doi:10.2498/iti.2013.0590

Applications of Geographic Information Systems and Geo-spatial Modeling in Environmental Epidemiology

John E. Vena, Sara E. Wagner and Stephen L. Rathbun
Department of Epidemiology and Biostatistics, University of Georgia, Athens, GA 30602, USA
E-mails: jvena@uga.edu, swagner@uga.edu, rathbun@uga.edu

Abstract

The ability of environmental epidemiology to determine the relationships between health and environmental insults has become exceedingly difficult. The multifactorial nature of disease and the diversity of the insults, which include biologic, physical, social and cultural factors, combined with genetic susceptibility, suggest the need to incorporate comprehensive perspectives of multidisciplinary epidemiologic investigation; utilize tools, such as geographic information systems (GIS) and other geospatial methods, which can integrate multi-level, spatial, and temporal factors and can help limit the potential for misclassification of exposure estimates; and to encourage collaborations and creativity in the field of environmental epidemiology. Examples of applications in cancer epidemiology will be presented. Statistical challenges to linking spatial pattern of cancer to radiation exposure will be discussed.

There has been much dialogue in the literature recently about the theoretical basis for investigation, suggesting that current epidemiologic methods and philosophies miss the mark and perhaps even have reached their limits. Concerns have been expressed regarding the emphasis on molecular epidemiology and biology of the disease outcomes while missing the big picture involving individuals in their social, cultural and physical settings. The proponents of the eco-epidemiologic paradigm would suggest that the focus on the details associated with the individual could result in missing the social determinants and the more population-level determinants of disease. As the disciplines involved in studies are expanded and thus, the perspectives leading to understanding increase, concern is being expressed about the need for an integrative process enabling the construction of new descriptions of risk and disease. One of the fundamental advantages of incorporating geospatial techniques is the ability to address these population-level disease determinants, which may be ignored in standard non-spatial individual-level epidemiologic studies.

Presently, public health must confront unprecedented challenges, including dramatic global population growth, increased aging, and possibly irreversible changes in key environmental health determinants with reference to globalization and climate change. Advancement in epidemiologic methods has occurred, but the determinants of health at the community level have been ignored thus leading to simplistic formulations of multiple risk factors. New tools for assessing health may promise greater efficiency and effectiveness for public health. Epidemiologic investigators should marry the biopsychosocial model of disease with the environmental-social cause model to determine common final pathways, furthering our understanding of how the underlying environmental and genetic factors mediate risks and how these translate into health, disease and quality of life. Susser and Susser and Pearce have been important voices in the dialogue criticizing modern risk-factor epidemiology as a discipline too focused on individual risk factors and too disconnected from examination of the broader historical and social forces that determine population disease risk. Epidemiologists, indeed, must learn to encompass multiple levels of organization from the societal to the molecular. The eco-epidemiology paradigm proposed by the Sussers addresses the interdependence of individuals and their connection with the biological, physical, social and historical contexts in which they live. It encompasses the changeable contributions and effects at both macro and micro levels of organization. The emphasis on the time

dimension implies that health and disease, in fact, involve temporal dynamic processes. Therefore, one would aim to assess causal factors at different levels of organization, over both the life course of the individual and the history of populations. Substantial methodological and inferential barriers need to be overcome and available research designs and analytic techniques are not well suited to elucidating processes at multiple levels of organization. We propose that a multidisciplinary environmental epidemiology that embraces and utilizes the new and innovative approaches and data resources of GIS and geospatial methods from multiple epidemiologic perspectives can address these methodological concerns, providing an integrated process to better understand disease from a more through viewpoint.

We want to emphasize the need to examine applications of GIS and geospatial analytic methods in environmental epidemiology. Although GIS and other spatial methods are now being used more commonly in environmental epidemiology (Mayer 1983; Stallones et al. 1992; Briggs and Elliott 1995; Clarke et al. 1996; Croner et al. 1996), their use tends not to encompass the entire continuum of health. The ultimate purpose of public health is to directly impact the health of a community or individual. We propose that the examination of exposure-disease relationships from multiple perspectives (traditional, acute event, community) may provide a more comprehensive knowledge base. For example, GIS analyses, considering issues such as exposure, disease risk, and population composition, could be used to decide on an ideal location for placing an intervention within a community for optimizing positive health changes.

The importance of multidisciplinary efforts in environmental epidemiology has been reinforced since the early 1990's. There was clear recognition of the need for cooperation among the disciplines including industrial hygiene, statistics, risk assessment, meteorology, engineering, epidemiology and biologic monitoring of both exposure and outcome. Collaborative efforts are paramount when proposing new visions for scientific research. The geospatial perspectives proposed in this paper are an example of a creative and collaborative addition to existing epidemiologic research.

Dr. Saxon Graham commented on enhancing creativity in epidemiology. He suggested that individual epidemiologists need energy, dedication, and enthusiasm. He urged early contact with top-notch mentors and stressed the importance of finding kindred spirits to work with. The stimuli to creativity are injection of new ideas from new sources and exposure to a diversity of new concepts and fields. He summarized by emphasizing that creativity meant learning from each other. He stated that there was a need to share ideas and methods to be creative. Stimulus of the new often derives from new technology and methods. Creative production is generally the result of the innovative joining of two disparate elements already in the field or of elements in the primary field with elements from a new field. Great ideas are hatched by just exploring or branching off onto new paths. New issues in epidemiology require new approaches. These also involve individuals with perceptions and knowledge different from traditional epidemiology. Creativity in this new setting requires utilization of new technologies. Innovation in multidisciplinary research has progressed as the members of the team achieved understanding and sharing of ideas. There is risk as new procedures are included. New ideas, approaches and methods must stand up to criticism and scrutiny. However, competitiveness and being super-critical will certainly reduce the amount of creative works produced. Too much criticism likely has been one of the important factors impeding the progress in environmental epidemiology. We propose that by more formal and integrative collaboration between environmental epidemiologists, geographers, and related disciplines such as geospatial scientists and spatial biostatisticians that innovative scientific approaches will be developed. We also propose that multi-disciplinary collaborations and scientific creativity are necessary in achieving a new frontier in environmental epidemiology and we believe that incorporating new geospatial perspective into existing thought is an important starting point. We now would like to illustrate examples of our recent applications of spatial biostatistical methods in environmental epidemiology.

Early life exposures, including exposure to PAHs, may have particular importance in the etiology of breast cancer. Early age at exposure to ionizing radiation, for example, confers

increased risk of breast cancer when compared with later age at exposure. Several other established risk factors also indicate the importance of early life factors in the etiology of breast cancer. We conducted a population-based, case-control study of exposure to PAHs in early life in relation to the risk of breast cancer using TSP, a measure of ambient air pollution, as a proxy for PAHs exposure. We examined time periods that are thought to be critical exposure periods with regard to susceptibility to breast cancer: at the time of birth, at menarche, at the time when the participant first gave birth, and 20 and 10 years before interview. In postmenopausal women, exposure to high concentrations of TSP (>140 Mg/m3) at birth was associated with an adjusted odds ratio of 2.42 (95% confidence interval, 0.97-6.09) compared with exposure to low concentrations (<84 Mg/m3). However, in premenopausal women, where exposures were generally lower, the results were inconsistent with our hypothesis and in some instances were suggestive of a reduction in the risk of breast cancer. Our study suggests that exposure in early life to high levels of PAHs may increase the risk of postmenopausal breast cancer.

Incidence of breast cancer is particularly high in the Northeastern US, an area with heavy industrial and traffic emissions. Traffic emissions are the major source of air pollutions in urban areas, and they contain many potential carcinogens, e.g., polycyclic aromatic hydrocarbons (PAHs) and benzene. Using lifetime residential histories in our case-control study (26), exposure to traffic emissions was modeled for each woman using her residence as a proxy. Estimates were calculated for residence at menarche, her first birth, and 20 and 10 years before interview. Higher exposure to traffic emissions at the time of menarche was associated with increased risk of premenopausal breast cancer (OR 2.05, 95% CI 0.92–4.54, p for trend 0.03); and at the time of a woman's first birth for postmenopausal breast cancer (OR 2.57, 95% CI 1.16–5.69, p for trend 0.19). Statistically significant associations were limited to lifetime non-smokers; there was a significant interaction between exposure at time of menarche and smoking for premenopausal women. Our findings add to accumulating evidence that early life exposures impact breast cancer risk and provide indication of potential importance of traffic emissions in risk of breast cancer.

To compare clustering patterns of breast cancer cases and controls at each time period, the primary method used was based on the K-function . The K-function for a point process is defined as the number of events within distance h of an arbitrary event, divided by the overall intensity of events. Under the null hypothesis of complete spatial randomness, the expected value of K(h) is $2\pi h^2$. Geographic clustering will yield values of the K-function that are greater than this, since clustering will result in more pairs of points separated by a distance of h than would be expected in a random pattern. We used the difference between K-functions for cases and controls to compare two patterns (i.e., D(h) = Kcase (h) – Kcontrol (h)). Positive values of D(h) indicate spatial clustering of cases relative to the spatial clustering of controls. Under the null hypothesis of random labeling of cases and controls, the expected value of D(h) is zero, indicating that the K-functions of the cases and controls are the same. The test statistic, D(h), was calculated with confidence envelopes using the splancs library in S-plus. When the estimated function D(h) deviated from zero by greater than two standard deviations, we interpreted this as a statistically significant difference between the case and control patterns.We found that the evidence for clustered residences at birth and at menarche was stronger than that for first birth or other time periods in adult life. Residences for pre-menopausal cases were more clustered than for controls at the time of birth and menarche. We also identified the size and geographic location of birth and menarche clusters in the study area, and found increased breast cancer risk for pre-menopausal women whose residence was within the cluster compared to those living elsewhere at the time of birth.

Levels of byproducts that result from the disinfection of drinking water vary within a water distribution system. This prompted us to question whether the risk for rectal cancer also varies, depending upon one's long term geographic location within the system. Such a geographic distribution in rectal cancer risk would follow naturally from an association between level of byproduct and rectal cancer risk. We assessed the effects of estimated geographic variability in exposure to some of the components of the trihalomethane group of disinfectant byproducts (DBPs) on the odds

ratios and probabilities for rectal cancer in white males in a case control study conducted in Monroe County, Western New York State, U.S.A. Using a combination of case control methodology and spatial analysis, the spatial patterns of THMs and individual measures of tap water consumption provide estimates of the effects of ingestion of specific amounts of some DBPs on rectal cancer risk. Trihalomethane levels varied spatially within the county; although risk for rectal cancer did not increase with total level of trihalomethanes, increasing levels of the component bromoform (measured in ug/day) did correspond with an increase in odds ratios (OR = 1.85; 95% CI = 1.25 – 2.74) for rectal cancer. The highest quartiles of estimated consumption of bromoform (1.69–15.43 ug/day) led to increased risk for rectal cancer (OR = 2.32; 95% CI = 1.22–4.39). Two other THMs were marginally associated with an increase in risk – chlorodibromomethane (OR = 1.78, 95% CI = 1.00–3.19) and bromodichloromethane (OR = 1.15; 95% CI = 1.00–1.32). Levels of THMs in the water distribution system exhibited spatial variation that was partially due to variation in water age. We also observed a geographic pattern of increased risk of rectal cancer in areas with the highest levels of bromoform in the county (15).

In 1993, Vena uncovered associations between increased total intake of fluids and risk for urinary bladder cancer which found increased risk with higher intake of tap water. This study had an estimate of tap water consumption but did not model estimated exposure to specific DBPs. We then planned to use DBP data made recently available to refine the tap water exposure estimate for a geographic subset of data from the previous study of tap water intake and cancer risk with estimated exposure doses to disinfection by-products based on residence location. We examined the relation between the estimated concentrations in drinking water of disinfectant by-product trihalomethanes (THMs) and the risk for urinary bladder cancer in a case control study of 567 white males ages thirty-five to ninety years in western New York State. Higher estimated levels of consumption of THMs led to increased risk for cancer of the urinary bladder (Total 551 (a composite measure of THMs) OR = 2.34, 95 % CI = 1.01-3.66). Results were most significant for Bromoform (OR = 3.05, 95 % CI = 1.51-5.69), and risk was highest (OR = 5.85, 95% CI = 1.93-17.46) for those who consumed the greatest amount of water at points within the distribution system with the oldest post-disinfection tap water.

One of the challenges facing environmental epidemiology is the spatial misalignment of data on disease outcomes and exposure data. For example, in our ongoing investigation of the relationship between radionuclide exposure and cancer in Georgia, we have the geographic coordinates of cases of different types of cancer together with measurements of household radon gas and uranium in wells at different point locations. No data on disease-free controls are available, but information on populations at risk and social factors may be obtained either in census tracks or at the county level. Data on cancer cases and the two sources of radionuclide exposure are obtained at different sets of spatial locations, and so the data are spatially misaligned. Moreover, while our data on cancer cases and exposure are at point locations, data on populations at risk are summarized in regions (i.e., counties or census tracks), so the data are at collected at different spatial scales. In most other applications, data on disease cases are spatially aggregated by such units as zip codes, census tracks or counties.

Bayesian spatial hierarchical models are well suited to analysis of data that are spatial misaligned and collected at different spatial scales (Wikle et al. 2001; Wikle 2003; Wikle and Berliner 2005). These models are constructed in three levels, a process model, a data model, and a model for the prior distribution of the model parameters (Berliner 1996). The process model describes the joint distribution of what might be considered ideally observable data, in our case, the distribution of the locations of cancer cases and controls, their social and demographic characteristics, and their exposures to the two types of radionuclides. As the name implies, the process model may be constructed so as to reflect the physical and epidemiology processes that generate exposure levels and disease cases. The data model describes the distribution of the observable data conditional on the realization of the process model. It may be constructed based on what is known regarding sampling design and measurement error. Both the process and data models depend on unknown parameters, the distribution of which is described by the prior model under the Bayesian inferential paradigm. Statistical inference is based on the posterior

distribution which is proportional to the product of the probability distributions of the process, data and prior models. Monte Carlo Markov Chain (MCMC) algorithms may be constructed to sample from this posterior distribution. MCMC samples may then be used to obtain the desired estimates together with measures of uncertainty that reflect all sources of variation in model.

Bayesian spatial hierarchical modeling has been successfully described the link between leukemia and Rn exposure by Smith, and exposure to air pollution or soil contaminants and other health outcomes by Zhu, Greco, Kim and Lee among others. A Bayesian approach is taken to ensure that all sources of variation are taken into account when assessing the uncertainty of risk estimates as well as other model parameters and spatial predictions. By using flat, non-informative priors, Bayesian interval estimates may have good frequentist interpretations.

In our implementation of the Bayesian spatial hierarchical model, the process model describes the joint distribution of the radiation exposure at point locations, and a point process model for distribution of cancer cases and controls. A multivariate geostatistical model for radiation exposure is constructed under which mean exposure depends on geological strata, and spatial dependence is modeled using anisotropic versions of the Matérn-class covariance function, a flexible class of covariance functions that includes a parameter describing the smoothness of realizations of the random field.

Locations of controls and cases of the different types of cancer are modeled as a realization of a multivariate Poisson point process model. Our multivariate point process model is an extension of the bivariate point process model used by Diggle and Rowlingston to describe the spatial distribution of disease cases and controls around a point source of environmental pollution. Controls are assumed to be realized from a point process with intensity $\lambda_0(s)$ while cancer cases of type *i* are assumed to be realized from a point process with intensity $\lambda_0(s)\exp\{\beta_i^T x(s)\}$, where $x(s)$ is a vector of covariates including radon and uranium exposure as well as important confounding variables such as age, gender and race. Census-track data may be used to estimate the baseline intensity $\lambda_0(s)$, and the multinomial logistic regression with an offset of $\log\lambda_0(s)$ may be used to describe the pattern of different types of cancers among the cases. While our long-term goal calls for the implementation of a fully Bayesian approach, various versions of two-stage Bayesian approaches are being explored from those that do not allow feedback between the exposure and health outcome components, through those that allow feedback for spatial prediction of exposure but not exposure parameter estimates.

We performed descriptive mapping of groundwater Uranium (U) data from NURE with GIS ArcMAP software to determine relative concentrations throughout Georgia,[19] using the Geostatistical Analyst tool to interpolate groundwater U samples via kriging. The results show elevated concentrations of U, consistent with the high concentrations expected in the Piedmont region of the Southeast US. As a proof of concept, we have applied a version of our Bayesian geospatial model to a subregion of northern Georgia with a more tractable sample size. Current results are based on a two-stage MCMC approach under which a geostatistical model is fit to radionuclide data and exposure at the cancer case sites is predicted using Bayesian kriging. In the absence of control data, a multinomial logistic regression model was used to describe the relative risks of the different types of cancer. Here, it is not possible to obtain unbiased estimate odds ratios, estimating the odds ratios of each cancer relative to controls, but it is possible to obtain odds ratios of each cancer relative to one or more of the remaining cancers. Our results suggest that uranium exposure has a lower impact on bladder and kidney cancer than it does on colorectal, breast and lung cancer.

Remaining challenges that have yet to be addressed are the large sample sizes of the radionuclide exposure data sets, (n > 20,000 observations), the left-censoring of radon measurements at the detection limit, and baseline intensity modeling. Predictive process and fixed rank kriging models will be considered for reducing the scale of the data sets. The approach of De Oliviera may be used to handle the left-censored radon data. The frequentist approach of Diggle and coworkers for estimating individual-

level risk from spatially aggregated information on the population at risk in census tracks may be extended to allow for spatial prediction of mean exposure in the census tracts and Bayesian inference for model parameters.

References

[1] Berliner, L.M. (1996). Hierarchical Bayesian time series models. In *Maximum Entropy and Bayesian Methods*, K. Hansen and R. Silver (Eds.), Kluwer Academic Publishers, 15-22.

[2] Stein, M.L. (1999). *Interpolation of Spatial Data. Some Theory for Kriging.* Springer: New York.

[3] Wikle, C.K. (2003). Hierarchical models in environmental science. *International Statistical Review* **71,** 181-199.

[4] Wilke, C.K., and Berliner, L.M. (2005). Combining information across spatial scales. *Technometrics* **47,** 80-91.