagers or commuting workers who use Twitter on their lunch breaks. In both cases we might expect highly localized concentrations of use near schools and workplaces, at

| UPOP source       | Subtype                                      | TCU–MCU impact                                                                                      |
|-------------------|----------------------------------------------|-----------------------------------------------------------------------------------------------------|
| Sampling bias     | Representativeness of individuals as data authors | Individuals’ characteristics may not represent area’s population (classic representativeness biases or effects of mobile individuals crossing several areas) |
|                   | Representativeness of data individuals create  | Data created by individuals is a selective sampling in terms of what, where, and when they sample (e.g., wildlife sightings near roads) |
| Data authoring    | Sensor–subject displacement                   | Uncertainty of whether a data point is associated with the correct zone because of:                   |
| uncertainty       |                                              | (a) differences between the recorded location (MCU) of a sensor (e.g., camera) and the object of data collection (TCU) (e.g., photo subject), and |
|                   |                                              | (b) individuals’ communication georeferenced to one zone (MCU) references objects and events in another zone (TCU) |
|                   | Software code bias                            | Impact of software on individuals’ reported locations:                                               |
|                   |                                              | (a) use of point geometries to collect data about areal features—simplification of context, and     |
|                   |                                              | (b) limited place names and feature taxonomies in code alter how people use, interpret, and code point–area relationships |
| Interpretational  |                                              | Errors in inferred relationships between individuals’ MCU and areas (TCUs) due to incorrect assumptions of the process or the data under study (e.g., linking characteristics of mobile Twitter users with static census zones areas) |
specific times of day, and variable degrees of association to their proximate physical and social environments. The associations that commuting students and workers may differ substantially from those of local residents, which naïve mapping of geolocated tweets would miss. Also, evaluating relationships between Twitter-derived variables (e.g., sentiment, place characterization) and neighborhood characteristics such as demographic mix would be contaminated by these spatial biases. Heterogeneity in network access might induce similar biases in data derived from location-based services (Crang, Crosbie, & Graham, 2006).

UPOP arises further in the data creation process, especially with intentional authoring of geographic data. When geographic data are created in-situ, artefacts of technology use can introduce uncertainties that extend beyond well-known types of error that relate to technical limitations (e.g., multi-path GPS error) or human error. We highlight here three ways that technological artefacts in data authoring can lead to UPOP. First, with some forms of spatial data it is not always clear if a recorded location reflects where an individual was at a point in time or if it describes where an observation of a more distant object or phenomenon was captured. For example, the location and tag metadata of geotagged photos from sites such as Flickr, Instagram, and Geograph have been used to gain new insights on questions as varied as tourist travel patterns, vernacular place boundaries, and place perceptions (García-Palomares, Gutiérrez, & Mínguez, 2015; Hollenstein & Purves, 2012; Jankowski, Andrienko, Andrienko, & Kisilevich, 2010; Li & Goodchild, 2012). When these data are combined with areal spatial data such as census zones, inference errors can occur if the location where the photo was taken from lies in a different zone (i.e., MCU) from the zone where the photo’s subject is found (i.e., TCU). This sensor–subject displacement is most apparent with landmarks, vistas, and features that can be captured from afar, but is also possible with analyses that use fine-grained areal units (e.g., block level). While photo subjects can be inferred by inspecting image content or context (Crandall, Li, Lee, & Huttenlocher, 2016; Dunkel, 2015), mining text tag attributes (e.g., Feick & Robertson, 2014), or by direct calculation if sufficient metadata exist (e.g., focal length, sensor size, actual object size, etc.), considerable uncertainty can remain due to the ambiguous (e.g., multi-subject, place-based) nature of many images and the diversity of user folksonomies in composite VGI datasets. UPOP related to sensor–subject displacement can be more challenging to quantify with geodata that are by-products of communication. Georeferenced micro-blog, SMS, and instant messaging posts, for example, figure prominently in research on near real-time detection of natural disasters and humanitarian emergencies (Sakaki, Okazaki, & Matsuo, 2010). While people generally report more often and quickly on local events (Crooks, Croitoru, Stefanidis, & Radzikowski, 2013; Stephens & Poorthuis, 2015), UPOP can occur as posts are encoded with the author’s GPS or IP-based coordinates (MCU), rather than the event’s location (TCU). This is demonstrated well by biases in using Twitter to track rainfall (Kitamoto & Sagara, 2012), estimate perceived risk associated with the 2014 West African Ebola outbreak (Fung, Tse, Cheung, Miu, & Fu, 2014), and share information concerning natural disasters (Goodchild & Glennon, 2010; Shelton et al., 2014) and human emergencies such as football-related riots in Lexington, KY (Crampton et al., 2013).

Second, software bias has more subtle and opaque impacts on TCU–MCU correspondence and UPOP. Thatcher (2014), for example, highlights how the limited feature taxonomies and place name lists that developers embed in their mobile applications affect how people use apps for tasks such as navigation and search (Kitchin & Dodge, 2011). This type of software bias can also shape how people perceive and record information about their surroundings, as people reconcile nuanced and place-based human sensing with the exactness of spatial software data models. Information loss and MCU–TCU uncertainty can also be traced to spatial data models and particularly the overwhelming use of point features in web, mobile, and tracking applications to represent phenomena with more complex or indeterminate geometries. We highlight two examples here. First, people often use points to represent observations about areas, whether while recording landscape preferences in situ or when regional phenomena mined from text documents are georeferenced to point centroids (Brown & Pullar, 2012). This is especially problematic in multi-authored VGI and citizen science datasets, where uncertainty relates to area-to-point simplification and to differences in how individuals perceive their environments—both of which may result in individual data being associated with an incorrect areal context (Robertson, Feick, Sykora, Shankardass, & Shaughnessy, 2017). Second, new sources of GPS, telemetry, and sensor data now allow animal and human movement patterns to be documented at spatial and temporal resolutions that were
not possible previously (Batty et al., 2012; Long & Nelson, 2015). However, there is growing appreciation that chaining time-stamped points limits our understanding of movement behavior and that it is also necessary to examine these data in light of their relationships to area-based contextual influences that constrain or facilitate movement (Purves, Laube, Buchin, & Speckmann, 2014).

A third source of UPOP we term interpretational, which describes errors in the meaning conferred on spatial associations observed between individual and MCU data. As discussed above, in health geography studies spatial co-location is frequently used as a proxy for exposure (which depends on a multitude of factors such as location, time, and housing materials). In most geographic data utilizing point data, spatial association between event frequency (obtained via point-in-polygon counts) and environmental conditions in MCUs is taken to imply a spatial association. In this form of UPOP, we make incorrect inferences that result from mismatches between MCUs and TCUs. Returning to census data and geotagged social media data, it is not unusual for census zones in business and entertainment districts to show high counts of tweets and high tweet frequency on a per capita basis. Any associations made between Twitter message frequency and demographic variables such as age, education, and ethnic background measured for the census zones (e.g., people in professional and service industries tweet frequently) would be tenuous due to sampling bias and the fact that many of the individuals tweeting would not be counted as residents of these zones in the census (Robertson et al., 2017). All that can be concluded from such an analysis is that this type of user activity is associated with neighborhoods with said demographic profiles. Similar interpretational UPOP issues may result in analyses where systematic errors in individuals’ recorded locations cause them to be associated with an incorrect zone. For example, if tracking of mobile individuals is interrupted due to lack of network or satellite coverage, then points may be georeferenced where network connection is re-established and not represent the individual’s true context. We may see this when messages commuters send while on a subway are only georeferenced after they emerge from a station (Stockx, Hecht, & Schöning, 2014) or when a satellite-linked sensor mounted on a shark’s fin nears the water’s surface and can send a signal (e.g., see Domeier, Nasby-Lucas, & Lam, 2012). This irregular locational sampling is not of itself UPOP. However, in the first case there may be UPOP concerning the census zone a delayed message should be associated with, while in the latter example the information loss when a shark is deep in the water column for extended periods would hinder efforts to link real-time shark location points with grid cells that represent occurrence counts or marine habitat characteristics.

To summarize, the UPOP occurs any time sampled locations are measured, the measurements pertain to locations outside of the TCU, and research questions direct inferences from individual records to the areal units they fall within.

6 IDENTIFYING AND ADDRESSING THE UPOP

Many researchers have recognized potential inference errors and uncertainty arising from UPOP and have demonstrated approaches to mitigate aspects of it on results. At the most fundamental level, there is a renewed appreciation that even in the era of big data, the datasets we are working with are often partial and indicative of easily recorded activities and processes (Kitchin, 2014; Miller & Goodchild, 2015). Consider a study of GPS data from urban bike commuters and accident counts by neighborhood. While we can confidently explore variations in bike rider characteristics (e.g., age, length of commute) within the sample and by neighborhood, the dynamic of cycling makes it much more tenuous to link characteristics of bike riders and the neighborhood populations they are cycling through (e.g., are cycling accidents related to neighborhood age structures?). Kelley (2014, p. 17), for example, notes that “[t]here is no way to know, for certain, the connection between users and the geographies where they actively produce geosocial information.” Acknowledgment of this reality and, at least implicitly, of possible UPOP effects is evident in the ways that many have crafted research questions and data processing methods to avoid erroneous individual–area inferences.

For example, while fragments of people’s daily movements and interactions can readily be uncovered from a variety of digital traces (e.g., use of smart transit passes, social media), typically little is known of these individuals, including where they reside (Elwood & Leszczynski, 2013). To improve the likelihood that data points represent “local” residents and allow linkages to zonal socio-demographic variables to be explored, filtering is often used to
exclude individuals whose time-stamped data points fail to meet an arbitrary (e.g., 14 days) residency threshold (Li et al., 2013; Robertson & Feick, 2016). This binary separation of “locals” and “tourists” offers a reasonable first cut for reducing sampling bias UPOP and may also shed light on how differences in familiarity with an area may influence urban space use or place perceptions (Hollenstein & Purves, 2012; Jankowski et al., 2010). However, it is

FIGURE 4 Scatterplot of WalkScore and positive sentiment tweets at: (a) the census tract; and (b) the neighborhood scale
limited to scales of analyses where an MCU can be expected to capture the majority of individuals’ regular patterns of movement (e.g., community or commuter shed with motorized travel modes, neighborhood for pedestrian travel). Li et al. (2013) illustrate this scale sensitivity by restricting their analysis of socio-demographic characteristics of Twitter and Flickr users to the county level.

The concept of open and closed systems in ecology can be adapted to provide guidance on analysis scales that minimize UPOP uncertainty with mobile data. To paraphrase Wiens (1989), a closed system with respect to UPOP is a unit of geography large enough that it captures the majority of individuals’ movements, while an open system permits flows between units. This concept can be operationalized with MCUs that represent smaller units of geography by centering analyses on more functional portrayals of individual mobility. In this way, areas of dominant space use can be distinguished from the more occasional, and thereby shed light on individuals’ behavior and MCU representativeness.

For human-centered data, sustained-use personal activity zones have been used to distil more realistic views of a person’s TCU (Huang & Wong, 2016; Kwan, 2012; Robertson et al., 2017), while concepts such as spatial range and home range are used more commonly with animal data (Long & Nelson, 2015). In both cases, the temporal sensitivity is important given that individuals’ behavior often occurs across several functional zones within a day (e.g., residential, work, recreation) or season (e.g., animals’ winter and summer ranges) (Hickman, 2013; Long & Nelson, 2015). Through the use of these individual-centric and time-sensitive approaches, uncertainty related to sampling bias UPOP (i.e., representativeness) can be recast in terms of degrees of exposure to specific MCUs. This offers potential to diagnose sampling bias UPOP and to explore more nuanced analyses that respect the conditional and scale-sensitive nature of many types of individual–area associations. Scaling associations made at the individual level up to the population level remain an active spatial research challenge.

7 | CASE STUDY: GEOREFERENCED TWEETS IN THE CITY OF TORONTO, CANADA

As part of ongoing research into urban stress and geosocial data (Sykora et al., 2015), we collected georeferenced tweets for the City of Toronto during the years 2013 and 2014. Details of these specific data are reported elsewhere (Robertson et al., 2017). For purposes of illustrating UPOP, we examined the relationship between tweet sentiment and a widely used metric of the quality of the pedestrian environment, WalkScore. This analysis was designed to mimic studies that link geographic variables describing the environment to social media content/VGI co-occurring in space in an attempt to identify associations and/or causal links (e.g., Quercia, Ellis, Capra, & Crowcroft, 2012; Tasse & Hong, 2014). A preliminary analysis of the relationship between positive sentiment and WalkScore is given in Figure 4, which shows a positive relationship at both the census tract scale (n = 531, Figure 4a) and stronger at the neighborhood scale (n = 140, Figure 4b). Such a finding could be considered evidence of a causal mechanism, whereby more walkable neighborhoods contribute to well-being and emotional affect, which shows up in aggregate measures of social media sentiment. This pattern is backed up by the regression analyses in Table 3, which quantify the degree of association at both scales, achieving an $R^2$ of 0.59 at the neighborhood scale.

| Table 3 | Naïve regression model results: positive tweet sentiment and local walkability scores |
|---------|-------------------------------------------------------------------------------|
| Scale   | Model                          | Term               | Coefficient | Standard error | t Statistic |
| Census tract (n = 531) | % Positive $\sim$ Walkscore ($R^2 = .14$) | Intercept* | 0.4007184      | 0.0145753      | 27.49       |
|         |                                 | Walkscore* | 0.0018242      | 0.0002022      | 9.02        |
| Neighborhood (n = 140) | % Positive $\sim$ Walkscore ($R^2 = .59$) | Intercept* | 0.3340978      | 0.0142951      | 23.37       |
|         |                                 | Walkscore* | 0.0028059      | 0.0002006      | 13.98       |

*p < .05.
neighborhoods—perhaps due to higher overall well-being conditioned in part by their neighborhood’s access to amenities; (b) when people happen to be “in” walkable neighborhoods, they tend to tweet more positively than when they are in less walkable neighborhoods; or (c) more walkable neighborhoods attract people more likely to tweet positively.

As discussed above, one way to attempt to parse these alternative interpretations is to separate “local” tweeters (i.e., high MCU–TCU correspondence) from those who may be just passing through a neighborhood (i.e., low MCU–TCU correspondence). In previous work, we used spatial clustering and density ranking to estimate individuals’ likely

**FIGURE 5** Maps of: (a) positive sentiment tweets; and (b) predicted residential tweeters
home and work locales from their geosocial footprints (Robertson et al., 2017). In this example, we find the densest cluster of tweets for each individual and designate that as their predicted residential locale. Next, we identified individual tweets as coming from this residential locale or not, and then enumerated the proportion of residential tweets by both geographical units (MCUs), in this case census tracts and neighborhoods in the City of Toronto. The proportion of tweets that had positive sentiment (Figure 5a) and the proportion predicted to be in a home locale (Figure 5b) at the neighborhood scale are given in Figure 5. Here we see that the spatial patterns in Figures 5a and b do not wholly correspond, and that the central neighborhoods have many more non-resident tweets than outlying areas. This aligns with the nature of activities and amenities in these areas, which serve to draw people from around the city for leisure, entertainment, and work, some of which may be posting to social media. These activities are likely confounded with sentiment expressed on social media. When we factor this into our regression model, we see a dramatic change in results which shows the new variable (percent resident) with a large and significant impact on tweet sentiment and the effect of walkability is negligible (Table 4). We have some multicollinearity between these variables, making it impossible to disentangle the correct interpretation from the association. However, cursory exploration of some of the non-resident tweets in the central area backs up the interpretation that these are driving up aggregate sentiment.

**8 | CONCLUSIONS**

The UPOP is a problem that arises when individual data are incorrectly linked to contextual units used to identify group-level differences and spatial patterns in a geographic variable. This problem is of increasing concern as more individual-level data become available to researchers from low-cost location-tracking technologies and sensors. We have shown that UPOP can be framed within a typology of geographical analysis problems related to inference with spatial data that arise from the use of polygonal units to describe contextual environmental/geographic variables. UPOP is rooted in the concept of an unrepresentative spatial sample, which is based on the research objective being investigated. The same Twitter data which perhaps could lead to flawed conclusions between neighborhood walkability and mental health outcomes could be used for accurate mapping of weather observations or landmarks. As such, we position UPOP as a critical consideration of spatial sampling design. Implicit acknowledgment of this problem is evident from the literature by researchers who have devised methods and ad-hoc schemes to minimize its effect using large geographical units (Li & Goodchild, 2012) or by filtering tourists from locals (Robertson & Feick, 2016). In this article, we attempted to make explicit the sources, impacts, and potential remedies to this problem in order to provide a starting point for additional research.

The issues of interpretation of spatial association made evident in the case study highlight the difficulty in moving from correlation to causation in observational data, especially in terms of environmental–individual processes. New
sources of individual and point-based sensor/observational data, combined with increasingly available open data that describe a variety of environmental and socioeconomic conditions, provide more opportunities to make spatial associations between variables represented in disparate datasets. However, due to UPOP and UGCoP, such spatial associations may also be clouded by multiple, often conflicting, causal interpretations which may be impossible to untangle. This demonstrates the increasing need for and emphasis on careful spatial analysis and interpretation of spatial patterns and associations. While identifying explicit strategies to deal with UPOP and UGCoP as part of research planning and design is an ideal endpoint, further research is first needed to develop the tools and strategies for handling these issues in a variety of spatial research contexts. This is especially true in geography, where visualization of spatial patterns and associations can make powerful impacts, while masking underlying uncertainties inherent in the data.

ORCID

Colin Robertson  http://orcid.org/0000-0003-0998-2971

REFERENCES

Ahas, R., Aasa, A., Yuan, Y., Raubal, M., Smoreda, Z., Liu, Y., … Zook, M. (2015). Everyday space–time geographies: Using mobile phone-based sensor data to monitor urban activity in Harbin, Paris, and Tallinn. *International Journal of Geographical Information Science, 29*(11), 2017–2039.

Aldstadt, J., & Getis, A. (2006). Using AMOEBA to create a spatial weights matrix and identify spatial clusters. *Geographical Analysis, 38*(4), 327–343.

Baecke, P., & Bocca, L. (2017). The value of vehicle telematics data in insurance risk selection processes. *Decision Support Systems, 98*, 69–79.

Batty, M., Axhausen, K. W., Giannotti, F., Pozdnoukhov, A., Bazzani, A., Wachowicz, M., … Portugali, Y. (2012). Smart cities of the future. *European Physical Journal Special Topics, 214*(1), 481–518.

Boruff, B. J., Nathan, A., & Nijënstein, S. (2012). Using GPS technology to [re]-examine operational definitions of “neighbourhood” in place-based health research. *International Journal of Health Geographics, 11*, 22.

Brabham, D. C. (2012). The myth of amateur crowds: A critical discourse analysis of crowdsourcing coverage. *Information, Communication & Society, 15*(3), 394–410.

Brown, G. G., & Pullar, D. V. (2012). An evaluation of the use of points versus polygons in public participation geographic information systems using quasi-experimental design and Monte Carlo simulation. *International Journal of Geographical Information Science, 26*(2), 231–246.

Brunsdon, C., Fotheringham, S., & Charlton, M. (1998). Geographically weighted regression. *Journal of the Royal Statistical Society: Series D, 47*(3), 431–443.

Buytaert, W., Zulkaflı, Z., Grainger, S., Acosta, L., Alemie, T. C., Bastiaensen, J., … Foggin, M. (2014). Citizen science in hydrology and water resources: Opportunities for knowledge generation, ecosystem service management, and sustainable development. *Frontiers in Earth Science, 2*, 26.

Calabrese, F., Diao, M., Di Lorenzo, G., Ferreira, J., & Ratti, C. (2013). Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. *Transportation Research, Part C: Emerging Technologies, 26*, 301–313.

Chaix, B. (2009). Geographic life environments and coronary heart disease: A literature review, theoretical contributions, methodological updates, and a research agenda. *Annual Review of Public Health, 30*(1), 81–105.

Chaix, B., Méline, J., Duncan, S., Merrien, C., Karusisi, N., Perchoux, C., … Kestens, Y. (2013). GPS tracking in neighborhood and health studies: A step forward for environmental exposure assessment, a step backward for causal inference?. *Health & Place, 21*(Suppl. C), 46–51.

Chaix, B., Merlo, J., & Chauvin, P. (2005). Comparison of a spatial approach with the multilevel approach for investigating place effects on health: The example of healthcare utilisation in France. *Journal of Epidemiology & Community Health, 59*(6), 517–526.

Clark, W. A. V., & Avery, K. L. (1976). The effects of data aggregation in statistical analysis. *Geographical Analysis, 8*(4), 428–438.

Crampton, J. W., Graham, M., Poorthuis, A., Shelton, T., Stephens, M., Wilson, M. W., & Zook, M. (2013). Beyond the geotag: Situating ‘big data’ and leveraging the potential of the geoweb. *Cartography & Geographic Information Science, 40*(2), 130–139.
Haklay, M. (2013). Neogeography and the delusion of democratisation. *Environment & Planning A*, 45, 55–69.

Haklay, M., Singleton, A., & Parker, C. (2008). Web mapping 2.0: The neogeography of the Geoweb. *Geography Compass*, 2(6), 2011–2039.

Heuvelink, G. B. M., & Pebesma, E. J. (1999). Spatial aggregation and soil process modelling. *Geodema*, 89(1), 47–65.

Hickman, P. (2013). Third places and social interaction in deprived neighbourhoods in Great Britain. *Journal of Housing & the Built Environment*, 28(2), 221–236.

Hochstatter, T. G., Leonard, J., & McKinzie, C. W. (2016). U.S. Patent No. 20.160.180.56: Systems and methods for credit approval using geographic data. Alexandria, VA: U.S. Patent Office.

Holland, P. W., Glymour, C., & Granger, C. (1985). Statistics and causal inference. *ETS Research Report Series*, 1985(2), i–72.

Hollenstein, L., & Purves, R. (2012). Exploring place through user-generated content: Using Flickr tags to describe city cores. *Journal of Spatial Information Science*, 2010(1), 21–48.

Huang, Q., & Wong, D. W. (2016). Activity patterns, socioeconomic status and urban spatial structure: What can social media data tell us?. *International Journal of Geographical Information Science*, 30(9), 1873–1898.

Hystad, P., Demers, P. A., Johnson, K. C., Brook, J., van Donkelaar, A., Lamsal, L., . . . Brauer, M. (2012). Spatiotemporal air pollution exposure assessment for a Canadian population-based lung cancer case control study. *Environmental Health*, 11(1), 22.

Jacquez, G. M., Sabel, C. E., & Shi, C. (2015). Genetic GIScience: Toward a place-based synthesis of the genome, exosome, and behavome. *Annals of the Association of American Geographers*, 105(3), 454–472.

Jankowski, P., Andrienko, N., Andrienko, G., & Kisilevich, S. (2010). Discovering landmark preferences and movement patterns from photo postings. *Transactions in GIS*, 14, 833–852.

Jelinski, D. E., & Wu, J. (1996). The modifiable areal unit problem and implications for landscape ecology. *Landscape Ecology*, 11(3), 129–140.

Jones, K. (1991). Specifying and estimating multi-level models for geographical research. *Transactions of the Institute of British Geographers*, 16(2), 148–159.

Jones, M. C., Marron, J. S., & Sheather, S. J. (1996). A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association*, 91(433), 401–407.

Kelley, M. J. (2014). Urban experience takes an informational turn: Mobile internet usage and the unevenness of geosocial activity. *GeoJournal*, 79(1), 15–29.

Kestens, Y., Wasfi, R., Naud, A., & Chaix, B. (2017). “Contextualizing context”: Reconciling environmental exposures, social networks, and location preferences in health research. *Current Environmental Health Reports*, 4(1), 51–60.

Kie, J. G., Matthiopoulos, J., Fieberg, J., Powell, R. A., Cagnacci, F., Mitchell, M. S., . . . Moorcroft, P. R. (2010). The home range concept: Are traditional estimators still relevant with modern telemetry technology?. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 365(1550), 2221–2231.

Kitamoto, A., & Sagara, T. (2012). Toponym-based geotagging for observing precipitation from social and scientific data streams. In *Proceedings of the First ACM Multimedia Workshop on Geotagging and its Applications in Multimedia* (pp. 23–26). Nara, Japan: ACM.

Kitchin, R. (2014). *The data revolution: Big data, open data, data infrastructures and their consequences*. Thousand Oaks, CA: Sage.

Kitchin, R., & Dodge, M. (2011). *Code/space: Software and everyday life*. Cambridge, MA: MIT Press.

Klein, K. J., & Kozlowski, S. W. (2000). From micro to meso: Critical steps in conceptualizing and conducting multilevel research. *Organizational Research Methods*, 3(3), 211–236.

Kravets, N., & Hadden, W. C. (2007). The accuracy of address coding and the effects of coding errors. *Health & Place*, 13(1), 293–298.

Krieger, N., Waterman, P., Lemieux, K., Zierler, S., & Hogan, J. W. (2001). On the wrong side of the tracts? Evaluating the accuracy of geocoding in public health research. *American Journal of Public Health*, 91(7), 1114–1116.

Kulldorff, M., & Nagarwalla, N. (1995). Spatial disease clusters: Detection and inference. *Statistics in Medicine*, 14, 799–810.

Kwan, M.-P. (2012). The uncertain geographic context problem. *Annals of the Association of American Geographers*, 102(5), 958–968.

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), 274–282.
Li, L., & Goodchild, M. F. (2012). Constructing places from spatial footprints. In Proceedings of the First ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information (pp. 15–21). Redondo Beach, CA: ACM.

Li, L., Goodchild, M. F., & Xu, B. (2013). Spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr. Cartography & Geographic Information Science, 40(2), 61–77.

Li, S., Dragicevic, S., Anton, F., Sester, M., Winter, S., Coltekin, A., ... Cheng, T. (2016). Geospatial big data handling theory and methods: A review and research challenges. ISPRS Journal of Photogrammetry & Remote Sensing, 115, 119–133.

Llorente, A., Garcia-Herranz, M., Cebrian, M., & Moro, E. (2015). Social media fingerprints of unemployment. PLoS One, 10(5), e0128692.

Long, J. A., & Nelson, T. A. (2015). Home range and habitat analysis using dynamic time geography. Journal of Wildlife Management, 79(3), 481–490.

Malizia, N. (2013). The effect of data inaccuracy on tests of space-time interaction. Transactions in GIS, 17(3), 426–451.

Matthews, S. A. (2011). Spatial Polygamy and the heterogeneity of place: Studying people and place via egocentric methods. In L. M. Burton, S. P. Kemp, M. Leung, K. A. Matthews, & D. T. Takeuchi (Eds.), Communities, neighborhoods, and health: Expanding the boundaries of place (pp. 35–55). New York, NY: Springer.

Matthews, S. A., & Yang, T.-C. (2013). Spatial polygamy and contextual exposures (SPACES): Promoting activity space approaches in research on place and health. American Behavioral Scientist, 57(8), 1057–1081.

McGarigal, K., Wan, H. Y., Zeller, K. A., Timm, B. C., & Cushman, S. A. (2016). Multi-scale habitat selection modeling: A review and outlook. Landscape Ecology, 31(6), 1161–1175.

Miller, H. J., & Goodchild, M. F. (2015). Data-driven geography. GeoJournal, 80(4), 449–461.

Miller, H. J., & Wentz, E. A. (2003). Representation and spatial analysis in geographic information systems. Annals of the Association of American Geographers, 93(3), 574–594.

Mitchell, L., Frank, M. R., Harris, K. D., Dodds, P. S., & Danforth, C. M. (2013). The geography of happiness: Connecting Twitter sentiment and expression, demographics, and objective characteristics of place. PLoS One, 8(5), e64417.

Mooney, P., & Corcoran, P. (2012). The annotation process in OpenStreetMap. Transactions in GIS, 16(4), 561–579.

Mullen, W. F., Jackson, S. P., Croitoru, A., Crooks, A., Stefanidis, A., & Agouris, P. (2015). Assessing the impact of demographic characteristics on spatial error in volunteered geographic information features. GeoJournal, 80(4), 587–605.

Nelson, T. A., & Boots, B. (2008). Detecting spatial hot spots in landscape ecology. Ecography, 31(5), 556–566.

Nelson, T. A., & Robertson, C. (2012). Refining spatial neighbourhoods to capture terrain effects. Ecological Processes, 1, 3.

Openshaw, S. (1984). Ecological fallacies and the analysis of areal census data. GeoJournal, 7(3–4), 382–398.

Preece, J. (2016). Citizen science: New research challenges for human–computer interaction. International Journal of Human–Computer Interaction, 32(8), 585–612.

Puig, X., & Ginebra, J. (2015). Ecological inference and spatial variation of individual behavior: National divide and elections in Catalonia. Geographical Analysis, 47(3), 262–283.

Quercia, D., Ellis, J., Capra, L., & Crowcroft, J. (2012). Tracking “gross community happiness” from tweets. In Proceedings of the 15th ACM Conference on Computer Supported Cooperative Work (pp. 965–968). Bellevue, WA: ACM.

Quackenboss, J. J., Lebowitz, M. D., & Crutchfield, C. D. (1989). Indoor-outdoor relationships for particulate matter: Exposure classifications and health effects. Environment International, 15(1), 353–360.

Rainham, D., McDowell, I., Krewski, D., & Sawada, M. (2010). Conceptualizing the healthscape: Contributions of time geography, location technologies and spatial ecology to place and health research. Social Science & Medicine, 70(5), 668–676.

Regalia, B., McKenzie, G., Gao, S., & Janowicz, K. (2016). Crowdsensing smart ambient environments and services. Transactions in GIS, 20(3), 382–398.

Robertson, C., & Feick, R. (2016). Bumps and bruises in the digital skins of cities: Unevenly distributed user-generated content across US urban areas. Cartography & Geographic Information Science, 43(4), 283–300.

Robertson, C., Feick, R., Sykora, M., Shankardass, K., & Shaughnessy, K. (2017). Personal activity centres and geosocial data analysis: Combining big data with small data. In A. Bregt, T. Sarjakoski, R. van Lammeren, & F. Rip (Eds.), Societal Geo-innovation: Selected papers of the 20th AGILE Conference on Geographic Information Science (pp. 145–161). Cham, Switzerland: Springer.
Wang, H., Kifer, D., Graif, C., & Li, Z. (2016). Crime rate inference with big data. In Wan, N., Lin Kan, G., & Wilson, G. (2017). Addressing location uncertainties in GPS-based activity monitoring: A methodological framework. Transactions in GIS, 21(4), 764–781.

Wang, H., Kifer, D., Graif, C., & Li, Z. (2016). Crime rate inference with big data. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 635–644). San Francisco, CA: ACM.

Warf, B., & Arias, S. (2008). The spatial turn: Interdisciplinary perspectives. New York, NY: Taylor & Francis.
Wiens, J. A. (1989). Spatial scaling in ecology. *Functional Ecology*, 3(4), 385–397.

Yang, A., Fan, H., & Jing, N. (2016). Amateur or professional: Assessing the expertise of major contributors in OpenStreetMap based on contributing behaviors. *ISPRS International Journal of Geo-Information*, 5(2), 21.

**How to cite this article:** Robertson C, Feick R. Inference and analysis across spatial supports in the big data era: Uncertain point observations and geographic contexts. *Transactions in GIS*. 2018;22:455–476. [https://doi.org/10.1111/tgis.12321](https://doi.org/10.1111/tgis.12321)