Day Time, Night Time, Over Time: Geographic and Temporal Uncertainty When Linking Event and Contextual Data

David C Folch
Northern Arizona University
Christopher S Fowler (csfowler@psu.edu)
Pennsylvania State University https://orcid.org/0000-0001-8415-0441
Levon Mikaelian
Florida State University

Methodology

Keywords: Uncertainty, Space-time, Neighborhood effects

Posted Date: June 4th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-32174/v1

Version of Record: A version of this preprint was published at Environmental Health on May 4th, 2021.
See the published version at https://doi.org/10.1186/s12940-021-00734-x.

Abstract

Background: The growth of geolocated data has opened the door to a wealth of new research opportunities in the health fields. One avenue of particular interest is the relationship between the spaces where people spend time and their health outcomes. This research model typically intersects individual data collected on a specific cohort with publicly available socioeconomic or environmental aggregate data. In spatial terms: individuals are represented as points on map at a particular time, and context is represented as polygons containing aggregated or modeled data from sampled observations. Uncertainty abounds in these kinds of complex representations.

Methods: We present *four* sensitivity analysis approaches that interrogate the stability of spatial and temporal relationships between point and polygon data. *Positional accuracy* assesses the significance of assigning the point to the correct polygon. *Neighborhood size* investigates how the size of the context assumed to be relevant impacts observed results. *Life course* considers the impact of variation in contextual effects over time. *Time of day* recognizes that most people occupy different spaces throughout the day, and that exposure is not simply a function residential location. We use eight years of point data from a longitudinal study of children living in rural Pennsylvania and North Carolina and eight years of air pollution and population data presented at 0.5 mile (0.805 km) grid cells. We first identify the challenges faced for research attempting to match individual outcomes to contextual effects, then present methods for estimating the effect this uncertainty could introduce into an analysis and finally contextualize these measures as part of a larger framework on uncertainty analysis.

Results: Spatial and temporal uncertainty is highly variable across the children within our cohort and the population in general. For our test datasets, we find greater uncertainty over the life course than in positional accuracy and neighborhood size. Time of day uncertainty is relatively low for these children.

Conclusions: Spatial and temporal uncertainty should be considered for each individual in a study since the magnitude can vary considerably across observations. The underlying assumptions driving the source data play an important role in the level of measured

Introduction

Geographic context matters when considering health outcomes. Physical location has been tied to obesity (Cummins and Macintyre 2005; Inagami et al. 2006), stress (Steptoe and Feldman 2001; Stigsdotter et al. 2010), cancer (Freedman, Grafova, and Rogowski 2011; Eschbach, Mahnken, and Goodwin 2005), stroke (Lisabeth et al. 2006), dementia (Chen et al. 2017), and many other health conditions. Complicating matters is that context can be both generalizable and idiosyncratic in terms of contributing to the health outcomes of people whose lives intersect with a particular space at a particular time. Locations with high particulate matter (PM) are broadly harmful, especially to people who spend considerable time outdoors or are at elevated risk to PM. In contrast, lead pipes are generally not harmful as strategies have been developed to prevent lead from leaching into the water supply. For example, the

2014 Flint, Michigan water crisis was the result of public policy, changes in water sources, and aging infrastructure, among other factors; not simply the presence of lead pipes.

The role of space and time in health outcomes can be viewed through the three components of the classic epidemiological triad: host, agent, and environment. The host is the population at risk, the agent is what is causing harm and the environment is the milieu in which the two interact. While host and agent are reasonably straightforward concepts, *environment* has taken on many meanings that can include time, location, and interaction process. The importance of environment is clear, but its operationalization is opaque. Specifically, the interrelationship between the location and time *attributes* of the host and agent, and the space in which they interact is not straightforward.

These interrelationships are further complicated by uncertainty in the supporting data, which is generally not considered. Uncertainty can arise through the quality of a global positioning system (GPS) or address geocoder used to convert places to latitude/longitude coordinates, through the choice of distance buffers used to define interaction spaces, or because the duration of exposure in a particular location is not well specified. These all fall into the opaque environment dimension of the triad, which does not always receive a high level of attention in research design. While patient interaction protocols and blood toxicity measurements are critical to research design and analysis, so too are the details of geolocation when environment is tested as contributing to a particular health outcome. The purpose of this paper is to illustrate both the mechanisms and magnitude of uncertainty related to geolocation in environmental health measurements. We further demonstrate how this uncertainty can be quantified and incorporated into assessments of the impacts of environment on health.

We demonstrate the impact of uncertainty in four spatial and temporal inputs to the research process using a sensitivity analysis approach. The first is positional accuracy. A common scenario is the allocation of observations with a latitude/longitude attribute to some polygon, e.g., a census tract or zip code; this case assesses the significance of getting this assignment process correct. The second case is neighborhood size, which is typically the size of the area in which interactions are assumed to be relevant. Because *neighborhood* may be defined differently in different research contexts and may ultimately be chosen based on data availability rather than theory or empirical analysis, it is important to understand how size may impact observed results. Third is life course. Data in environmental health contexts are often a snapshot of an individual at some point in time; shift the analysis a few days, months, or years and the results might differ. When impacts are assumed to be based on exposure, how much do we need to understand about the consistency of exposure over time in particular places or the propensity of individuals to move, thus potentially altering their exposure over time? The fourth case is closely related to the third by considering time of day rather than longer temporal change. People tend to occupy different spaces throughout the day; exposure is not simply a function of where people live, but all the locations where they spend time. Since a significant portion of administrative data is reported based on place of residence, what impacts might changing context over the day introduce into our understanding of exposure?

We explore these issues by intersecting two space-time datasets. The first comes from the US Environmental Protection Agency's (EPA) Risk-Screening Environmental Indicators (RSEI) program for the years 2003 to 2009. The program takes toxic release events as raw input and computes toxicity scores aggregated to half-mile (0.805 km) cells covering the entire United States. It also includes population data for each cell imputed from Census data. The second is a unique longitudinal dataset on children living in North Carolina and Pennsylvania collected by the Family Life Project (FLP) between 2003 and 2009. We have multiple observations on each child's home and daycare locations over the first 60 months of life. These datasets allow us to study the spatial and temporal data quality characteristics of childhood toxicity exposure. We further analyze snapshots of the entire state population toxicity exposure to develop benchmarks for the FLP data.

Methods And Results

In this analysis we explore four forms of geographic and temporal uncertainty. We begin by investigating the importance that a person is geolocated to the correct cell. Next, we address the spatial scale of our assessment of toxicity exposure by considering various neighborhood sizes around each person. The temporal dimension in toxicity exposure first considers longitudinal change in exposure over years, looking separately at people who remain in the same place and those who move. Finally, we study cyclical time in the form of daytime versus nighttime exposure. In most cases we are able to compare the statewide population to the FLP population; the temporal cases highlight the value of the FLP's data collection model for these tests.

In each of the following sub-sections we document both the method used to apprehend geographic uncertainty and the results of that method applied to the FLP and RSEI data. The forms of geographic uncertainty present in our data emerge out of a data structure that seeks to match point-level information on individuals with gridded environmental data; a data structure relevant for many kinds of environmental health research. By combining specific methods and results in each sub-section we hope to emphasize the importance of the specific geographic problem being faced and the types of methodological solutions available to handle it.

Each measure presented examines the impact of uncertainty in the relationship between the geolocation of an individual within the dataset and their contextual data. They recognize that researchers may have to use data where accuracy is either of poor quality or where accuracy is unknown. When researchers cannot fix a problem it is still important to understand the degree to which that problem could impact downstream results. The methods in the sub-sections that follow allow for the quantification of uncertainty. In our concluding section we describe ways that researchers might use these measurements to test the robustness of their findings.

A note on representation in this section. The toxicity scores include many low values and a small number of extremely high values, which affects the visualization of results. In lieu of transforming the results into logarithms, we adopt a consistent set of 6 logical custom bins for all the measures. The first two bins, 0–

0.025 and 0.025–0.05 characterize very low and low levels of uncertainty; the third, 0.05–0.75 is moderate uncertainty; and the final three bins are high and very high levels. The highest bin for all the figures is 5 and above; the upper end of this last bin varies for each measure and is presented in the map's legend. This approach increases the ease with which we can compare magnitudes across measures, but it can obscure some nuances within each measure. The replication script available with this paper allows different visualizations to be constructed.

Positional Accuracy

Positional accuracy measures the importance of a point being located in the correct pixel. Depending on the method by which a point was acquired, accuracy can vary widely. Modern GPS units such as those built into phones can be accurate within +/- 4 meters, but that accuracy can shift depending on the presence of buildings and other obstacles. Incorrect use of cartographic projections, or even simple transcription errors can and do affect the accuracy of geolocated data. More common is geolocation based on geocoded addresses, which can vary widely in quality based on the degree to which an address matches information in the geocoder. The difference between a geocode matched to a rooftop and one based on interpolation of an address along a road segment may seem insignificant, but it can alter the location of a point enough to impact which pixel or administrative unit it is assigned to. Even greater uncertainty emerges when location is assigned using only the zip code or city information in the address; these assignments give an appearance of exactness (the same number of significant digits are reported as more precise results) while masking tremendous uncertainty about where within the administrative unit an individual is actually located.

If a pixel and its neighbors are relatively similar, then being incorrectly placed in a neighboring pixel will not have much impact on the results of subsequent analyses. In contrast, high variation among neighboring pixels implies that positional accuracy is more important. We measure the level of uncertainty associated with positional accuracy using the positional coefficient of variation (PCV), which we define for pixel i as:

$$PCV_i = SD_i^{neighborhood} / est_i,$$

where est_i is the estimate for pixel i and $SD_i^{neighborhood}$ is the standard deviation of the estimate for pixel i and its neighbors. As will be seen later, "neighborhood" can be defined in many ways. We define neighborhood here as i and its eight immediate neighbors. Pixels are 0.5 mile (0.805 km) on a side meaning this neighborhood of nine pixels is 1.5 miles (2.414 km) on a side. The data we are working with was generated based on GPS coordinates matched to geocoded address data, so the 1.5 mile (2.414 km) range of our neighborhood fully captures the level of uncertainty we have in our assigned point locations. A study with finer resolution environmental data or with point assignment based on administrative unit centroids might require different neighborhood extents than what we use here.

The maps in Fig. 2 show PCV for 2008 (only one year of results is presented here) with lighter colors indicating lower PCV and thus less importance of small errors in a point's location. In general, locations closer to a release have a high PCV, with the PCV decreasing with distance. In the RSEI model, toxicity is assumed to decrease rapidly, even exponentially, as distance increases from the release site. As a result pixels near major releases experience rapid change with distance and the precise placement of individuals in these cells will significantly impact reported contextual measures. Moreover, the RSEI model radius is noticeable in the maps; in this case as dark rings. The abrupt end to the estimated toxicity from an acute release can result in neighboring pixels on the edge of the 30 mile (48.280 km) radius from the event having extremely different values, and thus high PCV.

It is also possible to estimate the impact of PCV on all residents in the state and FLP homes in particular. The RSEI program estimates the number of people in each pixel based on US Census Bureau data, which allows us to sum people by PCV. We can further assign each FLP address to the pixel in which it falls to measure its PCV. These results are presented in Fig. 2 as population weighted histograms. Figure 2 has some interesting properties. First, while North Carolina and Pennsylvania are fairly different in terms of their toxicity context (Fig. 1), their statewide PCV distributions are remarkably similar. One explanation for this similarity is that the effect of uncertainty being measured in the PCV is likely conditional on model parameters that structure rates of decay and effects of distance in the RSEI model more than it is related to specific toxicity releases. Second, while PCV for North Carolina's FLP sample diverges little from the state population (in distribution and median values), the Pennsylvania FLP sample has a noticeably different distribution and median skewed towards lower PCV than the population as a whole (although the mean is higher).

Neighborhood Size

The concept of 'neighborhood' is contested within social science research (Galster 2019) because it has a sociological meaning, but it is rarely measured as an administrative unit for which data is collected and released. In many cases the neighborhood represents an important context in which we might theorize some relationship between environment and individual, but the precise measurement and definition of neighborhood depends on a unit of analysis that will be, at best, an approximation. Common neighborhood metrics in the U.S. context include census tracts, zip codes, and occasionally school districts. Research has shown that the size of the unit will condition what we observe as well as the variance among our observations (Fowler et al. 2019). Significantly, if we use smaller units to describe our neighborhood context we will observe more extreme results and more variation between units. Larger units will tend to offer less extreme results and lower variation due to smoothing. The implications of these choices can therefore be significant for many types of analysis.

Here we explore the implications of different assumptions about neighborhood size that might parallel decisions made about whether to use, for example, census tracts or zip codes as a measure of neighborhood. For airborne toxicity we would expect impacts from a release to be conditioned on

distance from the release. The challenge is then to operationalize “neighborhood” as a relevant geographic area surrounding an individual’s location. If the neighborhood assigned is too big, the variation among cells will smooth to the mean eliminating variance among individuals. If the neighborhood is too small, then minor perturbations in the spatial pattern can have an outsized effects, overstating the variance among individuals. This problem has been shown to be especially problematic with low probability events such as murder or some cancers where a single event could greatly spike the incidence for a small geographic area (Kulldorff et al. 2006). The TRI data, which underlies the RSEI data, has a similar issue where extreme toxicity events have an outsized impact the pixel values.

We measure the sensitivity of measured toxicity to neighborhood size by first computing some function for neighborhoods of different size around pixel i. The neighborhood coefficient of variation (NCV) is computed over these values as:

$$NCV_i = SD(f(neighborhood_{i1}), f(neighborhood_{i2}), \ldots, f(neighborhood_{is}) / f(i).$$

For the RSEI data, the function will be the average of the estimates in the neighborhood, where the neighborhoods are defined as square areas with the following numbers of pixels: 1, 9(3 × 3), 25(5 × 5),49(7 × 7) 81(9 × 9), and 361(19 × 19). The resulting NCV values are presented in Fig. 3 for 2008. In light colored areas, there is little difference in the average toxicity as neighborhood size changes; darker areas indicate greater sensitivity to neighborhood size. Figure 3 displays a similar pattern to that for PCV of high value rings, but in contrast tends to have low values near the very center of the release sites. One feature that emerges in Fig. 3 is the role played by model assumptions that assume a very rapid decay effect within the first half mile (0.805 km). The maps show small 9 × 9 squares that emerge because the specific pixel where a given high toxicity release occurred will have a much higher toxicity than even its nearest neighbors. NCV calculations that just have that cell in their largest neighborhoods will have a much higher standard deviation and create the dark purple hollow boxes seen on the maps.

The state maps of NCV have two significant implications for uncertainty in assessing environmental health impacts. First, the generally low values for NCV across most of both states even when calculated across large neighborhood size differentials suggests that the rate of variation from cell to cell in the RSEI model is generally fairly low. While the RSEI model reports results at a relatively fine grain of 0.5 × 0.5 miles (0.805 × 0.805 km), the effective variation is mostly at much larger scales. As a result, larger or smaller “neighborhoods” (at least as that term is employed in the social science literature) will largely produce the same result with this data. However, the spatial patterns that emerge in the NCV maps show that artifacts of assumptions made within the RSEI model play an outsized role in conditioning the variation in NCV that we observe. Put another way, the degree to which our assumptions about neighborhood size will matter is a function of assumptions about the form that distance decay takes in the original environmental impact model. This form of spatially structured uncertainty driven by model assumptions should act as a call for caution when linking environmental and health data sets together. The linkage among assumptions in different data sets can create complex forms of uncertainty in research results obtained with modeled data.

Second, in examining the histograms in Fig. 3 we can see that, while the Pennsylvania FLP sample embodies less uncertainty based on neighborhood size than Pennsylvania residents more generally, the reverse is true for North Carolina's FLP population. Consequently, we can conclude that decisions about neighborhood size in subsequent work will impact results more strongly in the North Carolina sample than they will in the Pennsylvania sample. While NCV is relatively small in most places, it represents a significant effect in a small portion of the observations.

Life Course

By necessity, much of the research in environmental health will be cross-sectional. A significant portion of this research will also require assumptions about duration or consistency of exposure to a particular environment or contextual factor. Uncertainty may arise when it is unclear how much the context in a particular location changes over time, or how the location of a particular individual changes over time. Our assessment of life course uncertainty captures change in residential context over time.

Change in context can happen via two channels: 1) a person stays in the same home as the neighborhood transitions around them and 2) a person moves homes resulting in a change in context. It is important to note that neighborhoods tend to be stable, so channel 1 change will likely happen slowly. In contrast, a residential move offers the *opportunity* for a large shift in residential context, but the family could move to a statistically similar location. We therefore define life course coefficient of variation (LCV) for pixel i based on the standard deviation of toxicity scores over n time periods divided by the median of those n scores:

$$LCV_i = SD(est_{i1}, est_{i2}, ..., est_{in}) / median(est_{i1}, est_{i2}, ..., est_{in}).$$

LCV gives us a sense of how important the year we choose for assessment of environmental toxicity will be in terms of conditioning the results we observe. It is important to note that the maps and statewide histograms in Fig. 4 only reflect channel 1 type change since we do not have information on residential moves in the statewide dataset. However, the stability of toxicity in a place is only one part of the story, we also need to consider the stability of the population in a particular place (channel 2 type change). The FLP data give us this opportunity, which is to explore the effects of residential movement on our population. For the FLP families we know their locations at different points in time; this gives multiple possible transitions for each family and clearly identifies where a family moved. This is reflected in the histograms in Fig. 4 where the statewide histograms are relatively similar, and quite different from the the FLP histograms. The FLP distributions reflect more high values, and the median and mean values are higher since they are eligible for channel 1 and 2 changes.

With the FLP data we can more closely examine how the environmental context differed for families that stayed in place as opposed to families that moved. Figure 5 compares the toxicity from the previous year to the toxicity in the present year for FLP families that moved and families that stayed in the same [illegible] year over year for stayers. Across all years the

correlation was 0.73 in North Carolina and 0.81 for Pennsylvanians. Year over year correlations were significant but much lower for families that moved in a given year; 0.39 for North Carolina and 0.66 for Pennsylvania. The important differences between states, but most importantly between households that moved and households that stayed in the same residence gives a sense of the magnitude of uncertainty introduced when no information is available about the life course of individuals' presumed exposure to a particular contextual effect. Notably, the direction of the effect of a move on toxicity level does not appear to be significant, with households almost equally as likely to increase their toxicity exposure in a new location as to decrease their exposure. Fig. 5: Experience of toxicity compared to previous years by whether household moved that year

Day-night

The fact that most administrative data counts individuals at their place of residence elides the importance of daytime location as an important factor determining exposure. Day-night uncertainty is similar to life course in that it accounts for locational variability over time, in this case over a single day. For working adults this is typically work and home, for older children it is school and home, and for young children it is day care and home. A substantial literature has begun to address this issue with regards to adults and the workplace (e.g., Delgado-Saborit et al. 2011). Another literature asks similar questions about differences in context for children at home and at school (e.g., Burgoine et al. 2015). Here we explore differences for children before they enter school, an area that has not been widely studied. We expect that differences between home and day care locations will be relatively small as child care will tend to be tightly constrained by both parents' income and the need to remain relatively close to very small children. Nevertheless, the FLP data gives us an opportunity to test this hypothesis with respect to toxicity exposure.

Since we do not have day-night data on the statewide population, we only conduct this analysis on the FLP children based on their home and child care locations. In this case, each observation is a point in time where we have contemporaneous home and child care location information. The figure shows unequivocally that, for our dataset, the difference in exposure between day time care and night time residence is minimal. Across all years the correlation was 0.79 in North Carolina and 0.89 for Pennsylvania. We draw two conclusions from these findings. First, we would expect parents to limit the distance between residence and child care facility for a variety of reasons, which reduces variation between daytime and night time exposure, meaning that a single measure is probably appropriate for this population. Nevertheless, we also see some of this correlation as a function of the relatively low spatial variation in toxicity highlighted in both PCV and NCV maps. A model with higher spatial resolution might find more difference than we express here.

Discussion

Location and time have become common attributes of health datasets. Many tools are available to convert participant addresses into latitude/longitude coordinates, and nearly every phone can report its coordinates using a built-in global positioning system (GPS). Geographic information systems (GIS) are becoming more accessible and user friendly, which is further unlocking spatial analysis tools for a wide audience. In contrast to the current ease of collecting an individual's space-time data, individualized exposure data remains elusive. As a person moves through their day (and life), minor variations in their environment will affect the magnitude of their exposure to pollutants. The ideal approach is then to measure a person's microenvironment using a portable sensor that captures the characteristics of their immediate surroundings (Steinle, Reis, and Sabel 2013). Portable sensors exist to measure air pollution (e.g., Steinle, Reis, and Sabel 2013), radio frequency electromagnetic fields (e.g., Frei et al. 2009), noise pollution (e.g., Smith, Neilsen, and Grimshaw 2017) and other harmful characteristics of the environment, but these need to be issued to participants and become an extra thing the person must carry around. In addition, there are no sensors for socioeconomic and other types of important microenvironment exposures. Myriad reasons preclude widespread precise space-time tracking of microenvironments, which leads to hybrid approaches that approximate this ideal data. In particular, a person's exposure can be approximated by intersecting their locations with data on the context of areas. The mixture of precise and generalized datasets, and the assumptions grounding their commingling, are complex and demand further examination.

The sensitivity analysis methods presented in the previous section focus on a specific but common case in health research: the researcher has a cohort of individuals being investigated and is interested in the role of context on the subjects. The researcher typically has extensive and specific data on each individual's attributes since the researcher is in control of the data collection protocol. In contrast, the context data is generalized or modeled from samples collected by someone else, and is typically provided by a governmental agency as polygons. The implicit goal is to develop new knowledge by intersecting the deep understanding the researcher has of her own cohort with data developed by domain experts in entirely different fields. While this juxtaposition of expertise can lead to groundbreaking insights, there are also potential pitfalls.

The assumptions driving environmental, demographic, and other datasets can have non-trivial impact on the data and analyses that depend on them. The RSEI model, for example, integrates data on toxic releases, toxicity of chemicals, and weather patterns, to name a few inputs, to ultimately generate a detailed nationwide dataset on air pollution exposure. Demographic data is also complicated in its sampling strategies, weighting techniques, and imputations (see for example Spielman, Folch, and Nagle 2014 regarding the American Community Survey). Location is embedded in many of these assumptions. For example, the previous section highlights the impact of the 30 mile (48.280 km) radius assumption in the RSEI data. In demographic data, the smaller the polygon the fewer samples available to support the estimates. This is an issue faced by the American Community Survey, which sees large increases in uncertainty as polygon size is reduced (Folch et al. 2016).

One approach to reducing the potential problems identified in the previous section is to increase the size of the polygons. The larger the polygon, the less likely an error of say 0.75 miles (1.207 km) would land the point in the “wrong” polygon. Similarly, larger polygons are more likely to capture a person’s home and work locations. This even extends to residential moves. According the US Census Bureau’s Current Population Survey, 64% of people who moved between 2017 and 2018 remained in the same county. While larger polygons potentially solve one problem, there is often a desire to use the smallest possible polygon in order to reasonably represent the interactions being studied. Crime, for example, varies greatly across most counties; therefore, the county average crime rate is unlikely to be representative of the likelihood any individual has of experiencing crime in their daily life. Therefore, there is no simple rule of thumb such as “use the largest (smallest) unit available.”

Since researchers are usually relying on secondary context data, they are forced to adapt their research design to the best available data. In terms of spatial scale, it is unlikely that the spatial units provided by some other entity will exactly match the research question(s) at hand. That being said, there is a continuum of spatial correspondence. On one end of the spectrum are clearly defined spatial units. For example, when access to particular treatments depends on the offerings of the county health department, then county of residence is a clear factor for the individuals being studied. The other end of the spectrum is context data provided as points. For example, air pollution monitoring station readings, Superfund sites, or crime incidents give the researcher the ultimate freedom to define the context for each of their participants. Most researchers use data that falls between these two extremes for very practical reasons: we are rarely lucky enough to have research relevant polygons and point data is typically too raw for non-experts to work with.

These challenges can be partially addressed by taking a critical view of geolocated data and seeking verification and validation of location information in the ways suggested here. Researchers must always seek to understand the cohort to understand how meaningful any single measure of location is likely to be with respect to the contextual phenomenon under consideration. Most geocoders provide a reliability estimate along with the latitude/longitude coordinates that can help with the exactness of coordinates. More broadly, collecting data on previous residence, place of work (school, day care, etc.), and time living at the location can provide important information about the significance of geolocated values. Similarly, context measures often have their own built in reliability measures that are overlooked far too frequently. Many public datasets come with measures of reliability such as margins of error (MOE). Failure to consider the range of contextual values that could fall within margins of error increases the supposed certainty of relationships while potentially introducing noise that can problematically shape the interpretation of outcomes.

Conclusion

The methods presented here are not an exhaustive set of metrics, they are exemplars of a broad sensitivity analysis approach to spatial and temporal uncertainty. The impact of any sensitivity analyses o observe. In our case we required methods that

intersect air pollution exposure data available as polygons and home and day care locations available as points. This polygon to point relationship is common in health and environmental research, meaning that these methods are useful for a many specific research questions. In particular, our measures (PCV, NCV, and LCV) do not rely on any specific domain expertise on environmental toxicity in order to estimate the structure of geographic and temporal uncertainty in the data. The visualizations do, however, identify artifacts from the environmental toxicity model that may introduce noise into our measures in a geographically systematic way.

This approach is an intermediate step for any space-based heath outcome research as it demonstrates the extent of uncertainty about the placement of a subject in some context space. The magnitude of the uncertainty and the overall goals of the project dictate how to proceed. In the example presented here, our concern is placing children in the RSEI space. We might, for example, run subsequent analyses on a subset of FLP points where the PCV, NCV, or LCV is particularly low and see if those points (where uncertainty is lower) produce substantially different results than when we observe all points together. Given relatively low values for NCV we might decide to generalize RSEI model findings from a 0.5 mile (0.805 km) grid to a 2.5 mile (4.023 km) grid–reducing the likelihood that points are assigned to the wrong polygon and limiting the uncertainty associated with PCV. We might choose to assign contextual values that are based on the average across several years of toxicity data to reduce the likelihood that extreme events in the toxicity data are driving outcomes. In an extreme situation, the results might cause a researcher to abandon a dataset altogether.

More broadly, the specific tests presented here are not appropriate for every analytic case. The cases of polygon to polygon or point to point are not considered. Linear data is also highly relevant today with increasing use of personal activity monitors that track a person's movement through space and time. Our intention here is not to be comprehensive, but to foster a conversation within health and environment research on what the implications of geographic uncertainty might be so that contextual effects can be better captured and understood.

Declarations

Ethics approval and consent to participate

This research was categorized as "exempt" by the Northern Arizona University Institutional Research Board (project number 1473861-1).

Consent for publication

Not applicable.

Availability of data and materials

The methodological code needed to replicate the study and an anonymized point dataset can be found at [insert URL to organize the public data is available at [insert

GitHub link].

Competing interests

The authors declare that they have no competing interests.

Funding

Research reported in this publication was supported by the Environmental Influences on Child Health Outcomes (ECHO) program of the National Institutes of Health under award number 4UH3OD023332-03.

Authors' contributions

DF and CF conceived of the project and methods. DF (point data), CF (polygon data) and LM (both) constructed the datasets. DF formalized the methods and CF implemented them in R. DF and CF wrote the manuscript with support from LM.

Acknowledgements

We would like to thank Clancy B. Blair for feedback on the study and the Family Life Project for access to the data.

Authors' information (optional)

Not applicable

References

Burgoine, Thomas, Andy P Jones, Rebecca J Namenek Brouwer, and Sara E Benjamin Neelon. 2015. “Associations Between BMI and Home, School and Route Environmental Exposures Estimated Using GPS and GIS: Do We See Evidence of Selective Daily Mobility Bias in Children?” *International Journal of Health Geographics* 14 (8): 1–12.

Chen, Hong, Jeffrey C Kwong, Ray Copes, Karen Tu, Paul J Villeneuve, Aaron Van Donkelaar, Perry Hystad, et al. 2017. “Living Near Major Roads and the Incidence of Dementia, Parkinson’s Disease, and Multiple Sclerosis: A Population-Based Cohort Study.” *The Lancet* 389 (10070): 718–26.

Couclelis, Helen. 2003. “The Certainty of Uncertainty: GIS and the Limits of Geographic Knowledge.” *Transactions in GIS* 7 (2): 165–75. *https://doi.org/10.1111/1467-9671.00138*.

Cummins, Steven, and Sally Macintyre. 2005. “Food Environments and Obesity—Neighbourhood or Nation?” *International Journal of Epidemiology* 35 (1): 100–104. *https://doi.org/10.1093/ije/dyi276*.

Delgado-Saborit, Juana Maria, Noel J Aquilina, Claire Meddings, Stephen Baker, and Roy M Harrison.

nic Compounds to Home, Work and Fixed Site

Outdoor Concentrations." *Science of the Total Environment* 409 (3): 478–88.

Environmental Protection Agency. 2015. "Find Out What's Happening in Your Neighborhood: Using EPA's Toxics Release Inventory." Toxics Release Inventory Program.

———. 2018. "EPA's Risk-Screening Environmental Indicators (RSEI) Methodology." RSEI Version 2.3.6. Office of Pollution Prevention; Toxics.

Eschbach, Karl, Jonathan D. Mahnken, and James S. Goodwin. 2005. "Neighborhood Composition and Incidence of Cancer Among Hispanics in the United States." *Cancer* 103 (5): 1036–44. *https://doi.org/10.1002/cncr.20885*.

Folch, David C, Daniel Arribas-Bel, Julia Koschinsky, and Seth E Spielman. 2016. "Spatial Variation in the Quality of American Community Survey Estimates." *Demography* 53 (5): 1535–54.

Fotheringham, AS, and DWS Wong. 1991. "The modifiable areal unit problem in multivariate statistical analysis." *Environment and Planning A*.

Fowler, Christopher S., Nathan Frey, David C. Folch, Nicholas Nagle, and Seth Spielman. 2019. "Who Are the People in My Neighborhood?: The 'Contextual Fallacy' of Measuring Individual Context with Census Geographies." *Geographical Analysis*, no. 2: 155–68.

Freedman, Vicki A., Irina B. Grafova, and Jeannette Rogowski. 2011. "Neighborhoods and Chronic Disease Onset in Later Life." *American Journal of Public Health* 101 (1): 79–86. *https://doi.org/10.2105/ajph.2009.178640*.

Frei, Patrizia, Evelyn Mohler, Georg Neubauer, Gaston Theis, Alfred Bürgi, Jürg Fröhlich, Charlotte Braun-Fahrländer, John Bolte, Matthias Egger, and Martin Röösli. 2009. "Temporal and Spatial Variability of Personal Exposure to Radio Frequency Electromagnetic Fields." *Environmental Research* 109 (6): 779–85.

Galster, George C. 2019. *Making Our Neighborhoods, Making Our Selves*. University of Chicago Press.

Inagami, Sanae, Deborah A. Cohen, Brian Karl Finch, and Steven M. Asch. 2006. "You Are Where You Shop: Grocery Store Locations, Weight, and Neighborhoods." *American Journal of Preventive Medicine* 31 (1): 10–17. *https://doi.org/https://doi.org/10.1016/j.amepre.2006.03.019*.

Kulldorff, Martin, Changhong Song, David Gregorio, Holly Samociuk, and Laurie DeChello. 2006. "Cancer Map Patterns: Are They Random or Not?" *American Journal of Preventive Medicine* 30 (2): S37–S49.

Kwan, MP. 2012a. "How GIS can help address the uncertain geographic context problem in social science research." *Annals of GIS*.

———. 2012b. "The uncertain geographic context problem." *Annals of the Association of American Geographers*.

Lisabeth, L., A. Diez Roux, J. Escobar, M. Smith, and L. Morgenstern. 2006. “Neighborhood Environment and Risk of Ischemic Stroke: The Brain Attack Surveillance in Corpus Christi (Basic) Project.” *American Journal of Epidemiology* 165 (3): 279–87. *https://doi.org/10.1093/aje/kwk005*.

Openshaw, Stan. 1984. “Ecological Fallacies and the Analysis of Areal Census Data.” *Environment and Planning A* 16 (1): 17–31.

Robertson, Colin, and Rob Feick. 2018. “Inference and analysis across spatial supports in the big data era: Uncertain point observations and geographic contexts.” *Transactions in GIS*, March. *https://doi.org/10.1111/tgis.12321*.

Robinson, W S. 1950. “Ecological Correlations and the Behavior of Individuals.” *American Sociological Review* 13 (3): 351–57.

Smith, Kieren H, Tracianne B Neilsen, and Jeremy Grimshaw. 2017. “Full-Day Noise Exposure for Student Musicians at Brigham Young University.” In *Proceedings of Meetings on Acoustics*, 30:1–11. 1. Acoustical Society of America.

Spielman, Seth E, David C Folch, and Nicholas N Nagle. 2014. “Patterns and Causes of Uncertainty in the American Community Survey.” *Applied Geography* 46: 147–57.

Steinle, Susanne, Stefan Reis, and Clive Eric Sabel. 2013. “Quantifying human exposure to air pollution-Moving from static monitoring to spatio-temporally resolved personal exposure assessment.” *Science of the Total Environment* 443: 184–93. *https://doi.org/10.1016/j.scitotenv.2012.10.098*.

Steptoe, Andrew, and Pamela J. Feldman. 2001. “Neighborhood Problems as Sources of Chronic Stress: Development of a Measure of Neighborhood Problems, and Associations with Socioeconomic Status and Health.” *Annals of Behavioral Medicine* 23 (3): 177–85. *https://doi.org/10.1207/S15324796ABM2303_5*.

Stigsdotter, Ulrika K., Ola Ekholm, Jasper Schipperijn, Mette Toftager, Finn Kamper-Jørgensen, and Thomas B. Randrup. 2010. “Health Promoting Outdoor Environments - Associations Between Green Space, and Health, Health-Related Quality of Life and Stress Based on a Danish National Representative Survey.” *Scandinavian Journal of Public Health* 38 (4): 411–17. *https://doi.org/10.1177/1403494810367468*.

Vernon-Feagans, Lynne, Martha Cox, Michael Willoughby, Margaret Burchinal, Patricia Garrett-Peters, Roger Mills-Koonce, Patricia Garrett-Peiers, Rand D Conger, and Patricia J Bauer. 2013. “The Family Life Project: An Epidemiological and Developmental Study of Young Children Living in Poor Rural Communities.” *Monographs of the Society for Research in Child Development*, i–150.

Figures

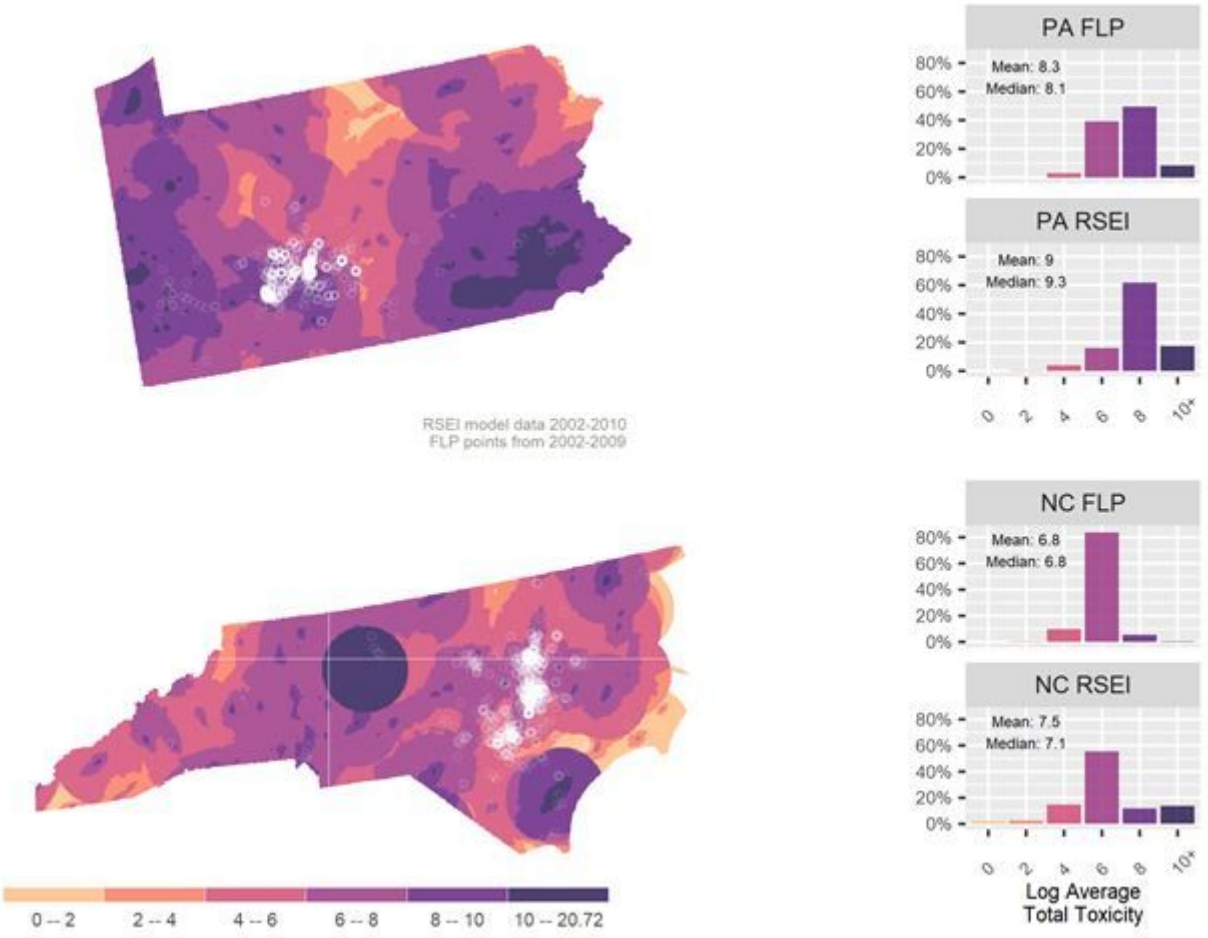

Figure 1

Log of Average Total Toxicity

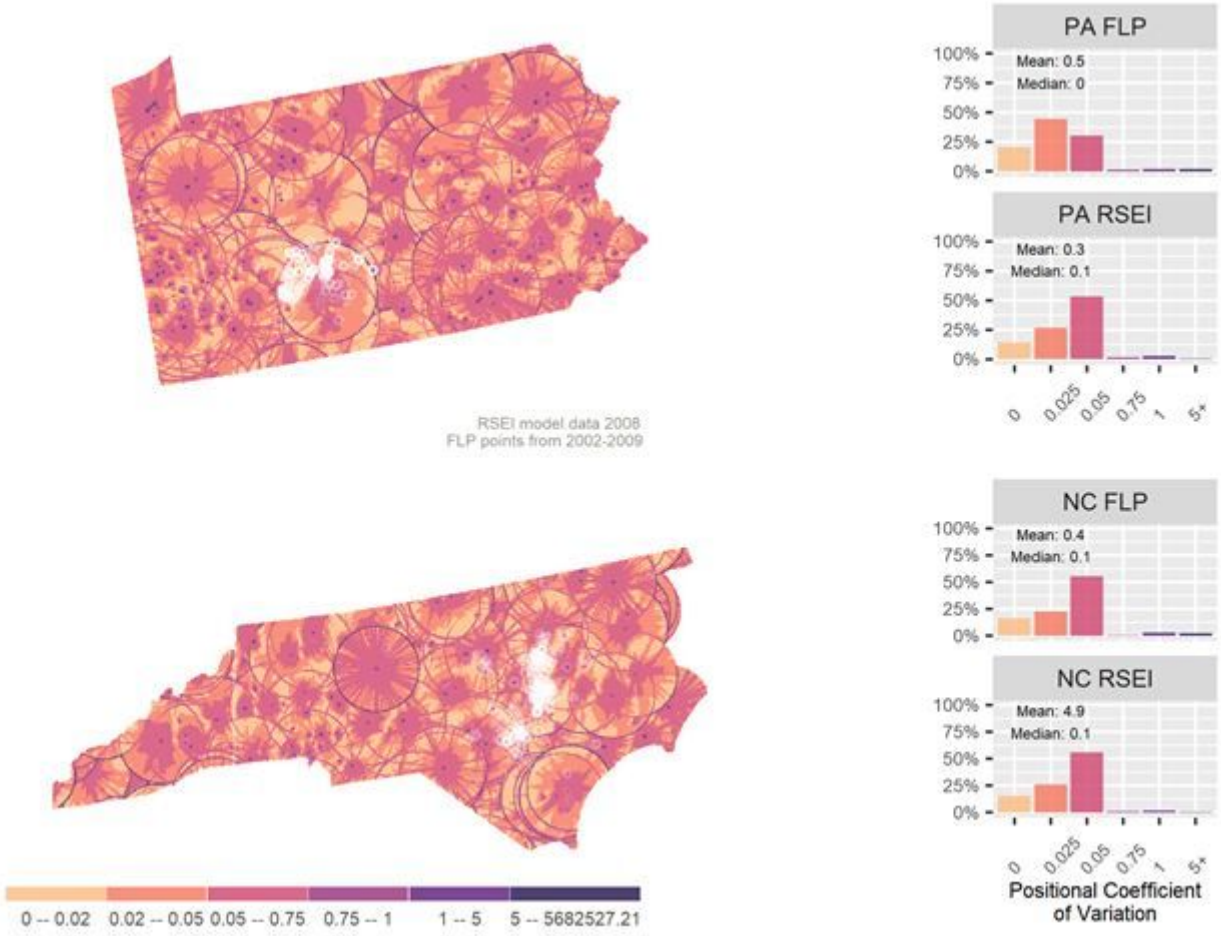

Figure 2

Positional Coefficient of Variation

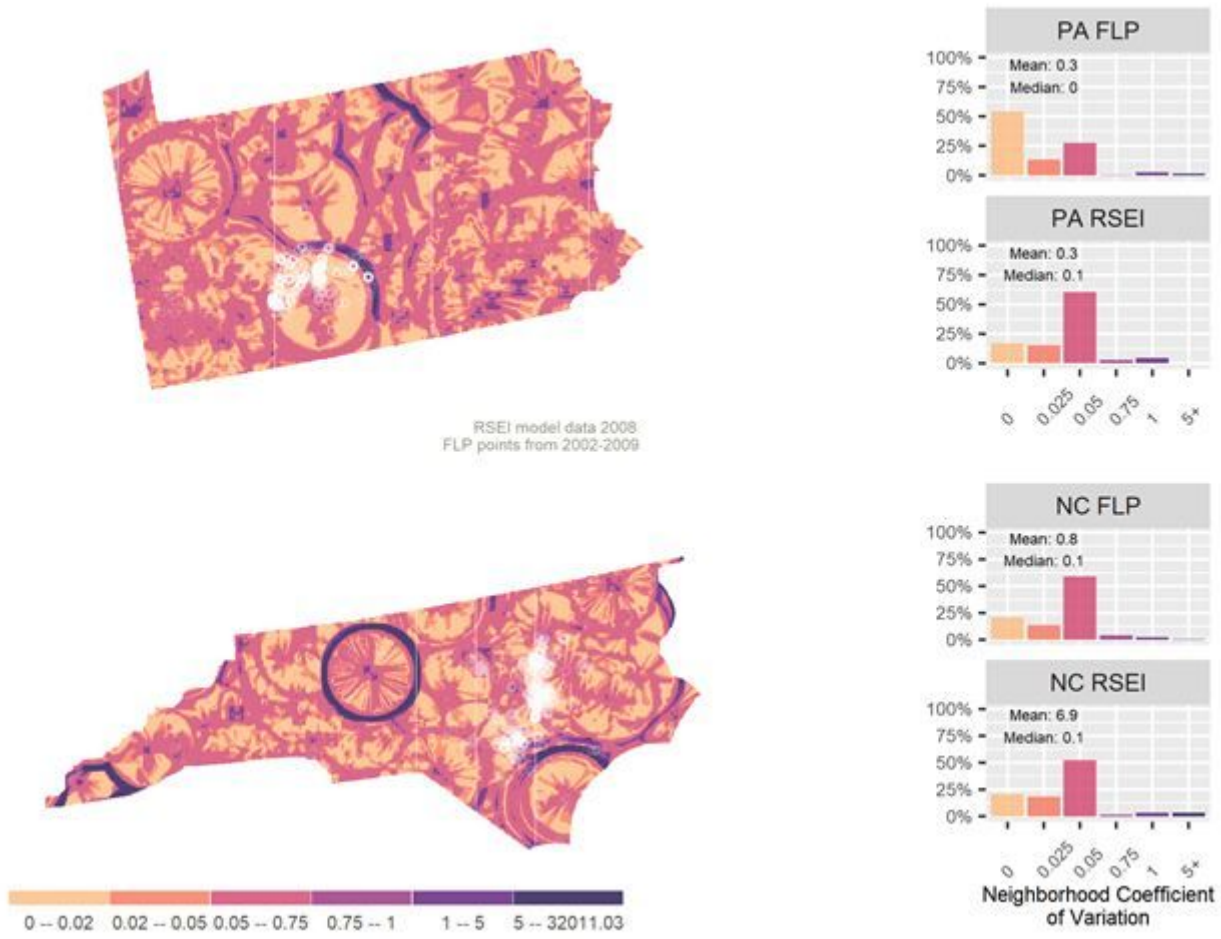

Figure 3

Neighborhood Coefficient of Variation

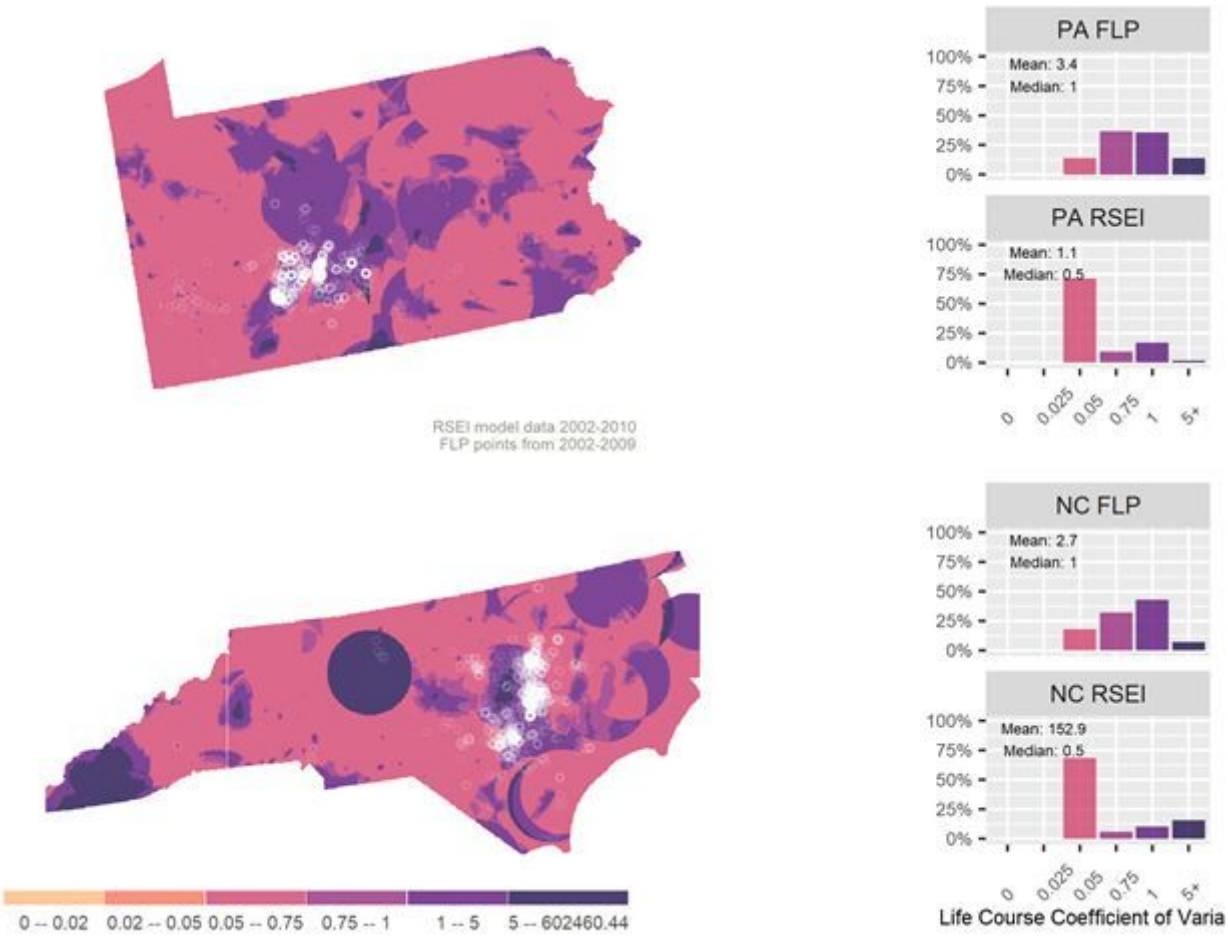

Figure 4

Life Course Coefficient of Variation

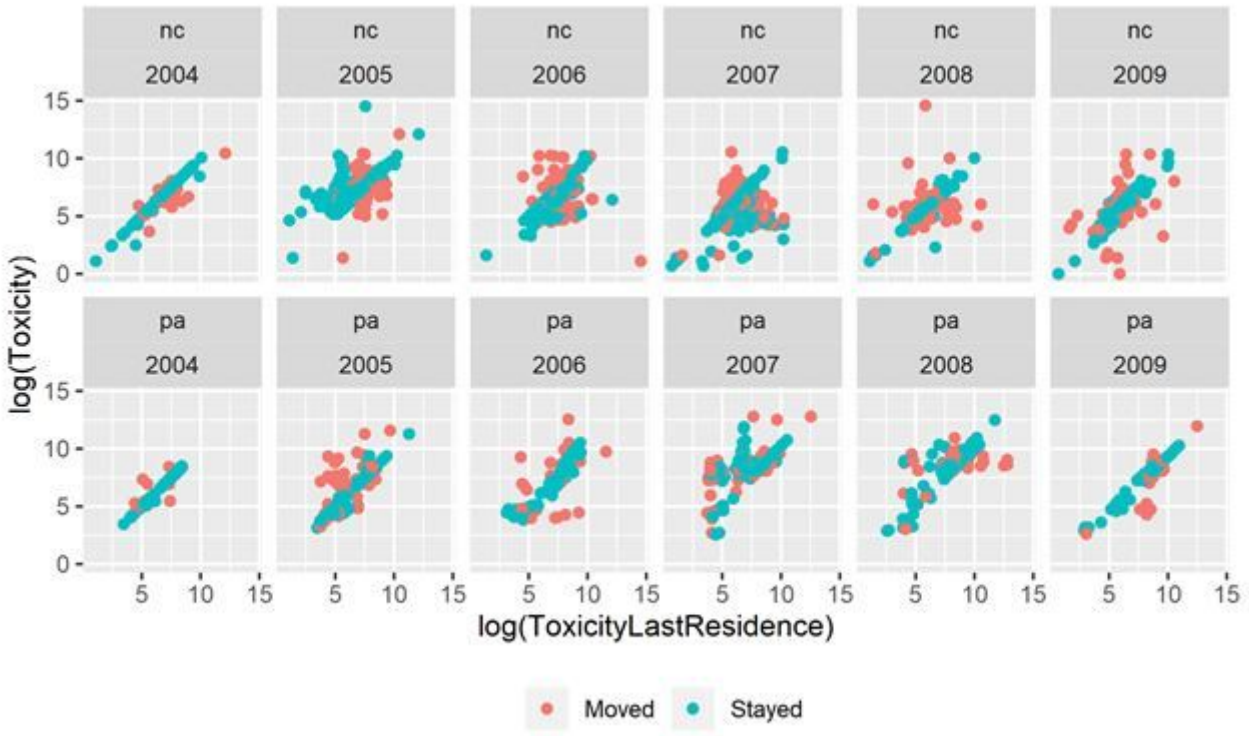

Figure 5

Experience of toxicity compared to previous years by whether household moved that year

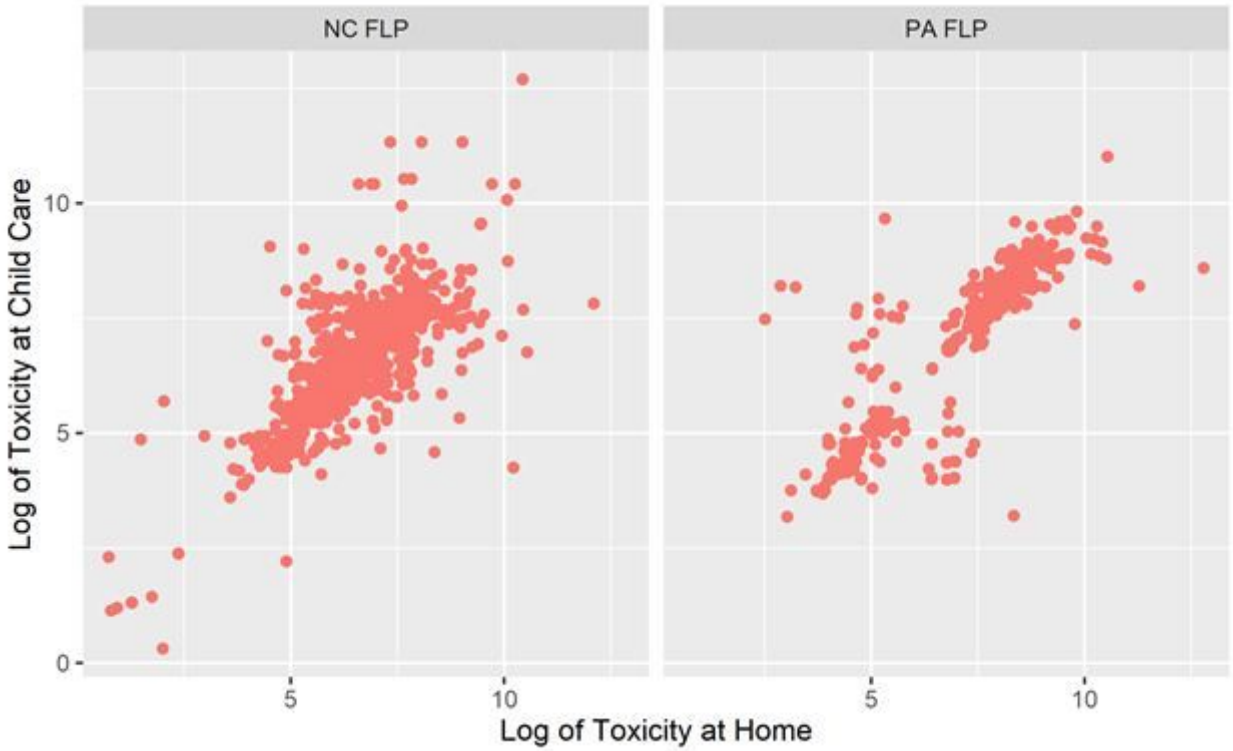

Figure 6