Exploring Spatiotemporal Patterns in County-Level Incidence and Reporting of Lyme Disease in the Northeastern United States, 1990-2000

Lance A. Waller  
Department of Biostatistics, Rollins School of Public Health  
Emory University, Atlanta GA

Brett J. Goodwin  
Department of Biology  
University of North Dakota, Grand Forks, ND

Mark L. Wilson  
Departments of Biology and Epidemiology  
University of Michigan, Ann Arbor, MI

Richard S. Ostfeld  
Institute of Ecosystem Studies, Millbrook NY

Ned Hayes  
U.S. Centers for Disease Control and Prevention

Stacie Marshall  
U.S. Centers for Disease Control and Prevention

Abstract

We present an exploratory analysis of reported county-specific incidence of Lyme disease in the northeastern United States for the years 1990-2000. We briefly review the disease ecology of Lyme disease and the use of risk maps to describe local incidence as estimates of local risk of disease. We place the relevant elements of local environmental and ecological variables, local disease incidence, and (importantly) local disease reporting in a conceptual context to frame our analysis. We then apply hierarchical linear models of increasing complexity to summarize observed patterns in reported incidence, borrowing information across counties to improve local precision. We find areas of increasing incidence in the central northeastern Atlantic coast counties, increasing incidence branching to the north and west, and an area of fairly stable and slightly decreasing reported incidence in western New York.

1. Background

Lyme disease is a zoonosis caused by the spirochete *Borrelia burgdorferi*, which cycles between tick vectors and dozens of species of vertebrate hosts. The first medical description of Lyme disease appears in Steere et al. (1977), although earlier reports
describe incident cases of erythema migrans, the localized rash serving as a primary symptom of Lyme disease (Scrimenti 1970). The United States Centers for Disease Control and Prevention (CDC) began systematic surveillance for the disease in 1982 (Schmid et al 1985), requiring (but not necessarily enforcing) reporting of newly diagnosed cases by local health departments and health-care providers. Surveillance data remained imprecise until the Council of State and Territorial Epidemiologists approved a standardized case definition in 1990, with nation-wide implementation in 1991 (Orloski et al. 2000). Currently, Lyme disease is one of the most common vector-borne disease in the United States with almost 90,000 cases reported between 1992 and 1998 (Orloski et al. 2000). Although cases have been reported from 49 states and the District of Colombia (Orloski et al. 2000), the bulk of reported cases occur in the northeastern US (CT, DE, DC, ME, MD, MA, NH, NJ, NY, PA, RI, and VT) with a second concentration in the midwestern US (MI, MN, and WI).

The ecology of Lyme disease depends on the life cycles and interactions between the causative agent, the vector, and the various hosts. Larval and nymphal tick vectors (the black-legged tick, *Ixodes scapularis*, in the eastern US) feed primarily on smaller vertebrates (Lane et al. 1991, Barbour and Fish 1993, Mather 1993). White-footed mice (*Peromyscus leucopus*) play a crucial role as both a preferred immature tick host and principal reservoir of *Borrelia burgdorferi* (Anderson and Magnarelli 1984, Donahue et al. 1997, Levine et al. 1985, Mather et al. 1989, Fish 1993, Mather and Ginsberg 1994, Keirans et al. 1996). Adult ticks feed primarily on white-tailed deer (*Odocoileus virginianus*) (Piesman and Spielman 1979, Wilson et al. 1990). Each generation of ticks must become infected via these vertebrate hosts since there is no transovarial
transmission between egg-laying females and their offspring (Piesman et al. 1986, Patrican 1997). Once infected, ticks may infect a human host during a blood meal (Barbour and Fish 1993).

There can be a great deal of spatial variability in Lyme disease incidence. The tick vectors are patchily distributed both regionally and locally, regardless of infection status (Kitron and Kazmierczak 1997, Wilson 1998). In particular, presence of *I. scapularis* is positively associated with sandy soils, woody or shrubby vegetation, and presence of deer (Ginsberg and Ewing 1989, Kitron et al. 1991, Kitron et al. 1992, Glass et al. 1994, Glass et al. 1995, Duffy et al. 1994). Countering this spatial variability in vector presence is the tendency of tick loads on mice to remain relatively constant even in the face of substantial variation in the densities of both mice and questing ticks (Goodwin et al. 2001). Host abundance and community composition also can vary dramatically in both space and time (van Buskirk and Ostfeld 1995, Giardina et al. 2000).

Spatial variability of Lyme disease hosts and vectors suggests construction of a *risk map*, i.e., a map of the potential risk of (human) infection. Such a map indicates areas where tick control, public education, or other interventions may be most beneficial (Orloski et al. 2000, Kitron 2000). Maps of vector abundance (perhaps as functions of environmental variables), vector infection rates, human-vector interactions, and reported human cases all address different components of the spatial pattern of risk of Lyme disease. Each map requires different data and reveals different elements of the interacting processes determining Lyme transmission. The literature contains examples of risk maps for Lyme disease for particular communities (Glass et al. 1995, Dister et al. 1997), particular states (Frank et al. 2002, Kitron and Kazmierczak 1997) and for the entire

While risk maps are potentially useful tools for public health practitioners, Lyme disease presents particular challenges in their creation, analysis, and interpretation. Surveillance data for Lyme disease can be problematic for a number of reasons: developing recognition of the signature symptoms by health-care providers over the study period, the associated potential lack of accurate diagnoses, imprecise serologic results, uneven case detection, reporting biases, and difficulty relating location of report to location of exposure (Kitron and Kazmierczak 1997). As a result, Lyme disease may be substantially under-reported, with the probability of reporting varying with stage of disease and age of the patient (Orloski et al. 1998, Naleway et al. 2002). Kitron (2000) provides a brief but thorough discussion of risk maps for vector-borne diseases which we expand on in discussions below. Despite these concerns, surveillance data show that both the incidence and geographic range of human cases has increased steadily, particularly in the northeastern US (White et al. 1991, Orloski et al. 2000), spreading outward from the initial diagnoses near Lyme, Connecticut.
In this study, we explore spatio-temporal patterns in temporal changes in reported incidence for human cases for the years 1990-2000. Our aim is primarily exploratory, since any observed patterns in reported incidence (and changes in reported incidence) reflect a combination of patterns in true incidence and local variations in diagnosis rates and reporting rates. The descriptive goal is important as it provides insight into both the evolving geographic coverage of the disease, but also into evolving coverage of the surveillance system monitoring the disease. While we cannot entirely separate the two components, we can draw conclusions on particular aspects of each, as will be seen below.

In particular, we report county-level trends in reported Lyme disease incidence (rates per 100,000 persons) in the northeastern US for the years 1990-2000, using data from the CDC data repository. The goal of this study is a descriptive analysis of county-specific changes in annual rate of reported cases, in particular assessment of the spatial pattern of rate changes in attempt to identify areas experiencing fast growth (either in number of cases or in improved reporting practices). We also examine geographic patterns of missing county-level reports and discuss relationships between missing reports and observed patterns in the incidence of disease. Previous investigations by the CDC concentrate on the years 1994-2000 due to varying data availability prior to 1994. We include available data from 1990-1994 and explore its compatibility with the proposed models.

2. Lyme disease surveillance data

Since 1992, the CDC has compiled reported cases of Lyme disease based on the standardized case definition approved by the Council of State and Territorial
Epidemiologists (Orloski et al. 2000). Using this data base, we determined the number of cases of Lyme disease annual in each county in the US northeast (i.e., the states PA, MD, DE, DC, MA, NJ, NH, VT, NY, CT, RI, and ME) from 1990 to 2000. To standardize for differences in county population size, we convert raw case numbers into incidence proportions per 100,000 people at risk using the 1990 U.S. Census population sizes. For simplicity, we use the 1990 Census population sizes as denominators, noting that adjustments for population growth change results only slightly.

Before examining the data, some review of the nature of public health surveillance data is in order (Teutsch and Churchill 1994, Brookmeyer and Stroup 2004) is in order to provide a context for the analyses below. In particular, Lyme disease surveillance data comprise a mix of passive, active and laboratory surveillance (Orloski et al. 2000). Passive surveillance relies on health-care providers to report any new diagnoses to either local or state public health departments. The health department in turn assesses the report with respect to the standardized definition (perhaps including confirmatory laboratory tests), then electronically submits those cases meeting the standard to the CDC via the National Electronic Telecommunication System for Surveillance (NETSS). Active surveillance involves proactive contact of health-care providers by local or state health departments, requesting information regarding any incident diagnoses of Lyme disease, followed by case assessment and reporting to the CDC. Finally, laboratory surveillance requires diagnostic laboratories to report all positive Lyme disease test results directly to the state or local health department. Because laboratory results typically include only limited patient information, the health department often must follow up with health-care providers to verify that the cases meet
the standardized case definition before reporting the case to the CDC. Not all states include active and laboratory surveillance efforts, and those in the study area that do (Connecticut, Maryland, Massachusetts, New Jersey, New York, and Rhode Island) only include them for part of the study period (Orloski et al. 2000).

For any particular year, a number of counties do not provide reports regarding Lyme disease. We term such occurrences as “no reports” (NRs). The NRs may represent an absence of cases, an absence of diagnoses, or an absence of reports and we consider a variety of options regarding the interpretation of the NRs below. Table 1 gives the number of counties with NRs by year. In each year and each state, some reported cases were attributed to the state, but not attributed to a specific county, and we ignore these cases in our analysis below. In a typical year, we find approximately 45-50 NRs from the 245 counties in the study area. The higher numbers of NRs associated with 1992 (99 counties) and 1993 (106 counties) are primarily due to Pennsylvania (67 counties) reporting cases only at the state level for these two years.

<table>
<thead>
<tr>
<th>Year</th>
<th>Total number of cases reported</th>
<th>Number of Counties with (non-zero) reports</th>
<th>Number of counties with “no report” (NR)</th>
<th>Cases identified to state level, but not county</th>
</tr>
</thead>
<tbody>
<tr>
<td>1990</td>
<td>6,108</td>
<td>189</td>
<td>56</td>
<td>130</td>
</tr>
<tr>
<td>1991</td>
<td>7,355</td>
<td>195</td>
<td>49</td>
<td>455</td>
</tr>
<tr>
<td>1992</td>
<td>8,041</td>
<td>146</td>
<td>99</td>
<td>1,333</td>
</tr>
<tr>
<td>1993</td>
<td>6,829</td>
<td>139</td>
<td>106</td>
<td>1,235</td>
</tr>
<tr>
<td>1994</td>
<td>11,454</td>
<td>200</td>
<td>45</td>
<td>81</td>
</tr>
<tr>
<td>1995</td>
<td>10,380</td>
<td>197</td>
<td>48</td>
<td>237</td>
</tr>
<tr>
<td>1996</td>
<td>15,023</td>
<td>201</td>
<td>44</td>
<td>296</td>
</tr>
<tr>
<td>1997</td>
<td>11,280</td>
<td>199</td>
<td>46</td>
<td>184</td>
</tr>
<tr>
<td>1998</td>
<td>15,111</td>
<td>201</td>
<td>44</td>
<td>114</td>
</tr>
<tr>
<td>1999</td>
<td>14,480</td>
<td>192</td>
<td>53</td>
<td>55</td>
</tr>
<tr>
<td>2000</td>
<td>15,621</td>
<td>189</td>
<td>56</td>
<td>156</td>
</tr>
</tbody>
</table>

Table 1. Numbers of reported cases of Lyme disease in the northeaster United States, 1990-2000.
Figure 1 provides maps of the total county-specific number of reported cases per million population (left map) and the county-specific number of years with “no report” (NRs). We shade each map by quintile so roughly 1/5 of the counties appear in each grayscale. For visual comparison, darker greyscales indicate higher quintiles of cases and lower quintiles of NRs. The visual similarity in pattern roughly suggests a pattern of consistent reporting mirroring that of (reported) disease incidence.

In general, as one might expect, we note higher numbers of reported cases per 100,000 persons at risk during the study period in counties along the central Atlantic coast, with reduced numbers radiating out from the locations of the initial documented cases in Connecticut. We note a pocket of reduced number of reports in the burroughs of New York City, and in western Massachusetts.

Figure 1 also appears to reveal some isolated pockets of higher reporting coverage that appear ahead of the generally perceived wavefront of incidence, as indicated by higher numbers of years reporting cases. These areas generally correspond to counties with higher population density than their surrounding neighbors (e.g., counties near Ithaca, Rochester, and Buffalo, New York). This may be due to the increased number of individuals at risk, but may also reflect communication patterns between physicians and subsequent raised awareness (making cities “closer” in a communication sense than they appear in a mileage sense).

[Figure 1 about here.]

3. Elements of a risk map
Any spatial summary (map or analysis) of disease incidence requires a proper context for interpretation. As mentioned in the introduction above, multiple spatial processes contribute to the observed pattern of Lyme disease incidence (in humans). Figure 2 illustrates these components and focuses the purpose of this report.

Beginning on the left hand side of Figure 2, we observe the environmental composition providing a background setting for the disease process. These factors (e.g., type of soil, land use, forest, etc.) typically are available for mapping by site surveys or remote sensing. The factors describe the habitats for hosts (deer mice, white tailed deer, and humans) as well as the vector (*I. scapularis*), and define areas where the hosts and (infected) vectors are likely to interact allowing transmission. People, mice, deer, and ticks moving through this environment interact, at times allowing transmission of infection to humans at points indicated in the middle map.

The (unobserved) middle map in Figure 2 represents the location of infection for all true cases. Coupled with information regarding the number of people at any location at a given time, this map would provide information regarding the true risk of infection as a function of location and time. However, we often do not observe the true incidence map, but rather a “filtered version” represented by the two schematic maps on the right hand side of Figure 2. The shaded map represents the probability of a given case being diagnosed and reported as a function of county where darker shading represents counties with higher probabilities of diagnosis/reporting. Diagnosis and reporting are two separate processes, each with its own potential for geographic variation, but we aggregate as a surveillance reporting filter in Figure 1. The result of this “filter” is the observed incidence pattern on the right where we only observe the filled circles (we also show the
“lost” cases via empty circles in Figure 2, but these would not be observed). Location uncertainty also impacts reporting since the reported residential locations may not correspond to infection locations, as mentioned in the introduction, but we ignore such errors here for clarity.

[Figure 2 about here.]

Figure 2 is helpful not only in providing a view of the complexity of the situation, but also in focusing attention and aiding interpretation. An accurate risk map (for human infection) seeks an unbiased estimate of the center map, but descriptions of the components of this risk map are also of interest. Some previous risk maps of Lyme disease focus only on mapping habitat/location of the vector (e.g., Glass et al. 1995), others only the reported outcomes (e.g., US CDC 1999). The former limits attention to the left-hand maps in Figure 2, the latter the map on the far right.

Analysts typically view the reporting (shaded) “filter” map as a nuisance, a source of bias between the observed and true incidence maps, but we suggest this map merits attention in its own right as a summary of spatial pattern and variation in the disease surveillance system itself. That is, estimates of this shaded map provide valuable information regarding the surveillance process, identifying areas with adequate reporting and others needing improvement.

Closer examination of the maps of incidence counts and no reports (Figure 1) provides insight into patterns in the typically unobserved “reporting map” conceptually defined in Figure 2. More specifically, consider the five burroughs of New York City
and counties in western Massachusetts which reflect concentrations of NRs within the general “core” area of reporting. In both, reduced reporting also appears to play a role in defining the pattern of reported cases.

For this report, our approach is to view temporal changes in the spatial pattern on the righthand side of Figure 2 as a means to gain partial insight into both the reported incidence map and the shaded reporting map. While we cannot completely separate the two (which would provide clearer insight into the true risk map in the center of Figure 2), we do find interesting patterns suggesting evolution of both incidence and reporting over the study period which provide an important step into understanding both the spread of infection and local aspects of the surveillance system itself.

4. Statistical models to describe patterns

As a descriptor of the evolution of Lyme disease incidence and reporting in the northeastern United States, and to address the issues above, we consider local linear regressions of the natural logarithm of the annual crude incidence ratios for each county. For simplicity (and supported by general diagnostics), we assume log-incidence ratios follow normal (Gaussian) distributions, and explore temporal trends in the reported ratio of cases to the 1990 county-level population size for each county.

We ignore NRs in our local regressions of log incidence rates, but then incorporate them as missing observations in a broader Bayesian context, allowing posterior prediction of the missing values rather than simply ignoring them. Some NRs, particularly those near the periphery of the study area, may reflect zero counts that were not reported, while likely values of NRs within prevalent areas are more difficult to
ascertain. In any case, assumptions of ‘missing at random’ or ‘missing completely at random’ appear overly simplistic. Finally, recall that some NR patterns, such as those in Pennsylvania, are direct results of evolving reporting practices and requirements.

To begin, let $Y_i$ and $n_i$ denote the number of reported cases and the 1990 population size, respectively, for county $i$ in year $t$ (we omit a time subscript on $n_i$ since we use only the 1990 population sizes, noting little impact on inference when adjusting for yearly population projections). For each county we fit (via least squares) the local linear regression model

$$\log\left(\frac{Y_i}{n_i}\right) = \beta_0 + \beta_1 t + \epsilon_{it}$$

where the error term, $\epsilon_{it}$, follows a normal (Gaussian) distribution with variance $\sigma^2_i$. For the local models, we assume a constant variance within (but not necessarily between) counties. For comparison, we also fit a single regression to the county-specific data within each state (now assuming constant variance across each state).

Not surprisingly, there is considerable variation in county-specific reported Lyme disease incidence and we next consider random effects models wherein both the slope and intercept vary by county. We use a hierarchical Bayesian formulation allowing either spatially unstructured (exchangeable) or spatially structured slopes or intercepts. First, for comparison with the local (maximum likelihood) regression models, we consider a Bayesian version of the overall regression model above where we assign diffuse prior distributions to both $\beta_0$ and $\beta_1$, as a comparison to the least squares (maximum likelihood) results for the model above.

Next, we generalize the model to

$$\log\left(\frac{Y_i}{n_i}\right) \mid u_{0i}, u_{1i} \sim N(\mu_{0i}, \sigma^2),$$
\[ \mu_{it} = \beta_0 + \beta_1 t + u_{0i} + u_{1i}, \]

where \( \mu_{it} \) represents the expected local log incidence rate (given the values of the local random effects, \( u_{0i} \) and \( u_{1i} \)), and we now assume a constant variance across the entire study area in order to link regression models across counties. Retaining diffuse priors for \( \beta_0 \) and \( \beta_1 \), we assign exchangeable priors

\[ u_{0i} \sim N(0, \sigma_{u0}^2), \text{ and} \]
\[ u_{1i} \sim N(0, \sigma_{u1}^2) \]

to the county-specific random effects which serve as local adjustments to the overall intercept and slope. We complete the model by assigning conjugate inverse-Gamma hyperpriors to \( \sigma_{u0}^2 \) and \( \sigma_{u1}^2 \). The result is a model with county-specific adjustments centered around an overall slope and intercept. The impact of the local random effects yield posterior estimates of the local intercept, \( (\beta_0 + u_{0i}) \), and slope, \( (\beta_1 + u_{1i}) \), representing compromises between the overall values across all counties and the values suggested by each set of county-specific data, i.e., the model “borrows strength” across all counties to improve estimates of the county-specific intercepts and slopes.

Since the elements of a Lyme disease risk map outlined in Figure 2 involve spatial patterns, we also consider borrowing strength locally rather than globally by replacing one or both of the exchangeable priors with a spatially correlated prior, effectively yielding either intercepts or slopes that are compromises between values supported by the local data and those supported by spatially neighboring counties. A popular family of such spatial prior distributions is the set of conditionally autoregressive (CAR) priors (Besag, York and Mollié 1991, Wakefield et al. 2000, Waller and Gotway 2004, Section 9.5), and we consider those defined by
\[ u_{ij} | u_{ij}, j \neq i \sim N\left( \frac{\sum_j w_{ij} u_{ij}}{\sum_j w_{ij}}, \frac{1}{\sigma_{u1}^2 \sum_j w_{ij}} \right), \text{ and} \]
\[ u_{0i} | u_{0j}, j \neq i \sim N\left( \frac{\sum_j w_{ij} u_{0j}}{\sum_j w_{ij}}, \frac{1}{\sigma_{u0}^2 \sum_j w_{ij}} \right) \]

where \( w_{ij} = 1 \) if regions \( i \) and \( j \) share a common border, and 0 otherwise. By convention, \( w_{ii} = 0 \). Besag (1974) shows this set of conditional distributions defines a valid joint normal distribution, although the use of adjacency weights defines a singular precision matrix yielding an improper prior distribution (Besag and Kooperberg 1995, Besag et al. 1995, Waller and Gotway 2004, Section 9.5). However, Besag et al. (1995) show that the impropriety is due to the joint distribution being defined for pairwise contrasts between local random effect values, and the addition of a “sum to zero” constraint on the random effects allows proper posterior inference for all model parameters. We complete the model by defining conjugate inverse-Gamma hyperprior distributions for \( \sigma_{u0}^2 \) and \( \sigma_{u1}^2 \) (Kelsall and Wakefield, 1999). Such models are very common in spatial smoothing of regional disease rates (disease mapping), but most models only use random intercepts with some notable exceptions (Knorr-Held and Besag 1998, Schootman and Sun 2004). Our interests here involve changes over time, motivating consideration of the CAR prior for the slope parameters.

5. Results

Figure 3 reveals the crude county-specific 1990-2000 time trends in \( \log(Y_{it}/n_i) \) by state with state-specific trends indicated by dashed grey lines. The plots suggest the adequacy of a linear model of log incidence and indicate a fair amount of relatively symmetric variation around the general trend observed in each state. Gaps in reporting are also apparent.
For comparison with results from more complicated models, Figure 4 provides a map of the ordinary least squares (OLS) estimates of county-specific time trends ($\beta_{1i}$) in the log incidence proportion, ignoring NRs. The general pattern is one of higher slopes (faster increases) in counties with the highest incidence rates, and small but negative slopes roughly in a band across upstate New York. We also note three counties (two in northern New York and one in western Pennsylvania) have zero or one reported values making local regression estimation impossible without additional assumptions regarding the NRs.

The results in Figure 4 build estimates based on local data alone. In moving toward compromises between local, global, and neighboring data, we next consider global OLS estimates based on all counties in the study area. Table 2 provides summary statistics indicating a small, positive increase in the log incidence proportion. This small positive slope is offset by considerable variation between counties, as illustrated in Figure 3, resulting in an $R^2$ of only 0.02, a situation we hope to improve through the use of county-specific covariates.
<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>s.e.</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_0$</td>
<td>3.931</td>
<td>0.063</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>0.071</td>
<td>0.010</td>
</tr>
</tbody>
</table>

Table 2. Ordinary Least Squares estimates of the global intercept and slope.

Moving to Bayesian formulations, Table 3 presents posterior inference for the same model with diffuse priors on the (still global) intercept and slope, and an inverse Gamma ($0.5, 0.005$) hyperprior on the overall variance $\sigma_Y^2$. We see very close agreement with the OLS (equivalent to maximum likelihood) estimates in Table 2. Again the slight positive global increase in incidence is offset by the large variation across counties.

<table>
<thead>
<tr>
<th></th>
<th>2.5 %ile</th>
<th>Median</th>
<th>97.5 %ile</th>
</tr>
</thead>
<tbody>
<tr>
<td>$[\beta_0</td>
<td>Y]$</td>
<td>3.807</td>
<td>3.931</td>
</tr>
<tr>
<td>$[\beta_1</td>
<td>Y]$</td>
<td>0.050</td>
<td>0.072</td>
</tr>
<tr>
<td>$[1/ \sigma_Y^2</td>
<td>Y]$</td>
<td>0.414</td>
<td>0.440</td>
</tr>
</tbody>
</table>

Table 3. Posterior inference for global intercept and slope.

We next add random effects as described in the preceding section. Figure 5 provides the 95% interval estimates (confidence intervals for OLS estimates and posterior credible intervals for the Bayes’ estimates) for each model considered namely: OLS, the global “general Bayes” model summarized in Table 3, the random effects model with exchangeable (spatially unstructured) priors on both intercepts and slopes (“Exchangeable”), the random effects model with exchangeable intercepts and spatially structured slopes (“Spatial slope”), and the random effects model with spatially structured
intercepts and slopes ("Both spatial"). For the random effects models, the intervals in Figure 5 represent the posterior variation associated with the global values $\beta_0$ and $\beta_1$, and do not incorporate the additional variation associated with the random effects. While the mean values of the estimated global intercept and global slope drop with the addition of random effects, the overall slope estimates remains “significantly” positive (the 95% interval estimate excludes zero). We also note that the model with the exchangeable intercept random effects and spatially structured slope random effect (the “spatial slope” model) increases the precision of the overall estimate of the time trend as evidenced by the narrower credible interval in Figure 5, suggesting some benefit to borrowing strength from neighboring counties. The model with spatial priors on random effects (the “both spatial” model) associated with both the intercept and the slope suggests even more precision in the estimation of the global parameters, but we reserve judgment on this feature for now.

[Figure 5 about here.]

Figure 6 provides maps of the posterior median local slope estimates, $(\beta_1 + u_{1i})$, $i=1,\ldots, 245$, for each of the three random effects models, using the same shading categories we used for the local OLS estimates in Figure 4. One immediate impact of the “borrowed strength” of incorporating all available data is the fact that no estimates occur in the lowest or highest categories, reflecting the influence of data from other counties and an overall “shrinkage” of slopes and intercepts toward global or local means. This is perhaps most notable for Aroostook County, Maine (the northernmost county on the
map) where the local OLS estimate (based on two data points) is highly positive, but the compromise estimate between either all (exchangeable) or neighboring (spatial) data results in a small negative slope estimate. Figure 7 displays the same data but shades counties by quintiles within each map so roughly one-fifth of the counties are found within each greyscale, allowing a better assessment of spatial variation within each map. All models mirror the initial pattern exhibited in Figure 4, namely, the greatest increases in the (log) incidence proportions occur in counties near the focus of the epidemic, not near the edges as we might expect if we were capturing a strong “wavefront” signal. The model with exchangeable intercept and spatial slope random effects exhibits the greatest amount of spatial smoothing among the posterior slope estimates and suggests a general spatial pattern with arms of increasing local incidence surrounding a broad area of slightly decreasing incidence in upstate New York for the time period considered.

[Figures 6 and 7 about here.]

Curiously, posterior estimates of the local slope ($\beta_1 + u_{1i}$) shown in Figure 7 for the model with spatial intercept and spatial slope random effects do not differ markedly from those in the exchangeable (non-spatial) model. To see why this may be, Figure 8 illustrates the local log incidence rates and the fitted linear models under OLS and each of the random effects models for New York. Here we see the impact of adding random effects reflected in the “pinching” of the intercept estimates, and the adjustment to local slopes. In the case of the model with spatially structured intercept and slope random effects the intercept “pinch” is severe resulting in slopes quite removed from the local
data values, resulting in what appears to be a worse visual “fit” to the data. In contrast, the model with exchangeable intercept and spatial slope random effects incorporates the spatial smoothing of slope estimates observed in Figures 6 and 7 yet maintains a fair amount of fidelity to the original data. Based on this assessment, the model with exchangeable intercepts and spatial slopes seems to provide the best qualitative fit to the data.

6. Discussion

The concepts and analyses above contribute several topics to the growing literature of risk maps for Lyme disease. First, the elements of a Lyme disease risk map outlined in Figure 2 synthesize a number of concepts raised in part by previous risk maps, and provide a conceptual mechanism for comparing existing maps and proposing new analytic models. Second, our statistical models provide insight into patterns of local reported incidence rates over the period 1990-2000. We see fairly linear changes in the local (log) incidence rates for most counties, with random variation of county-specific slopes and intercepts around central statewide rates. Local patterns include higher reported rates in the central northeastern Atlantic counties, near the area of the original identification of the disease, with continued increases over time along two “arms”, one to the north and one to the south, skirting the Adirondack Mountains in eastern New York State.

The collection of small-magnitude negative slopes in western New York raises several issues. Had a “wavefront” of reported incidence swept through the area during the study?
period, we might well expect a series of zero reported counts (or NRs) followed by a sudden increase generating steeper slopes in counties just behind the wavefront. No such pattern appears, in fact, the estimated local slopes in western New York suggest that any initial spread of disease through these counties occurred prior to the study period, since fairly stable and slightly decreasing (log) reported incidence rates appear in this area.

The relationship between incidence and NRs (“no reports”) illustrated in Figure 1 suggests a connection between the spread of the disease and the spread of reporting practices. We do not explicitly include this relationship in the models presented here, but the conceptual framework of Figure 2 coupled with the hierarchical structure of the models considered suggest modeling mechanisms for including spatially heterogeneous and perhaps spatially correlated reporting filters, provided we obtain some data on reporting accuracy. Such surveillance of disease surveillance is an area of growing interest (Buenconsejo 2004) and could be linked into the model in future work.

The results above represent our first steps toward more thorough analyses of the data, moving from summaries of observed patterns to prediction of patterns as a function of local covariates. Future work will involve such environmental and ecological covariates (e.g., percent forest cover, soil types, forest fragmentation, deer density, small mammal diversity, etc.), building on elements of the underlying disease ecology represented in the interaction between the maps on the left side of Figure 2.

More formal modeling will require more formal model choice and fit assessments. Due to the exploratory nature of the present results, the simple linear models above provide insight into fairly consistent descriptions of incidence patterns across models, with the hierarchical structure allowing ready prediction of missing reports and inclusion
of spatial correlation through the prior distributions for the random effects. The somewhat unusual behavior of the model with spatially correlated intercepts and slopes merits closer scrutiny to determine whether its origin is due to some quirk of the specific data considered or lies deeper within the structure of the model.

Acknowledgments

This research was supported in part by the National Center for Ecological Analysis and Synthesis (a center funded by NSF grant DEB-94-21535, University of California, Santa Barbara, the California Resources Agency, and the California Environmental Protection Agency) and by National Institute of Environmental Health Sciences grant R01-ES07750 (LAW). The views and opinions expressed reflect those of the authors and do not necessarily represent those of NCEAS, NSF, NIH, NIEHS, or CDC. The authors thank the many State and local health departments and the CDC for their collection and maintenance of the surveillance data and Les Real for many helpful comments and suggestions.

References


Figure 1. Maps of numbers of Lyme disease cases per 100,000 population (1990 U.S. Census) reported to the U.S. Centers for Disease Control (left) and number of years with no (county-specific) reports (right).
Figure 2. Conceptual elements of a Lyme disease risk map. Moving from left to right we find the geographic habitat structure for both hosts and vector, next the geographic distributions of various hosts and the vector, then the unobserved true incidence pattern which is “filtered” by county-specific reporting rates, yielding the observed pattern of reported locations (grey circles) and the hidden pattern of unreported locations (open circles).
Figure 3. County-specific log incidence rates by state (1990-2000). Thin lines connect county-specific annual log incidence rates with gaps representing “no reports”. Thick segments connect annual state-wide log incidence rates.
Figure 4. Ordinary least squares estimates of county-specific time trends in log incidence proportions, 1990-2000. Three counties (indicated by stripes) do not have enough reports to enable estimation of a time trend.
Figure 5. Ninety-five percent estimation intervals (confidence intervals for ordinary least squares and credible sets for the others) for each of five models fit to the 1990-2000 county-specific log incidence proportions (see text for details).
Figure 6. Posterior median estimates of the slope for models using spatially unstructured intercept and slope random effects, spatially unstructured intercept and spatially structured slope random effects, and spatially structured intercept and slope random effects, using same categories as the local OLS slopes in Figure 4).
Exchangeable intercept and slope (centered)

Exchangeable intercept
Spatial slope (centered)

Spatial intercept and slope (centered)

Legend:
-0.5 - -0.25
-0.25 - 0
0 - 0.25
0.25 - 0.5
0.5 - 0.75
Figure 7. Posterior median slope estimates, now shaded by quintiles within each map (roughly one fifth of the counties are in each interval, but interval boundaries are different between maps).
Figure 8. County-specific fitted values for New York. Grey lines connect reported log incidence proportions within a county, black lines illustrate fitted relationships based on posterior median estimates of the associated slope and intercept.
NY log rate trends: OLS

NY log rate trends: Exchangeable, midpoint

NY log rate trends: Spatial slope, midpoint

NY log rate trends: Double, midpoint