Published in final edited form as:

Environ Justice. 2013 June ; 6(3): 81–87. doi:10.1089/env.2013.0010.

Utilizing Exploratory Spatial Data Analysis to Examine Health and Environmental Disparities in Disadvantaged Neighborhoods

Kristin M. Osiecki,

Ph.D. Candidate at the School of Public Health, Division of Environmental and Occupational Health Sciences, at the University of Illinois at Chicago (UIC)

Dr. Seijeoung Kim,

Assistant professor of health policy and administration at the UIC School of Public Health

Ms. Ifeanyi B. Chukwudozie, and

Project coordinator at the UIC Institute for Health Research and Policy

Dr. Elizabeth A. Calhoun

Professor of health policy and administration at the UIC School of Public Health

Abstract

Health disparities research has focused primarily on racial and socioeconomic differences in health outcomes. Although neighborhood characteristics and the concept of built environment have been shown to affect individual health, measuring the effects of environmental risks on health has been a less developed area of disparities research. To examine spatial associations and the distribution of geographic patterns of sociodemographic characteristics, environmental cancer risk, and cancer rates, we utilized existing data from multiple sources. The findings from our initial analysis, which concerned with proximity to environmental hazards and at-risk communities, were consistent with results of previous studies, which often reported mixed relationships between health disparity indicators and environmental burden. However, further analysis with refined models showed that several key demographic and subdomains of cancer risk measures were shown to have spatial components. With the application of exploratory spatial data analysis, we were able to identify areas with both high rates of poverty and racial minorities to further examine for possible associations to environmental cancer risk. Global spatial autocorrelation found spatial clustering with percent black, percent poverty, point and non-point cancer risks requiring further spatial analysis to determine relationship of significance based on geography. This methodology was based upon particular assumptions associated with data and applications, which needed to be met. We conclude that careful assessment of the data and applications were required to properly interpret the findings in understanding the relationship between vulnerable populations and environmental burden.

Address correspondence to: Kristin M. Osiecki, University of Illinois at Chicago, School of Public Health, Division of Environmental and Occupational Health Sciences, 2121 W. Taylor Street, Chicago, IL 60612, kosiec3@uic.edu.

The authors have no conflicts of interest or financial ties to disclose.

INTRODUCTION

Health disparity research examines numerous factors such as socioeconomic and demographic variables to understand poor health outcomes including racial minorities living in poverty. Recently, a focus has been placed on the built environment taking into consideration the surroundings of a neighborhood that may contribute to disparity. Environmental health factors are of great interest because certain cancers had associations to environmental cancer risk. Epidemiological studies confirm a relationship between the location of pollution sources and incidences of multiple cancer types.^{1,2,3,4,5} The examination of environmental risk levels, human exposure and proximity to such risks in air, water, and soil, and availability of resources to mitigate the effects of these environmental risks health influence health outcomes.^{6,7} Moving forward with social research, there are numerous considerations that require an understanding of limitations and assumptions of both the data and geographical information system (GIS) software to ensure the correct interpretation of outcomes.

Traditional statistical results often fail to provide insights into complex relationships between geography, sociodemographic characteristics, and environmental exposures. For instance, multivariate regression captures the strength and the significance of statistical relationships between the dependent variable and other explanatory factors, however; important local variations between dependent and exploratory variables may not have been understood or looked at the high dimensionality of the datasets.⁸ Therefore, there is a need for other methods to describe these interactions of interest.⁹

Early environmental justice research explored a number of hazardous sites in disadvantaged communities. The premise was that more hazardous sites translated to higher exposure to environmental risk factors. The U.S. Environmental Protection Agency (EPA) acknowledged that past methods analyzed the proximity to sources of environmental hazards. But, mapping sites to evaluate spatial clusters lacked evidence between increased exposure and risk. In addition, mapping environmental injustice did not measure the correspondence between the location of potential environmental burdens, exposures, and health effects.¹⁰ National-Scale Air Toxics Assessment (NATA) cancer risk data was an opportunity to understand exposure risk in relation to health outcomes despite limitations within this assessment.

Geostatistical modeling involves multivariate data which needs an underlying joint multivariate distribution for valid inference.¹¹ Traditional data analysis methods for multivariate visualization such as tables and scatter plots, commonly used to examine health disparities and environmental health data, have been found to be limited in their ability to represent very large datasets.¹² GIS visualization methods have an advantage of assisting in identification and further exploration of certain patterns, thereby generating new analysis that can be created in an easily understood form.¹³ Various components of an analysis design coupled with the use of domain expertise through interactive exploration can develop into multivariate spatial patterns and the data can be allowed to show the obvious for hypotheses development.¹⁴

Exploratory data analysis (EDA) look at data such as correlations and measures of fit but also needs to be carefully investigated because these results become invalid when there is spatial dependence.¹⁵ Exploratory spatial data analysis (ESDA) focuses on spatial aspects of the data to find possible spatial patterns and outliers.¹⁶ Spatial methodologies find apparent spatial relationship; however, there are several problems inherent with social data that need to be considered. For example, spatial health research uses data that is collected for a purpose not specific for spatial analysis and is sometimes sampled in a systematic way from a spatially distributed population.¹⁷ Because there are limitations regarding inference from analyzing spatial patterns, researchers need to understand spatial systems, the selection of and specification of spatial weights and the subjectivity of the methods themselves.¹⁸

Visualization methods have an advantage of identifying and exploring certain patterns, thereby generating new analysis that could be designed in an easily understood form.¹⁹ ESDA tools such as maps, scatterplots, and parallel coordinate plots present information in a seeable manner to discover these patterns. Various components of an analysis design coupled with the use of domain expertise through interactive exploration could develop into multivariate spatial patterns and the data could show the obvious for hypotheses development.²⁰ Spatial analysis software has a statistical pattern recognition approach and is implemented in which a spatial cluster statistic or autocorrelation statistic is used to quantify a relevant aspect of a spatial pattern.²¹ However, the term “cluster” in health research is often generic that it fails to describe spatial variation without a precise description of the statistical test, heterogeneous population sizes, spatial autocorrelation, and non-uniform risks in social science.²²

In our study, limitations and assumptions associated with the EPA NATA environmental health risk assessment data posed challenges in reviewing the relationship between cancer risk and health disparity measures. The EPA provided data at the census tract level, yet stipulated that NATA assessments were not a definitive means to identify specific risk values within a census tract and that these results were more meaningful at the state or national level.²³ Current research suggested that smaller units such as tract or block group measures were: 1) most attuned to capturing economic deprivation, 2) meaningful across regions and over time, and 3) easily understood, and hence based on readily interpretable variables.²⁴

METHODS

First, we performed EDA using variables of interest that best captured vulnerable populations and environmental burden within Cook County, IL. This list was extensive utilizing health disparity indicators such as race/ethnicity and socioeconomic status (SES) and NATA environmental cancer risk data divided into six categories. STATA software was used for descriptive statistics to examine the distributions of dependent and independent variables, correlations between continuous demographic variables, the total cancer risk derived from NATA and cancer incidence rates, and bivariate models to explore relationships between the outcome measures and demographic and environmental risk factors.

We then used ArcGIS and OpenGeoDA for EDA visualization in identifying variables and areas warranting further examination. Based on the results of EDA and parallel coordinate plots, we tailored the list to key variables to model a series of global spatial autocorrelations to determine whether clustering existed and if so, to move forward with future local autocorrelation and spatial regression. Census tracts that showed possible correlation were then brushed and linked to box plot maps for visual comparison. In OpenGeoDA, brushing and linking was a technique that allowed us to highlight points of interest on a graph and then links or identifies these points on corresponding maps and charts.

We created choropleth maps in ArcGIS software based on the variables of interest outlined above to visualize descriptive statistics. ESDA including scatterplots and parallel coordinate plots looked at multivariate and bivariate relationships as a precursor to investigating spatial randomness in OpenGeoDA software. Although these tools provided additional insight into the data, there were no definitive relationships between variables of interest when we looked at the entire geographical area. Therefore, we decided to narrow the scope of our geography by examining census tracts located in the West and South regions of the City of Chicago.

We used spatial weights based on a queen matrix to take into consideration contiguity issues. Spatial autocorrelation, Global Moran's I with permutation inference, examined negative and positive spatial correlation with a standardized z -value. The goal of positive and negative spatial autocorrelation was to investigate similar and dissimilar values in relation to location. The spatial autocorrelation statistic captured both attribute and location similarity but it was important that the z -statistic was not interpreted as statistical significance. This method investigated clustering to decide if a spatial relationship exists and if so, to justify further research with local spatial autocorrelation and spatial regression.

RESULTS

We explored the data based upon the variables of interest that were found to be correlated with multiple socio-economic and demographic attributes. This included percent of Hispanic and black populations, percent of poverty and median household income, as well as percent rented housing units and percent of population without a high school diploma. These factors were then examined in relation to the six categories of NATA cancer risk and cancer rates at the census tract level. Initially, it appeared that there was no association between health disparity indicators and environmental burden. This corresponded to past research that had mixed results when looking at proximity to environmental hazards and at-risk communities.

Instead of looking at the data from a broad perspective to narrow down the list, we decided to start with three key variables: percent poverty, percent black residents, and NATA total cancer risk. By linking the upper outlier of percent poverty gave us an opportunity to identify these census tracts which were all located within the City of Chicago.

The choropleth maps (Figure 1) zoomed into these areas to see if these tracts were concentrated in certain Chicago Community Areas (CCAs). The 77 CCAs had unique neighborhood characteristics and have defined census data to correspond with these boundaries. We identified 18 of 77 CCAs predominantly on the south and west sides that

contain at least one census tract within their boundary. There was a visual pattern concentrated on the west and south sides with a few tracts scattered toward the north and southwest.

A parallel coordinate plot was then generated using the seventy census tracts in the highest outlier category with high percent poverty and high percent black population to look at the different types of cancers and cancer risk exposures from point and non-point sources. Figure 2 showed total cancer incidence rate, and the rates of breast and lung cancer, point source cancer risk and non-point source cancer risk. These two cancer risk categories were based on the premise that at-risk neighborhoods have higher environmental burden due to the proximity to hazard. Figure 1 showed that point source cancer risk appeared to be higher in eight of the CCAs, which were all known disadvantaged neighborhoods including North Lawndale. Also, there were census tracts of higher risk adjacent to these CCAs that needed to be further examined to determine if they are also areas with health disparity. The non-point source cancer risk map yielded a more prominent pattern within the majority of the eighteen CCAs and with neighborhoods adjacent to CCAs on the west side of Chicago.

The global spatial autocorrelation looked at a pattern as whole or “clustering” and a general approach to similarity and dissimilarity. In Figure 3, the positive spatial autocorrelation in the upper right quadrant looked at similarity of neighbors while negative spatial correlation in the lower left corner looked at the dissimilarity of neighbors. This was not a definitive outcome showing a spatial relationship but an indicator that we could move onto the next diagnostic test of local autocorrelation.

We conducted global spatial autocorrelation on percent poverty for census tracts in Cook County, IL to determine if there was clustering and if so, did the upper outliers correspond to the seventy census tracts identified through ESDA. With a Moran’s I score of 0.66, we observed clustering and with the linking technique we saw that the census tracts from the west and south side were in the upper right quadrant of the scatterplot (Figure 3). This indicated a positive and significant clustering of like values. We then examined global spatial autocorrelation with non-point source cancer risk for census tracts in Cook County, and found a weaker possibility of clustering with a Moran’s I score of 0.49. There was no clustering of the lung cancer incidence rates and the linking of the high-high quadrant showed no patterns.

DISCUSSION

The use of EDA/ESDA with health disparities and the built environment may provide additional insights into identifying at-risk neighborhoods with vulnerable populations and increased environmental burden. Our findings showed that some demographics and subdomains of cancer risk measures had possible spatial components. With the application of ESDA, we were able to identify seventy census tracts with both the high rates of poverty and racial minorities. The Global Moran’s I results further showed clustering for percent poverty and non-point cancer risk. These areas were predominantly poor neighborhoods in Chicago. Additional investigation will be required as to the reason for the high rate and risk in these census tracts when compared to other census tracts with lower rates and risks.

GIS methodologies have become popular in health disparities research because of the ability to conduct spatial analysis, however, a challenge with geographical aspects of data is the determination of appropriate boundaries of study. For the use of GIS to generate a reliable data for testing hypotheses in population health, however, there needs to be an understanding of the GIS methods used in and providing justification for the geographic level of study chosen.²⁵ Choosing levels and geospatial units to analyze depends on several factors, including the research objective, the causal model selected, the exposures and health outcomes of interest, and the extent to which data are available.²⁶ The availability of data is often the determining factor in decisions about geospatial issues especially with a driving force being sociodemographic variables from the census. In deciding the scale of analysis or the level of aggregation in a study, there tends to be a trade-off of the specificity of the study and the precision of the study.²⁷ The trend seems to be that the larger an area is, the less the specificity and the relevance of the findings of the study to the local populations but the higher the precision and the reduction of bias.²⁸

The reconciliation of various geographical boundaries presented numerous concerns. In the Chicago land area, data were collected by various geographical boundaries including: county, census tract, zip code, and CCAs. There were techniques within GIS programs to reconcile these boundaries such as clipping, intersection, and dissolving, however there was an assumption of homogeneity when manipulating these boundaries. We utilized crosswalk files from the U.S. Department of Housing and Urban Development which provided ratios to distribute various land categories. This proportionately estimated the type of land used within the geographic area but does not discern by locality.

Spatial patterns on GIS maps helped formulate hypotheses for explaining geographic patterns, but such approaches are hardly sufficient to explain complex interactions among spatial data.²⁹ We identified several issues in mapping environmental health disparities including the lack of comprehensive hazard databases; inadequacy of exposure indices, risk assessment methodologies, and insufficient health effects data.³⁰ When a disproportionate environmental burden based on race and/or income was found, it was critical to demonstrate the disproportionate effects of pollution rather than just the disproportionate distribution of pollution sources.³¹

GIS was based on spatial models that apply to static spatial systems,³² which made it difficult to represent human mobility and temporal change in cancer, environmental, and socioeconomic data.³³ Environmental health factors added another layer of complexity in determining exposure and risk especially when arbitrary boundaries could not take into effect the migration of contaminants. GIS had great promise in health disparities research as we investigated the physical environment to determine if hazardous exposures impact health in at-risk populations, however; exercising caution was needed especially with defining assumptions and limitations in interpreting outcomes.

CONCLUSION

ESDA identified seventy census tracts in the upper outlier for both percent poverty and percent blacks in eighteen CCAs in the City of Chicago. The Global Moran's I results

further showed clustering for percent poverty and non-point cancer risk. These areas were predominantly poor neighborhoods in the west and south side of Chicago with non-point cancer risk located on the west, north and south sides of Chicago. The North Lawndale neighborhood and the adjacent area had the most census tracts with high non-point cancer risk and cancer incidence. The next step was to continue with ESDA to look at local spatial autocorrelation to confirm our potential variables for spatial regression.

As we move forward with spatial analysis in health disparities research, it is important to understand the applicability of GIS software with social science data. In addition, researchers need to discern parametric approaches in traditional statistical methods and nonparametric approaches in spatial analysis; and, the limitations and assumptions associated with both. Methodologies to address these issues will ensure appropriate interpretations of outcomes.

Acknowledgments

Funding provided by NIH National Institute on Minority Health and Health Disparities (NIMHD) P60MD003424-S1.

References

1. Hockman, Elaine; Morris, Charles. Progress Towards Environmental Justice: A Five-year Perspective of Toxicity, Race and Poverty in Michigan, 1990–1995. *Journal of Environmental Planning and Management*. Mar.1998 41:157–176.
2. Bullard, Robert; Johnson, Glen. Environmentalism and Public Policy: Environmental Justice: Grassroots activism and Its Impact on Public Policy Decision-Making. *Journal of Social Issues*. Sep. 2000 100:555–578.
3. Boeglin, Michael; Wessels, Dennise; Henshel, Diane. An Investigation of the Relationship Between Air Emissions of Volatile Organic Compounds and the Incidence of Cancer in Indiana Counties. *Environmental Research*. Feb.2006 100:242–254. [PubMed: 15992796]
4. Clapp, Richard; Jacobs, MM.; Loechler, Edward. Environmental and Occupational Causes of Cancer New Evidence. *Review of Environmental Health*. Jan.2008 23:1–37.
5. Boffetta, Paolo; Nyberg, Fredrik. Contribution of Environmental Factors to Cancer Risk. *British Medical Bulletin*. Dec.2003 68:71–94. [PubMed: 14757710]
6. Cohen, Larry; Chehimi, Sana; Chavez, Vivian. *Prevention is Primary: Strategies for Community Wellbeing*. Jossey-Bass; 2010.
7. Kawachi, Ichiro; Berkman, Lisa. *Neighborhoods and Health*. Oxford University Press Inc; 2003.
8. Gilbert, Angela; Chakraborty, Jayajit. Using Geographically Weighted Regression for Environmental Justice Analysis: Cumulative Cancer Risks from Air Toxics in Florida. *Social Science Research*. Aug.2011 40:273–286.
9. Guo, Diansheng; Gahegan, Mark; MacEachren, Alan; Zhou, Biliang. Multivariate Analysis and Geovisualization with an Integrated Geographic Knowledge Discovery Approach. *Cartography and Geographic Information Science*. Dec.2005 32:113–132. [PubMed: 19960118]
10. Wilkinson, P.; Grundy, C.; Landon, M.; Stevenson, S. GIS in Public Health. In: Gattrell, Anthony; Taylor, Loytonen, editors. *GIS and Health: GISDATA*. 6. Taylor and Francis; 2004. p. 179-189.
11. Young, Linda; Gotway, Carol. Using Geostatistical Methods in the Analysis of Public Health Data: The Final Frontier? *Geostatistics for Environmental Applications*. 2010; 16:89–98. *geoENV VII* —.
12. Chen, Jin; MacEachren, Alan; Guo, Diansheng. Supporting the Process of Exploring and Interpreting Space-Time Multivariate Patterns: The Visual Toolkit. *Cartography and Geographic Information Science*. Dec.2009 35:33. [PubMed: 19960096]

13. Chen, Jin; MacEachren, Alan; Guo, Diansheng. Supporting the Process of Exploring and Interpreting Space-Time Multivariate Patterns: The Visual Toolkit. *Cartography and Geographic Information Science*. Dec.2009 35:33. [PubMed: 19960096]
14. Chen, Jin; MacEachren, Alan; Guo, Diansheng. Supporting the Process of Exploring and Interpreting Space-Time Multivariate Patterns: The Visual Toolkit. *Cartography and Geographic Information Science*. Dec.2009 35:33. [PubMed: 19960096]
15. Anselin, Luc. The Moran Scatterplot as an ESDA Tool to Assess Local Instability in Spatial Association. In: Fischer, Manfred, editor. *Spatial Analytical Perspective on GIS: GISDATA*. 4. Taylor and Francis Ltd; 1996. p. 121
16. Anselin, Luc. The Moran Scatterplot as an ESDA Tool to Assess Local Instability in Spatial Association. In: Fischer, Manfred, editor. *Spatial Analytical Perspective on GIS: GISDATA*. 4. Taylor and Francis Ltd; 1996. p. 121
17. Wilkinson, P.; Grundy, C.; Landon, M.; Stevenson, S. GIS in Public Health. In: Gattrell, Anthony; Taylor, Loytonen, editors. *GIS and Health: GISDATA*. 6. Taylor and Francis; 2004. p. 179-189.
18. Wilkinson, P.; Grundy, C.; Landon, M.; Stevenson, S. GIS in Public Health. In: Gattrell, Anthony; Taylor, Loytonen, editors. *GIS and Health: GISDATA*. 6. Taylor and Francis; 2004. p. 179-189.
19. Guo, Diansheng; Gahegan, Mark; MacEachren, Alan; Zhou, Biliang. Multivariate Analysis and Geovisualization with an Integrated Geographic Knowledge Discovery Approach. *Cartography and Geographic Information Science*. Dec.2005 32:113–132. [PubMed: 19960118]
20. Guo, Diansheng; Gahegan, Mark; MacEachren, Alan; Zhou, Biliang. Multivariate Analysis and Geovisualization with an Integrated Geographic Knowledge Discovery Approach. *Cartography and Geographic Information Science*. Dec.2005 32:113–132. [PubMed: 19960118]
21. Goovaerts, Pierre; Jacquez, Geoffrey. Detection of Temporal Changes in the Spatial Distribution of Cancer Rates Using Local Moran's I and Geostatistically Simulated Spatial Neutral Models. *Journal of Geographical Systems*. May.2005 7:137–159. [PubMed: 16710441]
22. Goovaerts, Pierre; Jacquez, Geoffrey. Detection of Temporal Changes in the Spatial Distribution of Cancer Rates Using Local Moran's I and Geostatistically Simulated Spatial Neutral Models. *Journal of Geographical Systems*. May.2005 7:137–159. [PubMed: 16710441]
23. U.S. Environmental Protection Agency, Office of Air Quality, Planning and Standards Research. An Overview of Methods for EPA's National-Scale Air Toxics Assessment. 2011.
24. Wilkinson, P.; Grundy, C.; Landon, M.; Stevenson, S. GIS in Public Health. In: Gattrell, Anthony; Taylor, Loytonen, editors. *GIS and Health: GISDATA*. 6. Taylor and Francis; 2004. p. 179-189.
25. Krieger, Nancy. Place, Space, and Health: GIS and Epidemiology. *Epidemiology*. Jul.2003 14:384–385. [PubMed: 12843759]
26. Gilboa, Suzanne; Mendola, Pauline; Olshan, Andrew; Harness, Catherine; Loomis, Dana; Langlois, Peter; Savitz, David; Herring, Amy. Comparison of Residential Geocoding Methods in Population-Based Study of Air Quality and Birth Defects. *Environmental Research*. Jun.2006 101:256–262. [PubMed: 16483563]
27. Wilkinson, P.; Grundy, C.; Landon, M.; Stevenson, S. GIS in Public Health. In: Gattrell, Anthony; Taylor, Loytonen, editors. *GIS and Health: GISDATA*. 6. Taylor and Francis; 2004. p. 179-189.
28. Wilkinson, P.; Grundy, C.; Landon, M.; Stevenson, S. GIS in Public Health. In: Gattrell, Anthony; Taylor, Loytonen, editors. *GIS and Health: GISDATA*. 6. Taylor and Francis; 2004. p. 179-189.
29. Ryttonen, Mika. Not all Maps are Equal: GIS and Spatial Analysis in Epidemiology. *International Journal of Circumpolar Health*. Mar.2004 63:9–24. [PubMed: 15139238]
30. Krieger, Nancy. Place, Space, and Health: GIS and Epidemiology. *Epidemiology*. Jul.2003 14:384–385. [PubMed: 12843759]
31. Krieger, Nancy. Place, Space, and Health: GIS and Epidemiology. *Epidemiology*. Jul.2003 14:384–385. [PubMed: 12843759]
32. Wilkinson, P.; Grundy, C.; Landon, M.; Stevenson, S. GIS in Public Health. In: Gattrell, Anthony; Taylor, Loytonen, editors. *GIS and Health: GISDATA*. 6. Taylor and Francis; 2004. p. 179-189.
33. Wilkinson, P.; Grundy, C.; Landon, M.; Stevenson, S. GIS in Public Health. In: Gattrell, Anthony; Taylor, Loytonen, editors. *GIS and Health: GISDATA*. 6. Taylor and Francis; 2004. p. 179-189.

FIG. 1.

Chloropleth maps for visualization of patterns. In the upper left corner, the map shows the seventy chosen census tracts located within Cook County, IL. The other three maps zoom into the south and west sides of Chicago showing the community areas that contain the seventy census tracts to examine patterns of total cancer rate (upper right corner), point source cancer risk (lower left corner), and non-point cancer risk in relation to the surrounding area. A color version of this figure is available in the online article at www.liebertpub.com/env.

FIG. 2.

Parallel coordinate plot for the seventy census tracts with highest percent poverty and highest percent black population in relation to total cancer incidence rate, breast cancer incidence rate, lung cancer incidence rate, point source cancer risk, and non-point source risk. A color version of this figure is available in the online article at www.liebertpub.com/env.

FIG. 3.

Box plot (hinge 1.5) percent poverty in Global Moran's I with upper outliers in the high-high quadrant highlighted with the corresponding census tract location on the Cook County, IL map. A color version of this figure is available in the online article at www.liebertpub.com/env.