Alleviating Ecological Bias in Poisson Models using Optimal Subsampling: The Effects of Jim Crow on Black Illiteracy in the Robinson Data

Adam N. Glynn[*] Jon Wakefield[†]

April 3, 2013

Summary

In many situations data are available at the group level but one wishes to estimate the individual-level association between a response and an explanatory variable. Unfortunately this endeavor is fraught with difficulties because of the ecological level of the data. The only reliable solution to such ecological inference problems is to supplement the ecological data with individual-level data. In this paper we illustrate the benefits of gathering individual-level data in the context of a Poisson modeling framework. Additionally, we derive optimal designs that allow the individual samples to be chosen so that information is maximized. The methods are illustrated using Robinson's classic data on illiteracy rates. We show that the optimal design produces accurate inference with respect to estimation of relative risks, with ecological bias removed.

Keywords: Ecological bias; Combining information; Sample design.

[*] Department of Government, Harvard University, 1737 Cambridge Street, Cambridge, MA 02138. aglynn@fas.harvard.edu

[†] Department of Statistics and Department of Biostatistics, University of Washington, Seattle, WA 98195.

1 Introduction

Ecological inference, the attempt to make inference about individuals with aggregate data, is well known to be problematic with biased estimates being the usual consequence of relying on aggregate data. The description of the problem has a long history with an early influential paper being that of Robinson (1950) and there is now a vast literature on characterization of the different forms of bias (Piantadosi et al., 1988; Greenland and Morgenstern, 1989; Greenland, 1992; Greenland and Robins, 1994; Richardson, 1992; Wakefield and Salway, 2001). The only solution to the ecological inference problem is to obtain individual level data, with one possible route of analysis being to make inference based on this data alone. However, it is inefficient to ignore the ecological data during analysis, since such data are usually aggregated from a large number of individuals and therefore provide a great deal of information, though not directly on the parameters of interest. Furthermore, the ecological data may inform the sampling of individual level data, effectively reducing the cost of data collection. In this paper, we consider these issues in the context of a Poisson modeling framework which is widely applicable for a rare binary response; many binary disease outcomes may be modeled within this framework.

We specifically address the following questions:

- How can we correct ecological bias by combining ecological data and subsample data?
- If the ecological data are readily available, as is frequently the case, can we design an optimal subsample of individual level data in order to maximize the information about parameters?

In its most inclusive definition, ecological inference is usually an attempt to estimate individual level parameters with data that have been aggregated above the individual level (to give ecological data). Not surprisingly, this is a difficult inference problem and within the research community two extreme positions have often been taken: those who disdain to use any ecological inference and advocate inference based on the sampling of individuals (Freedman et al., 1998), and those who attempt ecological inference through model assumptions or a description of the possible biases (King, 1997). With respect to this latter stance, many models have been proposed for inference (Goodman, 1953; King, 1997; King et al., 2004; Wakefield, 2004) though it has been recognized by some authors that the assumptions required for valid ecological inference are not checkable from the available ecological data (Freedman et al., 1998; Wakefield, 2004). A position between these extremes is to recognize that individual-level data is required to unlock the ecological information and a number of different approaches have been suggested (Prentice and Sheppard, 1995; Haneuse and Wakefield, 2008; Jackson et al., 2006; Glynn et al., 2008; Jackson et al., 2008; Wakefield and Haneuse, 2008). In this paper, we discuss how samples may be optimally drawn when the estimation of within-area risks is the objective. We begin by describing a motivating example.

2 Ecological Bias in the Estimation of the Effect of Jim Crow Laws

We consider the canonical example of ecological bias in the data on black illiteracy rates in the US in the 1930s. More specifically, we consider the effect of Jim Crow laws on black illiteracy by using the original data from Robinson (1950). These data contain a binary illiteracy indicator and the race/nativity (coded as foreign-born white, black, native-born white), at the level of the individual, across the United States, along with the presence/absence, in each state, of Jim Crow segregation laws for education. In the original paper, Robinson demonstrated the very different correlations that result depending on the level of spatial aggregation of the data. The data have generated a great deal of interest, including the recent set of articles in the International Journal of Epidemiology (Wakefield, 2009; Subramanian et al., 2009b; Oakes, 2009; Firebaugh, 2009; Subramanian et al., 2009a).

The estimation of the association between illiteracy and Jim Crow laws is an example of an association at the level of the group. Such associations are often of interest, with a particular example being a *contextual effect*. It has been shown (Greenland, 2001) that the ecological inference problem persists even when the goal is to estimate the association between an outcome and a *contextual variable*. A contextual variable represents a characteristic of individuals in a shared neighborhood or group and the estimation of the associations with multiple levels of variable is important in many disciplines including epidemiology (Greenland, 2001), public health (Diez-Roux, 1998) and sociology (Blalock, 1984).

We first observe that, even with the full individual level data, estimating the effects of Jim Crow laws on black illiteracy is difficult because the Jim Crow states tend to be quite different from the non-Jim Crow states. For example, the Jim Crow states had an average black population proportion of 20%, while the average black population proportion among the non-Jim Crow states was only 1.5%. Similarly, the average population proportion of foreign-born whites was quite different between the Jim Crow and non-Jim Crow states (3.4% and 17%, respectively). These discrepancies could prove problematic for inference about the effect of Jim Crow laws, because they may reveal important differences between the states that would have existed even if there had been an earlier attempt by the federal government to reign-in these laws. As in the paper of Oakes (Oakes, 2009), we will attempt to minimize these observed discrepancies (and also unobserved discrepancies) by limiting the Jim Crow states under analysis to those that were not part of the Confederacy during the Civil War: Kansas and Wyoming.

In order to estimate the effects of Jim Crow laws in Kansas and Wyoming, we match these states to non-Jim Crow states on the basis of the sizes of the black population and of the foreign-born white population. On the basis of these variables, Indiana is the only reasonable match for Kansas. Both Kansas and Indiana have a black population of 3.5%, and foreign-born white percentages of 4.6% and 4.9%, respectively. All other non-Jim Crow states with black populations of around 3.5% have much higher percentages of foreign-born whites (e.g., Michigan was 20.9% foreign-born white). Additionally, Nevada is the best match for Wyoming. Both states have a black population of 0.6%, and foreign-born white percentages of 15.6% and 13.1%, respectively. Nebraska and Colorado would also be reasonable matches for Wyoming, having foreign-born whites at 10.8% and 10.4%, respectively, but they both have slightly higher black populations (1.0% and 1.1%, respectively). This restriction to the data from four states only limits the generality of the results but maximizes the internal validity of the analysis. Also, to the extent that we expect the effects of Jim Crow laws to be larger in the Confederate states, an analysis of Kansas and Wyoming may provide a lower bound on the effect for those states. Appendix A contains the full individual-level data for these

In an ecological analysis only the state-level illiteracy rates would be available, and not the break-down by race. Figure 1(a) presents the results of a straightforward empirical analysis using only the ecological data on illiteracy rates (without the racial breakdown). In the figure we plot the log of the state illiteracy rate on the vertical axis, with the Jim Crow state indicator on the horizontal axis. From Appendix A these rates are 1.5% and 1.3% for Indiana and Nevada (the states without Jim Crow laws) and 0.9% and 0.8% for Kansas and Wyoming (the states with Jim Crow laws). Because we have matched the Indiana/Kansas pair and the Nevada/Wyoming pair on percent black and percent foreign-born white (the ecological covariate data), we simply compare state (log) illiteracy rates within the pairs (model based estimates would be very similar because of the matching). Surprisingly, Jim Crow laws appear to *decrease* illiteracy rates for both the Indiana/Kansas pair and the Nevada/Wyoming pair.

We now move to an analysis with the individual-level data. With the individual-level data, the analyst can use black illiteracy rates as the outcome and can use the native-born white illiteracy rates as a baseline of the overall level of illiteracy in the state. An intuitive approach that simplifies the presentation involves taking the ratio of the black and native-born white illiteracy rates as the outcome. Model-based approaches that control for native born white illiteracy will arrive at qualitatively similar conclusions. Figure 1(b) presents this analysis, and the relationship between Jim Crow laws and black illiteracy is reversed in comparison with Figure 1(a). In this, more reliable, individual-level analysis, Jim Crow laws appear to increase black illiteracy rates in both the Indiana/Kansas and Nevada/Wyoming pairs (the sample sizes are extremely large here, so a discussion of standard errors is unnecessary). Examination of the raw data in Appendix A clarifies why we have an example of the ecological fallacy in this example. In the states with Jim Crow laws the illiteracy rates of the three races are smaller than the illiteracy rates in the matching states without Jim Crow laws in five out of six cases (the only comparison for which this is not true is for blacks in Wyoming, in which the illiteracy rate is 4.2% as compared to 1.5% in Nevada). The state illiteracy rates are dominated by the white illiteracy rates and these are much greater in the states without Jim Crow laws. To summarize: the aggregation has obscured the within-state information on race-specific illiteracy rates.

We conclude that attempts to estimate group-level associations using only ecological data can produce biased results, agreeing with previous discussions (Greenland, 2001). Unfortunately, while using only the ecological data can lead to the wrong answer, using only individual-level data is often costly. In many cases, the full individual-level information is unavailable to the analyst, and therefore sampling will be required. However, in order to achieve reasonable margins of error for the state-level estimates produced in Figure 1(b), we would need to sample large numbers of individuals (especially if the sample is not stratified on racial category). Fortunately, as we will demonstrate in the remainder of this paper, such a large sample is unnecessary because 1) the ecological data can be combined with the sampled individual data to produce more precise estimates and 2) the existence of the ecological data allows efficient sampling designs to be constructed.

3 Combined Inference and Optimal Subsample Design

Within each generic ecological unit (e.g., state), we denote the individual binary outcome as y_j (e.g., whether or not each individual is illiterate), and we denote the covariate vector for each individual as $\boldsymbol{x}_j$, for $j = 1, ..., n$ individuals. For example, x_j could represent the pair of indicators $[x_{1j}, x_{2j}]$, where $x_{1j} = 1$ if individual j is black and $x_{2j} = 1$ if individual j is foreign-born white. With this notation, the ecological data (e.g., the state illiteracy total in a generic area) can be written as $y_+ = \sum_{j=1}^n y_j$, where n is the number of individuals in the ecological unit. The *sampled* individual level data, with sample size k, can be written as $\{\boldsymbol{y}^{(s)}, \boldsymbol{x}^{(s)}\}$, so that $\boldsymbol{y}^{(s)} = (y_1, ..., y_k)$ and $\boldsymbol{x}^{(s)} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_k)$ for $j = 1, ..., k$. So note that we have reserved the first k indices for the sampled individuals. Similarly, the unsampled individual level data, with sample size $n - k$, can be written as $\{\boldsymbol{y}^{(-s)}, \boldsymbol{x}^{(-s)}\}$ with $\boldsymbol{y}^{(-s)} = (y_{k+1}, ..., y_n)$ and $\boldsymbol{x}^{(-s)} = (\boldsymbol{x}_{k+1}, ..., \boldsymbol{x}_n)$. Furthermore, it is straightforward to derive the ecological total for the *unsampled* individuals by subtracting from the overall total, i.e. $y_+^{(-s)} = \sum_{j=1}^n y_j - \sum_{j=1}^k y_j$. To analyze the combined ecological and individual data we use likelihood inference to estimate the regression parameters, which we write as $\boldsymbol{\beta}$. If we assume independence between the individuals conditional on the covariates (which is a standard assumption in regression, corresponding to assuming that the errors within an area are uncorrelated), then we can

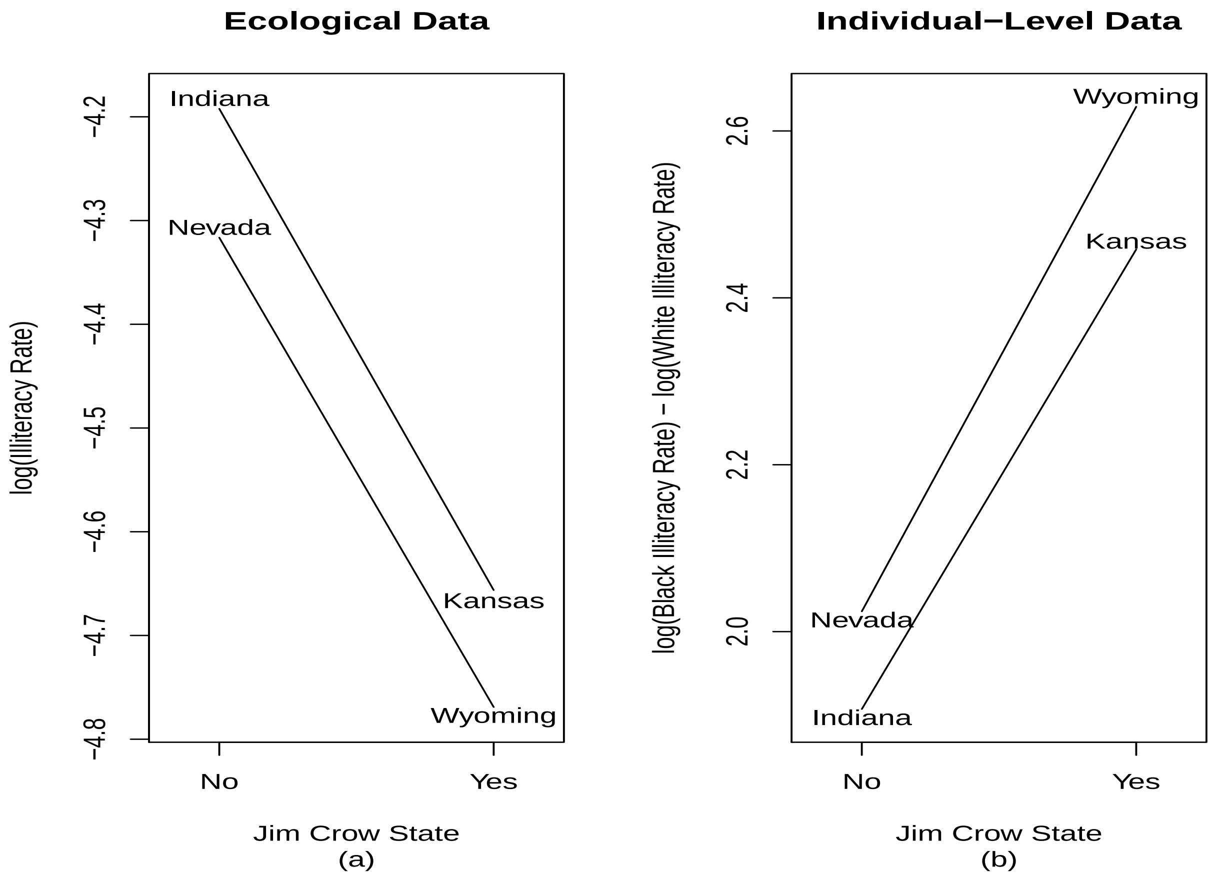

Figure 1: Analysis of the effect of Jim Crow laws on black illiteracy using ecological and individual data. Panel (a) shows the analysis using the ecological (i.e., state-level) data. In this analysis, Jim Crow laws appear to *decrease* (log) illiteracy for both the Indiana/Kansas (matched) pair (3.5% black) and the Nevada/Wyoming (matched) pair (0.6% black). Panel (b) shows the analysis using the individual-level data. With the data at this level, it is possible to measure black illiteracy for each state and also to use native-born white illiteracy as a baseline for each state. For this example, we perform the analysis by taking the (log) ratio between black and white illiteracy for each state, and Jim Crow laws now appear to *increase* black illiteracy.

write the joint likelihood for the sample data and ecological data in in the following manner:

$$\begin{aligned} f(\boldsymbol{y}^{(s)}, y_+ | \boldsymbol{x}, \boldsymbol{\beta}) &= f(\boldsymbol{y}^{(s)}, y_+^{(-s)} | \boldsymbol{x}, \boldsymbol{\beta}) \\ &= f(\boldsymbol{y}^{(s)} | \boldsymbol{x}, \boldsymbol{\beta}) \times f(y_+^{(-s)} | \boldsymbol{x}, \boldsymbol{\beta}) \\ &= f(\boldsymbol{y}^{(s)} | \boldsymbol{x}^{(s)}, \boldsymbol{\beta}) \times f(y_+^{(-s)} | \boldsymbol{x}^{(-s)}, \boldsymbol{\beta}) \end{aligned} \tag{1}$$

where $f(\boldsymbol{y}^{(s)}|\boldsymbol{x}^{(s)}, \boldsymbol{\beta}) = \prod_{j=1}^{k} f(y_j|\boldsymbol{x}_j, \boldsymbol{\beta})$. Hence, the likelihood component of the generic area consists of two terms, one for the individual-level data and one for the ecological data on the unsampled individuals.

For some probability models $f(y_+^{(-s)}|\boldsymbol{x}^{(-s)}, \boldsymbol{\beta})$ may have a simplified form. For example, if $f(y_j|x_j, \boldsymbol{\beta})$ is a Gaussian distribution, then $f(y_+^{(-s)}|\boldsymbol{x}^{(\boldsymbol{-s})}, \boldsymbol{\beta})$ will be Gaussian also. Similarly, if $f(y_j|x_j, \boldsymbol{\beta})$ is a Poisson distribution, then $f(y_+^{(-s)}|\boldsymbol{x}^{(-s)}, \boldsymbol{\beta})$ will be Poisson. The binomial distribution does not share this property. In the Gaussian/Poisson situations, inference is more straightforward and optimal sampling strategies may be more simply determined. For other cases such as a binomial logistic regression model, $f(y_+^{(-s)}|\boldsymbol{x}^{(-s)}, \boldsymbol{\beta})$ must be written as a convolution likelihood, however. A number of numerical methods are available for analysis, so combined likelihood inference is generally possible with ecological and sample data (Wakefield, 2004), though it is less straightforward than in the Gaussian or Poisson cases.

Given the combined data likelihood, it is sometimes possible to derive the optimal design of the sample, conditional on the ecological data. The qualification is that we require knowledge of the $\boldsymbol{x}$ variables for all individuals in the area. Such information will often be available from the census for demographic variables such as age, gender and race and may be approximated in cases when incomplete information only is available. The overall strategy we take is to define the expected *information* in a potential sample, conditional on the ecological data. The optimal design is then that which maximizes the information. The expected information is calculated from the following conditional likelihood:

$$\begin{aligned} f(\boldsymbol{y}^{(s)}|y_+,\boldsymbol{x},\boldsymbol{\beta}) &= \frac{f(\boldsymbol{y}^{(s)},y_+|\boldsymbol{x},\boldsymbol{\beta})}{f(y_+|\boldsymbol{x},\boldsymbol{\beta})} \\ &= \frac{f(\boldsymbol{y}^{(s)},y_+^{(-s)}|\boldsymbol{x},\boldsymbol{\beta})}{f(y_+|\boldsymbol{x},\boldsymbol{\beta})} \\ &= \frac{f(\boldsymbol{y}^{(s)}|\boldsymbol{x},\boldsymbol{\beta}) \times f(y_+^{(-s)}|\boldsymbol{x},\boldsymbol{\beta})}{f(y_+|\boldsymbol{x},\boldsymbol{\beta})} \\ &= \frac{f(\boldsymbol{y}^{(s)}|\boldsymbol{x}^{s},\boldsymbol{\beta}) \times f(y_+^{(-s)}|\boldsymbol{x}^{(-s)},\boldsymbol{\beta})}{f(y_+|\boldsymbol{x},\boldsymbol{\beta})} \end{aligned} \tag{2}$$

The form of the expected information from this likelihood has previously been derived when $f(y_j|\boldsymbol{x}_j,\boldsymbol{\beta})$ was Gaussian (Glynn et al., 2008). In Appendix B, we present the expected information from this likelihood when $f(y_j|\boldsymbol{x}_j,\boldsymbol{\beta})$ follows a Poisson distribution. The latter is often used as an approximation to the less tractable binomial distribution, when the outcome of interest is rare.

4 Application: Estimating the Effects of Jim Crow Laws

In Section 2 we showed that using ecological data to estimate the effects of Jim Crow laws results in biased estimates. In this section, we use combined inference on the basis of (1), and optimal sampling design on the basis of (2) to demonstrate that small samples of individual-level data can produce the accurate estimates summarized in Figure 1(b).

In order to reproduce the results in Figure 1(b), we need to estimate the ratio between black and white illiteracy rates within each state. This would be possible by taking large random samples of size k_i, within each state $i = 1, ..., 4$, from among individuals $j = 1, ..., n_i$. We let Y_{ij} indicate whether individual j in state i is illiterate and x_{i1j} indicates whether this individual is black and x_{i2j} indicates whether this individual is foreign-born white. Because the outcome (illiteracy) is relatively rare, we use the Poisson model

$$Y_{ij}|x_{i1j},x_{i2j},\boldsymbol{\beta}_i \sim \text{Poisson}\left[\,\exp(\beta_{0i}+\beta_{1i}x_{i1j}+\beta_{2i}x_{i2j})\,\right], \tag{3}$$

where $\boldsymbol{\beta}_i = (\beta_{0i}, \beta_{1i}, \beta_{2i})$. Consequently, in area i: $\exp(\beta_{0i})$ is the risk of illiteracy for a native-born white individual, $\exp(\beta_{1i})$ is the relative risk of illiteracy for a black individual, when compared to a native-born white, and $\exp(\beta_{2i})$ is the relative risk of illiteracy for a foreign-born white individual, when compared to a native-born white. In Figure 1(b) we plot empirical estimates of the log of the ratios (β_{1i}) on the vertical axis; in this figure we have plotted four estimates of this quantity, one for each state. The effects of Jim Crow laws in each of the two matched states are empirically estimated by the differences between these (log) rates for each of the matched pairs.

If we attempt to estimate the β_{1i} using the Poisson model and data from a random sample of individual level data, then the standard errors will in general be large because of the rareness of the outcome. In fact, for small simple random samples within each state, it will often be impossible to estimate the standard error, because we are quite likely to obtain a sample from one of the states without any black respondents. Even with samples gathered within race strata, the rarity of illiteracy will lead to large standard errors. The use of ecological data for combined estimation reduces the standard errors (Appendix B presents the analytical formula for the extra expected information provided by the combined approach). We first illustrate the benefits of adding ecological data to the sample data, when the latter are a random sample.

Table 1 compares the combined (sample plus ecological) approach to the sample only approach by presenting the ratio of standard errors under the two approaches for the estimation of β_{1i}, $i = 1, ..., 4$. The numbers in the table are the ratio of the standard errors using combined estimation (sample and ecological data) to the standard errors using only the individual sample data. All of the entries are less than 1, and are usually considerably less than 1, showing the benefits of augmenting the individual-level data with the ecological data. The first row presents the ratio of standard errors for random samples stratified on state and all three racial categories (e.g., 300 observations within a state are allocated as 100 native-born white, 100 foreign-born white, and 100 black). These results show that the addition of ecological data in combined estimation can dramatically reduce standard errors. The second row presents the ratio of standard errors for random samples stratified on state and only the native-born white and black racial categories (e.g., 300 observations within a state are allocated as 150 native-born white and 150 black). This comparison is more favorable to the sample-only approach although the combined approach still reduces standard errors by at least 50%.

	200 Respondents per State				300 Respondents per State			
Groups	Indiana	Kansas	Wyoming	Nevada	Indiana	Kansas	Wyoming	Nevada
All	0.18	0.21	0.43	0.55	0.09	0.09	0.36	0.51
Black/White	0.39	0.42	0.44	0.50	0.39	0.41	0.43	0.46

Table 1: Ratios of standard errors for estimators of log relative risks in each of four states, β_{1i}, $i = 1, ..., 4$, based on stratified random samples stratified on state and race with equal within-strata sample sizes. The numbers in the table are the ratio of the standard errors using combined estimation (sample and ecological data) divided by the standard errors using only the sample. The first row presents the ratio of standard errors for random samples stratified on state and all three racial categories (e.g., 300 observations within a state are allocated as 100 native-born white, 100 foreign-born white, and 100 black). The second row presents the ratio of standard errors for random samples stratified on state and only the native-born white and black racial categories (e.g., 300 observations within a state are allocated as 150 native-born white and 150 black).

We have illustrated that combined estimation with ecological data reduces the variance for these two types of stratified samples, when the ecological data is used to inform the sampling design, the improvement can be more dramatic. In order to derive the optimal sampling design, we must consider the expected information contained in any sample, conditional on the ecological data. The closed form expression for expected information in the sample conditional on the ecological data is presented in Appendix B as equation (6).

We make two observations about the expected information in the sample, conditional on the ecological data. First, the information quantity does not include an intercept term because conditioning on the ecological data (the state illiteracy rate and the proportion of the population that is white, black, and foreign born white) has removed this from the expression. This has important consequences for our sampling design, because it means that we only need to sample from two of the three racial categories in order to estimate the illiteracy rates for all three categories. Second, as in other non-linear design problems, the expected information is a function of the parameters to be estimated: the parameter of interest (the log ratio between black and native-born white illiteracy rates) and a nuisance parameter (the log ratio between foreign-born white and white illiteracy rates). Therefore, we can only determine a range of optimal designs, and each is dependent on the ratios in illiteracy rates between racial groups.

In order to pick the optimal design for this example, we first specify a range of plausible values for β_{1i} and β_{2i}. For this analysis, we consider values of β_{1i}, β_{2i} of (0,0) and (2.5,2.5) with these pairs of values providing extreme cases of null associations and very strong associations. The numbers reported in rows 1 and 2 of Table 2 correspond to the optimal design (the percentages to sample within each racial group) for within-state samples when the log of the relative risks are both 2.5 (i.e. the log illiteracy rates for blacks and foreign-born whites are 2.5 greater than the log illiteracy rates for native-born whites). Specifically, we set $\beta_{1i} = \beta_{2i} = 2.5$ and for all possible allocations of observations to black and foreign-born white respondents, we calculate the expected information for the parameter of interest β_{1i}, while taking account of the uncertainty due to the nuisance parameter β_{2i} (see Appendix B for details). Because we are conditioning on the ecological data, there is no need to sample native-born whites.

When we reduce the sizes of the log rates β_{1i} and β_{2i}, the optimal allocation to foreign-born whites decreases in the direction of the ecological proportions of foreign-born whites. The numbers reported in the bottom half of Table 2 correspond to the optimal design for within state samples when the log ratios are both 0 (i.e. the illiteracy rates for blacks and foreign-born whites are the same as the illiteracy rates for native-born whites). This optimal sampling design (with the allocation to foreign-born whites equal to the ecological proportions of foreign-born whites) is consistent with the optimal Gaussian linear design (Glynn et al., 2008), as one would expect because (3) is effectively a linear model when $\beta_{1i} = \beta_{2i} = 0$. It is also intuitive that if there are no race effects we would just randomly sample individuals without regard to race.

In order to assess the benefits of optimally designed combined inference, we compare its performance to combined inference using the native-born white/black samples used in the second row of Table 1 with 300 observations in each state with 50% of the sample allocated to native-born white individuals and 50% to black individuals. Standard errors from this combined analysis are presented in the first row of Table 3. For combined inference with optimally chosen samples, the design presented in the first two rows of Table 2 was used.

	Indiana	Kansas	Wyoming	Nevada
Log Illiteracy Relative Risk of 2.5				
% of Sample Black	69	71	37	32
% of Sample Foreign-Born White	31	29	63	68
Log Illiteracy Relative Risk of 0				
% of Sample Black	95	95	87	84
% of Sample Foreign-Born White	5	5	13	16

Table 2: Optimal sampling design to estimate the log ratios between black and native-born white illiteracy rates within each state conditional on the ecological data and assuming log ratios in illiteracy rates between the racial groups are 2.5 (top half of table) and 0 (bottom half of table). As the log ratios in illiteracy rates approach 0, the % of the sample foreign-born white approaches the ecological proportions: Indiana 5%, Kansas 5%, Wyoming 13%, Nevada 16%.

Standard errors from this analysis are presented in the second row of Table 3.

The benefit of the optimally designed samples over the random samples is clear. There is improvement for all states but the benefits are most apparent for the Indiana/Kansas pair where the standard errors for the optimal approach are 6–7 times smaller than the random sample case.

	Indiana	Kansas	Wyoming	Nevada
Ecological Data plus Random Sample	2.96	3.38	3.43	4.03
Ecological Data plus Optimal Sample	0.42	0.57	1.99	3.93

Table 3: Comparison of standard errors using designs with 300 observations per state and utilizing the ecological data. Each row gives standard errors for a combined ecological and individual sample. In the first row, 50% black and 50% native-born white samples are sampled within each state. In the second row the samples are optimally chosen (assuming $\beta_{1i} = \beta_{2i} = 2.5$). These optimal allocations were presented in the top half of Table 2.

5 Discussion

In this paper we have shown that a small amount of individual-level data can alleviate the ecological bias that arises when within-area risks are being estimated. Further, we have demonstrated that the ecological data both allows an optimal design to be developed and is beneficial to estimation. Within-area sampling to remove ecological bias is not a new idea. An elegant method to deal with ecological bias based on samples of covariates was described by Prentice, Sheppard and co-authors in the context of a dietary study (Prentice and Sheppard, 1995; Sheppard and Prentice, 1995; Sheppard et al., 1996). A rationale for this *aggregate data* design was that when random samples are taken it is likely that few, often zero, disease cases will be sampled. Hence, the individual covariate information only is used in the analysis, along with the ecological outcomes. Many authors have subsequently suggested ways in which inference from combined samples can be carried out (Haneuse and Wakefield, 2008; Jackson et al., 2006; Glynn et al., 2008; Jackson et al., 2008; Wakefield and Haneuse, 2008). Little work is available on optimal design, however, though for a Gaussian outcome results have been derived (Glynn et al., 2008). In this paper we have extended this work to the Poisson framework, which is applicable in many epidemiological situations.

In the illiteracy example we considered we illustrated the benefits of both optimal, as opposed to random, sampling and supplementing individual-level data with ecological information. In this example, it was not required to sample one of the three race groups at the individual level. For categorical covariates this result is true in general, which may be useful when one of the groups of interest is difficult to sample/reach.

We have derived optimal designs in the situation in which the covariate distribution is known in each area, which will often be the case if the covariates consist of demographic variables such as gender, age and race. In other situations one may have more limited information (such as the average of a covariate and a measure of the spread). In this situation one may posit a distribution for the covariate and derive the optimal design based on this assumed form. The optimality of the subsequent design obviously depends on the closeness of the assumed distribution to the true distribution. One would expect that a reasonably informed choice would lead to improved efficiency over random sampling, but this requires further investigation.

Funding

The second author was supported by grant R01 CA095994 from the National Institutes of Health.

Appendix A: Illiteracy Data in Four States

	Indiana			**Nevada**		
	Illiterate	Total		Illiterate	Total	
Native-Born White	22,510	2,501,111	0.9%	130	65,000	0.2%
Foreign-Born White	13,536	134,020	10.1%	909	12,120	7.5%
Black	5,605	93,417	6.0%	7	467	1.5%
	41,651	2,728,548	1.5%	1,046	77,587	1.3%

Table 4: Data for states without Jim Crow laws.

	Kansas			**Wyoming**		
	Illiterate	Total		Illiterate	Total	
Native-Born White	7,001	1,400,200	0.5%	381	127,000	0.3%
Foreign-Born White	4,113	69,712	5.9%	811	19,310	4.2%
Black	3,228	54,712	5.9%	47	1,119	4.2%
	14,342	1,524,624	0.9%	1,239	147,429	0.8%

Table 5: Data for states with Jim Crow laws.

Appendix B: Combined Inference and Design for a Poisson Loglinear Model

In a Poisson loglinear regression model we have individual likelihood in a generic area

$$\boldsymbol{y}^{(s)}|\boldsymbol{x}^{(s)}, \boldsymbol{\beta} \;\sim\; \prod_{j=1}^{k} \text{Poisson}\,[\,\exp(\boldsymbol{x}_j\boldsymbol{\beta})\,]$$

where $\boldsymbol{y}^{(s)} = (y_1, ..., y_k)$ are the individual responses in the ecological sample and $\boldsymbol{x}^{(s)} = (\boldsymbol{x}_1^{(s)}, ..., \boldsymbol{x}_k^{(s)})$ are the associated covariates. The ecological likelihoods for unsampled and all individuals take the form:

$$\begin{aligned} y_+^{(-s)}|\boldsymbol{x}^{(-s)}, \boldsymbol{\beta} &\sim \text{Poisson}\left[\sum_{j=k+1}^{n} \exp(\boldsymbol{x}_j\boldsymbol{\beta})\right], \\ y_+|\boldsymbol{x}, \boldsymbol{\beta} &\sim \text{Poisson}\left[\sum_{j=1}^{n} \exp(\boldsymbol{x}_j\boldsymbol{\beta})\right]. \end{aligned}$$

The combined likelihood (1) for ecological and sampled data can be written as the following:

$$f(\boldsymbol{y}^{(s)}|\boldsymbol{x},\boldsymbol{\beta}) = f(\boldsymbol{y}^{(s)}, y_+|\boldsymbol{x},\boldsymbol{\beta}) = \prod_{j=1}^{k} \text{Poisson}\,[\,\exp(\boldsymbol{x}_j\boldsymbol{\beta})\,] \times \text{Poisson}\left[\,\sum_{j=k+1}^{n} \exp(\boldsymbol{x}_j\boldsymbol{\beta})\,\right],$$

and forms the basis for maximum likelihood estimation. Suppose we wish to subsample within the generic area, if we write

$$\lambda_j = \exp(\boldsymbol{x}_j\boldsymbol{\beta}), \quad \lambda_+ = \sum_{j=1}^{n} \lambda_j, \quad \lambda_+^{(-s)} = \sum_{j=k+1}^{n} \lambda_j,$$

then the observed information in the area can be written as

$$\begin{aligned} \frac{\partial^2 \ell}{\partial \boldsymbol{\beta}} &= -\sum_{j=1}^{k} \boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{x}_j \exp(\boldsymbol{x}_j\boldsymbol{\beta}) \\ &+ \left(\frac{y_+^{(-s)}}{\lambda_+^{(-s)}} - 1\right) \sum_{j=k+1}^{n} \boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{x}_j \exp(\boldsymbol{x}_j\beta) - \frac{y_+^{(-s)}}{(\lambda_+^{(-s)})^2} \left(\sum_{j=k+1}^{n} \boldsymbol{x}_j^{\mathrm{T}} \exp(\boldsymbol{x}_j\boldsymbol{\beta})\right) \left(\sum_{j=k+1}^{n} \boldsymbol{x}_j \exp(\boldsymbol{x}_j\boldsymbol{\beta})\right). \end{aligned}$$

and because $E[y_+^{(-s)}] = \lambda_+^{(-s)}$, the expected information $-\mathrm{E}\left[\frac{\partial^2 \ell}{\partial \boldsymbol{\beta}}\right]$ is the following

$$\boldsymbol{I}(\boldsymbol{\beta}) = \sum_{j=1}^{k} \boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{x}_j \exp(\boldsymbol{x}_j\boldsymbol{\beta}) + \frac{1}{\lambda_+^{(-s)}} \left(\sum_{j=k+1}^{n} \boldsymbol{x}_j^{\mathrm{T}} \exp(\boldsymbol{x}_j\boldsymbol{\beta})\right) \left(\sum_{j=k+1}^{n} \boldsymbol{x}_j \exp(\boldsymbol{x}_j\boldsymbol{\beta})\right). \tag{4}$$

where the first term on the right hand side represents the information from the sample and the second term represents the additional information provided by the ecological data. Recall that the variance of $\widehat{\boldsymbol{\beta}}$ can be calculated by inverting this matrix, i.e. $\text{var}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{I}(\boldsymbol{\beta})^{-1}$, which fits with the concept of information (lower variance corresponding to greater information). Hence, it makes sense to define an optimal design as one that maximizes information (minimizes variance).

We now derive the conditional likelihood (2) in order to find an optimal sampling design based on the expected information of the conditional likelihood. This likelihood conditions on the ecological total, so that we maximize information given we have the ecological data. Now we suppose we have a sampling frame for the entire design matrix (i.e. if we know all possible values of $\boldsymbol{x}$ before the sampling of individual data), (2) simplifies to

$$f(\boldsymbol{y}^{(s)}|y_+,\boldsymbol{x},\boldsymbol{\beta}) \quad = \quad \frac{\prod_{j=1}^{k} \text{Poisson}\,[\,\exp(\boldsymbol{x}_j\boldsymbol{\beta})\,] \times \text{Poisson}\left[\,\sum_{j=k+1}^{n}\exp(\boldsymbol{x}_j\boldsymbol{\beta})\,\right]}{\text{Poisson}\left[\,\sum_{j=1}^{n}\exp(\boldsymbol{x}_j\boldsymbol{\beta})\,\right]}.$$

Suppose we wish to subsample within the generic area. The observed information in the generic area can be written as

$$\begin{aligned}
\frac{\partial^2\ell}{\partial\boldsymbol{\beta}} \quad &= \quad -\sum_{j=1}^{k}\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{x}_j\exp(\boldsymbol{x}_j\boldsymbol{\beta}) \\
&+ \quad \left(\frac{y_+^{(-s)}}{\lambda_+^{(-s)}}-1\right)\sum_{j=k+1}^{n}\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{x}_j\exp(\boldsymbol{x}_j\boldsymbol{\beta}) - \frac{y_+^{(-s)}}{(\lambda_+^{(-s)})^2}\left(\sum_{j=k+1}^{n}\boldsymbol{x}_j^{\mathrm{T}}\exp(\boldsymbol{x}_j\boldsymbol{\beta})\right)\left(\sum_{j=k+1}^{n}\boldsymbol{x}_j\exp(\boldsymbol{x}_j\boldsymbol{\beta})\right) \\
&- \quad \left(\frac{y_+}{\lambda_+}-1\right)\sum_{j=1}^{n}\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{x}_j\exp(\boldsymbol{x}_j\boldsymbol{\beta}) - \frac{y_+}{\lambda_+^2}\sum_{j=1}^{n}\boldsymbol{x}_j^{\mathrm{T}}\exp(\boldsymbol{x}_j\boldsymbol{\beta})\sum_{j=1}^{n}\boldsymbol{x}_j\exp(\boldsymbol{x}_j\boldsymbol{\beta}) \\
&= \quad \frac{y_+^{(-s)}}{\lambda_+^{(-s)}}\left[\sum_{j=k+1}^{n}\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{x}_j\exp(\boldsymbol{x}_j\boldsymbol{\beta}) - \frac{1}{(\lambda_+^{(-s)})}\left(\sum_{j=k+1}^{n}\boldsymbol{x}_j^{T}\exp(\boldsymbol{x}_j\boldsymbol{\beta})\right)\left(\sum_{j=k+1}^{n}\boldsymbol{x}_j\exp(\boldsymbol{x}_j\boldsymbol{\beta})\right)\right] \\
&- \quad \frac{y_+}{\lambda_+}\left[\sum_{j=1}^{n}\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{x}_j\exp(\boldsymbol{x}_j\boldsymbol{\beta}) - \frac{1}{\lambda_+}\left(\sum_{j=1}^{n}\boldsymbol{x}_j^{\mathrm{T}}\exp(\boldsymbol{x}_j\boldsymbol{\beta})\right)\left(\sum_{j=1}^{n}\boldsymbol{x}_j\exp(\boldsymbol{x}_j\boldsymbol{\beta})\right)\right]. \qquad (5)
\end{aligned}$$

Notice that we can factor $\exp(\beta_0)$ out of the top and bottom of all of the terms in (5), and consequently the observed information no longer depends on β_0. This is not surprising, because the ecological sums, upon which we condition, contain all the information about the overall average.

Furthermore, the two terms of the observed information (5) can be written as weighted sums of squares. We write

$$\overline{\boldsymbol{x}}_w = \frac{1}{\lambda_+}\sum_{j=1}^{n}\boldsymbol{x}_j\exp(\boldsymbol{x}_j\boldsymbol{\beta})$$

and

$$\boldsymbol{x}_{w,j} = \left(\lambda_j^{1/2}(\boldsymbol{x}_j - \overline{\boldsymbol{x}}_w)\right)_j$$

for individuals $j = 1, ..., n$ and

$$\overline{\boldsymbol{x}}_w^{(-s)} = \frac{1}{\lambda_+}\sum_{j=1}^{k} \boldsymbol{x}_j \exp(\boldsymbol{x}_j\boldsymbol{\beta})$$

and

$$\boldsymbol{x}_{w,j^{(-s)}} = \left(\lambda_j^{1/2}(\boldsymbol{x}_j - \overline{\boldsymbol{x}}_w^{(-s)})\right)_j$$

for rows $j = 1, ..., k$. Substitution of these terms into (5) gives the simplified form

$$\frac{\partial^2 \ell}{\partial \boldsymbol{\beta}} = \frac{y_+^{(-s)}}{\lambda_+^{(-s)}}\left(\boldsymbol{x}_w^{(-s)\mathrm{T}}\boldsymbol{x}_w^{(-s)}\right) - \frac{y_+}{\lambda_+}\left(\boldsymbol{x}_w^{\mathrm{T}}\boldsymbol{x}_w\right)$$

Therefore, the observed information can be seen as the combination of the weighted sums of squares of the "unsubsampled" covariates and the weighted sums of squares of the full data. From the design standpoint, we know $y_+^{(-s)}$ prior to subsampling, so we need to derive the expected information conditional on the ecological data. Conveniently, $y_+^{(-s)}|y_+$ follows a binomial distribution and hence $\mathrm{E}[y_+^{(-s)}|y_+] = \frac{y_+\lambda_+^{(-s)}}{\lambda_+}$. Therefore, the conditional Fisher's information can be written as,

$$\boldsymbol{I}(\boldsymbol{\beta}) = \frac{y_+}{\lambda_+}\left\{\left(\boldsymbol{x}_w^{\mathrm{T}}\boldsymbol{x}\right) - \left(\boldsymbol{x}_w^{(-s)\mathrm{T}}\boldsymbol{x}_w^{(-s)}\right)\right\} \tag{6}$$

If, as in the illiteracy example, the first element of $\boldsymbol{\beta}$ is the parameter of interest and the other elements are treated as nuisance parameters, then we maximize information for the parameter of interest by minimizing the row 1, column 1 element of $\boldsymbol{I}(\boldsymbol{\beta})^{-1}$.

It is important to note that although $\boldsymbol{I}(\boldsymbol{\beta})$ in (6) is a function of only the ecological outcome y_+ and not the individual-level outcome data, this information is a function of the unknown parameters $\boldsymbol{\beta}$. Therefore, we must specify potential values of $\boldsymbol{\beta}$ in order to derive the design that maximizes expected information. This difficulty is typical of non-linear design problems, however we also note that by conditioning on the ecological data (y_+), we have reduced the problem since without the ecological data, calculating the optimal design would require the specification of all parameters in $\boldsymbol{\beta}$. Conditional on the ecological data, we do not need to specify the baseline risk (β_0).

References

Blalock, H.M. 1984. “Contextual-effects models: theoretical and methodological issues.” *Annual Review of Sociology* 10:353–372.

Diez-Roux, A.V. 1998. “Bringing context back into epidemiology: variables and fallacies in multilevel analysis.” *American Journal of Public Health* 88:216–222.

Firebaugh, G. 2009. “Commentary: ‘Is the social world flat? W.S. Robinson and the ecological fallacy’.” *International Journal of Epidemiology* 38:368–370.

Freedman, D.A., S.P. Klein, M. Ostland, and M.R. Roberts. 1998. “Review of A Solution to the Ecological Inference Problem.” *Journal of the American Statistical Association* 93:1518–1522.

Glynn, A., J. Wakefield, M. Handcock, and T. Richardson. 2008. “Alleviating linear ecological bias and optimal design with subsample data.” *Journal of the Royal Statistical Society, Series A* 171:179–202.

Goodman, L.A. 1953. “Ecological Regression and the Behavior of Individuals.” *American Sociological Review* 18:663–664.

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

Greenland, S., and H. Morgenstern. 1989. “Ecological Bias, Confounding, and Effect Modification.” *International Journal of Epidemiology* 18 (1):269–274.

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

Greenland, Sander. 1992. “Divergent biases in ecologic and individual-level studies.” *Statistics in Medicine* 11:1209–1223.

Haneuse, S., and J. Wakefield. 2008. “The combination of ecological and case-control data.” *Journal of the Royal Statistical Society, Series B* 70:73–93.

Jackson, C., N. Best, and S. Richardson. 2008. “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* 171:159–178.

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

King, G. 1997. *A Solution to the Ecological Inference Problem*. Princeton: Princeton.

King, G., O. Rosen, and M. Tanner. 2004. “Information in Ecological Inference: An Introduction.” In *Ecological Inference: New Methodological Strategies* (G. King, O. Rosen, and M. Tanner, editors), Cambridge: Cambridge University Press.

Oakes, J.M. 2009. “Commentary: Individual, ecological and multilevel fallacies.” *International Journal of Epidemiology* 38:361–368.

Piantadosi, S., D.P. Byar, and S.B. Green. 1988. “The Ecological Fallacy.” *American Journal of Epidemiology* 127:893–904.

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

Richardson, S. 1992. “Statistical methods for geographical correlation studies.” In *Analysis of Survey Data* (P. Elliott, J. Cuzick, D. English, and R.Stern, editors), New York: Oxford University Press, pp. 181–204.

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

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

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

Subramanian, S.V., K. Jones, A. Kaddour, and N. Krieger. 2009a. "Response: The value of a historically informed multilevel analysis of Robinson's data." *International Journal of Epidemiology* 38:379–373.

Subramanian, S.V., K. Jones, A. Kaddour, and N. Krieger. 2009b. "Revisiting Robinson: the perils of individualistic and ecologic fallacy." *International Journal of Epidemiology* 38:342–360.

Wakefield, J. 2004. "Ecological Inference for 2x2 Tables (with discussion)." *Journal of the Royal Statistical Society - A* 167:385–445.

Wakefield, J., and S. Haneuse. 2008. "Overcoming eological bias using the two-phase study design." *American Journal of Epidemiology* 167:908–916.

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

Wakefield, J.C. 2009. "Multi-level modelling, the ecologic fallacy, and hybrid study designs." *International Journal of Epidemiology* 38:330–336.