May 2007

Ecologic Studies Revisited

Jon Wakefield
University of Washington, jonno@u.washington.edu

1 Introduction

Ecologic studies are characterized by being based on grouped data, with the groups often corresponding to geographical areas. Such studies have a long history in many disciplines including political science (39), geography (50), sociology (59) and epidemiology and public health (48). Here we concentrate on the latter and discuss why ecologic studies are widely-used, along with their unique drawbacks, namely the potential for *ecologic bias*, which describes the difference between ecologic and individual associations. Ecological data may be used for a variety of purposes including disease mapping (the geographical summarization of risk measures), and cluster detection (in which geographic anomalies are flagged); here we focus on geographical correlation studies in which the aim is to investigate associations between risk and exposure. In disease mapping ecologic bias is not a problem since prediction of area-level risk summaries is the objective, rather than the estimation of associations. Interesting within-area features may be masked by the process of aggregation, but although ecologic covariates may be used in disease mapping models to improve predictions, the coefficients are not of direct interest, (77) provides more discussion.

There are a number of reasons for the popularity of ecologic studies, the obvious

one being the wide and increasing availability of aggregated health and population data; exposure information is usually less readily available. If the exposure is an environmental pollutant, concentration information will rarely be aggregate in nature; it is more typical for measurements from a set of pollution monitors to be available. Nevertheless, we will still refer to such non-individual summaries as "ecologic". Improved ease of analysis also contributes to the widespread use of ecologic data. For example, geographical informations systems (GIS) allow the effective storage and combination of data sets from different sources and with differing geographies (13, 14, 47, 58, 61), and recent advances in statistical methodology allow a more refined analysis of ecologic data, references (20) and (79) contain reviews.

There are numerous examples of ecologic studies in the public health and epidemiology literature. For example, Figure 1, reproduced from (45) and using data from (42) displays stomach cancer mortality in 1991–1993 versus infant mortality in 1921–1923, each measured in 27 countries. The suggested hypothesis is that the association is due to stomach cancer risk being related to *H. pylori* infection, transmitted in the same way as diarrheal diseases that contributed to diseases that caused the observed childhood mortality rates. The interpretation of the apparent association is complicated due to the potential for ecological bias, however. Specifically, the 27 countries differ in many respects in addition to their rates of stomach cancer and infant mortality. The variables representing these differences may be related to both rates, and so the observed ecologic association may be due to confounding. The three highlighted countries, Japan, Russia and Chile, "... share very little in terms of their current socio-environmental conditions, and historically they are very different countries culturally, economically,

and socially", (45); the implication being that confounding is not responsible for the simultaneous high values of the two rates. But confounding is harder to characterize in ecologic studies, since it consists of both within-area and between-area components. For example, within each country there will be variability in infant mortality rates and this may covary with confounders, as discussed in Section 3.2. For motivation we briefly describe three additional ecological associations. Mortality rates for cervix cancer and the percentage pap test rate, both by state, are presented in Figure 3 of (61) as an example of an exploratory spatial analysis, and to illustrate the flexibility of a GIS. In the context of income inequality and health Figure 1 of (70) presents life expectancy versus income inequality in 11 countries; the correlation between income inequality and health is -0.81, but it is noted that, "... data from aggregate-level studies of the effect of income inequality on health ... are largely insufficient to discriminate between competing hypotheses", which makes the point that the loss of information in ecologic studies leads to a fundamental identifiability problem, with many scientifically interesting models being indistinguishable from the observed aggregate data. Finally, the two plots of Figure 5 of (44) show the percentage of individuals with forced vital capacity less than 85%, versus two measures of particulate matter ($< 2.1\ \mu m$ and $2.1 - 10\ \mu m$) for 22 US and Canadian communities. These plots are based on semi-ecological data in that individual-level data on outcome and confounders are supplemented with ecological exposure information. Such studies are less susceptible to ecological bias due to the increase in information when compared to a pure ecologic study, see Section 3.4.

The paucity of exposure data has recently lead a number of authors to combine ecologic population and health data with modeled exposure concentration sur-

faces; for a review of such modeling see (36). For example, Zidek and colleagues (81) examine the association between daily hospital admissions for respiratory disease and sulphate concentrations, while Carlin et al. (9) examine the relationship between pediatric asthma emergency room visits and ozone, the latter modeled using kriging within a GIS. In each of these examples great effort is placed on the modeling of the concentration surface without consideration of ecologic bias.

The structure of this review is to provide an illustrative ecologic study in Section 2, before cataloging a number of sources of ecologic bias in Section 3. Section 4 describes approaches to combining ecologic and individual data, and Section 5 provides concluding comments.

2 Illustrative Example: SIDS Risk in North Carolina

We examine data on sudden infant death syndrome (SIDS) and race; these data are available at the individual level, thus allowing the implications of aggregation to be examined. Mortality and birth data were obtained from the North Carolina State Center for Health Statistics website (`www.schs.state.nc.us/SCHS/`). SIDS cases were extracted for the years 2001–2004, by race for each of the 100 counties of North Carolina, along with the number of live births. There were a total of 386 cases and Figure 2(a) shows the distribution of risk across the 100 area. Race was categorized as white/non-white with 220 white deaths. There were 473,484 live births over the four years, with 343,811 being white. Figure 2(b) shows the proportion of non-white births; across the counties the proportion non-white live births ranges between 0.006 and 0.733 with median 0.222, so that in the majority of areas there are more white births than non-white births.

The mortality rates for non-whites and whites are 0.00128 and 0.00064, giving a relative risk of 2.0 with asymptotic 95% confidence interval (1.64,2.45).

We now assume that ecological data only are available. An ecologic dataset would consist of the proportion non-white, $\overline{x}$, along with the number of SIDS deaths, y, and the total births, n, in each area. The top map in Figure 3 displays the proportion non-white, with areas of relative high frequency in the north-east and south, though these are not reflected in the risk map in the bottom figure. A naive ecologic model is given by

$$\text{Ecologic Risk } = e^{\alpha^e + \beta^e \overline{x}} \tag{1}$$

and fitting this model gives the ecologic relative risk $e^{\beta^e} = 0.89$ (0.44–1.79) so that the risk decreases as the proportion non-white increases. The fitted curve is superimposed on the scatterplot of y versus x in Figure 2(c). If this point estimate was assumed to apply at the individual-level we would conclude that non-white babies are at lower risk than white babies, the opposite of that found in the individual-level analysis, thus providing an example of the ecological fallacy. The source of the fallacy will be returned to after we discuss sources of ecological bias.

3 Ecologic Bias

There is a vast literature describing sources of ecological bias, see for example, (25, 28, 29, 40, 48, 51, 52, 56, 57, 68, 74, 75, 77). The fundamental problem with ecological inference is that the process of aggregation reduces information, and this information loss usually prevents identification of association of interest in the underlying individual-level model. Ecologic bias is relative to a particular individual-level model. When trying to understand ecologic bias it is beneficial

to specify an individual-level model, and aggregate to determine the consequences (64, 75, 76). If there is no within-area variability in exposures and confounders, then there will be no ecological bias; so ecological bias occurs due to within-area variability in exposures and confounders; though there are a number of distinct consequences of this variability. Ecologic bias is also referred to as aggregate, or cross-level, bias, the latter emphasizing the differing levels of the data and inference. Throughout we assume that at the individual level the outcome is a 0/1 disease indicator.

3.1 Pure Specification Bias

So-called pure specification bias, (27) (also referred to as model specification bias, (64)) arises because a nonlinear risk model changes its form under aggregation. We initially assume a single exposure x and the individual-level model

$$\text{Individual Risk} = e^{\alpha+\beta x}, \tag{2}$$

which is often used for a rare disease; e^{α} is the risk associated with $x = 0$ (baseline risk) and e^{β} is the relative risk corresponding to an increase in x of one unit. We concentrate on this model but will also comment on linear forms. The logistic model, which is often used for non-rare outcomes, is unfortunately not amenable to analytical study and so the effects of aggregation are difficult to discern (63).

We consider a generic area containing n individuals with exposures x_i, $i = 1, ..., n$. Aggregation of (2) yields:

$$\text{Ecologic Risk} = \frac{1}{n}\sum_{i=1}^{n} e^{\alpha+\beta x_i} \tag{3}$$

so that the ecologic risk is the average of the risks of the constituent individuals. We let $\overline{x}$ represent the proportion of exposed individuals, i.e. $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. A

naive ecologic model would assume

$$\text{Ecologic Risk} = e^{\alpha^e + \beta^e \overline{x}} \tag{4}$$

where the ecologic parameters, α^e, β^e have been superscripted with "e" to distinguish them from the individual-level parameters in (2). Model (4) is a *contextual effects* model since risk depends on the proportion of exposed individuals in the area, see Section 3.3 for further discussion. Interpreting e^{β^e} as an individual association would correspond to a belief that it is average exposure that is causative, and that individual exposure is irrelevant. The difference between (3) and (4) is clear, while the former averages the risks across all exposures, the latter is the risk corresponding to the average exposure. We have $e^{\beta} = e^{\beta^e}$ only when there is no within-area variability in exposure so that $x_i = \overline{x}$ for all $i = 1, ..., n$ individuals. Hence pure specification bias is reduced in size as homogeneity of exposures within areas increases; hence small areas are advantageous. Unfortunately data aggregation is usually carried out according to administration groupings and not in order to obtain areas with constant exposure.

For a binary exposure (2) can be written

$$e^{\alpha+\beta x} = (1 - x)e^{\alpha} + xe^{\alpha+\beta}$$

which is linear in e^{α} and $e^{\alpha+\beta}$. This form simply yields the aggregate form:

$$\text{Ecologic Risk} = (1 - \overline{x})e^{\alpha} + \overline{x}e^{\alpha+\beta} \tag{5}$$

showing that with a linear risk model there is no pure specification bias. If model (4) is fitted using a binary proportion, $\overline{x}$, there will be no correspondence between e^{β} and e^{β^e} since they are associated with completely different comparisons. The extension to general categorical exposures is straightforward, and the parameters

of the disease model are identifiable so long as we have the aggregate proportions in each category.

For a continuous exposure pure specification bias is dominated by the relationship between the within-area mean and variance of the exposure and will be small if the within-area variability is unrelated to the mean; if the variance increases with the mean (which will often be the case for environmental exposures) then overestimation of a harmful exposure ($\beta > 0$) will occur (74). Unfortunately this condition is impossible to assess without individual-level data on the exposure. If β is close to zero pure specification bias is also likely to be small (since then the exponential model will be approximately linear for which there is no bias), though in this case confounding is likely to be a serious worry.

With respect to pure specification bias will result unless we have a categorical variable and we know the within-area proportions in each category, except when the exposure is constant within areas, or the risk model is linear. If the exposure is heterogeneous within areas we need information on the variability within-each area in order to control the bias. Such information may come from a sample of individuals within each area; how to use this individual-level data is the subject of Section 4.

Example Revisited

Returning to the North Carolina example, the discrepancy between the individual-level relative risk estimate of 2.0, and the ecologic association derived from model (1) of 0.89, is explained by pure specification bias; we fitted the contextual effects model (4), and not the aggregate form (5). Unfortunately fitting the latter model produces an estimate of 0.91 for these data, the reason for this discrepancy is that

model (5) is very unstable statistically and produces a likelihood surface that is highly irregular. In particular an asymptotic confidence interval is not appropriate here. This phenomenon has been observed elsewhere (32) which suggests great care should be taken in fitting model (5).

3.2 Confounding

We assume a single exposure x, a single confounder z, and the individual-level model

$$\text{Individual Risk} = e^{\alpha+\beta x+\gamma z} \tag{6}$$

As with pure specification bias, the key to understanding sources of, and correction for, ecological bias is to aggregate the individual-level model to give

$$\text{Ecologic Risk} = \frac{1}{n}\sum_{i=1}^{n} e^{\alpha+\beta x_i+\gamma z_i}. \tag{7}$$

	Female	Male	
Unexposed	p_{00}	p_{01}	$1-\overline{x}$
Exposed	p_{10}	p_{11}	$\overline{x}$
	$1-\overline{z}$	$\overline{z}$	1.0

Table 1: Exposure and gender distribution in a generic area, $\overline{x}$ is the proportion exposed and $\overline{z}$ is the proportion male; $p_{00}, p_{01}, p_{10}, p_{11}$ are the within-area cross-classification frequencies.

To understand why controlling for confounding is in general impossible with ecologic data we consider the simplest case of a binary exposure and a binary confounder, which for ease of explanation we assume is gender. Table 1 shows the distribution of the exposure and confounder within a generic area. The complete within-area distribution of exposure and confounder can be described by three

frequencies, but the ecologic data usually consist of the proportion exposed, $\overline{x}$, and the proportion male, $\overline{z}$, only. From (7) the aggregate form is

$$\begin{aligned} \text{Ecologic Risk} &= p_{00}e^{\alpha} + p_{10}e^{\alpha+\beta} + p_{01}e^{\alpha+\gamma} + p_{11}e^{\alpha+\beta+\gamma} \\ &= (1-\overline{x}-\overline{z}+p_{11})e^{\alpha} \\ &+ (\overline{x}-p_{11})e^{\alpha+\beta} + (\overline{z}-p_{11})e^{\alpha+\gamma} + p_{11}e^{\alpha+\beta+\gamma} \end{aligned} \tag{8}$$

showing that the marginal prevalances, $\overline{x}, \overline{z}$, alone, are not sufficient to characterise the joint distribution unless x and z are independent, in which case z is not a within-area confounder. This scenario has been considered in detail elsewhere (41), where it was argued that if the proportion of exposed males (p_{11}) is missing it should be estimated by the marginal prevalences ($\overline{x} \times \overline{z}$); it is not possible to determine the accuracy of this approximation without individual-level data, however. This is a recurring theme in the analysis of ecologic data, bias can be reduced under model assumptions, but estimation is crucially dependent on the appropriateness of these assumptions, which are uncheckable without individual-level data.

We now turn to the situation in which we have a binary exposure and a continuous confounder. Let the confounders in the unexposed be denoted, z_i, $i = 1, ..., n_0$, and the confounders in the exposed, z_i, $i = n_0 + 1, ..., n_0 + n_1$. In this case the ecologic form corresponding to (6) is

$$\text{Ecologic Risk} = q_0 \times r_0 + q_1 \times r_1$$

where $q_0 = n_0/n$ and $q_1 = n_1/n$ are the probabilities of being unexposed and exposed, and

$$r_0 = \frac{e^{\alpha}}{n_0}\sum_{i=1}^{n_0} e^{\gamma z_i}, \quad r_1 = \frac{e^{\alpha+\beta}}{n_1}\sum_{i=n_0+1}^{n_0+n_1} e^{\gamma z_i}$$

so that r_0 and r_1 are the aggregated risks in the unexposed and exposed. This makes clear that we need the confounder distribution within each exposure category, unless z is not a within-area confounder. The requirement for stratum-defined exposure distributions is closely related to mutual standardization as described in (60), which requires exposure distributions to be standardized with respect to a confounder, if risk has been standardized to this confounder. Again it is clear that if we fit the model:

$$\text{Ecologic Risk } = e^{\alpha^e + \beta^e \overline{x} + \gamma^e \overline{z}}$$

where $\overline{z} = \frac{1}{n}\sum_{i=1}^{n} z_i$, then the coefficient β has no relation to β^e in the naive ecologic model.

Often an attempt to control for confounding via expected numbers, E, using the regression model:

$$\text{Ecologic Risk } = E \times e^{\alpha^e + \beta^e \overline{x}},$$

(17, 18, 22). This approach implicitly assumes that there is no within-area confounding, however, (77). For example, the expected numbers are often calculated on the basis of the age and gender distribution, but this only controls for between-area confounding, and will only provide confounder control if the within-area exposure distribution is the same across age and gender stratum, and for age in particular this will be unlikely to hold. Whenever an ecologic study is considered the ability to control for known confounders for the disease/exposure under investigation should be considered. For most chronic diseases known lifestyle risk factors include one or more of smoking, alcohol, and diet. In an ecologic study individual-level information on these variables is not available and it has become popular to attempt to control for these variables using area-level mea-

sures of socio-economic status, e.g. (46). While these measures may be strongly correlated with lifestyle variables, (19), they cannot pick up the subtleties of within-area confounding and so unless the association of interest is strong, ecologic results controlled for confounding in this way should be interpreted with great caution.

The extension to general exposure and confounder scenarios is obvious from the above. If we have true confounders that are constant within areas (for example, access to health care) then they are analogous to conventional confounders, since the area is the unit of analysis, and so the implications are relatively easy to understand and adjustment is straightforward.

Without an interaction between exposure and confounder the parameters of a linear model are estimable from marginal information only, though if an interaction is present within-area information is required.

3.3 Contextual Effects

A contextual variable represents a characteristic of individuals in a shared neighborhood and in some scenarios (for example, the measurement of health disparities) such effects are of great interest. For example the mean income in an area, in addition to individual income, has been hypothesized as being predictive of health (37). We consider the simple individual-level linear model

$$E[Y_i|x_i, \overline{x}] = \alpha + \beta_W(x_i - \overline{x}) + \beta_B\overline{x} \tag{9}$$

where β_B is the between-area (contextual) effect, and β_W is the within-area individual effect. The aggregate form is

$$E[\overline{Y}|\overline{x}] = \alpha + \beta_B\overline{x},$$

showing that both individual and contextual effects cannot be simultaneously estimated without individual-level data. In a non-linear model both effects may be estimable with ecologic data, but the amount of information concerning β_W is small, (64) and, more importantly, the above derivation with the linear model reveals that estimation is crucially dependent on the form of the non-linear model, and the form of the model is not checkable from the ecologic data only. Hence, while sensitivity analyses to identify both parameters may be carried out, inference is totally unreliable with ecologic data only. It has also been pointed out, (26), that when contextual effects are of interest they are susceptible to cross-level bias when estimated from ecologic data.

It has been argued, in dietary and environmental contexts, that the contextual exposure $\overline{x}$ may be a better estimate of exposure for an individual than x_i when individual-level measurement error is large. For example, (49) propose a design that combines individual-level regression with ecologic comparisons in order to attempt to combine the best aspects of each data source; individual-level analyses are free of ecologic bias but may have poor power and measurement error in exposures, each of which may be rectified in ecologic data.

In general, multi-level models have provided a popular framework for analyzing associations at different geographical scales (for example, to estimate neighborhood effects), but these models cannot control for confounding due to unmeasured variables, and the interpretation of parameters is not always straightforward. The usual interpretation of a parameter associated with a particular variable is revealed by increasing the variable by one unit, while keeping all other variables fixed. Consideration of model (9) illustrates the difficulties in applying this approach in cases in which the variable appears at more than one level. Suppose we

wish to interpret β_W; if we increase x_i by one unit, the mean also increasees by $1/n$. To interpret β_W we must keep the mean in the area constant, for example by reducing everyone else's x by $1/(n-1)$. Further discussion may be found in (27), and interpretation of more complex models is provided in (2,67).

3.4 Semi-Ecologic Studies

Table 2 summarizes four distinct scenarios in terms of data availability, (40,64). In a semi-ecologic study, sometimes more optimistically referred to as a "semi-individual study", (40), individual-level data are collected on outcome and confounders, with exposure information arising from another source. The Harvard six-cities study, (16), provides an example in which the exposure was city-specific and an average of pollution monitors over the follow-up of the study.

		Exposure	
		Individual	Ecologic
Outcome	Individual	Individual	Semi-Ecologic
	Ecologic	Aggregate	Ecologic

Table 2: Study designs by level of outcome and exposure data.

We consider the risk for an individual in confounder stratum c; under aggregation we have

$$\text{Semi-Ecologic Risk in stratum } c \;=\; e^{\alpha+\gamma_c}\sum_{i=1}^{n_c} e^{\beta x_{ci}}$$

where x_{ci} are the exposures of individuals within stratum c, $i = 1, ..., n_c$, and γ_c is the baseline risk in stratum c. A naive semi-ecologic model is:

$$\text{Semi-Ecologic Risk in stratum } c = e^{\alpha^e+\gamma_c^e+\beta^e x} \qquad (10)$$

where x is some summary exposure measure. Kunzli and Tager (40) argue that semi-ecologic studies are free of ecologic bias, but there are two possible sources of bias here; the first is that we have pure specification bias because we have not acknowledged within-area variability in exposure, and the second is that we have not allowed the exposure to vary by confounder stratum so we have not controlled for within-area confounding. In an air pollution study in multiple cities x may correspond to a monitor average or an average over several monitors. In this case (10) will provide an approximately unbiased estimate of β if there is small exposure variability in cities and if this variability is similar across confounder stratum.

Semi-ecologic studies frequently have survival as an endpoint but there has been less focus on the implications of aggregation in the context of survival models, but (1, 33) discuss some of the implications.

3.5 Spatial Dependence and Hierarchical Modeling

When data are available as counts from a set of contiguous areas we might expect residual dependence in the counts, particularly for small-area studies, due to the presence of unmeasured variables with spatial structure. The use of the word "residual" here acknowledges that variables known to influence the outcome have already been adjusted for in the mean model. Analysis methods that ignore the dependence are strictly not applicable, with inappropriate standard errors being the most obvious manifestation. A great deal of work has focused on models for spatial dependence (3, 5, 10–12, 15, 38, 43); (55) provides an excellent review of this literature. With respect to ecological bias the most important message is that unless the mean model is correct, adjustment for spatial dependence is a

pointless exercise (77).

In a much-cited book (39) a hierarchical model was proposed for the analysis of ecologic data in a political science context, as "a solution to the ecological inference problem". Identifiability in this model is imposed through the random effects prior, however, and it is not possible to check the appropriateness of this prior from the ecological data alone (23, 75).

4 Combining Ecologic and Individual Data

As we saw in Section 3 the only solution to the ecologic inference problem that does not require uncheckable assumptions is the supplementation of ecologic-level with individual-level data. We stress that ecologic data can also supplement already available individual data, in order to improve power. Here we briefly review some of the proposals for such an endeavor. The obvious approach is to collect a random sample of individuals within areas. For a continuous outcome, Raghunathan et al. (54) show that moment and maximum likelihood estimates of a common within group correlation coefficient will improve when aggregate data are combined with individual data within groups, and Glynn et al. (24) derive optimal design strategies for the collection of individual-level data when the model is linear. With a binary non-rare outcome the benefits have also been illustrated (69, 75).

For a rare disease few cases will be present in the individuals within the sample, and so only information on the distribution of exposures and confounders will be obtained via a random sampling strategy (which is therefore equivalent to using a survey sample of covariates only). This prompted the derivation of the so-called aggregate data method of Prentice and Sheppard (53, 65, 66), Table 2. Inference

proceeds by constructing a model based on the sample of $m \leq n$ individuals in each area and estimates the mean (which is given by (3) for the case of a single exposure), based on the empirical averages. This is an extremely powerful design since estimation is not based on any assumptions with respect to the within-area distribution of exposures and confounders (though this distribution may not be well characterized for small samples, (62)). Ecologic bias is reduced to a greater extent than in the semi-ecologic study since within-area variability in exposures and confounders is acknowledged.

An alternative approach is to assume a parametric distribution for the within-area distribution of exposures and confounders, (57, 76) though this implicitly assumes that a sample of these is available; see also (34, 35). As an example, if we assume that exposures in an area are normally distributed with mean $\overline{x}$ and variance s^2 then the implied ecologic risk is $e^{\alpha+\beta\overline{x}+\beta^2 s^2/2}$, and this model may be fitted to ecologic data, if $\overline{x}$ and s^2 are available in each area, (4). More recently an approach has been suggested that takes the mean as a combination of these two approaches, with the parametric approach dominating for small samples (when the aggregate data method can provide unstable inference), (62).

A different approach in the context of a rare disease is outcome dependent sampling, which avoids the problems of zero cases encountered in random sampling. For the situation in which ecologic data are supplemented with individual case-control information gathered within the constituent areas, inferential approaches have been developed, (30–32). The case-control data remove ecologic bias while the ecologic data provide increased power and constraints on the sampling distribution of the case-control data, which improves the precision of estimates.

Two-phase methods have a long history in statistics and epidemiology (7,8,78,

80) and are based on an initial cross-classification by outcome and confounders and exposures; this classification providing a sampling frame within which additional covariates may be gathered via the sampling of individuals. Such a design may be used in an ecologic setting, where the initial classification is based on one or more of area, confounder stratum, and possibly error-prone measures of exposure, (72).

In all of these approaches it is clearly vital to avoid response bias in the survey samples, or selection bias in outcome-dependent sampling, and establishing a relevant sampling frame is essential.

5 Concluding Remarks

The use of ecological data are ubiquitous. This article has concentrated on area-aggregated data, but many other variables can be collapsed over. For example, it is common practice to collapse continuous age into 5-year age bands; this results in a loss of information, but within each age bands the changes in risk are small and so ecologic bias will be ignorable.

A sceptic might conclude from the litany of potential biases described in Section 3 that ecologic inference should never be attempted, but this would be too pessimistic a view. A useful starting point for all ecologic analyses is to write down an individual-level model for the outcome-exposure association of interest, including known confounders. Ecologic bias will be small when within-area variability in exposures and known confounders is small, and for small-area studies in particular this may be approximately true. A serious source of bias is that due to confounding, since ecologic data on exposure are rarely stratified by confounder strata within areas. If a small area study has been carried out with a correctly

aggregated individual-level model, then parameter estimates can be cautiously interpreted at the individual-level and compared with other studies at the individual level, and hence add to the totality of evidence for a hypothesis. When comparing ecologic and semi-ecologic estimates with individual-level estimates it is clearly crucial to have a common effect measure (e.g. a relative risk or a hazard ratio). So, for example, it will be difficult to compare an ecologic correlation coefficient, which is a measure that is often reported, with an effect estimate from an individual-level study.

Less well-designed ecologic studies can be suggestive of hypotheses to investigate, if strong ecologic associations are observed. An alternative to the pessimistic outlook expressed above is that when a strong ecological association, such as that observed in Figure 1, is seen an attempt should be made to explain how such a relationship could have arisen, if it is not due to the ecologic predictor.

There are a number of issues that we have not discussed. Care should be taken in determining the effects of measurement error in an ecologic study since the directions of bias may not be predictable. For example, in the absence of pure specification and confounder bias for linear and log-linear models, if there is non-differential measurement error in a binary exposure there will be overestimation of the effect parameter, in contrast to individual-level studies, (6). We refer interested readers to alternative sources, (21, 71), for other issues such as consideration of migration, latency periods, and the likely impacts of inaccuracies in population and health data.

Studies that investigate the acute effects of air pollution are another common situation in which ecologic exposures are used. For example, daily disease counts in a city are often regressed against daily and/or lagged concentration measure-

ments taken from a monitor, or the average of a collection of monitors to estimate the acute effects of air pollution. If day-to-day exposure variability is greater than within-city variability then we would expect ecologic bias to be relatively small.

With respect to data availability, exposure information is generally not aggregate in nature (unless the "exposure" is a demographic or socio-economic variable), and in an environmental epidemiological setting the modeling of pollutant concentration surfaces will undoubtedly grow in popularity. However, an important insight is that in a health-exposure modeling context it may be better to use measurements from the nearest monitor, rather than model the concentration surface, since the latter approach may be susceptible to large biases, particularly when, as is usually the case, the monitoring network is sparse(73). A remaining challenge is to diagnose when the available data are of sufficient abundance and quality to support the use of complex models.

In Section 4 we described a number of proposals for the combination of ecologic and individual data. Such endeavors will no doubt increase and will hopefully allow the reliable exploitation of ecologic information.

Summary Points

1. Ecologic bias, defined as the difference between associations obtained from individual and ecologic data, occurs because of within-group variability in exposures and/or confounders.
2. To understand the implications of the use of ecologic data in any setting it is useful to first write down the individual-level model that would be fitted it individual-level data were available. Aggregation of an individual-level model allows the characterization of ecologic bias and reveals the individual-

level data that would reduce the chance of ecologic bias.

3. Ecologic bias can only be safely removed by combining ecologic- and individual-level data.

4. Semi-ecologic studies are less susceptible to ecologic bias, since some components of bias are not possible, but again the implications of aggregation should be carefully examined.

Mini Glossary

1. *Ecological bias:* The difference between associations at the individual and ecologic level.

2. *Ecological fallacy:* The result of ecologic bias in which incorrect individual-level inference is drawn from ecologic data.

3. *Pure (or model) specification bias:* Non-linear individual models do not retain the same form under aggregation (unless there is no within-area variability in exposure) and so using an ecologic model that is of the same form as the individual-level model will lead to bias.

4. *Semi-ecologic studies:* Studies in which individual-level data is available on outcome and confounders, with an ecologic exposure assessment.

Acknowledgments

This work was supported by grant R01 CA095994 from the National Institutes of Health.

LITERATURE CITED

1. M. Abrahamowicz, R. du Berger, D. Krewski, R. Burnett, G. Bartlett, R.M. Tamblyn, and K. Leffondré. Bias due to aggregation of individual covariates in the cox regression model. *American Journal of Epidemiology*, 160:696–706, 2004.
2. K. Berhane, W.J. Gauderman, D.O. Stram, and D.C. Thomas. Statistical issues in studies of the long-term effects of air pollution: The southern california children's health study. *Statistical Science*, 19:414–434, 2004.
3. J. Besag, J. York, and A. Mollié. Bayesian image restoration with two applications in spatial statistics. *Annals of the Institute of Statistics and Mathematics*, 43:1–59, 1991.
4. N. Best, S. Cockings, J. Bennett, J. Wakefield, and P. Elliott. Ecological regression analysis of environmental benzene exposure and childhood leukaemia: sensitivity to data inaccuracies, geographical scale and ecological bias. *Journal of the Royal Statistical Society, Series A*, 164:155–174, 2001.
5. N.G. Best, K. Ickstadt, and R.L. Wolpert. Ecological modelling of health and exposure data measured at disparate spatial scales. *Journal of the American Statistical Association*, 95:1076–1088, 2000.
6. H. Brenner, D. Savitz, K.-H. Jockel, and S. Greenland. Effects of nondifferential exposure misclassification in ecologic studies. *American Journal of Epidemiology*, 135:85–95, 1992.
7. N. E. Breslow and K. C. Cain. Logistic regression for two-stage case-control data. *Biometrika*, 75:11–20, 1988.
8. N.E. Breslow and N. Chatterjee. Design and analysis of two-phase studies with binary outcome applied to wilms tumour prognosis. *Applied Statistics*,

48:457–468, 1999.

9. B.P. Carlin, H. Xia, O. Devine, P. Tolbert, and J. Mulholland. Spatio-temporal hierarchical models for analyzing atlanta pediatric asthma er visit rates. In C. Gatsonis, R.E. Kass, B. Carlin, A. Carriquiry, A. Gelman, I. Verdinelli, and M. West, editors, *Case Studies in Bayesian Statistics, Volume IV*, pages 303–320, New York, 1999. Springer.
10. O.F. Christensen and R. Waagepetersen. Bayesian prediction of spatial count data using generalised linear mixed models. *Biometrics*, 58:280–286, 2002.
11. D. Clayton, L. Bernardinelli, and C. Montomoli. Spatial correlation in ecological analysis. *International Journal of Epidemiology*, 22:1193–1202, 1993.
12. N. Cressie and N.H. Chan. Spatial modelling of regional variables. *Journal of the American Statistical Association*, 84:393–401, 1989.
13. E.K. Cromley. GIS and disease. *Annual Review of Public Health*, 24:7–24, 2003.
14. C.M. Croner. Public health, GIS and the internet. *Annual Review of Public Health*, 24:57–82, 2003.
15. P.J. Diggle, J.A. Tawn, and R.A. Moyeed. Model-based geostatistics (with discussion). *Applied Statistics*, 47:299–350, 1998.
16. D. Dockery, CA. III Pope, X. Xiping, J. Spengler, J. Ware, M. Fay, B. Ferris, and F. Speizer. An association between air pollution and mortality in six U.S. cities. *N Engl J Med*, 329:1753–9, 1993.
17. N. Eaton, G. Shaddick, H. Dolk, and P. Elliott. Small-area study of the incidence of neoplasms of the brain and central nervous system among adults in the West Midlands. *British Journal of Cancer*, 75:1080–1083, 1997.
18. P. Elliott, G. Shaddick, I. Kleinschmidt, D. Jolley, P. Walls, J. Beresford,

and C. Grundy. Cancer incidence near municipal solid waste incinerators in Great Britain. *British Journal of Cancer*, 73:702–707, 1996.

19. P. Elliott and J. C. Wakefield. Bias and confounding in spatial epidemiology. In P. Elliott, J. C. Wakefield, N. G. Best, and D. Briggs, editors, *Spatial Epidemiology: Methods and Applications*, pages 68–84. Oxford University Press, Oxford, 2000.

20. P. Elliott, J. C. Wakefield, N. G. Best, and D. J. Briggs. *Spatial Epidemiology: Methods and Applications.* Oxford University Press, Oxford, 2000.

21. P. Elliott and J.C. Wakefield. Small-area studies of environment and health. In Barnett V., Stein A., and Turkman K.F, editors, *Statistics for the Environment 4: Health and the Environment*, pages 3–27. John Wiley, New York., 1999.

22. P. Elliott, A.J. Westlake, I. Kleinschmidt, M. Hills, L. Rodrigues, P. McCabe, K. Marshall, and C. Rose. The small area health statistics unit: a national facility for investigating health around point sources of environmental pollution in the united kingdom. *Journal of Epidemiology and Community Health*, 46:345–9, 1992.

23. D. A. Freedman, S. P. Klein, M. Ostland, and M. R. Roberts. A solution to the ecological inference problem (book review). *Journal of the American Statistical Association*, 93:1518–1522, 1998.

24. A. Glynn, J. Wakefield, M. Handcock, and T. Richardson. Alleviating linear ecological bias and optimal design with subsample data. *Journal of the Royal Statistical Society, Series A*, 2005. To appear.

25. S. Greenland. Divergent biases in ecologic and individual level studies. *Statistics in Medicine*, 11:1209–1223, 1992.

26. S. Greenland. Ecologic versus individual-level sources of bias in ecologic estimates of contextual health effects. *International Journal of Epidemiology*, 30:1343–1350, 2001.

27. S. Greenland. A review of multilevel theory for ecologic analyses. *Statistics in Medicine*, 21:389–95, 2002.

28. S. Greenland and H. Morgenstern. Ecological bias, confounding and effect modification. *International Journal of Epidemiology*, 18:269–274, 1989.

29. S. Greenland and J. Robins. Ecological studies: biases, misconceptions and counterexamples. *American Journal of Epidemiology*, 139:747–760, 1994.

30. S. Haneuse and J. Wakefied. Hierarchical models for combining ecological and case-control data. *Biometrics*, 63:128–136, 2007.

31. S. Haneuse and J. Wakefield. The combination of ecological and case-control data. Under revision, 2005.

32. S. Haneuse and J. Wakefield. Geographic-based ecological correlation studies using supplemental case-control data. *Statistics in Medicine*, 2007. Under revision.

33. S. Haneuse, J. Wakefield, and L. Sheppard. The interpretation of exposure effect estimates in chronic air pollution studies. *Statistics in Medicine*, 2007. To appear.

34. C. Jackson, N. Best, and S. Richardson. Hierarchical related regression for combining aggregate and individual data in studies of socio-economic disease risk factors. *Journal of the Royal Statistical Society, Series A*, 2007. To appear.

35. C.H. Jackson, N.G. Best, and S. Richardson. Improving ecological inference using individual-level data. *Statistics in Medicine*, 25:2136–2159, 2006.

36. M. Jerrett, A. Afrain, P. Kanaroglou, B. Beckerman, D. Potoglou, T. Sahsuvaroglu, J. Morrison, and C. Giovis. A review and evaluation of intraurban air pollution exposure models. *Journal of Exposure Analysis and Envirnomental Epidemiology*, 15:185–204, 2005.

37. K. Judge, J. Mulligan, and M. Benzeval. Income inequality and population health. *Social Science and Medicine*, 46:565–579, 1998.

38. J.E. Kelsall and J.C. Wakefield. Modeling spatial variation in disease risk: a geostatistical approach. *Journal of the American Statistical Association*, 97:692–701, 2002.

39. G. King. *A Solution to the Ecological Inference Problem.* Princeton University Press, Princeton, 1997.

40. N. Künzli and I.B. Tager. The semi-individual study in air pollution epidemiology: a valid design as compared to ecologic studies. *Environmental Health Perspectives*, 10:1078–1083, 1997.

41. V. Lasserre, C. Guihenneuc-Jouyaux, and S. Richardson. Biases in ecological studies: utility of including within-area distribution of confounders. *Statistics in Medicine*, 19:45–59, 2000.

42. D.A. Leon and G.D. Smith. Infant mortality, stomach cancer, stroke, and coronary heart disease: ecological analysis. *British Medical Journal*, 320:1705–1706, 2000.

43. B.G. Leroux, X. Lei, and N. Breslow. Estimation of disease rates in small areas: A new mixed model for spatial dependence. In M.E. Halloran and D.A Berry, editors, *Statistical Models in Epidemiology. the Environment and Clinical Trials*, pages 179–192. Springer, New York, 1999.

44. M. Lippmann and R.B. Schlesinger. Toxicological bases for the setting of

health-related air pollution standards. *Annual Review of Public Health*, 21:309–333, 2000.

45. J. Lynch and G. Davey Smith. A life course approach to chronic disease epidemiology. *Annual Review of Public Health*, 26:1–35, 2005.

46. R. Maheswaran, S. Morris, S. Falconer, A. Grossinho, I. Perry, J. Wakefield, and P. Elliott. Magnesium in drinking water supplies and mortality from acute myocardial infaction in North West England. *Heart*, 82:455–60, 1999.

47. S.L. McLafferty. GIS and health care. *Annual Review of Public Health*, 24:25–42, 2003.

48. H. Morgenstern. Ecologic study. In P. Armitage and T. Colton, editors, *Encyclopedia of Biostatistics Vol. 2*, pages 1255–76. John Wiley and Sons, New York, 1998.

49. W. Navidi, D. Thomas, D. Stram, and J. Peters. Design and analysis of multilevel analytic studies with applications to a study of air-pollution. *Environmental Health Perspectives*, 102, Suppl. 8:25–32, 1994.

50. S. Openshaw. *The Modifiable Areal Unit Problem.* CATMOG No. 38, Geo Books, Norwich., 1984.

51. S. Piantadosi, D.P. Byar, and S.B. Green. The ecological fallacy. *American Journal of Epidemiology*, 127:893–904, 1988.

52. M. Plummer and D. Clayton. Estimation of population exposure. *Journal of the Royal Statistical Society, Series B*, 58:113–126, 1996.

53. R.L. Prentice and L. Sheppard. Aggregate data studies of disease risk factors. *Biometrika*, 82:113–25, 1995.

54. T.E. Raghunathan, P.K. Diehr, and A.D. Cheadle. Combining aggregate and individual level data to estimate an individual level correlation coefficient.

Journal of Educational and Behavioral Statistics, 28:1–19, 2003.

55. S. Richardson. Spatial models in epidemiological applications. In P.J. Green, N.L. Hjort, and S. Richardson, editors, *Highly Structured Stochastic Systems*, pages 237–259. Oxford Statistical Science Series, Oxford, 2003.

56. S. Richardson and C. Montfort. Ecological correlation studies. In P. Elliott, J. C. Wakefield, N. G. Best, and D. Briggs, editors, *Spatial Epidemiology: Methods and Applications*, pages 205–220. Oxford University Press, Oxford, 2000.

57. S. Richardson, I. Stucker, and D. Hémon. Comparison of relative risks obtained in ecological and individual studies: some methodological considerations. *International Journal of Epidemiology*, 16:111–20, 1987.

58. T.C. Ricketts. Geographic information systems and public health. *Annual Review of Public Health*, 24:1–6, 2003.

59. W. S. Robinson. Ecological correlations and the behavior of individuals. *American Sociological Review*, 15:351–57, 1950.

60. P.R. Rosenbaum and D.B. Rubin. Difficulties with regression analyses of age-adjusted rates. *Biometrics*, 40:437–443, 1984.

61. G. Rushton. Public health, GIS, and spatial analytic tools. *Annual Review of Public Health*, 24:43–56, 2003.

62. R. Salway and J. Wakefield. A hybrid model for reducing ecological bias. *Biostatistics*, 2007. To appear.

63. R.A. Salway and J.C. Wakefield. Sources of bias in ecological studies of non-rare events. Submitted for publication, 2004.

64. L. Sheppard. Insights on bias and information in group-level studies. *Biostatistics*, 4:265–278, 2003.

65. L. Sheppard, R. L. Prentice, and M. A. Rossing. Design considerations for estimation of exposure effects on disease risk, using aggregate data studies. *Statistics in Medicine*, 15:1849–1858, 1996.

66. L. Sheppard and R.L. Prentice. On the reliability and precision of within- and between-population estimates of relative rate parameters. *Biometrics*, 51:853–863, 1995.

67. L. Sheppard and J. Wakefield. Discussion of: Statistical issues in studies of the long-term effects of air pollution: The southern california children's health study. *Statistical Science*, 19:438–441, 2004.

68. D. G. Steel and D. Holt. Analysing and adjusting aggregation effects: The ecological fallacy revisited. *International Statistical Review*, 64:39–60, 1996.

69. D.G. Steele, E.J. Beh, and R.L. Chambers. The information in aggregate data. In G. King, O. Rosen, and M. Tanner, editors, *Ecological Inference: New Methodological Strategies*. Cambridge University Press, Cambridge, 2004.

70. A. Wagstaff and E. van Doorslaer. Income inequality and health: what does the literature tell us? *Annual Review of Public Health*, 21:543–567, 2000.

71. J. Wakefield and P. Elliott. Issues in the statistical analysis of small area health data. *Statistics in Medicine*, 18:2377–2399, 1999.

72. J. Wakefield and S. Haneuse. Ecological two-phase studies. *Submitted*, 2007.

73. J. Wakefield and G. Shaddick. Health-exposure modelling and the ecological fallacy. *Biostatistics*, 7:438–455, 2006.

74. J. C. Wakefield. Sensitivity analyses for ecological regression. *Biometrics*, 59:9–17, 2003.

75. J. C. Wakefield. Ecological inference for 2×2 tables (with discussion).

Journal of the Royal Statistical Society, Series A, 167:385–445, 2004.

76. J. C. Wakefield and R. E. Salway. A statistical framework for ecological and aggregate studies. *Journal of the Royal Statistical Society, Series A*, 164:119–137, 2001.

77. J.C. Wakefield. Disease mapping and spatial regression with count data. *Biostatistics*, 8:158–183, 2007.

78. A.M. Walker. Anamorphic analysis: sampling and estimation for covariate effects when both exposure and disease are known. *Biometrics*, 38:1025–1032, 1982.

79. L.A. Waller and C.A. Gotway. *Applied Spatial Statistics for Pubic Health Data.* Wiley, 2004.

80. J.E. White. A two stage design for the study of the relationship between a rare exposure and a rare disease. *American Journal of Epidemiology*, 115:119–128, 1982.

81. J.V. Zidek, R. White, N.D. Lee, W. Sun, and R.T. Burnett. Imputing unmeasured explanatory variables in environmental epidemiology with application to health impact analysis of air pollution. *Environmental and Ecological Statistics*, 5:99–115, 1998.

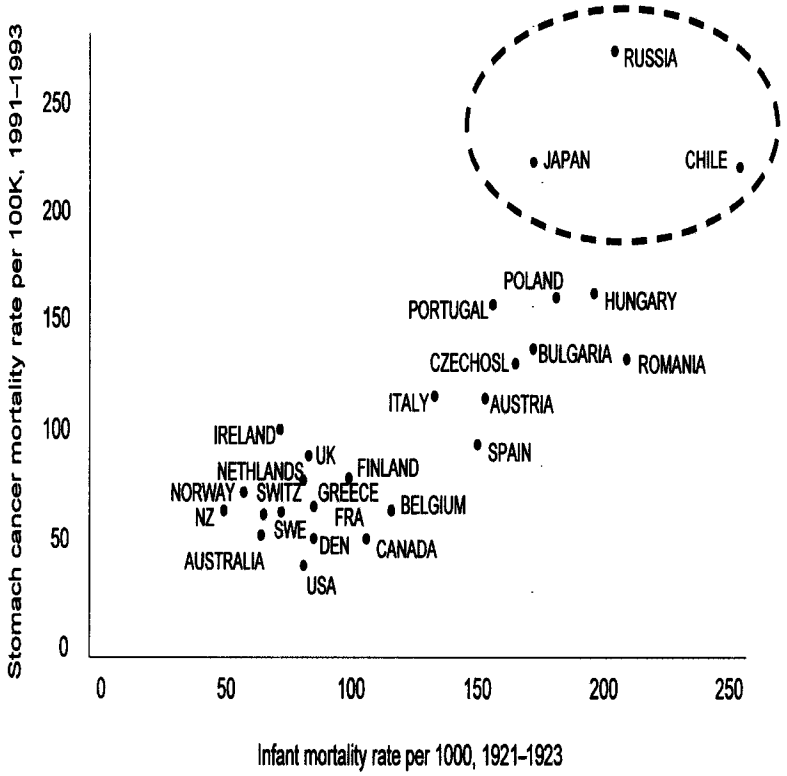
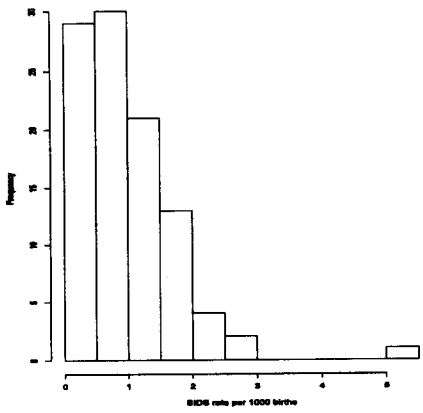
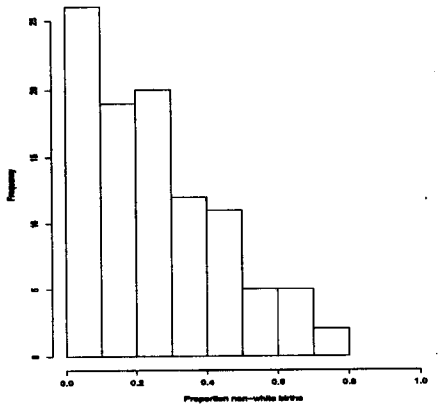


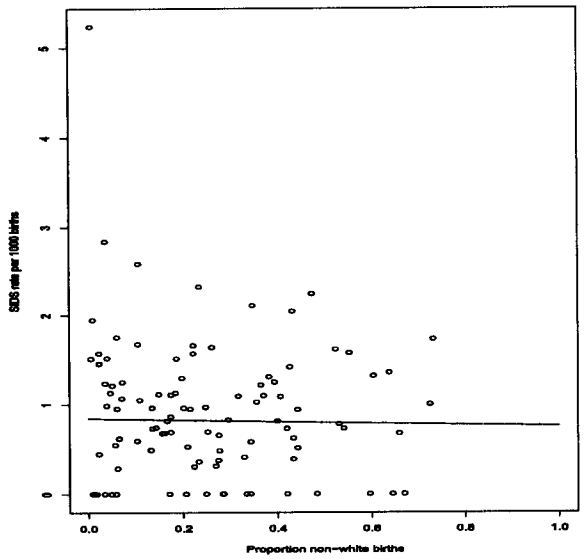
Figure 1: Stomach cancer mortality in 1991–1993 versus infant mortality rate in 1921–1923 in 27 countries. Reprinted, with permission, from the Annual Review of Public Health, Volume 26 (c)2005 by Annual Reviews www.annualreviews.org.



(a) SIDS risk (× 1000)



(b) Non-white proportion



(c) Risk versus proportion non-white

Figure 2: Proportion non-white births and risk of SIDS (×1000) across 100 counties of North Carolina, in the years 2001–2004.

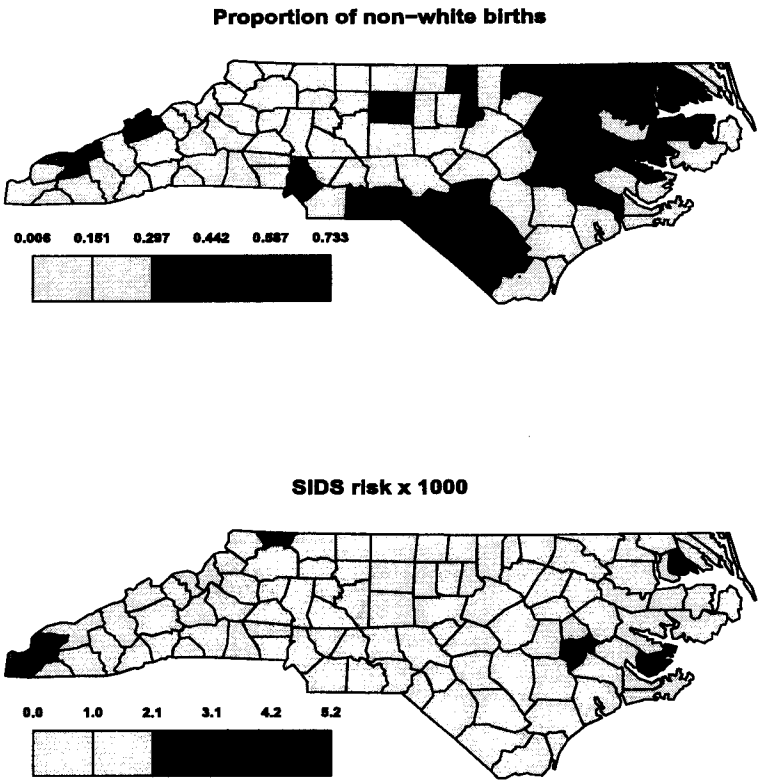


Figure 3: Maps of proportion non-white and risk across 100 counties of North Carolina.