NISS

Bayes Methods for Combining Disease and Exposure Data in Assessing Environmental Justice

Lance A. Waller, Thomas A. Louis, and
Bradley P. Carlin

Technical Report Number 45
July, 1996

National Institute of Statistical Sciences
19 T. W. Alexander Drive
PO Box 14006
Research Triangle Park, NC 27709-4006
www.niss.org

Bayes Methods for Combining Disease and Exposure Data in Assessing Environmental Justice

Lance A. Waller
University of Minnesota

Thomas A. Louis
University of Minnesota

Bradley P. Carlin
University of Minnesota

Research Report 96-006
July 1996

Submitted to *Environmental and Ecological Statistics*

UNIVERSITY OF MINNESOTA
Division of Biostatistics
School of Public Health
A460 Mayo Building, Box 303
420 Delaware St. SE
Minneapolis, MN 55455-0378
(612) 624-4655

Bayes Methods for Combining Disease and Exposure Data in Assessing Environmental Justice

LANCE A. WALLER, THOMAS A. LOUIS, AND BRADLEY P. CARLIN[1]

Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, MN 55455, USA

July 25, 1996

Environmental justice reflects the equitable distribution of the burden of environmental hazards across various sociodemographic groups. The issue is important in environmental regulation, siting of hazardous waste repositories, and prioritizing remediation of existing sources of exposure. We propose a statistical framework for assessing environmental justice. The framework includes a quantitative assessment of environmental equity based on the cumulative distribution of exposure within population subgroups linked to disease incidence through a dose-response function. This approach avoids arbitrary binary classifications of individuals solely as "exposed" or "unexposed." We present a Bayesian inferential approach, implemented using Markov chain Monte Carlo methods, that accounts for uncertainty in both exposure and response. We illustrate our method using data on leukemia deaths and exposure to toxic chemical releases in Allegheny County, Pennsylvania.

Keywords: environmental equity, hierarchical model, Markov chain Monte Carlo, regulation

[1]Lance A. Waller is Assistant Professor, Thomas A. Louis is Professor and Head, and Bradley P. Carlin is Associate Professor, all in the Division of Biostatistics, School of Public Health, Box 303 Mayo Memorial Building, University of Minnesota, Minneapolis, MN 55455. Dr. Waller has developed methodology for spatial epidemiologic modeling and environmental applications of geographic information systems. Drs. Louis and Carlin have made contributions to empirical Bayes and Bayesian inference, especially with regard to biostatistical applications.

This research was supported in part by University of Minnesota Graduate School Grant-in-Aid of Research, Artistry and Scholarship #15196 (LAW), Environmental Protection Agency EPA CR 819638 to the National Institute of Statistical Sciences (LAW), National Institute of Allergy and Infectious Diseases (NIAID) FIRST Award 1-R29-AI33466 (BPC), and NIAID Contract NO1-AI05073 (TAL). The leukemia data were supplied by the Allegheny County Health Department, Pittsburgh, Pennsylvania. The Allegheny County Health Department specifically disclaims responsibility for any analyses, interpretations or conclusions.

1 Introduction

In recent years, researchers and the general public have questioned whether environmental exposures and risks are equitably distributed over population subgroups. The phrases "environmental justice" and "environmental equity" are used to describe situations where the risk of adverse outcomes due to environmental exposures is equitably distributed across subpopulations. These subpopulations may be defined by variables which are either demographic (e.g. socioeconomic status) or geographic (e.g. proximity to a hazardous waste site).

Environmental justice is becoming an important part of the federal environmental regulatory process. The United States Environmental Protection Agency established the Office of Environmental Justice in November of 1992 (Sexton, Olden and Johnson, 1993). In February 1994, President Clinton signed the Environmental Justice Executive Order, which requires an assessment of environmental justice in regulatory decisions made by every federal agency involved with environmental and public health.

Wagener and Williams (1993) divide environmental justice into three components providing a natural starting point for discussions of possible statistical analyses. The three components correspond to the geographic distributions of (a) the exposure to an environmental pollutant, (b) the health status of the population with respect to both disease incidence and access to health care, and (c) various subgroups of the population subject to potentially increased risk. These distributions lead to three primary questions of interest in studies of environmental justice, namely: (1) are members of a particular subpopulation subject to disproportionately high exposure, (2) are they experiencing a disproportionate number of adverse outcomes, and (3) is their risk of particular outcomes unduly increased by the exposure. These three questions directly correspond to the Environmental Protection Agency's paradigm of exposure assessment,

effects assessment, and risk characterization (Sacks and Steinberg, 1994, pg. 4). Appropriate statistical collection and analyses of environmental heath data play essential roles in addressing these issues.

Data regarding the demographic distribution of the population at risk typically are available from census data. Researchers often wish to stratify by various aspects, such as age, gender, race, or ethnicity. These variables and others, such as the percentage of the population with household income below certain thresholds, are available for census subregions (tracts, or block groups). These data may be analyzed as reported, or used to provide a sampling frame for prospective studies in a particular area.

The issue of environmental exposure is complex. It is essential to distinguish between the amount of a given substance present at a particular location (ambient exposure) and the dose actually received by a person at the same location. The dose received depends on many factors such as occupation, hand-mouth contact, and aerosol penetration rates. In a large-scale monitoring situation, a reasonable goal is to estimate the ambient exposure. Specific studies in industrial hygiene can then be used provide insight into average delivered dose levels associated with certain levels of ambient exposure and daily activities.

The remaining issue is the relation of subgroup exposure to the observed distribution of health events. The geographic distribution of health events is usually estimated from disease registries. In the United States, these registries are managed by the Centers for Disease Control and Prevention or by individual states. Due to confidentiality concerns, the data often are available only as census district counts.

Quality and availability of exposure data vary widely. The United States Environmental Protection Agency maintains the Toxic Chemical Release Inventory (TRI). It contains industry-reported releases of over 300 toxic chemicals to the air, land, and water. Transportation of

these chemicals to off-site facilities is also reported. The TRI is not a complete assessment of toxic releases since not all facilities handling toxic chemicals are required to report releases. Nevertheless, TRI data are publicly available and are often used in assessments of environmental justice (e.g. Bowen et al. 1995, Glickman and Hersh 1995). We illustrate the methods proposed below with TRI data from Allegheny County, Pennsylvania.

In the following, we distinguish between the terms "exposure inequity" and "risk injustice": exposure inequity refers to differences in exposure distributions, while risk injustice refers to differences in adverse outcomes due to exposure inequity. We differentiate between the terms to emphasize that we do not presume that differential exposures automatically result in differential adverse effects (i.e. risk injustice). From a public health perspective, risk injustices typically involve adverse health effects, but a more general view may encompass economic effects of exposure inequities as well.

Section 2 motivates our methods by an example from TRI-reported releases in Allegheny County, Pennsylvania for 1990. In Section 3 we outline the methods we propose, and show how they summarize risk injustice. Section 4 illustrates the approach with a dataset on leukemia cases and TRI site exposure in Allegheny County, highlighting the sampling-based Bayesian analytical methods we employ. Finally, Section 5 discusses our findings and suggests directions for future research.

2 Motivating example

Glickman and Hersh (1995) illustrate the use of geographical information systems (GIS) to assess environmental justice in Allegheny County, Pennsylvania. To motivate the methods described below, we use 1990 census tract data from Allegheny County and reported releases of toxic

chemicals from the 1990 TRI.

Glickman and Hersh (1995) distinguish between *proximity-based* and *risk-based* assessments of environmental justice. Proximity-based assessments use distance as a surrogate for exposure so that populations near releases are assumed to receive higher exposure than populations farther away. Risk-based assessments incorporate more data such as groundwater flow, location of occupation and residence, time spent at work and at home, and so on to portray a more accurate estimate of individual risk. For illustrative purposes we adopt a proximity-based assessment of environmental justice in Allegheny County for 1990.

TRI data include substance released, amount released, latitude, longitude, and street address for the reporting facilities. For 1990 there were 423 reported releases of TRI chemicals in Allegheny County. Since releases are reported by chemical or compound, a single location often is associated with several releases.

Latitude and longitude data were inconsistent with the boundaries of Allegheny County for 34 of the 423 releases. Of these, 17 appeared to have transposed latitude and longitude values (the negative sign associated with western hemisphere longitudes is implicit and not stored in the longitude field). The 406 locations with consistent latitude and longitude values correspond to the 84 unique release locations shown in Figure 1. We illustrate our methods using these 84 sites. The sites are found throughout the county but, as one might expect, concentrate along the three rivers (the Allegheny in the northeast, the Monongahela in the southeast, and the Ohio in the west).

A more complete assessment of environmental justice near TRI sites in Allegheny County requires assessing the accuracy of the release locations. One validity assessment checks the latitude and longitude values through "geocoding" (i.e., abstracting latitude and longitude from the recorded street address). Most modern geographic information systems (GIS's) include some

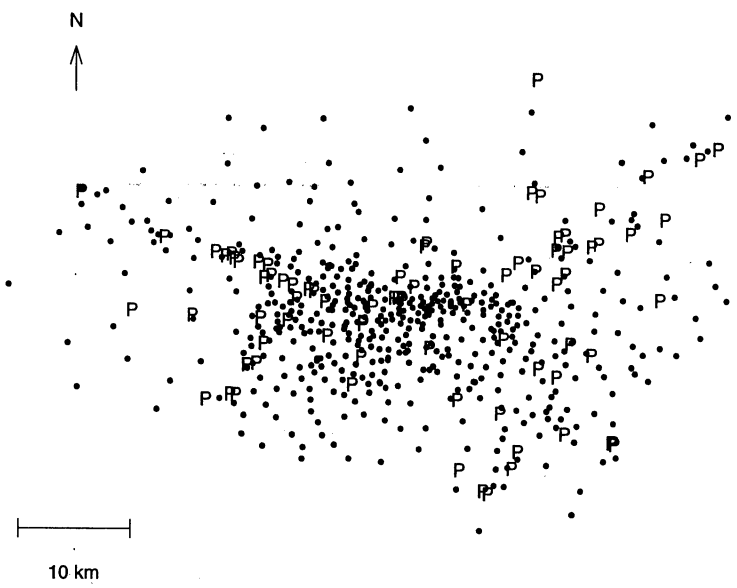

Figure 1: 1990 census tract centroids for Allegheny County, Pennsylvania. "P" indicates reported locations of 84 toxic chemical release sites from the 1990 Toxic Chemical Release Inventory.

form of automated geocoding although some effort is required to match a large percentage (say 95%) of addresses.

Table 1 gives summary data on the racial and ethnic composition of the Allegheny County population as reported in the 1990 Census. We see that the largest groups in terms of population size and in terms of proportion of population in census tracts are those with census designation "black" or "white". (We use the term "black" rather than "African American" for consistency with the census data.) These two groups comprise most of the census tract populations for Allegheny County in 1990. Our illustrative analysis compares the distribution of proximity to TRI sites between these two groups.

Glickman and Hersh (1995) make an interesting conceptual claim in their discussion of environmental justice and statistical uncertainty: namely, that when both exposure and demograph-

	Mean	Range
Population size	2678.26	(0, 8523)
Population per racial/ethnic groups		
White	2343.59	(0, 8297)
Black	299.70	(0, 4632)
American Indian	2.91	(0, 29)
Asian/Pacific Islander	26.99	(0, 484)
Other	5.06	(0, 56)
Hispanic	17.50	(0, 140)
Percent of tract population		
White	84.06%	(0.77, 100)
Black	14.62%	(0, 98.79)
American Indian	0.14%	(0, 4.76)
Asian/Pacific Islander	0.98%	(0, 21.61)
Other	0.20%	(0, 1.53)
Hispanic	0.70%	(0, 5.17)

Table 1: Population summary statistics for 1990 census tracts in Allegheny County, Pennsylvania

ics are known without error, no statistical evaluation is necessary; the situation is equivalent to a census, rather than a random sample. We agree with this viewpoint if the goal is purely descriptive, as when the question being asked is, "What is the distribution of exposure?" If, however, the questions concern causal evaluations, such as, "Is this distribution of exposure consistent with an equitable spatial-temporal exposure process," or "Does this distribution of exposure result in an inequitable pattern of adverse health effects," then statistical inference will have an important role. Of course, in many settings personal exposure will itself be estimated with considerable uncertainty, and so even the descriptive goal will require a statistical evaluation.

3 Statistical assessment of environmental justice

Suppose we have two population subgroups ("black" and "white") and let $G_i(\cdot)$ denote the cumulative distribution function (CDF) of exposure to a TRI site for group i, $i = B, W$. This distribution is constructed from regional and group-specific data. Inequity in exposure is defined

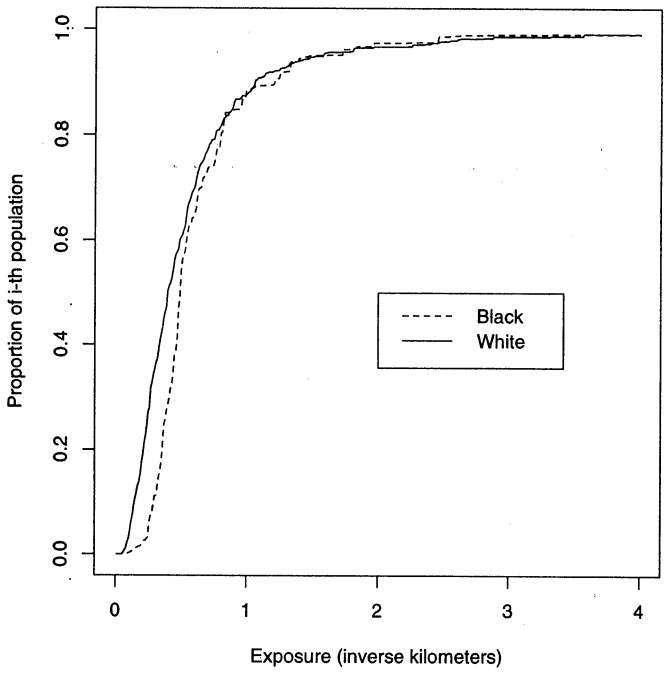

Figure 2: Cumulative distribution of exposure for Allegheny County, Pennsylvania 1990. Exposure is defined as the inverse distance in kilometers from the nearest TRI site to the centroid of a census tract. Separate cumulative distribution functions are shown for census designated "black" and "white" populations.

by differences between G_B and G_W.

For our proximity-based study we define exposure as the inverse distance (in km^{-1}) from a census tract centroid to the nearest TRI site. For simplicity, we assume that all individuals within a tract receive the same exposure. We define G_B and G_W as the empirical cumulative distribution function (CDF) of the inverse distance of census tract centroids to the nearest TRI site. Even though we use *empirical* CDF's, we treat the distributions as known population quantities in the analysis below since the "steps" in the functions are determined by census geography and not sampling uncertainty. (This issue is discussed more fully in Section 5 below.)

Figure 2 illustrates G_B and G_W for the 1990 Allegheny County data. We see a clear separation of the two curves for exposure values between 0.2 and 1 km^{-1}. We also see that picking

a single exposure value as a dichotomous exposed/unexposed threshold could result in different estimates of the effect magnitude. For example, Glickman and Hersh (1995) considered two threshold values, 0.5 and 1 mile. They report greater observed inequity for the 1 mile limit (corresponding to approximately 0.62 km^{-1}) than for the 0.5 mile limit (corresponding to approximately 1.25 km^{-1}). Figure 2 captures these results with a greater vertical distance between $G_B(x)$ and $G_W(x)$ when $x = 0.62$ than when $x = 1.25$ km^{-1}. Similar use of dichotomous proximity-based estimates of exposure appear in other assessments of environmental justice (Anderton et al., 1994; Bowen et al., 1995).

Creating exposure zones within a given distance of fixed locations is straightforward in a GIS. However, the dependence of results on the chosen threshold distance is a drawback. We prefer a summary of inequity that depends on the entire CDF of exposure (x) within each subpopulation. Rather than choosing an exposure threshold, we propose quantifying exposure inequity by a distance measure $D(G_W, G_B)$. There are many possible measures of differences between two CDF's, such as the familiar Kolmogorov-Smirnoff-type statistic

$$D^*(G_W, G_B) = \sup_x \{G_W(x) - G_B(x)\}.$$

Alternatively, a weighted, integrated difference of G_B and G_W yields a summary measure having general form

$$D_\theta(G_W, G_B) = \int y(x;\theta)\, d\{G_B(x) - G_W(x)\}^p \qquad (1)$$

for some function of exposure $y(x;\theta)$, parameterized by θ, and some power p. This family

includes the familiar Cramér-von Mises-type statistic

$$D(G_W, G_B) = \int \{G_W(x) - G_B(x)\}^2 \, dx \, .$$

The function $y(x;\theta)$ translates exposure differences into risk differences. Letting $y(x;\theta) = x$ in equation (1) produces the difference in expected exposure. If $y(x;\theta)$ computes the risk of an adverse outcome associated with exposure x and the dose-response relation is the same in the two groups, then $D_\theta(G_W, G_B)$ assesses differences in environmental justice. For illustration, the "one-hit" model gives a simple dose-response relationship. In this case, $y(x;\theta)$ results from an underlying model of carcinogenesis where a single exposure affects the probability of cell transformation from a normal to a diseased state, and is defined by

$$y(x;\theta) = 1 - \exp(-\theta x), \ \theta > 0. \tag{2}$$

In the special case where $p = 1$, it is easy to see what is assessed by $D_\theta(G_W, G_B)$. Following integration by parts, $D_\theta(G_W, G_B)$ can be written as

$$\int_0^\infty \left[\frac{\partial y(x;\theta)}{\partial x}\right] \{G_W(x) - G_B(x)\} \, dx. \tag{3}$$

We see that $D_\theta(G_W, G_B)$ represents the area under the curve

$$f(x;\theta) = [\partial y(x;\theta)/\partial x]\{G_W(x) - G_B(x)\} \, .$$

Figure 3 shows $y(x;\theta)$ and Figure 4 shows $f(x;\theta)$ for the 1990 Allegheny County data with the one-hit model and $\theta = 0, 2$, and 20. The one-hit model with $\theta = 0$ corresponds to a disease

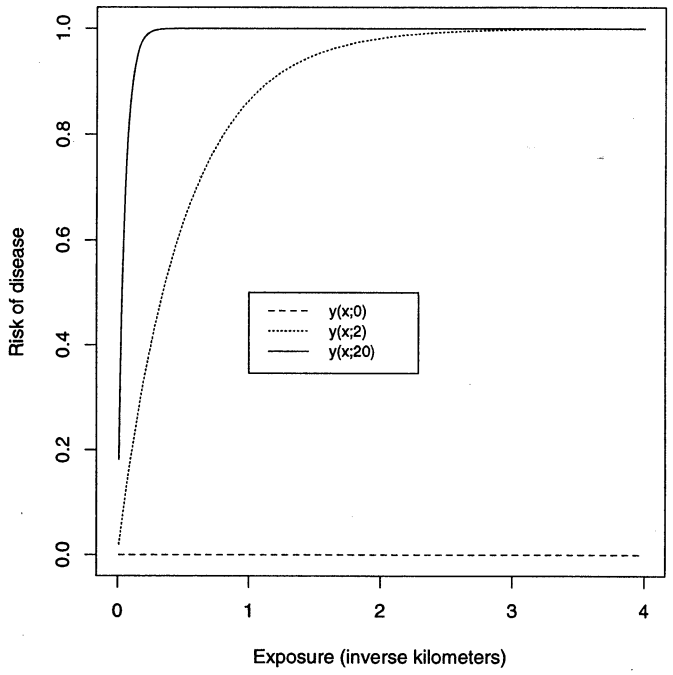

Figure 3: Dose response function $y(x;\theta) = 1 - \exp(-\theta x)$ from a one-hit model for $\theta = 0, 2, 20$.

unrelated to the exposure of interest. In this case, $D_\theta(G_W, G_B)$ is identically zero, and the exposure inequity observed in Figure 2 does not translate into a risk injustice for the disease of interest. In contrast, $\theta = 2$ corresponds to considerable risk injustice, since the area of greatest exposure inequity (x between 0.2 and 1 km^{-1}) corresponds to the area of greatest increase in disease risk with respect to x. When $\theta = 20$, only a small risk injustice remains, since the risk of disease is near 1.0 for $x > 0.5$ km^{-1}. For such a dose-response relationship, differences in exposure CDF's for $x > 0.5$ km^{-1} do not translate into sizable risk differences.

A potential problem in interpreting an estimate of D_θ for any given dataset is its somewhat arbitrary scale. To remedy this, we might switch to a *relative* scale by defining

$$R_\theta(G_W, G_B) = 100 \times \frac{D_\theta(G_W, G_B)}{T_{\theta,W}(G_W)},$$

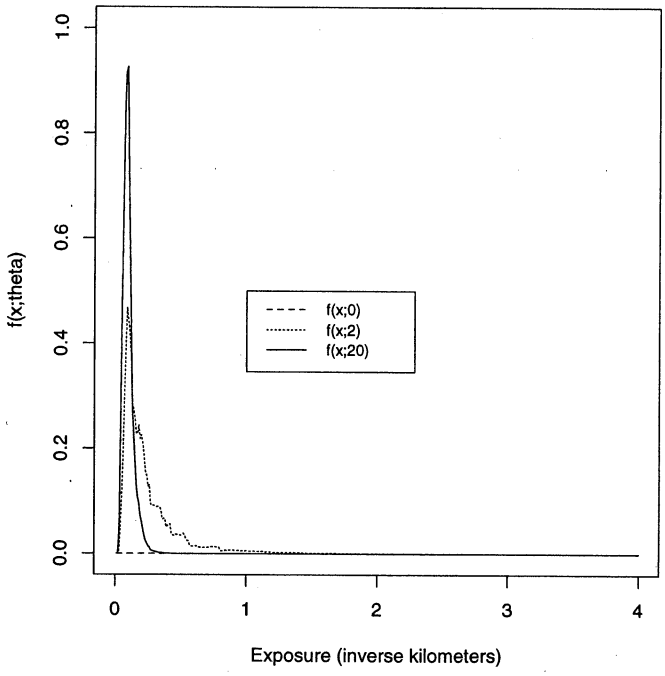

Figure 4: The "injustice function" $f(x;\theta) = \frac{\partial}{\partial x}[y(x;\theta)]\{G_W(x) - G_B(x)\}$ for $\theta = 0, 2, 20$. Similar line types in Figures 3 and 4 indicate the same value for θ.

where $T_{\theta,W}(G_W) \equiv \int y(x;\theta)\, dG_W(x)$, the "total" risk of disease for group W, marginalizing over the potential exposure values. $R_\theta(G_W, G_B)$ thus gives the percent increase in risk associated with group B, an easily interpreted measure of environmental inequity.

4 Data analysis

Using the development above, inference regarding environmental justice reduces to inference about D_θ and R_θ. We adopt a Bayesian approach, in order to obtain the full posterior distributions of these quantities rather than mere point estimates and associated asymptotic standard error estimates. Let Z_j be a random variable representing the number of leukemia cases occurring in census tract j, $j = 1, \ldots, J$. Similarly, let n_j denote the number of individuals at risk in

tract j. Our dataset, obtained from the Allegheny County (Pennsylvania) Health Department, provides disease incidence and population counts for all $J = 499$ census tracts in the county from 1985-1992.

For our analysis, Z_j includes counts of both acute myelogenous leukemia, AML, and acute lymphocytic leukemia, ALL. The two types of leukemia differ in etiology but are both of interest in environmental health. Most cases of leukemia in children are ALL. Childhood malignancies are often considered "sentinel health events" due to shorter latent periods between exposure and onset, and the shorter time for change of residence among the susceptible population. Toxicology studies link AML to exposure to benzene (a TRI-reported contaminant). We combine two outcomes of interest (ALL and AML) to illustrate our approach. A more focused assessment could consider only particular releases and outcomes.

Turning to the exposures, the locations of 84 waste sites in the county were obtained from the EPA Toxic Release Inventory. Let x_j denote the TRI site exposure of an individual residing in tract j, again defined as the reciprocal of the distance from the geographical center of the tract to the nearest site. We treat the x_j as fixed covariates, observed without error. Due to the presence of leukemia cases in tracts having $x_j \approx 0$, we generalize the simple one-hit model (2) slightly to

$$y(x;\theta,\gamma) = 1 - \gamma \exp(-\theta x),\ \theta > 0,\ 0 < \gamma < 1\,, \tag{4}$$

so that the "background" disease rate is $1 - \gamma$. Thus we have the binomial model $Z_j \sim Bin(n_j, y(x_j;\theta,\gamma))$, $j = 1,\ldots J$.

To complete the Bayesian model specification, we require prior distributions for θ and γ. For simplicity, we adopt familiar distributional forms for each, namely $\theta \sim Gamma(\alpha,\beta)$ and

$\gamma \sim Beta(a, b)$. Writing $\mathbf{z} = (z_1, \ldots, z_J)'$, the posterior distribution for $(\theta, \gamma)'$ is thus

$$p(\theta, \gamma|\mathbf{z}) \propto \left\{ \prod_{j=1}^{J} [1 - \gamma \exp(-\theta x_j)]^{z_j} [\gamma \exp(-\theta x_j)]^{n_j - z_j} \right\} \theta^{\alpha-1} \exp(-\theta/\beta) \gamma^{a-1} (1-\gamma)^{b-1} ,$$

up to an unknown constant of proportionality. Finally, we set $\alpha = \beta = a = b = 1$, thus determining an $Exponential(1)$ prior for θ and a $Uniform(0, 1)$ prior for γ, both very vague specifications designed to let the data (rather than prior information) dominate the posterior.

To obtain the standardized posterior, we use a Markov chain Monte Carlo (MCMC) method called the Hastings algorithm (Hastings, 1970; see Carlin and Louis, 1996, Sec. 5.4.3 for a more elementary description). This algorithm operates by alternately drawing "candidate" values θ^* and γ^* and accepting or rejecting them according to a criterion based on the unnormalized posterior. We used $Gamma(\alpha^*, \beta^*)$ and $Beta(a^*, b^*)$ candidate distributions for θ^* and γ^* respectively, with parameters chosen by trial-and-error to produce acceptance rates near 50%, a value that should provide good mobility in the Markov chain (Gelman et al., 1996).

Running five initially overdispersed parallel sampling chains for 1000 iterations each, we found algorithm convergence to be almost immediate, requiring a "burn-in" period of no more than 10 iterations. Figures 5(a) and (b) give kernel marginal posterior density estimates for θ and γ, respectively, based on the remaining (post-convergence) samples $\{(\theta^{(g)}, \gamma^{(g)}),\ g = 1, \ldots, G = 5(990) = 4950\}$. The estimated posterior mean for γ is 0.999677, implying a background leukemia probability of $1 - E(\gamma|\mathbf{z}) = 0.000323$. The estimated posterior mean for θ is 2.5765×10^{-5}, suggesting an increased leukemia rate at the highest exposures in our dataset ($x = 4.0$) of 0.000426.

An advantage of our MCMC approach is that posterior samples $D_\theta^{(g)}$ and $R_\theta^{(g)}$ may be obtained by simple transformation of the $(\theta^{(g)}, \gamma^{(g)})$ samples. For example, taking the derivative

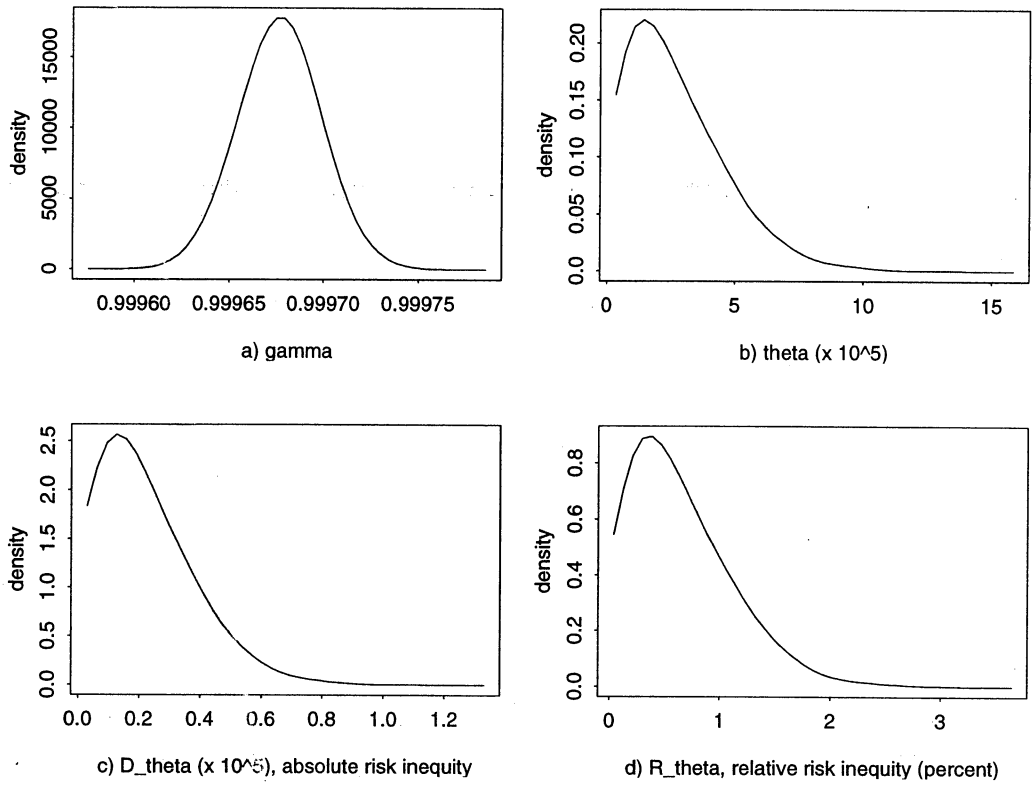

Figure 5: Estimated posterior distributions, Allegheny county dataset

of $y(x;\theta,\gamma)$ with respect to x, equation (3) gives

$$D_\theta^{(g)} = \int_0^\infty \gamma^{(g)}\theta^{(g)} \exp(-\theta^{(g)}x)\,\{G_W(x) - G_B(x)\}\,dx\ ,$$

a simple one-dimensional integral that can be evaluated using a grid-based (e.g. trapezoidal) rule. A similar expression is available for $T_{\theta,W}^{(g)}$, whence $R_\theta^{(g)} = 100D_\theta^{(g)}/T_{\theta,W}^{(g)}$.

Kernel estimates of the marginal posterior densities for D_θ and R_θ obtained using this approach are shown in Figures 5(c) and (d), respectively. Note that both are highly skewed, suggesting that the Gaussian approximation normally used to obtain traditional confidence intervals would be badly misleading in this case. R_θ has estimated posterior mean 0.626, with corresponding 95% equal-tail credible interval (.0284, 1.72). While this interval does exclude

zero, implying a statistically significant increase in leukemia risk for blacks, from a practical standpoint our results strongly suggest a high degree of environmental equity: the increase in risk among blacks is probably less than 1%, and is very unlikely to be more than 2%. While this result may be somewhat surprising in view of the relative affluence of suburban whites residing in Allegheny County, it is apparently due to the wide dispersion of TRI sites throughout the county, as well as the presence of both white and black socioeconomically depressed neighborhoods near many urban waste sites. Our findings also generally agree with those of Glickman and Hersh (1995), though these were somewhat more equivocal due to their strong dependence on the radius chosen to distinguish "exposed" and "unexposed" tracts, a difficulty our method was designed to avoid.

5 Discussion

The approach presented in this paper offers a family of quantifications of environmental justice with several advantages. The choice of the dose-response function $y(x;\theta)$ allows assessment of both exposure inequity and risk injustice. The method uses the whole distribution of exposures observed within population subgroups, rather than assessing differences based on particular threshold values. Finally, our Bayesian approach produces the entire posterior distribution of D_θ, rather than summary statistics alone, the sole output of a classical approach. The fact that Bayesian procedures implemented with vague priors like ours produce effective frequentist inferences has been well-documented (see e.g. Carlin and Louis, 1996, Chapter 4), so the investigator need not worry about lurking "subjectivity" in the results.

In order to outline the methodology, we have made several simplifying assumptions. However, the method is flexible and allows for generalization. For example, the dose-response models we

have illustrated allow the probability of response to approach 1.0 as exposure grows to infinity, but we could easily place an upper bound on this growth by adding a third parameter to model (4). Also, while the leukemia data represent all cases reported by physicians, whites represent 93.8% of reported cases of *ALL* (122 of 130) and 95.6% of *AML* (307 of 321). For comparison, the 1990 Census reports whites representing 88.7% of the combined black and white populations of Allegheny County. The difference may be due to differing age structures within the two subpopulations, but perhaps also to underreporting or underdiagnosis among some sociodemographic groups. Due to the small number of incident cases among one of the subpopulations, our data are insufficient to enable estimation of separate dose-response curves $y_i(x; \theta_i, \gamma_i)$ for each demographic group, $i = B, W$. Measured discrepancies in disease outcome (as well as exposure) among groups could well magnify environmental inequity, as captured in the posterior distributions of D_θ and R_θ like those plotted in Figures 5(c) and (d). Differential response between children and adults with similar exposures could also be accommodated in this way.

Another advantage of a Bayesian framework is the allowance for multiple levels of uncertainty inherent in data encountered in environmental justice studies. For example, disease rates in the various tracts might be spatially correlated, due to the effect of variables other than demographic group. We thus might want to add a random effect ϕ_j to our model for the response probability y, and assign a spatial smoothing prior to the vector $(\phi_1, \ldots, \phi_J)'$ that encourages the fitted rates in adjacent counties to be similar. Such models are often used in conjunction with a Poisson approximation to the binomial likelihood in disease mapping problems (Besag, York and Mollié, 1991; Clayton and Bernardinelli, 1992), and recently have been extended to the spatio-temporal case (Waller et al., 1997). Another possibility is to think of the G_i not as known population quantities to be computed from census data, but as additional unknowns to

be estimated. A Bayesian would likely do this nonparametrically, using recent developments in the theory and implementation of Dirichlet process priors (Escobar and West, 1995). This modification would likely have little impact on our point estimates for D_θ and R_θ, but would widen the corresponding interval estimates due to the added uncertainty in the G_i now accounted for by the model. Finally, if the exposure value x were itself measured with error (as opposed to our proximity-based approach), as would likely be the case with atmospheric measurements, an error distribution for x can be added to the hierarchy and the analysis may proceed, accounting for the errors-in-covariates (Bernardinelli et al., 1996). In all of these examples, MCMC methods continue to enable ready computation of the required posterior distributions for D_θ and R_θ.

Data availability and format will lead to further research into appropriate use of the methodology. Cox and Piegorsch (1996) and Piegorsch and Cox (1996) note that environmental health studies can include data from several sources and of many types. Exposure data for risk-based (rather than proximity-based) assessments are generally gathered from point releases or sampling stations. Typically, one uses geostatistical methods to impute exposure between stations. Demographic variables such as income or race are available as regional data. Disease incidence data are available as regional counts, but death certificates may provide addresses (points) for deaths. As mentioned above, demographic and disease incidence data collected in the United States are typically regional counts rather than points due to confidentiality requirements. A proper synthesis of these data is central to valid and efficient assessments of environmental justice.

A particular problem is the compatibility of the demographic regional data and continuous exposure data. In the sections above, we assume that exposure data and demographic data are available for the same regions, but this may not be the case. Exposure data may be available for regions from an entirely different partition of the study area. For instance, water quality may

be defined for various watersheds that have little to do with political boundaries such as census regions. Also, dose-response effects may be related to particular spatial scales and aggregation at other scales may hide true relationships (Zimmerman, 1993). The integration of data collected on such "misaligned" regional systems has been explored in the geography literature by Flowerdew and Green (1992). An EM algorithm is used for such "areal interpolation" problems. We are currently extending our formulation above to address such incompatibilities.

Geographic information systems (GIS) provide powerful tools for integrating spatially referenced data over a given geographic area. Unfortunately, the statistical analysis capabilities of GIS lag behind capabilities for data management and presentation (Meyers, 1993). Appropriate methods for statistical analysis of mapped data are needed to realize the full potential of GIS as an analytic tool. Particular problems, such as the assessment of environmental justice, provide opportunities for collaboration between geographers, statisticians, epidemiologists, and environmental scientists. Tools like GIS provide the means to quickly provide maps illustrating differences, and appropriate statistical methods are needed to evaluate and quantitatively assess the magnitude of apparent differences.

References

Anderton, D.L., Anderson, A.B., Oakes, J.M., and Fraser, M.R. (1994) Environmental equity: the demographics of dumping. *Demography*, **31**, 229–248.

Bernardinelli, L., Pascutto, C., Best, N.G. and Gilks, W.R. (1996) Disease mapping with errors in covariates. To appear *Statistics in Medicine.*

Besag, J., York, J.C., and Mollié, A. (1991) Bayesian image restoration, with two applications in spatial statistics (with discussion). *Annals of the Institute of Statistical Mathematics*,

43, 1–59.

Bowen, W.M., Salling, M.J., Haynes, K.E., and Cyran, E.J. (1995) Toward environmental justice: Spatial equity in Ohio and Cleveland. *Annals of the Association of American Geographers*, **85**, 641–663.

Carlin, B.P. and Louis, T.A. (1996) *Bayes and Empirical Bayes Methods for Data Analysis*, London: Chapman and Hall.

Clayton, D.G. and Bernardinelli, L. (1992) Bayesian methods for mapping disease risk. In *Geographical and Environmental Epidemiology: Methods for Small-Area Studies*, P. Elliott, J. Cuzick, D. English, and R. Stern, eds. London: Oxford University Press.

Cox, L.H. and Piegorsch, W.W. (1996) Combining environmental information I: Environmental monitoring, measurement and assessment. *Environmetrics*, **7**, 299–308.

Escobar, M.D. and West, M. (1995) Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, **90**, 577–588

Flowerdew, R. and Green, M. (1992) Developments in areal interpolation methods and GIS. *Annals of Regional Science*, **26**, 67–78.

Gelman, A., Roberts, G.O., and Gilks, W.R. (1996) Efficient Metropolis jumping rules. In *Bayesian Statistics 5*, eds. J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith. Oxford: Oxford University Press, pp. 599–607.

Glickman, T.S. and Hersh, R. (1995) Evaluating environmental equity: the impacts of industrial hazards on selected social groups in Allegheny County, Pennsylvania. Discussion paper 95-13. Washington DC: Resources for the Future.

Hastings, W.K. (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97–109.

Horm, John W., Asire, A.J., Young, J.L. Jr., Pollack, E.S. (eds.) (1984) *SEER Program: Cancer Incidence and Mortality in the United States 1973-1981.* Bethesda, MD: National Institute of Health Publication No. 85-1837.

Meyers, D.E. (1993) Spatial statistics. In *Environmental Modeling with GIS*, M.F. Goodchild, B.O. Parks, and L.T. Steyaert, eds. New York: Oxford University Press.

Piegorsch, W.W. and Cox, L.H. (1996) Combining environmental information II: Environmental epidemiology and toxicology. *Environmetrics*, **7**, 309–324.

Sacks, J. and Steinberg, L.J. (1994) Environmental equity: statistical issues, report of a forum. National Institute of Statistical Sciences Technical Report No. 11.

Sexton, K. Olden, K., and Johnson, B.L. (1993) 'Environmental justice': the central role of research in establishing a credible scientific foundation for informed decision making. *Toxicology and Industrial Health*, **9**, 685–727.

Wagener, D.K. and Williams, D.R. (1993) Equity in environmental health: data collection and interpretation issues. *Toxicology and Industrial Health*, **9**, 775–795.

Waller, L.A., Carlin, B.P., Xia, H. and Gelfand, A. (1997) Hierarchical spatio-temporal mapping of disease rates. To appear *Journal of the American Statistical Association.*

Zimmerman, R. (1993) Social equity and environmental risk. *Risk Analysis*, **13**, 649–666.