

RESEARCH

Open Access



Spatial measurement errors in the field of spatial epidemiology

Zhijie Zhang^{1,2*}, Justin Manjourides³, Ted Cohen^{4,5,6}, Yi Hu^{1,2} and Qingwu Jiang^{1,2}

Abstract

Background: Spatial epidemiology has been aided by advances in geographic information systems, remote sensing, global positioning systems and the development of new statistical methodologies specifically designed for such data. Given the growing popularity of these studies, we sought to review and analyze the types of spatial measurement errors commonly encountered during spatial epidemiological analysis of spatial data.

Methods: Google Scholar, Medline, and Scopus databases were searched using a broad set of terms for papers indexed by a term indicating location (space or geography or location or position) and measurement error (measurement error or measurement inaccuracy or misclassification or uncertainty): we reviewed all papers appearing before December 20, 2014. These papers and their citations were reviewed to identify the relevance to our review.

Results: We were able to define and classify spatial measurement errors into four groups: (1) pure spatial location measurement errors, including both non-instrumental errors (multiple addresses, geocoding errors, outcome aggregations, and covariate aggregation) and instrumental errors; (2) location-based outcome measurement error (purely outcome measurement errors and missing outcome measurements); (3) location-based covariate measurement errors (address proxies); and (4) Covariate-Outcome spatial misaligned measurement errors. We propose how these four classes of errors can be unified within an integrated theoretical model and possible solutions were discussed.

Conclusion: Spatial measurement errors are ubiquitous threat to the validity of spatial epidemiological studies. We propose a systematic framework for understanding the various mechanisms which generate spatial measurement errors and present practical examples of such errors.

Keywords: Spatial epidemiology, Environmental epidemiology, GIS, Geographical epidemiology, Measurement error, Misclassification

Background

Studying the distribution of health-related events in specified population, over time and across space is the business of epidemiologists [1]; however, until a decade ago, the use of spatial data was mainly descriptive. As the rapid development of spatial information techniques [e.g., geographic information systems (GIS), remote sensing (RS), and global positioning systems (GPS)], the availability of spatially referenced health data and associated risk factors in digital format has increased greatly [2–4]. This has been accompanied by the appearance of

new software packages for spatial data analysis. Together, these developments have created unprecedented new opportunities for researchers to investigate the association of geographically indexed health events with various demographic, environmental, behavioral, socioeconomic, and genetic risk factors to explore and explain geographic variation in disease risk. Researchers are becoming more familiar with space-related epidemiological studies (hereafter referred to as “spatial epidemiology”) [5–7].

As access to spatial data and spatial analytic approaches advances, so does the need to address sources of bias in spatial epidemiology. In most studies of spatial epidemiology, the data are assumed to be reliable and free of measurement errors. But in practice, this is often not the case. Measurement errors can affect the data through several

*Correspondence: epistat@gmail.com

¹ Department of Epidemiology and Biostatistics, School of Public Health, Fudan University, Shanghai 200032, China

Full list of author information is available at the end of the article

mechanisms and many different stages of data collection and analysis (the term “measurement errors” is most commonly used for continuous variables and “misclassification” for categorical variables, but for convenience we will use the term of measurement errors in the remainder of the paper since misclassification can be considered a special case of measurement errors). Spatial measurement errors can be random (such as errors originating from a GPS device) or there may be quantifiable factors contributing to the errors (such as errors due to imperfect sensitivity or specificity of a diagnostic test, latency periods of diseases, multiple addresses, etc.). While the concept of measurement errors is often discussed in classical epidemiology [8–10], it has not yet received much discussion as it relates to spatial epidemiologic studies [11].

Spatial epidemiologic data is like all epidemiological data, except it has at least one additional attribute describing the spatial or geographical location for each observation. Hence, some measurement errors encountered in spatial epidemiology are analogous to those previously described in classical epidemiology, but the spatial component of the data introduces additional measurement errors which need to be described, categorized, and accounted for Elliot et al. and Beale et al. [3, 4, 11] described this issue with regard to bias and confounding, but there has not been a systematic investigation of location-based measurement errors from the spatial perspective. A rigorous discussion and classification of these types of spatial errors is vital to ensure valid inference can be obtained from spatial epidemiologic studies [4].

Methods

Searches for relevant studies were carried out using two academic databases (Medline and Scopus) between the available earliest date and December 20, 2014. Databases were searched using a broad search term purposefully relating to “space” (space or geography or location or position) as well as a term relating to “measurement error” (measurement error or measurement inaccuracy or misclassification or uncertainty). The search strategy was (spatial [Title/Abstract] OR space [Title/Abstract] OR geographic * [Title/Abstract] OR geography [Title/Abstract] OR locational [Title/Abstract] OR location [Title/Abstract] OR positional [Title/Abstract] OR position [Title/Abstract] AND (measurement error * [Title/Abstract] OR measurement inaccurac * [Title/Abstract] OR misclassification * [Title/Abstract] OR uncertain * [Title/Abstract]) AND (english [Language])). We manually reviewed the abstracts to judge the relevance with spatial measurement errors and then we reviewed the cited references in the selected papers to identify additional potential articles. Those studies focusing on statistical methods or epidemiologic issues were also

included in this review, while studies focusing mainly on technical problems such as equipment usage were deemed out of scope for the current review.

Conceptual framework for classifying sources of spatial measurement errors

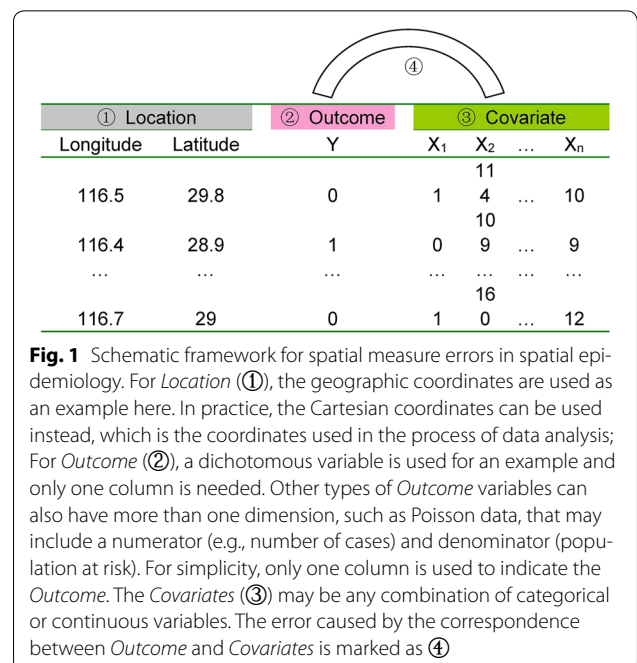
To discuss the spatial measurement errors systematically, we suggest the following conceptual framework as a helpful way to effectively organize the topic. We recognize that there is no way to completely capture the intricacies of all possible routes of measurement errors, however we present a basic structure, that can be expanded upon, to begin the process of systematically categorizing the types and structure of these potential errors. Spatial epidemiologic studies include two additional variables recording the locations/positions of observations in the leftmost two columns, while the other outcome and covariate variables are often similar to those used in classical epidemiology (Fig. 1).

The following four types of spatial measure errors are divided accordingly from the point of application:

1. Pure spatial location measurement errors;
2. Location-based outcome measurement errors;
3. Location-based covariate measurement errors; and;
4. Covariate-Outcome spatial misaligned measurement errors.

Theoretical model to characterize spatial measurement errors

These sources of spatial measurement errors can be integrated in a mathematical framework. Spatial



epidemiology studies include two variables representing outcome locations (L), outcome measurements (Y), and covariate measures (X) that may include spatial information as well. Consider one example where the unit of analysis is the individual, and a researcher is trying to identify associations between cancer status and exposure to contaminated drinking water. In this setting L would represent the individual's location, Y would be that individual's case or control status, and X could represent an individual's spatially varying exposure (e.g., amount of exposure to contaminated water). In another setting, one may be interested in relating deaths in a zip code to air pollution concentrations measured at an air pollution monitoring site. In this second scenario, L would be the common location assigned to everyone in the same zip code, Y would be the number of deaths observed, and X would be the spatially varying measure of air pollution that would be assigned to all individuals within a certain radius of the monitoring site. Suppose (L_0, Y_0, X_0) are the true measures of the outcome location, outcome, and covariate measures, respectively. In practice this information is measured with some errors. Let (L, Y, X) are the observed values of the outcome location, outcome, and covariate measures, which are the true value plus measurement errors Δ , $\Delta = (\Delta_L, \Delta_Y, \Delta_X)$. Each of these individual measurement errors may be a function of location, outcome, or exposure, or a combination of these measures. We present a simple linear combination framework for these potential errors below, but stress that the relationships between truth and measurements may be more complex. In the situation where more complex relationships may exist, these models can include more advanced modeling terms such as interactions between measurements, polynomials, and indicator random variables, or even distributional assumptions on the multiplicative coefficients. The basic formulation for our measurement error framework is given as follows:

$$\begin{aligned} L &= L_0 + \Delta_L, & \text{where } \Delta_L &= \gamma_0 L_0 + \gamma_1 Y + \gamma_2 X + \varepsilon_L \\ Y &= Y_0 + \Delta_Y, & \text{where } \Delta_Y &= \gamma_3 L + \gamma_4 Y_0 + \gamma_5 X + \varepsilon_Y \\ X &= X_0 + \Delta_X, & \text{where } \Delta_X &= \gamma_6 L + \gamma_7 Y + \gamma_8 X_0 + \varepsilon_X \end{aligned}$$

In the above equations, we define ε_L , ε_Y , and ε_X as the residual measurement errors remaining that is not directly related to the linear combination of the outcome, exposure, or covariates.

Results

7131 papers in Medline and 7595 papers in Scopus were found, where 149 papers were identified as closely related with the topic discussed in this study and were read thoroughly. 97 literatures were cited here.

Pure spatial location measurement errors

Pure spatial location measurement errors are introduced by inaccuracies in the positioning of spatial/geographical locations that will affect the outcome and covariate simultaneously. These errors can be broken down to instrumental errors (e.g., global positioning systems errors) and non-instrumental errors (e.g., multiple address, geocoding errors, outcome and covariate aggregations). We demonstrate how even pure spatial location measurement errors can lead to mismeasurement of both the outcomes and covariates, and accordingly, in this situation, the observed data is of the form (L, Y, X) , where no y s are guaranteed to be non-zero.

Instrumental errors

Instrumental errors are caused by inaccuracies in the tools used to measure the spatial locations (e.g., GPS). It is a worldwide, satellite-based, radio-navigation system developed by the U.S. Department of Defense (DOD) and may be the most widely used tool to obtain the geographical locations in spatial epidemiologic studies. A ground-based GPS receiver calculates the time it takes for individual signals to arrive from at least three satellites to the receiver to compute a two-dimensional location (latitude and longitude), given an assumed height. And the detection of satellite signals of four satellites can determine three-dimensional positions and time, whereas five or more can provide greater precision [12]. The accuracy of these GPS methods have been well reported [13]. It has been shown that positional errors resulting from inaccuracies in GPS system influence spatial analytic methods by inflating standard errors of estimates, which results in reduced power to detect spatial clustering and spatio-temporal trends. For example, Armstrong et al. [14] used the classical North Humber-side leukemia and lymphoma case-control data to quantify the effects of increasing levels of positional error perturbation on the deterioration of the power of the Cuzick-Edwards test for spatial clustering; Cassa et al. [15] added artificial clusters of various shapes and sizes to data on residence locations of individuals making hospital emergency department visits for respiratory illness. These locations were then perturbed at various levels according to a bivariate normal distribution with standard deviation inversely proportional to the local population density. The ability of the spatial scan statistic to detect the clusters was quantified and shown to decline as the level of perturbation increased; Olson et al. [16] used essentially the same baseline data as Cassa et al. [15], but moved the locations to zip code or census tract centroids rather than perturbing them according to a normal distribution, and obtained the qualitatively similar results; Gabrosek and

Cressie [17] and Cressie and Kornak [18] investigated, via simulation, the impact of circular uniform and normal perturbation of location errors on kriging approach and trend estimation and the latter authors showed how spatial autocorrelation in a geostatistical model attenuates as the perturbation level increases.

However, purely location-based measurement errors are inevitable. For example, GPS accuracy is affected by many factors such as the atmospheric conditions affecting the velocity of GPS signals, obstructions such as concrete and steel within many buildings preventing reception of GPS signals, and multi-path errors arising from the reflection of satellite signals from other surfaces including buildings, vegetation, the ground or water [19]. The actual errors may range from only centimeters to hundreds meters [12, 20]. In practical applications, the main consideration for addressing this type of errors is to balance these instrumental errors with the study scale.

In the presence of these instrumental errors, we observe the outcome (Y) accurately ($\gamma_3 = \gamma_4 = \gamma_5 = 0$). The errors in the measured locations are not related to the outcome, or the covariates ($\gamma_1 = \gamma_2 = 0$). As some covariates may depend on the mis-measured locations (i.e. $PM_{2.5}$ exposure at an individual's home address), observed covariates can be biased, meaning non-zero contributions of γ_6 and γ_8 . The observed data takes the form (L, Y_0, X) .

Non-instrumental errors

Multiple addresses Multiple addresses indicate that the studied cases/individuals have several locations where exposure or onset of disease may have occurred. Accordingly, multiple addresses raise the possibility of positional uncertainty since it may be difficult to ascertain the actual location for the outcome(s) and relevant covariate measurements. At the same time, capturing address history may provide the opportunity to adjust for this uncertainty when attempting to link exposures with outcomes or when seeking to identify spatial clusters of cases.

For example, studies of environmental exposures and adverse birth outcomes often rely on maternal address at birth obtained from the birth certificate to classify exposure. Although the gestational age of interest is often early pregnancy, maternal addresses are not available for women who move during pregnancy when using maternal addresses abstracted from birth certificates. Chen et al. [21] explored the extent of ambient air pollutant exposure measurement error due to maternal residential mobility during pregnancy among a subgroup within a New York birth cohort, but no significant impact was found, which may mainly be because of limited population mobility; A study of breast cancer in Upper Cape Cod, Massachusetts demonstrated potential exposure uncertainty introduced by a mobile population [22]. The authors conducted three

separate analyses on a sample of breast cancer cases: (1) all cases were considered (2) only cases that have lived at the same location for at least 15 years were considered and (3) only cases that have lived at the same location for at least 20 years were considered. The overall results associating location and disease (across three separate cluster detection techniques) showed decreasing p values as the lag increases. This result provides evidence of a spatial association that is stronger when only those cases that have not recently moved are included.

Several other studies have attempted to recognize the influence of multiple addresses, generally by only analyzing residential locations during a pre-specified time interval. One such study found that the most likely cluster of lymphoma in a case-control study was found by examining addresses with a 20 year lag period compared to the 5, 10, 15 years lag period [23]. Han et al. [24] used kernel density estimation methods to identify clustering of breast cancer using residential histories. Sabel et al. [25] examined clustering of Amyotrophic Lateral Sclerosis in Finland based on place of birth and place of death, respectively, showing incomplete agreement. Gallagher et al. [26] used residential histories to assess the affect of drinking water exposure to breast cancer, by examining any previous address where a study participant was exposed to public drinking water impacted by effluent.

Another potential solution for using multiple addresses involves using weighted distance-based methods to account for multiple addresses. By basing tests on the distances between locations of interest, researchers can give increased weight to addresses that may be more important (or informative) and less weight to addresses that may be less important (or less informative). Weighted distance based cluster detection methods have been shown to improve power to detect clustering in mobile populations [27].

In the presence of multiple addresses, independent of the outcome or covariate, we observe the outcome (Y) accurately ($\gamma_3 = \gamma_4 = \gamma_5 = 0$). The errors at measured locations are not influenced by the outcome, or the covariates ($\gamma_1 = \gamma_2 = 0$). As multiple addresses could affect the measured covariate (as is the case in Gallagher et al.) [26], γ_6 and γ_8 may be non-zero. The observed data takes the form (L, Y_0, X) . If the multiple addresses are not independent of the outcome or covariate (i.e. moving closer to elderly people moving to care facilities, or people moving as a result of a shift in socioeconomic status), then we can no longer assume that the corresponding γ terms are zero, and the observed data may take the form (L_0, Y_0, X) , (L, Y_0, X_0) , or (L_0, Y_0, X_0) .

Geocoding errors Geocoding requires the matching of an address of interest to an address-ranged street segment georeferenced within a street-line database, followed by

interpolation of the position of the address along that segment. This is always done through automated geocoding techniques widely used in geographical analyses. Unfortunately, geocoding results obtained by these automated procedures are well-known to contain positional errors of several hundred meters or more (defined here as the (vector) difference between the location of an address ascertained by automated geocoding and its corresponding ground-truth location) [28–41]. For example, in the study on a four-county area of upstate New York, Cayo and Talbot [31] found that 10 % of a sample of rural addresses geocoded with errors of more than 1.5 km, and 5 percent geocoded with errors exceeding 2.8 km; In a case study of Orange County, Florida, Zandbergen and Green [42] determined the effect of positional errors in geocoding on the analysis of exposure to traffic-related air pollution of children at school locations through comparisons with a parcel database and digital orthophotography, suggesting that typical geocoding is insufficient for fine-scale analysis of school locations; Mazumdar et al. [43] found, via simulation, that the magnitude of the odds ratio (OR) measuring the relationship between covariates (environmental exposure) and the outcome (disease) generally declined with decreasing geocoding accuracy; Zimmerman et al. [40] sought to model the probability distribution of positional errors associated with automated geocoding, where the mixtures of bivariate t distributions with few components appear to be flexible enough to fit many positional error datasets associated with geocoding.

A special case for geocoding errors is when no locations can be found by geocoding. The standard approach for handling these missing geocodes is to discard those observations and analyze only complete observations. Reich et al. [44] proposed a hierarchical Bayesian spatial model to handle missing observation locations. Through a simulation study, it was found that this method may allow for more reliable epidemiological analysis. The authors also applied this method to a study of the relationship between fine particulate matter and birth outcomes in southeast Georgia. Oliver et al. [45] described geographic bias in GIS analyses with unrepresentative data owing to missing geocodes, using as an example a spatial analysis of prostate cancer incidence among whites and African Americans in Virginia, 1990–1999. They found that cluster maps showed patterns that appeared markedly different, depending upon whether one used all cases or those geocoded to the census tract. Geocoding errors will also result in exposure measurement error, which will depend on the spatial variation of the exposure being studied [35].

In the presence of geocoding errors, the errors in the measured locations are not influenced by the outcome ($\gamma_1 = 0$), but may be influenced by certain covariates

(i.e. rural vs urban) leading to a potentially nonzero contribution of γ_2 . If we consider the outcome (Y) to be a count of the number of cases of a specific disease in a census tract or zip code, then the mismeasurement of locations of cases could influence these values and (Y) may also be inaccurate ($\gamma_4 = \gamma_5 = 0$). Some covariates may still depend on the location (which may be measured with error) resulting in non-zero contribution of γ_6 ($\gamma_7 = \gamma_8 = 0$). The observed data takes the form (L, Y, X).

Outcome aggregations Aggregation is a manipulation of data that is widely used in spatial epidemiologic studies either because of the availability of associated data or the need to protect confidentiality. Aggregation is usually performed at the level of particular administrative units. Upon aggregation of outcomes, researcher will need to specify a location to represent the aggregated outcome for the spatial data analysis. The centroid of the aggregated unit (e.g., postal/zip codes, census tracts, dissemination areas, blocks or block groups) is frequently used as the address proxy for sample unit locations, but this will always serve as a source of errors. Additionally, aggregation masks the original detailed outcome information; this may be even more complicated if the original (i.e. disaggregated) outcome included measurement errors.

Waller and Jacquez [46, 47] demonstrated empirically that the effect of aggregation on tests for focused spatial clustering and space–time clustering reduced power, and that the amount of power reduction was directly related to the level of aggregation. Ozonoff et al. [48] reported that increasing levels of aggregation led the spatial scan statistic to not only lose power to detect disease clusters but also to increase the false detection rate. Berke and Waller [49] used regional aggregated count data to investigate the measurement error effect of West Nile virus infections among dead birds sampled from the 30 public health units of southern Ontario in 2005 on semivariogram, Moran's I statistic and the spatial scan test. They found that no serious spatial bias was introduced by the use of an imperfect diagnostic test as long as the imperfection itself was spatially unbiased.

Aggregation is closely related with the well-known modified area unit problems (MAUPs), which remains an unsolved problem in geography. But if the aggregation bias includes the original outcome measurement errors, then Bayesian approaches with probability formulas and prior information on the potential measurement error probabilities of outcomes may be a promising approach for mitigating this issue [50, 51].

In the presence of outcome aggregations, the errors associated with the measured locations are not influenced by the outcome, or the covariates ($\gamma_1 = \gamma_2 = 0$). But

the outcome (Y) is not accurate for either the locations or the measurement ($\gamma_5 = 0$), which results in the inaccuracy of some covariates ($\gamma_6 = \gamma_7 = 0$). The observed data takes the form (L, Y, X) .

Covariate aggregation Covariate aggregation means that the aggregated covariates were used in place of individual covariates. For example, socio-economic covariates (e.g., gross domestic product, GDP) in aggregated units are sometimes used to represent the socio-economic status of individuals.

The impact of covariate aggregation in classical epidemiological studies has been well described, but in spatial epidemiologic studies the impact of such aggregation has not been explored in great detail. Many studies reviewed have simply combined the covariates at different scales in the process of spatial analysis, ignoring completely the potential effects of covariate aggregation [52–54]. For example, Raso et al. [55] used an integrated approach for risk profiling and spatial prediction of coinfection with *Schistosoma mansoni* and hookworm for western Côte d'Ivoire through combining demographic, environmental, and socioeconomic data, where the Normalized difference vegetation index and land surface temperature data with a spatial resolution of 1 km and the Rainfall estimate data with an 8 km spatial resolution were used.

Hierarchical modeling and analysis for spatial data, which considers the spatial relationships and hierarchical structures is one promising approach for dealing with covariate aggregation [56]. For example, Yang et al. [57] used a hierarchical multi-level model to explore the risk factors of schistosomiasis japonica by nesting the individual variables such as gender, age and occupation within the village level variables such as type of *S. japonicum* endemic area, drinking water source, and sewage treatment. However, methods to correct the covariate aggregation from different spatial resolutions as described above (i.e. not the obvious different scales such as county, village, individuals) and the associated impacts on the study results remained unexplored.

In the presence of covariate aggregations, the errors in the measured locations are not influenced by the outcome, or the covariates ($\gamma_1 = \gamma_2 = 0$). The covariate (X) is not accurate for either the locations or the measurement ($\gamma_7 = 0$), while the outcome (Y) is accurate ($\gamma_3 = \gamma_4 = \gamma_5 = 0$). The observed data takes the form (L, Y_0, X) .

Location-based outcome measurement errors

Pure outcome location measurement errors mean that the outcome measurements have errors in some locations (e.g., a case is misclassified as noncase), which will

only affect the outcome. It becomes more complicated because of its location attributes compared to non-spatial situations and may include two different types: purely outcome measurement errors, and missing outcome measurements.

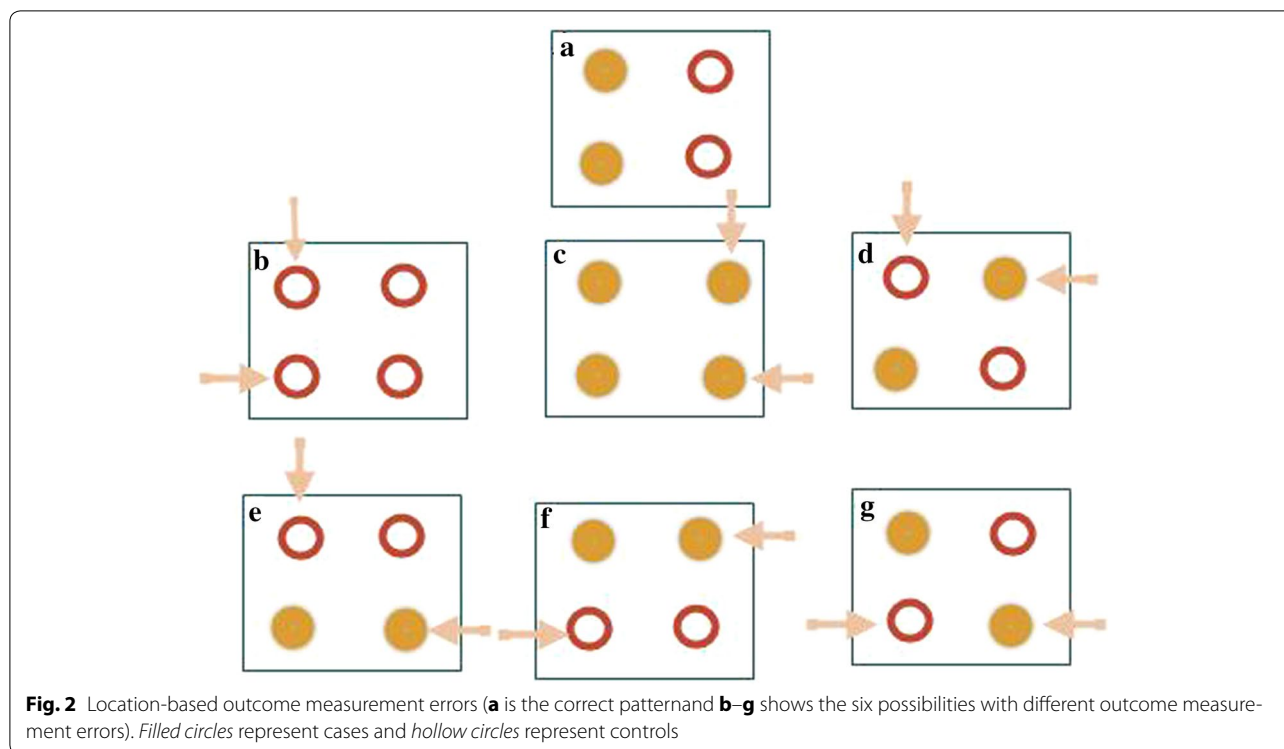
In the presence of purely outcome-based measurement errors or missing outcome measurements [58], the errors in the measured locations are assumed to be zero ($\gamma_0 = \gamma_1 = \gamma_2 = 0$). The outcome (Y) is not accurate for the measurement ($\gamma_3 = \gamma_5 = 0$), while the covariate is accurate ($\gamma_6 = \gamma_7 = \gamma_8 = 0$). The observed data takes the form (L_0, Y, X_0) , as there is only mismeasurement of the outcome.

Purely outcome measurement errors

In order to describe this form of measurement errors we introduce a simple example with a binary outcome. Suppose there are four locations and the true outcome includes two cases and two controls as shown in Fig. 2a. Say there are two measurement errors. Figure 2b–g shows six possibilities with different spatial patterns: one two-case misclassification, one two-control misclassification, and four one-case and one-control misclassifications.

In schistosomiasis studies, the fecal examination is always used as the “gold standard” test to diagnose a disease. However, as it is more difficult, and potentially more time consuming, to prove the complete absence of eggs in a clean stool sample than it is to prove the presence of eggs in a contaminated sample, especially in the low endemic regions, there could be many false negatives. In this situation, we may expect preferential misclassification of $(L, Y_0 = 1, X)$ to $(L, Y = 0, X)$. Another example follows from the surveillance of the highly pathogenic avian influenza (HPAI) H5N1. Many farmers are reluctant to report H5N1 cases among their livestock because of the considerations of economic loss [59, 60]. Hence, many locations with HPAI H5N1 cases are subsequently misclassified as locations without cases. These results are then linked with spatial location such as residential or farm address for spatial epidemiological studies, where the location-based outcome measurement errors occur. This will bias the results of spatial epidemiologic studies.

Li et al. [61] found that the misclassification of birth defects can elevate the state-wide congenital anomaly reporting rate from 1.1 to 1.8 % of live births, and after removing the misclassified data geographic clustering in congenital anomaly reports disappeared. Bihmann et al. [62] explored Bayesian logistic regression to adjust the outcome misclassification and concluded that adjustment for misclassification must be included to produce unbiased regression estimates.



Missing outcome measurements

Outcome measurements may be completely missing at some locations for various reasons such as non-responses, uncorrected recording errors, and loss of records. Zukovi and Hristopoulos [63] attempted to address the problem of missing values estimation on two-dimensional grids by means of spatial classification methods based on spin models. The “spin” variables provide an interval discretization of the process values, and the spatial correlations are captured in terms of interactions between the spins. The spins at the unmeasured locations are classified by means of the “energy matching” principle: the correlation energy of the entire grid (including prediction sites) is estimated from the sample-based correlations. They also compared the spin-based methods with standard classifiers such as the k -nearest neighbor, the fuzzy k -nearest neighbor, and the support vector machine (SVM), finding that the spin-based classifiers provide competitive choices. While in classical statistics, the techniques such as multiple imputation have been widely studied [64], the extension of these approaches to address missing data in spatial analyses has not received wide attention. Within the framework proposed in this work, the true and unobserved Y would be replaced by Y_0 , estimated via a function dependent on both spatial and non-spatial terms. In Manjourides et al. [58], the missing outcomes were imputed based on both

the distance from the nearest health center and a disease-specific covariate that was not spatially dependent. Assuming the covariates are measured with no errors, non-zero gammas could be present in the estimation of the distance from the nearest health center (nonzero γ_0 , γ_1) and in the estimation of the outcome (non zero γ_3 , γ_4). If we assume that the location and covariate are both measured without errors, then following the reasoning from multiple imputation, we could set γ_4 equal to the average of the observed Y s times an indicator that is 1 when Y is unobserved and 0 otherwise.

Location-based covariate measurement errors

Pure covariate location measurement errors occur when imprecise locations are used for covariate measurements. Covariate measurement errors have been commonly observed in public health studies. Li et al. [65] showed that ignoring this type of measurement errors would result in attenuated regression coefficients and in inflated variance components. For spatial covariate measurement errors, the imprecise address proxies are the major issue, while the traditional nonspatial covariate measurement errors has been studied widely and will not be discussed further [66, 67].

Imprecise address proxies occur when a single measure (e.g., central monitor data or spatial average estimates using data from multiple monitors) is used to

characterize a covariate such as ambient pollutant levels across a study area [68]. Only under the strong assumption of spatial homogeneity, will the problem of address proxies be avoided.

Few studies have reported and compared the positional discrepancies between address proxies and the exact address they are supposed to represent. Bow et al. [69] determined the locations in meters for both the street address (location of residence) and postal code location for each cardiac catheterization case in an urban Canadian City and found that 87.9 % of the postal code locations were within 200 m of the true address location (straight line distances) and 96.5 % were within 500 m of the address location, suggesting in this case the postal code locations may be a reasonably accurate proxy for address location. However, Healy et al. [70] quantified the magnitude of distance errors and accessibility misclassification that result from several commonly-used address proxies in spatial epidemiologic studies. They found that using address proxies based on large aggregated units such as centroids of census tracts or dissemination areas can result in large positional discrepancies (with median errors of 343 and 2088 m in urban and rural areas, respectively) and the commonly used proxies for residential address such as postal code centroids can also have large positional discrepancies (median errors of 109 and 1363 m in urban and rural areas, respectively), and are prone to misrepresenting accessibility in small towns and rural Canada. Lo Iacono et al. [71] found that mapping owner addresses as a proxy for horse location significantly underestimates the risk of an outbreak of African horse sickness (AHS) in Great Britain. Duncan et al. [72] used three different “neighborhood” definition, including specific home addresses, census block groups, and census tracts, to explore how neighborhood definition influences the measurement of youths’ spatial accessibility to tobacco retailers and found that measurements of neighborhood exposures likely vary depending on the definition of “neighborhood” selected. Accordingly, address proxies should be used with caution in spatial epidemiologic researches, which can lead to the associated nonspatial covariate measurement errors and further bias the results.

In the presence of location-based covariate measurement errors, we observe the outcome (Y) accurately ($\gamma_3 = \gamma_4 = \gamma_5 = 0$) and the measured location errors are zero ($\gamma_0 = \gamma_1 = \gamma_2 = 0$). The errors are completely contained in the covariates which are not accurate ($\gamma_6 \neq 0$, $\gamma_7 \neq 0$, $\gamma_8 \neq 0$). The observed data takes the form (L_0, Y_0, X) .

Covariate-Outcome spatially misaligned measurement errors

Covariate-Outcome spatial misaligned measurement errors comes from the process of alignment between

covariate and outcome, which is mainly caused by the inconsistent measurement or usage of location data. The outcome and the covariates are often observed at different locations or aggregated over different geographical units. Such data are said to be spatially misaligned. For example, a point-to-point misalignment problem arises when relating air quality measurements, observed at monitors (points), and birth weights observed at the residential locations of the mothers (different points). In many spatial epidemiologic studies, the locations of covariates and health outcomes do not match.

Standard regression methods cannot be applied to such misaligned data. To overcome this problem, several methods have been proposed. Most approaches involve directly using predictions from statistical exposure models that incorporate spatial structure [73–75] such as kriging and its extensions [76], Gaussian process (GP) modeling and Bayesian smoothing [77, 78], penalized regression splines [73, 79], and kernel smoothing [80], among others. The most common strategy is to employ one of the previously mentioned models to predict covariates at locations with outcomes and then estimate a regression parameter of interest using the predicted covariates. For example, Higgins et al. [81] used polynomial regression to generate covariate predictions when outcomes and covariates were misaligned. Waller and Gotway [82] used kriging to predict exposures and used resampling to account for the uncertainty in using the predictions in place of the true values. Kunzli et al. [74] assigned exposure values for subject-specific locations derived from a geostatistical model and used weighted least squares in the subsequent health effects model with the weights specified as the inverse of the standard errors (SEs) from the exposure kriging model. For a similar problem, Madsen et al. [83] considered both a generalized least squares (GLS) estimator with a bootstrap type variance estimator and a maximum likelihood approach that jointly fits the exposure and health models.

Such covariate predictions contain measurement errors since the predicted values will not equal the true exposures [84, 85]. Using predictions rather than true exposures in health modeling introduces two sources of measurement error—*Berkson-like errors* arising from smoothing the true covariate surface and *classical-like errors* coming from estimating the covariate model parameters, which have been widely discussed in environmental epidemiology [86]. For example, when assessing health effects of particulate matter (PM) constituents, it is necessary to effectively estimate exposure and to account for exposure errors induced by spatial misalignment to avoid bias [87]. After characterizing the spatial misalignment using geostatistical methods, Goldman et al. [88] found that errors due to spatial misalignment

resulted in average risk ratio reductions of <16 % for secondary pollutants (O3, PM2.5 sulfate, nitrate and ammonium) and between 43 and 68 % for primary pollutants (NO_x, NO₂, SO₂, CO, PM2.5 elemental carbon) while pollutants of mixed origin (PM10, PM2.5, PM2.5 organic carbon) had intermediate impacts. Sheppard et al. [89] found that measurement errors resulting from spatial misalignment led to an attenuation of acute health effect estimates of 7.7 % for PM2.5 mass. Ong et al. [90] found that relative to geographically corrected data, spatial misaligned information produced a modest bias in the aggregated number of facilities at risk but generated a substantial number of false positives and negatives.

Some methods to correct for spatial misalignment have been proposed. Lopiano et al. [91] developed an approach for an REML-based estimated generalized least squares (RBGLS) estimator accounting for the misalignment error structure and estimating covariance parameters using likelihood-based methods. They also provide insights into when it is important to fully account for the covariance structure induced from the different error sources. These researchers also developed another pseudo-penalized quasi-likelihood algorithm to account for spatial misaligned errors and showed by simulation that the method performs well in terms of coverage for 95 % confidence intervals [92]. Szpiro et al. [93] characterized the measurement errors by decomposing it into Berkson-like and classical-like components and developed two correction approaches of the parametric bootstrap and the “parameter bootstrap” [86] and also proposed another robust method for the spatially misaligned errors to correct finite-sample bias and correctly estimate standard errors. Gryparis et al. [84] developed a generalized linear model framework for spatial misaligned measurement error modeling; Chang et al. [94] addressed the challenge of exposure measurement

errors due to spatial misalignment through measurement error modeling and developed a Bayesian framework to fully account for uncertainty in the estimation of model parameters. Bayesian hierarchical models which account for uncertainties due to spatial misalignment were also applied to spatial correlated exposures measured with errors by Smith et al. [95], and Molitor et al. [96].

In the presence of Covariate-Outcome spatial misaligned measurement errors, the covariate is not accurate ($\gamma_6 \neq 0, \gamma_7 \neq 0, \gamma_8 \neq 0$). The mismeasured locations and the outcome (Y) are only affected by themselves ($\beta_0 \neq 1, \beta_4 \neq 1$). The observed data takes the form (L, Y, X).

Discussion

Like measurement errors in classical epidemiology, spatial measurement errors are also ubiquitous in spatial epidemiology. Some types of errors have been widely recognized and extensively studied in other disciplines. For example, geocoding errors arising from non-instrumental measurement errors are widely discussed in the field of geography and concern with spatial misaligned measurement errors have been raised in air pollution studies. However, we are not aware of previous efforts to systematically review and classify the types of spatial measurement errors.

We proposed a classification framework for spatial measurement errors which includes four categories (see summaries in Table 1). We then integrated these with a unified theoretical model; while we illustrated each type of errors as an isolated effect, in practice, many measurement errors can occur simultaneously. For simplicity, we have described only non-differential errors since differential errors (i.e. those errors in the probability of errors differs by location) will cause even greater mischief.

In this effort to review spatial measurement errors, we hope to attract the other spatial epidemiology

Table 1 Summaries of spatial measurement errors

Classes	Subclass	Observed data	Nonzero γ s
Pure spatial location measurement errors	Instrumental errors (e.g., global positioning systems errors)	(L, Y ₀ , X)	$\gamma_0, \gamma_6, \gamma_8$
	Non-instrumental errors: multiple address	(L, Y ₀ , X)	$\gamma_0, \gamma_6, \gamma_8$
	Non-instrumental errors: geocoding errors	(L, Y, X)	$\gamma_0, \gamma_2, \gamma_3, \gamma_6$
	Non-instrumental errors: outcome aggregations	(L, Y, X)	$\gamma_0, \gamma_3, \gamma_4, \gamma_8$
	Non-instrumental errors: covariate aggregations	(L, Y ₀ , X)	$\gamma_0, \gamma_6, \gamma_8$
Location-based outcome measurement errors	Purely outcome measurement errors	(L ₀ , Y, X ₀)	γ_4
	Missing outcome measurement	(L ₀ , Y, X ₀)	γ_4
Location-based covariate measurement errors	Location-based covariate measurement errors	(L ₀ , Y ₀ , X)	$\gamma_6, \gamma_7, \gamma_8$
Covariate-Outcome spatial misaligned measurement errors	Covariate-Outcome spatial misaligned measurement errors	(L, Y, X)	$\gamma_0, \gamma_4, \gamma_6, \gamma_7, \gamma_8$

$L = L_0 + \Delta_L$, where $\Delta_L = \gamma_0 L_0 + \gamma_1 Y + \gamma_2 X + \epsilon_L$; $Y = Y_0 + \Delta_Y$, where $\Delta_Y = \gamma_3 L + \gamma_4 Y_0 + \gamma_5 X + \epsilon_Y$; $X = X_0 + \Delta_X$, where $\Delta_X = \gamma_6 L + \gamma_7 Y + \gamma_8 X_0 + \epsilon_X$. (L_0, Y_0, X_0), (L, Y, X) and ($\Delta_L, \Delta_Y, \Delta_X$) are the true measures, observed values and measurement error of the outcome location, outcome, and covariate measures, respectively

researchers' attention to this common challenge. We believe there is opportunity for methodologists to develop new approaches for addressing spatial measurement errors, and provide some useful tools for applied spatial epidemiologists. We know our classification may be not perfect (e.g., geocoding errors can further result in the exposure/covariate measurement errors, suggesting there might be somewhat "overlapping") [97], but we are confident that this first attempt to add structure to this issue will generate greater discussion, consideration, and solutions to these very practical problems.

Conclusion

Spatial measurement errors are ubiquitous threat to the validity of spatial epidemiological studies. We propose a systematic framework in this study for understanding the various mechanisms which generate spatial measurement errors and present practical examples of such errors and potential solutions.

Authors' contributions

ZJ, JM and TC designed this study and carried out the initial literature research. YH and QJ participated in the literature review and discussion of the key ideas of the proposed framework. All authors read and approved the final manuscript.

Author details

¹ Department of Epidemiology and Biostatistics, School of Public Health, Fudan University, Shanghai 200032, China. ² Key Laboratory of Public Health Safety, Ministry of Education, Shanghai 200032, China. ³ Department of Health Sciences, Northeastern University, Boston, MA 02115, USA. ⁴ Department of Epidemiology and the Center for Communicable Disease Dynamics, School of Public Health, Harvard University, Boston, MA 02115, USA. ⁵ Division of Global Health Equity, Brigham and Women's Hospital, Boston, MA 02115, USA. ⁶ Harvard Medical School, Boston, MA 02115, USA.

Acknowledgements

This research was supported by a Foundation for the Author of National Excellent Doctoral Dissertation of PR China (FANEDD) (201186), the Talent Programs for Fostering Outstanding Youth of Shanghai (XYQ2013071), the National Science Fund for Distinguished Young Scholars (81325017) and Chang Jiang Scholars Program (T2014089).

Competing interests

The authors declare that they have no competing interests.

Received: 5 October 2015 Accepted: 15 June 2016

Published online: 01 July 2016

References

- Buettner P, Muller R. *Epidemiology*. Oxford: Oxford University Press; 2011.
- Zandbergen PA. A comparison of address point, parcel and street geocoding techniques. *Comput Environ Urban Syst*. 2008;32:214–32.
- Beale L, Abellan JJ, Hodgson S, Jarup L. Methodologic issues and approaches to spatial epidemiology. *Environ Health Perspect*. 2008;116:1105–10.
- Elliott P, Wartenberg D. Spatial epidemiology: current approaches and future challenges. *Environ Health Perspect*. 2004;112:998–1006.
- Hu Y, Gao J, Chi M, Luo C, Lynn H, Sun L, Tao B, Wang D, Zhang Z, Jiang Q. Spatio-temporal patterns of *Schistosomiasis japonica* in lake and marshland areas in China: the effect of snail habitats. *Am J Trop Med Hyg*. 2014;91:547–54.
- Hu Y, Xiong C, Zhang Z, Luo C, Ward M, Gao J, Zhang L, Jiang Q. Dynamics of spatial clustering of schistosomiasis in the Yangtze River Valley at the end of and following the World Bank Loan Project. *Parasitol Int*. 2014;63:500–5.
- Schrader M, Haufe T, Zhang Z, Davis GM, Jopp F, Remais JV, Wilke T. Spatially explicit modeling of schistosomiasis risk in Eastern China based on a synthesis of epidemiological, environmental and intermediate host genetic data. *PLoS Negl Trop Dis*. 2013;7:e2327.
- Rothman KJ, Greenland S, Lash TL. *Modern epidemiology*. 3rd ed. Pennsylvania: Lippincott Williams & Wilkins; 2012.
- Gustafson P. *Measurement error and misclassification in statistics and epidemiology: impacts and Bayesian adjustments*. London: Chapman & Hall; 2004.
- Buonaccorsi JP. *Measurement error: models, methods and applications*. Boca Raton, FL: Chapman & Hall/CRC; 2009.
- Elliott P, Wakefield P, Best N, Briggs D. *Spatial epidemiology: methods and applications*. USA: Oxford University Press; 2001.
- Dana P. *Global Positioning system overview*. The Geographer's Craft Project. Boulder, Colorado, USA: Department of Geography, The University of Colorado at Boulder; 2000.
- Goldberg DW. *A geocoding best practices guide*. Springfield: North American Association of Central Cancer Registries; 2008.
- Armstrong MP, Rushton G, Zimmerman DL. Geographically masking health data to preserve confidentiality. *Stat Med*. 1999;18:497–525.
- Cassa CA, Grannis SJ, Overhage JM, Mandl KD. A context-sensitive approach to anonymizing spatial surveillance data. *J Am Med Inform Assoc*. 2006;13:160–5.
- Olson KL, Grannis SJ, Mandl KD. Privacy protection versus cluster detection in spatial epidemiology. *Am J Public Health*. 2002;2006:96.
- Gabrosek J, Cressie N. The effect on attribute prediction of location uncertainty in spatial data. *Geogr Anal*. 2002;34:262–85.
- Cressie N, Kornak J. Spatial statistics in the presence of location error with an application to remote sensing of the environment. *Stat Sci*. 2003;18:436–56.
- Rainham D, Krewski D, McDowell I, Sawada M, Liekens B. Development of a wearable global positioning system for place and health research. *Int J Health Geogr*. 2008;7:59.
- Schipperijn J, Kerr J, Duncan S, Madsen T, Klinker CD, Troelsen J. Dynamic accuracy of GPS receivers for use in health research: a novel method to assess GPS accuracy in real-world settings. *Front Public Health*. 2014;2:21.
- Chen L, Bell EM, Caton AR, Druschel CM, Lin S. Residential mobility during pregnancy and the potential for ambient air pollution exposure misclassification. *Environ Res*. 2010;110:162–8.
- Ozonoff A, Webster T, Vieira V, Weinberg J, Ozonoff D, Aschengrau A. Cluster detection methods applied to the Upper Cape Cod cancer data. *Environ Health*. 2005;4:19.
- Wheeler DC, De Roos AJ, Cerhan JR, Morton LM, Severson R, Cozen W, Ward MH. Spatial-temporal analysis of non-Hodgkin lymphoma in the NCI-SEER NHL case-control study. *Environ Health*. 2011;10:63.
- Han D, Rogerson PA, Bonner MR, Nie J, Vena JE, Muti P, Trevisan M, Freudenheim JL. Assessing spatio-temporal variability of risk surfaces using residential history data in a case control study of breast cancer. *Int J Health Geogr*. 2005;4:9.
- Sabel CE, Boyle PJ, Löytönen M, Gatrell AC, Jokelainen M, Flowerdew R, Maasilta P. Spatial clustering of amyotrophic lateral sclerosis in Finland at place of birth and place of death. *Am J Epidemiol*. 2003;157:898–905.
- Gallagher LG, Webster TF, Aschengrau A, Vieira VM. Using residential history and groundwater modeling to examine drinking water exposure and breast cancer. *Environ Health Perspect*. 2010;118:749–55.
- Manjourides J, Pagano M. Improving the power of chronic disease surveillance by incorporating residential history. *Stat Med*. 2011;30:2222–33.
- Dearwent SM, Jacobs RR, Halbert JB. Locational uncertainty in georeferencing public health datasets. *J Exposure Sci Environ Epidemiol*. 2000;11:329–34.
- Krieger N, Waterman P, Lemieux K, Zierler S, Hogan JW. On the wrong side of the tracts? Evaluating the accuracy of geocoding in public health research. *Am J Public Health*. 2001;91:1114.
- Bonner MR, Han D, Nie J, Rogerson P, Vena JE, Freudenheim JL. Positional accuracy of geocoded addresses in epidemiologic research. *Epidemiology*. 2003;14:408–12.

31. Cayo MR, Talbot TO. Positional error in automated geocoding of residential addresses. *Int J Health Geogr*. 2003;2:10.
32. McElroy JA, Remington PL, Trentham-Dietz A, Robert SA, Newcomb PA. Geocoding addresses from a large population-based study: lessons learned. *Epidemiology*. 2003;14:399–407.
33. Whitsel EA, Rose KM, Wood JL, Henley AC, Liao D, Heiss G. Accuracy and repeatability of commercial geocoding. *Am J Epidemiol*. 2004;160:1023–9.
34. Yang D-H, Bilaver LM, Hayes O, Goerge R. Improving geocoding practices: evaluation of geocoding tools. *J Med Syst*. 2004;28:361–70.
35. Ward MH, Nuckols JR, Giglierano J, Bonner MR, Wolter C, Airola M, Mix W, Colt JS, Hartge P. Positional accuracy of two methods of geocoding. *Epidemiology*. 2005;16:542–7.
36. Whitsel EA, Quibrera PM, Smith RL, Catellier DJ, Liao D, Henley AC, Heiss G. Accuracy of commercial geocoding: assessment and implications. *Epidemiol Perspect Innov*. 2006;3:8.
37. Kravets N, Hadden WC. The accuracy of address coding and the effects of coding errors. *Health Place*. 2007;13:293–8.
38. Schootman M, Sterling DA, Struthers J, Yan Y, Laboube T, Emo B, Higgs G. Positional accuracy and geographic bias of four methods of geocoding in epidemiologic research. *Ann Epidemiol*. 2007;17:464–70.
39. Strickland MJ, Siffel C, Gardner BR, Berzen AK, Correa A. Quantifying geocode location error using GIS methods. *Environ Health*. 2007;6:1–8.
40. Zimmerman DL, Fang X, Mazumdar S, Rushton G. Modeling the probability distribution of positional errors incurred by residential address geocoding. *Int J Health Geogr*. 2007;6:1.
41. Jacquez GM. A research agenda: does geocoding positional error matter in health GIS studies? *Spat. Spatio-Temporal Epidemiol*. 2012;3:7–16.
42. Zandbergen PA, Green JW. Error and bias in determining exposure potential of children at school locations using proximity-based GIS techniques. *Environ Health Perspect*. 2007;115:1363–70.
43. Mazumdar S, Rushton G, Smith BJ, Zimmerman DL, Donham KJ. Geocoding accuracy and the recovery of relationships between environmental exposures and health. *Int J Health Geogr*. 2008;7:13.
44. Reich BJ, Chang HH, Strickland MJ. Spatial health effects analysis with uncertain residential locations. *Stat Methods Med Res*. 2014;23:156–68.
45. Oliver MN, Matthews KA, Siadaty M, Hauck FR, Pickle LW. Geographic bias related to geocoding in epidemiologic studies. *Int J Health Geogr*. 2005;4:29.
46. Jacquez GM, Waller LA. The effect of uncertain locations on disease cluster statistics. Michigan: Arbor Press: Chelsea; 2000.
47. Waller LA. Statistical power and design of focused clustering studies. *Stat Med*. 1996;15:765–82.
48. Jefery C, Ozonof A, Manjourides JD, White LF, Pagano M. Effect of spatial resolution on cluster detection: a simulation study. *Int J Health Geogr*. 2007;6:52.
49. Berke O, Waller L. On the effect of diagnostic misclassification bias on the observed spatial pattern in regional count data—a case study using West Nile virus mortality data from Ontario, 2005. *Spat Spatio-Temporal Epidemiol*. 2010;1:117–22.
50. Wang X-H, Zhou X-N, Vounatsou P, Chen Z, Utzinger J, Yang K, Steinmann P, Wu X-H. Bayesian spatio-temporal modeling of *Schistosoma japonicum* prevalence data in the absence of a diagnostic “gold” standard. *PLoS Negl Trop Dis*. 2008;2:e250.
51. Mwalili SM, Lesaffre E, Declerck D. A Bayesian ordinal logistic regression model to correct for interobserver measurement error in a geographical oral health study. *J R Stat Soc Ser C (Appl Stat)*. 2005;54:77–93.
52. Hu Y, Zhang ZJ, Chen Y, Wang Z, Gao J, Tao B, Jiang QW, Jiang QW. Spatial pattern of schistosomiasis in Xingzi, Jiangxi Province, China: the effects of environmental factors. *Parasit Vectors*. 2013;6:15.
53. Zhang ZJ, Carpenter TE, Lynn HS, Chen Y, Bivand R, Clark AB, Hui FM, Peng WX, Zhou YB, Zhao GM. Location of active transmission sites of *Schistosoma japonicum* in lake and marshland regions in China. *Parasitology*. 2009;136:737–46.
54. Yang G-J, Vounatsou P, Zhou X-N, Tanner M, Utzinger J. A Bayesian-based approach for spatio-temporal modeling of county level prevalence of *Schistosoma japonicum* infection in Jiangsu province, China. *Int J Parasitol*. 2005;35:155–62.
55. Raso G, Vounatsou P, Singer BH, Eliézer KN, Tanner M, Utzinger J. An integrated approach for risk profiling and spatial prediction of *Schistosoma mansoni*–hookworm coinfection. *Proc Natl Acad Sci*. 2006;103:6934–9.
56. Banerjee S, Carlin BP, Gelfand AE. Hierarchical modeling and analysis for spatial data. 1st ed. USA: Chapman and Hall/CRC; 2003.
57. Yang J, Zhao Z, Li Y, Krewski D, Wen SW. A multi-level analysis of risk factors for *Schistosoma japonicum* infection in China. *Int J Infect Dis*. 2009;13:e407–12.
58. Manjourides J, Lin HH, Shin S, Jeffery C, Contreras C, Cruz JS, Jave O, Yagui M, Asencios L, Pagano M, Cohen T. Identifying multidrug resistant tuberculosis transmission hotspots using routinely collected data. *Tuberculosis*. 2012;92:273–9.
59. Zhang Z, Chen D, Chen Y, Davies TM, Vaillancourt J-P, Liu W. Risk signals of an influenza pandemic caused by highly pathogenic avian influenza subtype H5N1: spatio-temporal perspectives. *Vet J*. 2012;192:417–21.
60. Zhang Z, Chen D, Chen Y, Liu W, Wang L, Zhao F, Yao B. Spatio-temporal data comparisons for global highly pathogenic avian influenza (HPAI) H5N1 outbreaks. *PLoS One*. 2010;5:e15314.
61. Li J, Robbins S, Lamm SH. The influence of misclassification bias on the reported rates of congenital anomalies on the birth certificates for West Virginia—a consequence of an open-ended query. *Birth Defects Res A*. 2013;97:140–51.
62. Bihmann K, Toft N, Nielsen SS, Ersbøll AK. Spatial correlation in Bayesian logistic regression with misclassification. *Spat Spatio-Temporal Epidemiol*. 2014;9:1–12.
63. Žuković M, Hristopulos DT. Classification of missing values in spatial data using spin models. *Phys Rev E*. 2009;80:011116.
64. Carpenter J, Kenward M. Multiple imputation and its application. UK: Wiley; 2013.
65. Li Y, Tang H, Lin X. Spatial linear mixed models with covariate measurement errors. *Stat Sin*. 1077;2009:19.
66. Sheppard L, Burnett RT, Szpiro AA, Kim S-Y, Jerrett M, Pope CA III, Brunekreef B. Confounding and exposure measurement error in air pollution epidemiology. *Air Qual Atmos Health*. 2012;5:203–16.
67. Rhomberg LR, Chandalia JK, Long CM, Goodman JE. Measurement error in environmental epidemiology and the shape of exposure-response curves. *Crit Rev Toxicol*. 2011;41:651–71.
68. Huang Y-L, Batterman S. Residence location as a measure of environmental exposure: a review of air pollution epidemiology studies. *J Exposure Sci Environ Epidemiol*. 1999;10:66–85.
69. Bow CJD, Waters NM, Faris PD, Seidel JE, Galbraith PD, Knudtson ML, Ghali WA. Accuracy of city postal code coordinates as a proxy for location of residence. *Int J Health Geogr*. 2004;3:5.
70. Healy MA, Gilliland JA. Quantifying the magnitude of environmental exposure misclassification when using imprecise address proxies in public health research. *Spat Spatio-Temporal Epidemiol*. 2012;3:55–67.
71. Iacono GL, Robin CA, Newton JR, Gubbins S, Wood JL. Where are the horses? With the sheep or cows? Uncertain host location, vector-feeding preferences and the risk of African horse sickness transmission in Great Britain. *J R Soc Interface*. 2013;10:20130194.
72. Duncan DT, Kawachi I, Subramanian SV, Aldstadt J, Melly SJ, Williams DR. Examination of how neighborhood definition influences measurements of youths' access to tobacco retailers: a methodological note on spatial misclassification. *Am J Epidemiol*. 2014;179:373–81.
73. Gryparis A, Coull BA, Schwartz J, Suh HH. Semiparametric latent variable regression models for spatiotemporal modelling of mobile source particles in the greater Boston area. *J R Stat Soc Ser C (Appl Stat)*. 2007;56:183–209.
74. Künzli N, Jerrett M, Mack WJ, Beckerman B, LaBree L, Gilliland F, Thomas D, Peters J, Hodis HN. Ambient air pollution and atherosclerosis in Los Angeles. *Environ Health Perspect*. 2005;113:201–206.
75. Shaddick G, Wakefield J. Modelling daily multivariate pollutant data at multiple sites. *J R Stat Soc Ser C (Appl Stat)*. 2002;51:351–72.
76. Cressie NAC. Statistics for spatial data. New York: Wiley; 1993.
77. Gaudard M, Karson M, Linder E, Sinha D. Bayesian spatial prediction. *Environ Ecol Stat*. 1999;6:147–71.
78. Banerjee S, Carlin BP, Gelfand AE. Hierarchical modeling and analysis for spatial data. New York: Chapman & Hall; 2004.
79. Kammann EE, Wand MP. Geoadditive models. *J R Stat Soc Ser C (Appl Stat)*. 2003;52:1–18.
80. Hobert JP, Altman NS, Schofield CL. Analyses of fish species richness with spatial covariate. *J Am Stat Assoc*. 1997;92:846–54.
81. Higgins KM, Davidian M, Giltinan DM. A two-step approach to measurement error in time-dependent covariates in nonlinear mixed-effects

- models, with application to IGF-I pharmacokinetics. *J Am Stat Assoc.* 1997;92:436–48.
82. Waller LA, Gotway CA. *Applied spatial statistics for public health data.* New York: Wiley; 2004.
 83. Madsen L, Ruppert D, Altman NS. Regression with spatially misaligned data. *Environmetrics.* 2008;19:453.
 84. Basagaña X, Aguilera I, Rivera M, Agis D, Foraster M, Marrugat J, Elosua R, Künzli N. Measurement error in epidemiologic studies of air pollution based on land-use regression models. *Am J Epidemiol.* 2013;178:1342–6.
 85. Szpiro AA, Sheppard L, Lumley T. Efficient measurement error correction with spatially misaligned data. *Biostatistics.* 2011;1:kxq083.
 86. Bergen S, Sheppard L, Sampson PD, Kim S-Y, Richards M, Vedal S, Kaufman JD, Szpiro AA. A national prediction model for PM_{2.5} component exposures and measurement error—corrected health effect inference. *Environ Health Perspect.* 2017;2013:121.
 87. Bell ML, Ebisu K, Peng RD. Community-level spatial heterogeneity of chemical constituent levels of fine particulates and implications for epidemiological research. *J Exposure Sci Environ Epidemiol.* 2011;21:372–84.
 88. Goldman GT, Mulholland JA, Russell AG, Srivastava A, Strickland MJ, Klein M, Waller LA, Tolbert PE, Edgerton ES. Ambient air pollutant measurement error: characterization and impacts in a time-series epidemiologic study in Atlanta. *Environ Sci Technol.* 2010;44:7692–8.
 89. Sheppard L, Slaughter JC, Schildcrout J, Liu LS, Lumley T. Exposure and measurement contributions to estimates of acute air pollution effects. *J Exposure Sci Environ Epidemiol.* 2005;15:366–76.
 90. Ong P, Graham M, Houston D. Policy and programmatic importance of spatial alignment of data sources. *Am J Public Health.* 2006;96:499.
 91. Lopiano KK, Young LJ, Gotway CA. Estimated generalized least squares in spatially misaligned regression models with Berkson error. *Biostatistics.* 2013;14:737–51.
 92. Lopiano KK, Young LJ, Gotway CA. A pseudo-penalized quasi-likelihood approach to the spatial misalignment problem with non-normal data. *Biometrics.* 2014;70:648–60.
 93. Szpiro AA, Paciorek CJ. Measurement error in two-stage analyses, with application to air pollution epidemiology. *Environmetrics.* 2013;24:501–17.
 94. Chang HH, Peng RD, Dominici F. Estimating the acute health effects of coarse particulate matter accounting for exposure measurement error. *Biostatistics.* 2011;12:637–52.
 95. Smith BJ, Zhang L, Field RW. Iowa radon leukaemia study: a hierarchical population risk model for spatially correlated exposure measured with error. *Stat Med.* 2007;26:4619–42.
 96. Molitor J, Jerrett M, Chang C-C, Molitor N-T, Gauderman J, Berhane K, McConnell R, Lurmann F, Wu J, Winer A. Assessing uncertainty in spatial exposure models for air pollution health effects assessment. *Environ Health Perspect.* 2007;115:1147–53.
 97. Lane KJ, Scammell MK, Levy JI, Fuller CH, Parambi R, Zamore W, Mwamburi M, Brugge D. Positional error and time-activity patterns in near-highway proximity studies: an exposure misclassification analysis. *Environ Health.* 2013;12:75.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at
www.biomedcentral.com/submit

