The Clustering of Infected SIV Cells in Lymphatic Tissue

Cavan REILLY, Timothy SCHACKER, Ashley T. HAASE, Steve WIETGREFE, and David KRASON

Here we investigate the clustering of simian immunodeficiency virus (SIV)–infected cells in a lymphatic tissue sample taken from a rhesus macaque to test a spatial proximity model of the spread of infection. We see that standard methods for analysis of the clustering of point processes are not entirely satisfactory for this application, so we define a novel statistic to understand the clustering in the data. This statistic examines how events spread out from certain points deemed cluster centers. Using this statistic, we can demonstrate the statistical significance of the clustering and examine over what distances this clustering is witnessed. We use Bayesian methods to fully assess the uncertainty in the estimation of this statistic by positing a model for the process. (We assume the process is a nonhomogeneous Poisson process with an intensity that is a linear combination of Gaussian densities.) We see that the distances at which clustering is present are consistent with a simple model of SIV spread within the lymph node.

KEY WORDS: Bayesian inference; Cluster analysis; point processes; Ripley's K function; SIV pathogenesis.

1. INTRODUCTION

Research over the nearly 3 decades of the AIDS epidemic has provided a clearer picture of the pathogenesis of human immunodeficiency virus (HIV), yet questions remain regarding such fundamental issues as how virus spreads and infection is propagated in the host. After the virus gains access to a host, it travels through the blood and eventually reaches other compartments. One particularly important compartment is the lymphatic system. Because AIDS is fundamentally a disease of the immune system, the mechanisms by which HIV gains access to the lymphatic tissues and the manner in which it replicates therein are crucial components to the developing picture of the pathogenesis of HIV (see Haase 1999 for a recent review).

The spread of infection is mediated by the ability of the virus to gain access to cells and its ability to use these cells for producing more virus. HIV can infect various cells, but its primary targets are activated CD4 + T lymphocytes. These cells are the most efficient hosts for HIV production. Much research has been conducted in vitro to investigate the mechanisms by which HIV gains access to a cell's reproductive machinery and exploits this machinery for its own ends, yet the dynamics of infection in vivo have been much harder to investigate. Two basic models have been proposed for the spread of HIV in the host; following Grossman, Feinberg, and Paul (1998), we call these proximal activation and transmission (PAT) and long-range transmission (LRT). The PAT model of transmission is characterized by local spread of infection from one cell to other cells in the vicinity of that cell. The LRT model maintains that infection is sustained by a cell-free virus continually infecting new cells. Grossman et al. argued that both varieties of transmission occur during the course of infection, but because LRT needs a large quantity of cell-free virus to be self-sustaining, it can only be the dominant source of transmission during the final stages of disease.

The possibility of PAT was first noticed by Tenner-Racz et al. (1988), who reported foci of infected cells in lymphatic tissue. Further evidence was obtained by Wain-Hobson and colleagues (Cheynier et al. 1994), who noticed that HIV developed into genetically distinct quasi-species in spatially distant regions of spleen tissue obtained from HIV-infected individuals. Reinhart et al. (1998) observed a similar phenomenon in a variety of tissue samples taken from rhesus macaques infected with simian immunodeficiency virus (SIV), a virus very similar to HIV in pathogenesis. The PAT model implies that one should observe clustering of infected cells in tissue samples, and whereas some have claimed to witness such clustering (e.g., Tenner-Racz et al. 1988; Grossman et al. 1998), these tissue sample images have never been submitted to statistical analysis.

To examine the spread of virus in vivo, we devised an experiment that hopefully would allow us to document the propagation of infection during the initial stages of the disease. The basic idea was to infect rhesus macaques with SIV and then examine tissue samples from these animals over the course of a few weeks. If in the early stages of infection, infection spreads according to the PAT model, then we would expect to see infected "progeny" clustered about infected "parent cells." Later we refer to the model of local spread instead of the PAT model, because we have not attempted to determine whether the infected cells are activated. (This is an important component of the PAT model as put forth by Grossman et al. 1998 about which we say more later.)

1.1 Clustering of Random Point Patterns

Assessing the extent of clustering present in a realization of a point process is a common problem. Many methods have been proposed to test the hypothesis that the dataset is a realization of a homogeneous Poisson process. Although each has its merits and shortcomings, the most widely used and supported method is based on the K function (see Ripley 1976). The estimated K function plots how many points are separated

Cavan Reilly is Assistant Professor, Division of Biostatistics; Timothy Schacker is Associate Professor, Division of Infectious Diseases; Ashley T. Haase is Regents Professor and Head, Department of Microbiology; Steve Wietgrefe is Scientist, Department of Microbiology; and David Krason is Senior Research Fellow, Division of Infectious Diseases, University of Minnesota, Minneapolis, MN 55455 (E-mail: cavanr@biostat.umn.edu). This work was supported by Great Lakes Center for AIDS research grant 1P30-CA79458-01. The authors thank Jim Neaton for helpful discussions.

Journal of the American Statistical Association
December 2002, Vol. 97, No. 460, Applications and Case Studies
DOI 10.1198/00

by a given distance against distance. If more points are separated by some distance than is compatible with the sample realization being from a homogeneous Poisson process, then one concludes that there is clustering.

For the tissue sample analyzed here, we know the date of the initial infection; hence the distances at which we find clustering are informative about the time lag between initial infection and distribution of the virus to the lymphatic tissues under the local spread model. We can make this connection because the life cycle of HIV (and SIV) is relatively well understood; the half life of productively infected T cells is 1 to 2 days (Perelson et al. 1996). Hence our goal is not simply to determine whether there is clustering in the tissue sample analyzed. We also want to assess the distances over which we find clustering in the data and to examine whether these distances are compatible with the local spread model. We see that the K function is not the most useful statistic for investigation of this aspect of the data, and we suggest another approach that tests the model more satisfactorily. This summary complements the K function in a way that is informative about the manner in which clustering is present.

1.1.1 Clustering and Models for Point Processes. Ironically, as far as we know, there have not been any applications of the usual methods of cluster analysis (and its close relative, mixture modeling) to the analysis of clustering of point processes, although there is clearly a close connection. We think that the analysis of point processes can benefit from this connection, because in the cluster analysis literature, clustering algorithms usually can be interpreted as procedures informed by model choices (see Banfield and Raftery 1993). Different algorithms lead to different cluster assignments for the data points, and the right choice of the cluster assignment rule can be determined by deciding which is the more appropriate model. Similarly, by assessing the extent of clustering for a point process in the context of a specific model, we can draw conclusions about the clustering in the data while incorporating our knowledge about the scientific context (which is incorporated in our model). The value of using a model is most readily appreciated when one realizes that there most likely is not a uniformly most powerful test of the hypothesis of complete spatial randomness, so we should construct our test to be sensitive to alternatives of substantive interest. As an added benefit, our test statistic is constructed so that departures from the null are informative about the spread of infection, whereas, as we argue later, departures from the null for the K function are not necessarily as informative. Additionally, with a specific model, we can use the tools of model selection and diagnostics to ascertain the quality of the fit. A number of other researchers have fit parametric models to two- and higher-dimensional inhomogeneous Poisson processes (see, e.g., Kooijman 1979; Lawson 1988; Ogata and Katsura 1988).

First, we explain in greater detail how the data was collected. Then we perform the standard analysis for assessing clustering. Next, we develop our new statistic and compare the results to the standard analysis. Finally, we describe a refined model-based analysis that does not use some shortcuts that we use to expedite estimation of the novel statistic. This analysis is somewhat complicated, but we describe the procedure in detail because the primary difficulties are related to fitting parametric models for inhomogeneous Poisson processes, and we think that this topic holds interest beyond the application to our specific problem.

2. DATA ACQUISITION AND STRUCTURE

Because animal models provide the easiest method for understanding the spread of disease immediately after infection, we analyzed the distribution of productively infected cells in axillary lymphatic tissue taken from a rhesus macaque infected intravaginally with SIV (the macaque equivalent to HIV) 12 days before the tissue sample was taken. The axillary lymph nodes are infected by systemic spread of SIV throughout the host's immune system. The use of SIV as an animal model for HIV is widespread because of extensive commonalities in viral replication and pathogenesis.

In these cross-sectional studies, productively infected cells had not been detected in the lymph nodes at 7 days but had been detected at the next time point of 12 days. Assuming a mean life cycle of 1.5 days would allow for three "generations" of infection in the lymph node if the virus reached the lymph nodes on about the seventh day. To identify productively infected cells, operationally defined as those with at least 20 copies of viral RNA, we hybridized a digoxygenin-labeled virus-specific probe to a 5-micron-thick section of the lymph node sample. After hybridization and washing, the viral RNA-positive cells that bound the digoxygenin-labeled probe were stained immunohistochemically with antibodies to digoxygenin. The stained cells vary in size; some are actually collections of five or six cells that are not differentiated individually due to the relatively low magnification used for this analysis.

Once the sample was on a slide under the microscope, an imaging program (Metamorph) was used to turn the image into data amenable to statistical analysis. To accomplish this, the user selects a threshold level that distinguishes stained cells from background. Then the program determines the area and centroid of viral RNA positive cells. Here we ignore the area (which measures the relative concentration of intramolecular viral RNA) and use the centroids as the locations of the diseased cells. Because the average lymphocyte has a diameter of about 10 microns, if two average-sized cells have centroids separated by about 10 microns, then the cells are basically touching one another (if we ignore the thickness of the sample).

As alluded to previously, we suppose that the infection propagates locally. To simplify the analysis, we further assume that the infection spreads to nearby cells in a spherically symmetric fashion. Under this model, we would expect to find approximate rings of infected cells formed around the parent cells in a cross-section. Actually, because a parent cell might have moved or died before the time the sample was taken, or the parent cell could be in an adjacent slice, we should not be surprised if the parent cell is not at the center of one of these rings. Because there is a considerable time lag between infection and the date the data were obtained, we actually could have the results of several generations of infected cells. If this is the case, then the spherical symmetry of the distribution of infected cells about a parent will be difficult to discern. In

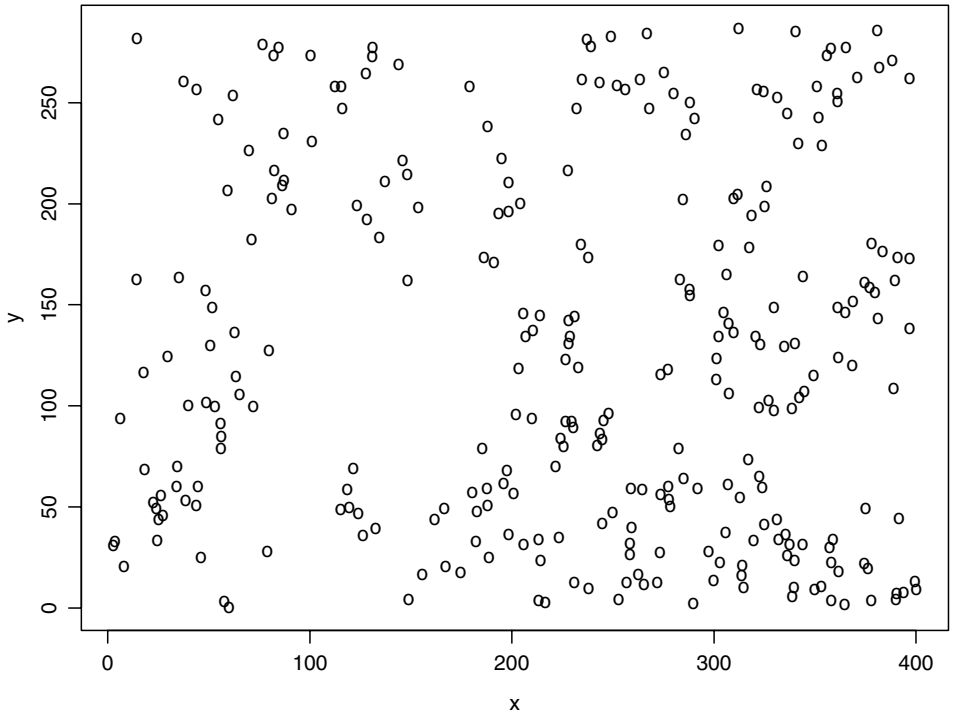

Figure 1. Location of SIV-Infected Cells. The axes are in microns.

addition, we would expect some cells to be newly infected and have not yet infected their neighbors. In short, whereas infected cells should form rings around certain parent cells under the local spread model, in actuality, observations from a tissue sample would not be expected to correspond exactly to this ideal model.

Figure 1 is a map of all events (the centroids of the stained cells). The dots represent infected cells. A cursory examination suggests clustering of the infected cells; moreover the shapes of these clusters do not appear greatly at odds with the spherical symmetry assumption.

3. ANALYSIS WITH THE *K* FUNCTION

To statistically investigate clustering, it is standard practice to estimate the K function. The K function is defined as

$$K(d) = \lambda^{-1}\mathrm{E}(\text{number of events within distance } d \text{ of an arbitrary event}),$$

where λ is the intensity of the process and the expectation is with respect to the measure induced by the point process; furthermore, $K(0) = 0$ rather than $K(0) = 1$. If the process is a homogeneous Poisson process in the plane, $K(d) = \pi d^2$, and if the process exhibits clustering at some distance, d^*, then $K(d) > \pi d^2$ for $d \geq d^*$. An easier function to visually interpret is $L(d) = \sqrt{K(d)/\pi}$, because $L(d) = d$ for a homogeneous Poisson process. If the intensity or the expectation is not constant over the sampled region (so that λ and the expectation are a function of location), then interpretation of K is difficult. For these reasons, if the K function indicates that the process is not a homogeneous Poisson process, then the K function may not be the most useful data summary.

Given a realization of a point process with N points, $s_1, \ldots, s_N$, where s_i is a point in the sample region A (later we use s to denote the collection of all these sites), it is standard to estimate $K(d)$ by

$$\widehat{K}(d) = \hat{\lambda}^{-1} \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} 1_{\{\|s_i - s_j\| \leq d\}} / N,$$

where $\hat{\lambda} = \frac{N}{|A|}$ with $|A|$ = the area of the region over which we have our sample. There has been considerable interest in examing the effects of edges, and although these can be substantial, we ignore this because we look at the K function over distances that are rather small compared to the size of the sample region. Monte Carlo is usually used to obtain an upper and lower bound for $\widehat{K}(d)$ under the assumption that the process is a homogeneous Poisson process, and these limits are then used to asses statistical significance of departures of $\widehat{K}(d)$ from πd^2 (see, e.g., Cressie 1993).

3.1 The Derivative of the *K* Function

One feature of the K function (and L) that makes it difficult to interpret is its cumulative nature. This poses a problem, because if there is clustering at some small distance, this raises the level of the K function (and L function) for all distances greater than this small distance, although the effect decays as distance increases. The result is that one must try to compare the slope of the L function to a $y = x$ line to determine whether there is clustering at higher distances as well. (Clearly, the problem is even worse for the K function.) Much like the practice of working with probability density functions rather than working with cumulative distribution functions, we prefer to work with the derivative of $L(h)$, which

is closely related to the pair-correlation function. (See Stoyan, Kendall, and Mecke 1995 for more on this function and other summaries of the second-order behavior of point processes.) We would expect $L'(h)$ to be 1 if the sample were from a homogeneous Poisson process. Here we simply estimate this derivative with discrete differences from binned estimates of K (using 100 bins), although we can imagine smoothed estimates that have lower mean integrated squared error and more refined methods for determining the number of bins. But we do not try to develop such estimates here.

3.2 Shortcomings of the K Function for the Current Application

Although the K function (and quantities derived from it) is certainly useful and informative, the continued reuse of the data can introduce misleading features for the current application. Suppose that the infection propagates locally in a spherical pattern infecting some constant proportion of all susceptible cells, as mentioned in Section 2. For concreteness, suppose that six cells are infected and form an approximate ring with a radius of 10 microns around the parent cell. Such a configuration would give rise to six pairs of cells separated by a distance of 10 microns from the parent. If these infected cells are spread out uniformly around the parent cell, then we would also have three pairs of infected cells separated by a distance of approximately 20 microns. If this pattern of contagion was present throughout the entire sampled region, then one could conclude that there is clustering at distances of 10 and 20 microns, although the clustering at 10 microns is the aspect of the data that is informative about the spread of infection (if we already assume the spread is in a spherically symmetric pattern). Furthermore, if the parent cell died before collection of the sample, then we would see the clustering only at 20 microns.

Figure 2(a) illustrates a hypothetical example of the sort of data that we could obtain from the spatial proximity model of viral spread, and 2(b) depicts the associated L function. This point pattern was generated by setting points at $(2, 2)$, $(2, -2)$, $(-2, 2)$, $(-2, -2)$, then generating 50 draws from a homogeneous Poisson process over annuli with inner radius .6 and outer radius 1.0 situated about these four parent points. We would like to find clustering at a distance of about .8, but this is completely lost due to the structure of the data. Instead, we witness clustering at distances less than .4 (which is sensible given the width of the annuli), a little clustering in the 1.5–1.8 range, and then a considerable drop off (spatial inhibition, due to the separation of the four rings) from 1.8 to 2.5. Ironically, we conclude there is clustering at short distances, but the only points separated by such short distances are actually drawn from a homogeneous Poisson process. This last phenomenon is due to the role of the estimate of λ for the process; the estimate of the intensity is too low for the points separated by short distances, because there is a great deal of area with no events taking place outside of the annuli. Hence there appears to be clustering at these distances relative to what occurs outside of the annuli. Whereas the K function is useful because it investigates clustering at all distances simultaneously, this same feature can make it misleading as a data summary if there is clustering at some distance.

3.3 The K Function for Our Dataset

If the virus gained access to this lymph node on the first day of infection and replicated there continuously with each generation replicating in 2 days, then after 12 days we would have clusters spreading out about 60 microns (because the average lymphocyte has a diameter of about 10 microns). Therefore, we do not expect any clustering beyond 60 microns, and hence we estimate $L'(h)$ up to 62 microns for our sample. The result is shown in Figure 3. Similar conclusions are reached if one uses the K function. Also shown in Figure 3 are 95% confidence bands (obtained via simulation) for $L'(h)$ under the assumption that our process is a realization from a homogeneous Poisson process. It appears that there is clustering in the 1.9–2.5 micron, 3.1–3.7 micron, 14.8–15.5 micron, and 25.3–26 micron ranges, with borderline results for the 7.4–8 micron, 8.7–9.3 micron, and 39.6–40.2 micron ranges.

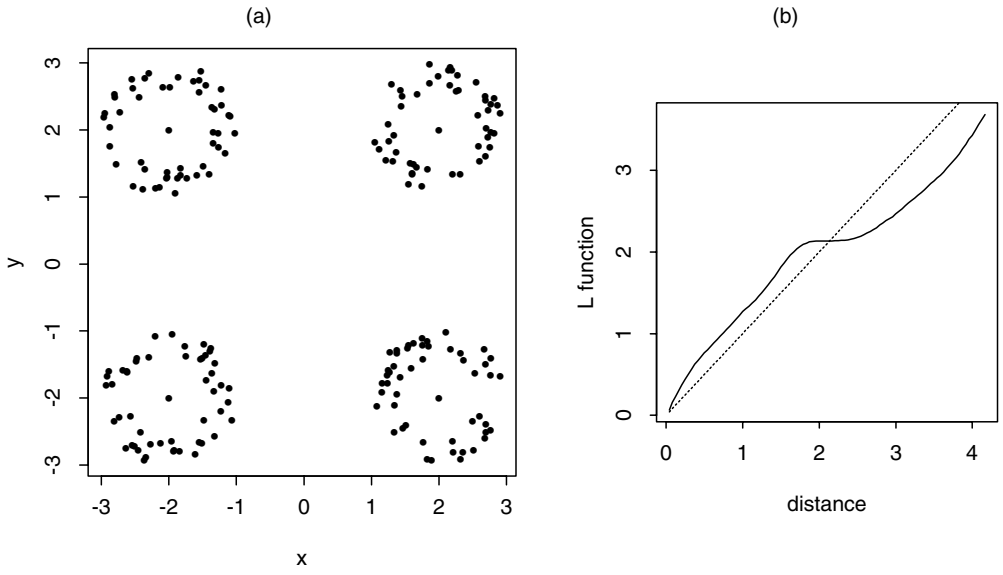

Figure 2. A Hypothetical Example Point Process (a) and Its L Function (b). Although the example is extreme, we see that the L function is potentially misleading when the process is not a homogeneous Poisson process.

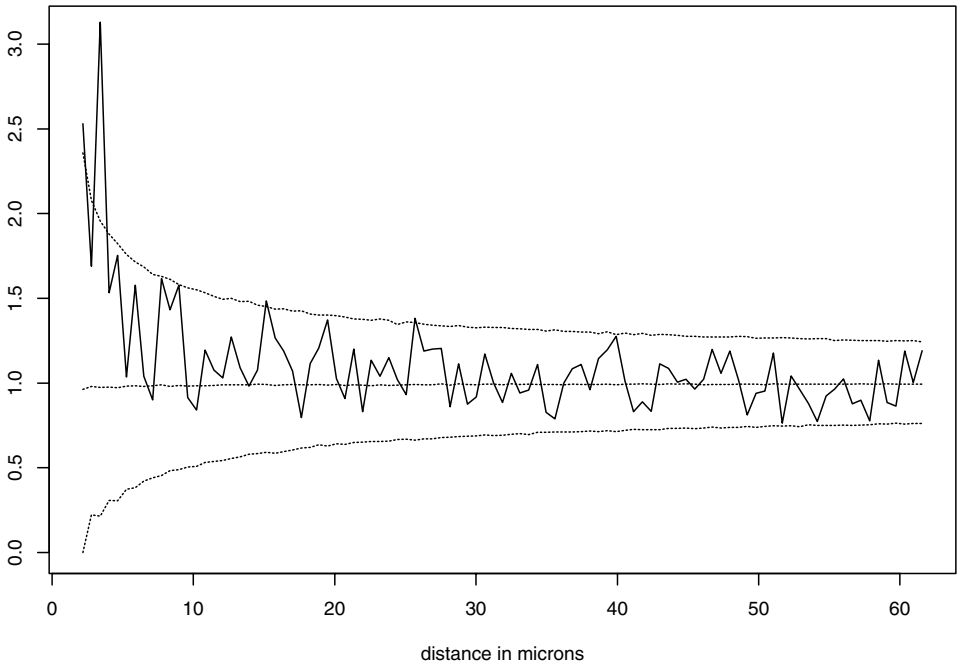

Figure 3. Derivative of the L Function for Our Dataset With 95% Confidence Bands (· · · · · ·) Under the Assumption That the Process is a Homogeneous Poisson Process. There appears to be clustering at short distances.

Clearly, we can not conclude that there is significant clustering at all of these distances (because this would not control the overall type I error), but it does appear that there is some clustering at the very short distances. Given that the borderline clustering occurs roughly at distances $8 * i$ for $i = 1, 2, 3, 5$, one may wonder whether these bumps are due to the phenomenon illustrated in the previous section.

4. AN ALTERNATIVE TO THE *K* FUNCTION FOR ASSESSING CLUSTERING

Rather than examine how all of the points are separated from one another, we really would like to examine how points spread out from the parent cells. This leads us to examine a different function,

$$H_u^*(d) = \text{number of clusters} \times \text{E(number of events separated by at least } d-\delta\text{, but no more than } d+\delta \text{ from the cluster center within an arbitrary cluster)},$$

where the expectation is with respect to the measure induced by the point process. The greatest difference between H_u^* and K is the designation of certain points in the sample region as cluster centers. (Such points need not actually be sites where there was an event.) A plot of $H_u^*(d)$ by d [for $d = \delta(2i-1)$ where $i = 1, \ldots, I$)], indicates at what distances clustering occurs and at what distances this clustering drops off. Also note that each point gets used only once in the construction of H_u^*; that is, when we estimate H_u^*, we assign points to clusters, find the distance from the cluster center to each point in that cluster, then sum over all clusters. Rather than work directly with $H_u^*(d)$, we work with a normalized version, $H^*(d)$, where the normalization ensures $\mathrm{E}H^*(d) = 0$ and $\mathrm{var}\, H^*(d) = 1$ for all d if the process is a homogeneous Poisson process. (The expectation is with respect to the measure induced on A under the homogeneous Poisson assumption conditional on the cluster centers and assignments.) Finally, we note δ is a user-defined tolerance that should be selected with reference to the application.

An approach similar in spirit to our proposed method has been developed by Besag and Newell (1991) in the spatial epidemiology literature. In that approach, one determines how far a disk must be extended around each point until a certain predetermined number of events, k, are contained in the disk. Although Besag and Newell have demonstrated the usefulness of their method, we prefer an approach that does not require the specification of the number of events in a cluster, k. In addition, our method does not require that an event be at the center of a cluster (although we do enforce this requirement for the simple estimate of the next section), and Besag and Newell recognize this as a weakness of their method.

Finally we note that there is a connection between the H^* function and Ripley's K function. Consider the multivariate point process obtained by joining the "offspring" process with the "parent" process. This parent process is a hidden, unobservable point process, whereas the observed data constitute a realization from the offspring process. In this context, the H^* function for distances smaller than the minimal distance between a cluster center and another point in any other cluster is nothing more than the cross-K function for this multivariate process (for the cross-K function, see, e.g., Cressie 1993). For larger distances, the cross-K function would also consider distances between cluster centers and points in other clusters, but the H^* function excludes these distances.

4.1 A Simple Method of Estimating H^*

A simple method of estimating H^* is as follows. Use a hierarchic clustering algorithm on the coordinates of the sampled points that assumes that the clusters are spherical but of

variable size (because we expect the clusters to be of this form) (see, e.g., Banfield and Raftery 1993). We determine the number of clusters using the approximate Bayes factor method of Banfield and Raftery (1993). Next, designate the point that minimizes the average distance to all other points within each cluster as the center of that cluster. (To allow for cluster centers that are not locations of events, one could use, for example, the centroid of the convex hull of the cluster.) Finally, find the distances within clusters from cluster centers to each point in the cluster and sum over all clusters to obtain Y_i for $i = 1, \ldots, I$. That is, if we let μ_k represent the cluster centers for $k = 1, \ldots, K$, and

$$a_{ki} = \{s \in A : \| s - \mu_k \| \in [d_i - \delta, d_i + \delta)\},$$

where $d_i = \delta(2i - 1)$ for $i = 1, \ldots, I$, then

$$Y_i = \sum_{j=1}^{N} 1_{\{s_j \in \bigcup_{k=1}^{K} a_{ki}\}}.$$

This yields an estimate of $H_u^*(d)$ for $d = \delta(2i - 1)$. Next, normalize the Y_is so that they have zero mean and unit variance under the assumption that we have a realization of a homogeneous Poisson process. Because $Y_i \sim \text{Poi}(\lambda_i)$ with $\lambda_i = \lambda |\bigcup_{k=1}^{K} a_{ki}|$, where λ is intensity of the process over the entire region A, this standardization is straightforward once we estimate λ with $N/|A|$. For now, we use the further simplification that $|\bigcup_{k=1}^{K} a_{ki}| = K|a_{1i}| = 4\pi K\delta^2(2i - 1)$, which is correct only if the annuli a_{ki} do not overlap and are strictly within the boundaries of A. In general, these annuli can overlap and intersect the boundary, making the exact calculation of $|\bigcup_{k=1}^{K} a_{ki}|$ more difficult, but at this point we ignore this potential complication. The normalized variables are then

$$Z_i = \frac{Y_i - \lambda_i}{\sqrt{\lambda_i}}.$$

We then plot Z_i by d_i to obtain our estimate of H^*.

We can assess the statistical significance of any departure of the set of Z_i from 0 as follows. Under the hypothesis that we have a realization from a homogeneous Poisson process, all of the Z_i's have zero mean, unit variance, and, provided that all of the cluster centers are sufficiently far apart, they are also independent. Hence we define the test statistic

$$T = \max_{1 \le i \le I} Z_i.$$

We then have

$$P_{H_0}\{T \le t\} = \prod_{i=1}^{I} \sum_{j=0}^{\lfloor \lambda_i + t\lambda_i^{\frac{1}{2}} \rfloor} \frac{\lambda_i^j}{j!} e^{-\lambda_i},$$

and so we can easily calculate the value of the maximum of H^*, which allows us to reject the null hypothesis of complete spatial randomness with any desired α level.

4.1.1 Application to Our Dataset. Figure 4 depicts the H^* function along with a horizontal line indicating the cutoff for statistical significance at the .05 level. This plot indicates that H^* exceeds the cutoff level for the maximum at distances in the 7.4–8.7 micron range and in the 14.9–16.1 micron range (which is approximately one and two cell diameters). By computing the joint distribution of the two largest values of H^* under the null hypothesis (which is elementary because the values of H^* are approximately independent although not identically distributed), we find that the p value associated with the largest and second largest values of

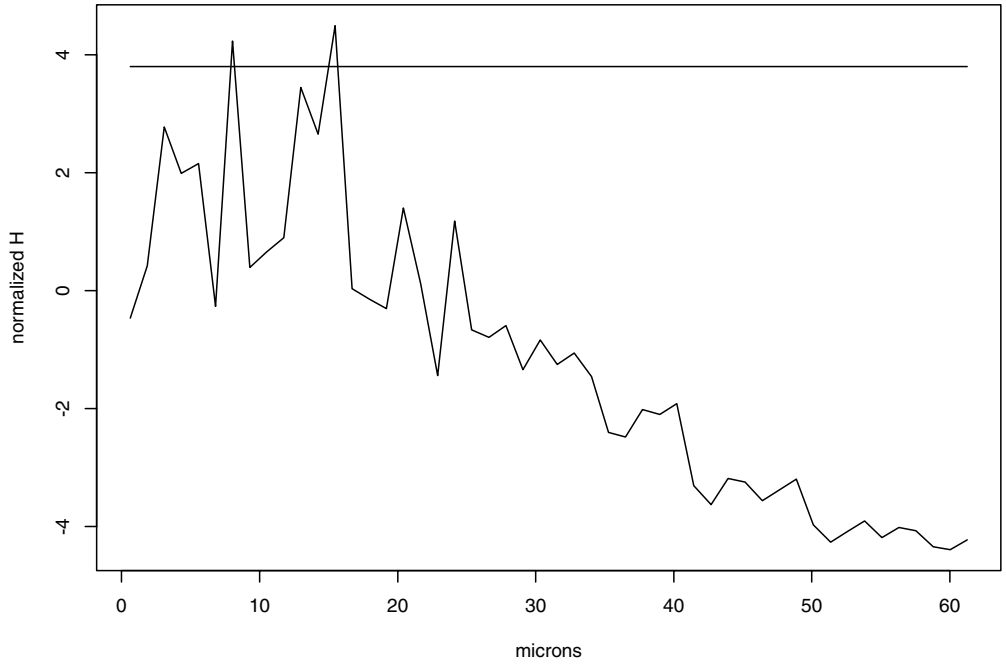

Figure 4. The H^* Function. The horizontal line indicates the cutoff level for the maximum for statistical significance at the .05 level. Note the clustering at distances of 8 and 15 microns.

H^* (simultaneously) is less than .001 (one can also use the parametric bootstrap). Furthermore, the clustering drops off at distances greater than about 30 or 40 microns. These findings are consistent with the life cycle of HIV and infection of a "parent" and two generations of progeny. As mentioned in Section 2, these findings are what we expect under the spatial proximity model.

4.2 Why the Simple Method Needs Refinement

Although the previous method is simple to implement, it has a number of shortcomings. First, the hierarchic clustering methods that we use are not appropriate for the data. The clustering algorithms available in commercial statistical software packages typically assume that the measurements (i.e., the coordinates of the points) are from a multivariate normal distribution, and those that do not explicitly make this assumption can usually be construed as making some sort of assumption along these lines (see Banfield and Raftery 1993). This assumption is also used for selecting the number of clusters. A more reasonable model than assuming that the coordinates are normal mixture deviates is to assume that our data are a realization of an inhomogeneous Poisson process. Later we present a method for clustering in this context.

Another problem with this method is the test is conditional on the output of the clustering algorithm. Because we use the data to do the clustering, there is uncertainty in the cluster assignments and the locations of the cluster centers; therefore, our p value needs to be adjusted to include this uncertainty.

Finally, the previous algorithm assumes that the cluster center is in the sample. As noted before, this is not necessarily the case. The parent cell could have moved, died, or be in an adjacent slice of tissue.

5. A BAYESIAN ALTERNATIVE TO H^*

If one treats the cluster centers as unknown parameters, then one is faced with the problem that with a single realization of the process (as is necessary due to the destructive nature of the sampling), there is not really any replication in the data that can be used to directly determine the location of these cluster centers. For this reason, we assume that our sample is a realization of an inhomogeneous Poisson process, and we parameterize this process in a manner that allows us to exploit replication present due to the parameterization. This modeling strategy is rooted in the observation (first reported in Bartlett 1964) that from a single realization of a point process, one can not distinguish inhomogeneity from clustering. In particular, we suppose that the intensity of the process is given by a linear combination of bivariate Gaussian densities with zero correlation and identical standard deviations in both directions (circularly symmetric),

$$\lambda(x, y) = \sum_{k=1}^{K} \frac{\omega_k}{\sigma_k^2} \exp\left\{-\frac{1}{2\sigma_k^2}\left[(x-\mu_{xk})^2 + (y-\mu_{yk})^2\right]\right\}.$$

Hence there are $4K$ parameters if there are K clusters. With this parameterization, the spread of infected cells about the cluster centers is informative about the location of the cluster centers.

Given values for all of these parameters, we then treat the points (μ_{xk}, μ_{yk}) as the cluster centers and assign points to clusters based on which cluster center to which they are closest. This clustering algorithm is basically a parametric version of the K-means approach to clustering (see MacQueen 1967). Finally, given the cluster centers and the cluster assignments, we construct the H^* function as before. To indicate that H^* depends on a vector of parameters, $\theta = (\omega_1, \sigma_1, \mu_{x1}, \mu_{y1}, \ldots, \omega_K, \sigma_K, \mu_{xK}, \mu_{yK})$, we now denote it by $H^*(d; \theta)$.

If we treat parameter estimation in a Bayesian fashion, then we can incorporate uncertainty about the locations of the cluster centers and the cluster assignments into the construction of the H^* function. Let $p(\theta|s)$ denote the posterior distribution of the vector θ given the collection of observed sites s. We then define the H function via

$$H(d) = \int H^*(d; \theta) p(\theta|s)\, d\theta.$$

Readers familiar with the theory of spatial point processes may wonder why we do not use one of the more common point process models that allow for clustering, such as the Poisson cluster process. First, because the likelihood for the Poisson cluster process model is not useful for computations (see, e.g., Ripley 1988), one must fit these models by matching the K function for the model to the observed K function. But because the K function does not uniquely determine the point process (for an example, see Baddeley and Silverman 1984), this is a rather questionable practice. Second, if we fit one of these models, then we cannot determine the locations of the centers of the clusters, but can determine only the number of clusters and the properties of these clusters (such as, e.g., the radii of the clusters), and this information is not sufficient to calculate H^*. Finally, we suspect that a very wide range of spatial point processes can be approximated by inhomogeneous Poisson processes; hence there is little lost in using this family, provided that the scientific context suggests a suitable parametric form for the intensity, as here.

5.1 Application to the Lymph Node Sample

Figure 5 displays a representation of the posterior mode of the intensity when there are seven clusters (We discuss choosing the number of clusters in Sec. 6.2.) The two closest cluster centers at the posterior mode are separated by 105 microns. Figure 6 shows the cluster assignments made using the algorithm outlined earlier based on the posterior mode for the intensity. Figure 7 shows the H function for the data along with 95% pointwise credible intervals for the values in the bins. The posterior probability that H is greater than 0 for the second bin is .978; hence this provides substantial evidence that there is clustering in the 6–12 micron range. In addition, the probability that the H function is greater than zero in the 18–24 micron range is .971, and the probability that H is positive at both distances (simultaneously) is .95. Note that we find clustering at the distances of approximately one and two cell diameters (but not at the intermediate distances). These findings are consistent with the simplified estimate in Section 4.1.1, and they are consistent with what we would expect to find under the spatial proximity model of viral spread.

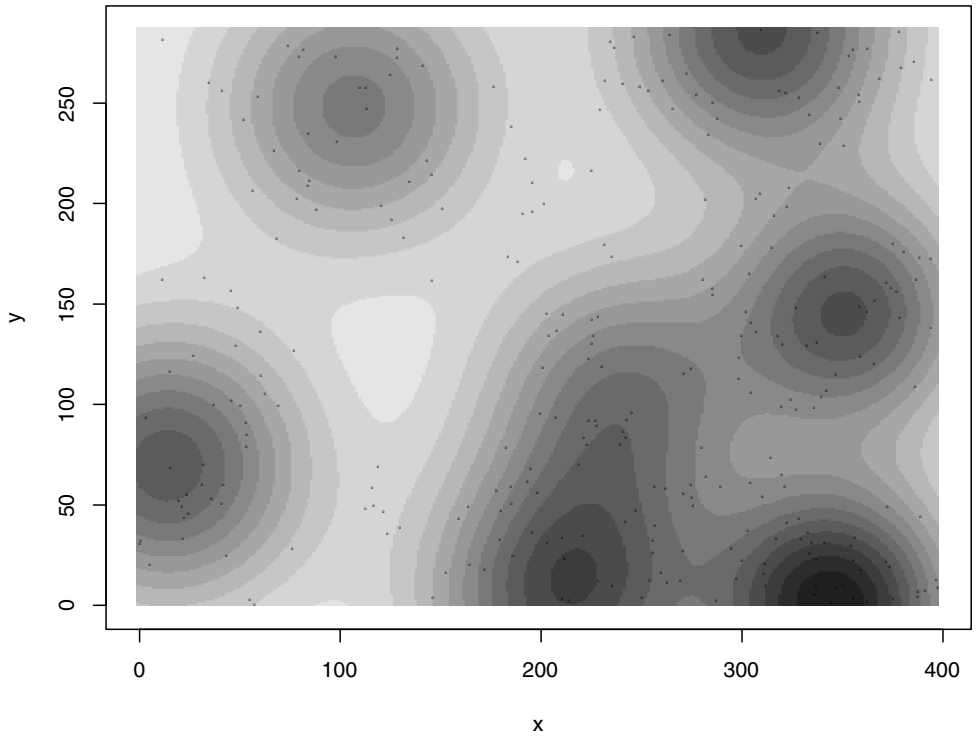

Figure 5. A Representation of the Intensity of the Inhomogeneous Poisson Process at the Posterior Mode Along With the Locations of SIV RNA–Positive Cells.

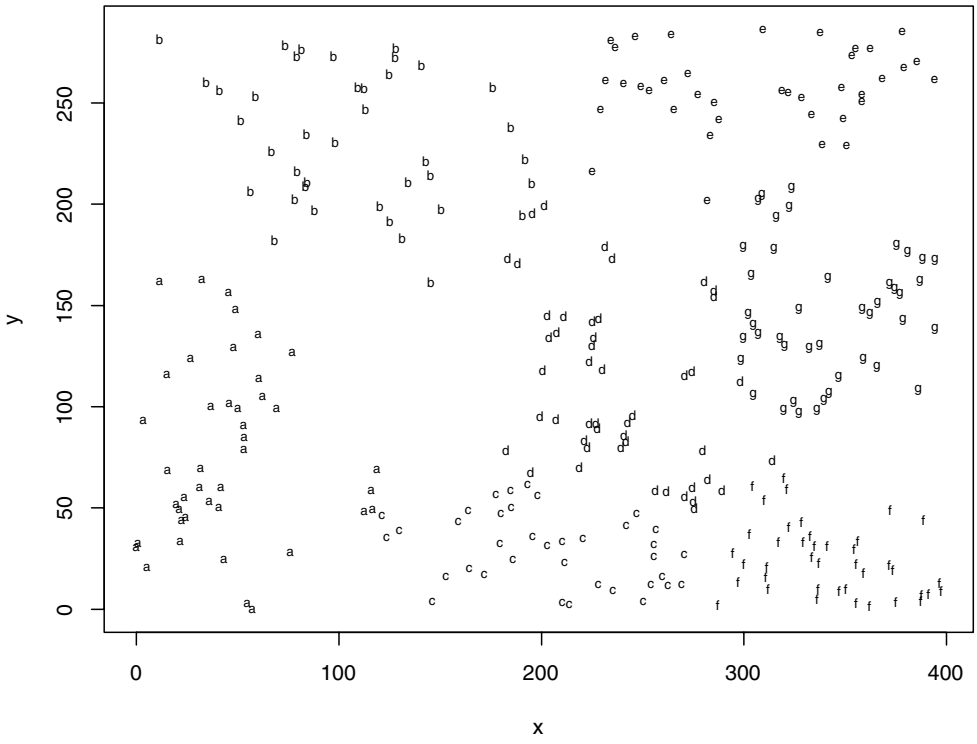

Figure 6. The Posterior Mode of the Cluster Assignments. Each letter represents one of seven clusters.

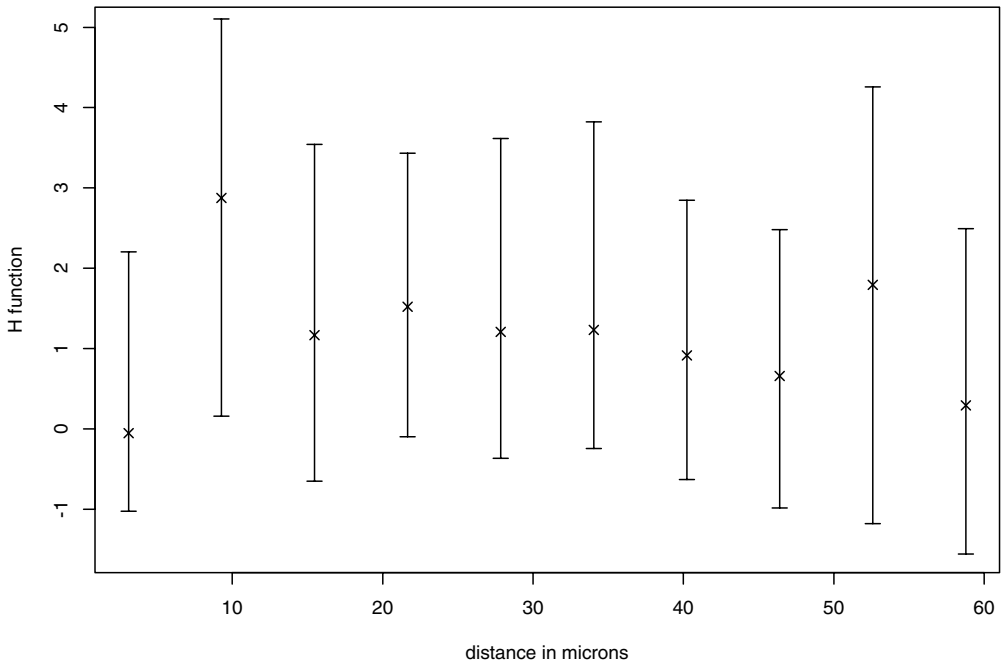

Figure 7. The H Function Averaging Over the Uncertainty in the Clustering and Cluster Centers. The bars represent the extent of 95% probability intervals.

6. COMPUTATIONAL ASPECTS OF CONSTRUCTION OF THE BAYESIAN ALTERNATIVE

We use simulation to calculate the H function. We first obtain draws from the posterior distribution $p(\theta|s)$, then calculate the H^* function for each draw. This provides samples from the posterior distribution of the H function. Obtaining simulations from the posterior $p(\theta|s)$ is somewhat complicated. For the likelihood portion of the posterior, as is usually done, we condition on the total number of points in the sample and use the Janossy density as our likelihood, $L(\theta)$. Hence

$$L(\theta) = \exp\Big\{-\int_A \lambda(s)\,ds\Big\}\prod_{i=1}^{N}\lambda(s_i),$$

where the integral is over the sample region, A, in the plane (see Daley and Vere-Jones 1988 for details on this likelihood). Given the smooth form of λ, Gaussian quadrature is the method of choice for carrying out this numeric integration.

6.1 Priors for the Model Parameters

Unfortunately, much as is the case for fitting mixture models, this likelihood is unbounded (see, e.g., Titterington, Smith, and Makov 1985). For this reason, we must provide some sort of prior structure for the parameters. Although we do not need informative priors for the locations of the cluster centers (i.e., we suppose that these priors are uniform over the sampled region), we must constrain the weights, ω_k, and the standard deviations, σ_k, so that they do not drift to infinity or 0. Rather than specifying priors for the values of these parameters (as done in Gelman and King 1990), we specify priors for the ratios of the largest weight to the smallest weight and a prior for the ratio of the largest standard deviation to the smallest standard deviation (which is in the spirit of Box and Tiao 1968, Gelman and Rubin 1992, and Belin and Rubin 1995, although in all of these analyses there were only two components and the ratio of the variances of the components was fixed). Furthermore, we take these priors to be independent gamma distributions,

$$\max_{i,j}\frac{\omega_i}{\omega_j} - 1 \sim \text{gam}(.7, 7.0)$$

and

$$\max_{i,j}\frac{\sigma_i}{\sigma_j} - 1 \sim \text{gam}(2.3, 23.0).$$

These priors assume that we expect both of these ratios to be 1.1, but the variance of the ratio of standard deviations is much smaller (by about a factor of 10) than the variance of the ratio of the weights. In practice, we find that we can specify weaker priors when we have a few clusters but must specify stronger priors as the number of clusters increases to prevent the standard deviations from drifting to 0. Hence we use priors sufficiently strong to obtain acceptable results for up to 12 clusters. In this way we use the same prior structure for all of the models when we select the number of clusters. It transpires that for the number of clusters we finally select (namely, seven), one can specify very vague priors (priors too vague to be useful when there are 12 clusters). Nonetheless, our final model with seven clusters uses the priors defined earlier.

6.2 Choosing the Number of Clusters

Although it is tempting to treat the number of clusters as a parameter and estimate it from the data, this is not identifiable as a parameter (jointly with θ) without a prior on the number of clusters (an approach that we do not favor) or some sort of prior on the cluster centers that specifies that they cannot

get too close. One may think that because the lymphocytes have an average diameter of 10 microns, cluster centers can not be closer than 10 microns. But in fact the tissue sample is 5 microns thick; hence cells can be closer because they do not both have centers in a plane parallel to the cross-section. Thus we do not favor putting a prior on the minimum distance between cluster centers. Rather than take either of these approaches, we adopt the method used by Gelman, Carlin, Stern, and Rubin (1995)—namely, we use as few clusters as is possible to make certain features of the posterior predictive distribution match the same features of the posterior distribution.

To this end, we define a set of realized residuals by dividing the rectangular region into an R by R array of equal-sized rectangles then counting how many sample points, s_i, are in each rectangle. This yields a collection of R^2 counts x_i. Next, we transform these counts into standardized residuals, as follows. Given a value for the parameter vector, θ, we integrate the intensity over each rectangle to get R^2 ξ_i's. We then standardize the counts with the integrated intensities to get a set of approximately standardized realized residuals for each parameter draw $r_i = (x_i - \xi_i)/\sqrt{\xi_i}$. (This is only an approximate standardization, because we have conditioned on N.) These residuals depend on θ, so we can simulate their posterior distribution by taking draws from the posterior of θ, say θ^ℓ, for $\ell = 1, \ldots, L$, and then calculating the residuals, r_i^ℓ, for all L. To obtain a similar set of residuals from the posterior predictive distribution, we first simulate a draw from the inhomogeneous Poisson process with intensity $\lambda(\theta^\ell)$ to obtain $\hat{s}^\ell$, and then compute the residuals as if the simulated data $\hat{s}^\ell$ were actual data (using the same θ^ℓ as the parameter vector when we compute the ξ_i's). We then compare the sets of residuals to determine whether the number of clusters is adequate. Note that for each parameter vector we get a set of realized residuals from the observations and a set of posterior predictive residuals (using the simulated data based on the parameter draws).

The choice of R is important for our ability to differentiate between models. If we choose this value too low, then we will miss differences on small scales, whereas if we choose it too large, then we find that all of the counts are 0s and 1s, which will not be very informative unless we consider the spatial distribution of these residuals. Because we prefer simple measures of model misfit (e.g., the maximum of the absolute value of the standardized residuals), we choose R so that the counts, x_i, range from about 5 to 10. If we set $R = 10$, then we achieve this aim.

After experimenting with several measures of misfit, we find that the maximum of the absolute value of the residuals appears to be the most useful approach for model discrimination with the range of number of clusters considered here (i.e., 4–12 clusters). As the number of clusters rises above 12, the clustering algorithm starts finding clusters with only several events in them; hence this set an upper bound on the number of clusters that we allowed. The method of Banfield and Raftery used for constructing H^* found 11 clusters, some of which were quite sparse. Figure 8 displays scatterplots of the realized residuals against the residuals under the posterior predictive distribution for several different numbers of clusters. We always find that the maximum of the absolute value of the realized residuals is on average larger than the maximum of the absolute value of the residuals in the posterior predictive distribution (as is expected), but the probability that the maximum of the absolute value of the realized residuals is larger than this statistic under the posterior predictive distribution varies with the number of clusters. If we have four clusters, then the probability that the maximum of the absolute value of the realized residuals is greater than the maximum of the absolute value of the posterior predictive residuals is .013. With six clusters, this probability rises to .023; with seven clusters, it becomes .091. As we increase the number of clusters, this probability always exceeds .05, but it never rises above .2 for the models that we considered. Because seven clusters gives a reasonable fit by this criterion, and because increasing the number of clusters above this number did not improve the fit substantially, we use the seven-cluster model for the purpose of evaluating the H function. Whereas the choice of the number of clusters does influence the exact values of various posterior probabilities, the qualitative features of H are not unduly sensitive to this choice.

6.3 Computational Details

We use the Metropolis algorithm to obtain simulations from the posterior distribution of the parameters, $p(\theta|s)$. The details and complications are much like those of fitting mixture models (see e.g., Gelman et al. 1995), although here there is another difficulty because we are fitting bivariate mixtures. When implementing Markov chain Monte Carlo for unidimensional mixtures, one must use some sort of convention to uniquely identify the different components of the parameter vector. For example, consider a two-component univariate Gaussian mixture model where each component is characterized by a mean and standard deviation. If θ_1 is the mean for component 1 and θ_2 is the mean for component 2, then one can uniquely identify the vector θ by requiring $\theta_1 \leq \theta_2$. If one fails to implement this constraint, then the Markov chain will travel back and forth between two (equivalent) modes, one mode where $\theta_1 < \theta_2$ and the other where $\theta_2 < \theta_1$. This can cause difficulties in terms of the chain converging to its equilibrium distribution, but these difficulties are purely for semantic reasons. There is a similar problem with bivariate mixtures, but the effects are harder to overcome because the plane cannot be ordered like the real line. That is, to identify the vector θ, we must use some constraint—for example, that the x-components of the cluster centers are increasing. The problem is that if two clusters have nearly identical x-components but widely separated y-components, then there will be two modes in the parameter space that are nearly equivalent as the y-components switch. In such a case, the marginal distribution of the y-components of the cluster centers will be bimodal. If one attempts to identify θ with a constraint on, say, the standard deviations of the clusters then the problem is even worse, because then both the x and y components can be bimodal. In practice, our Metropolis algorithm (which uses a multivariate normal jumping kernel with covariance matrix given by the negative inverse of the observed information at the posterior mode scaled to obtain a 20%–30% acceptance rate) moves fluidly between these modes, but long run times (in terms of

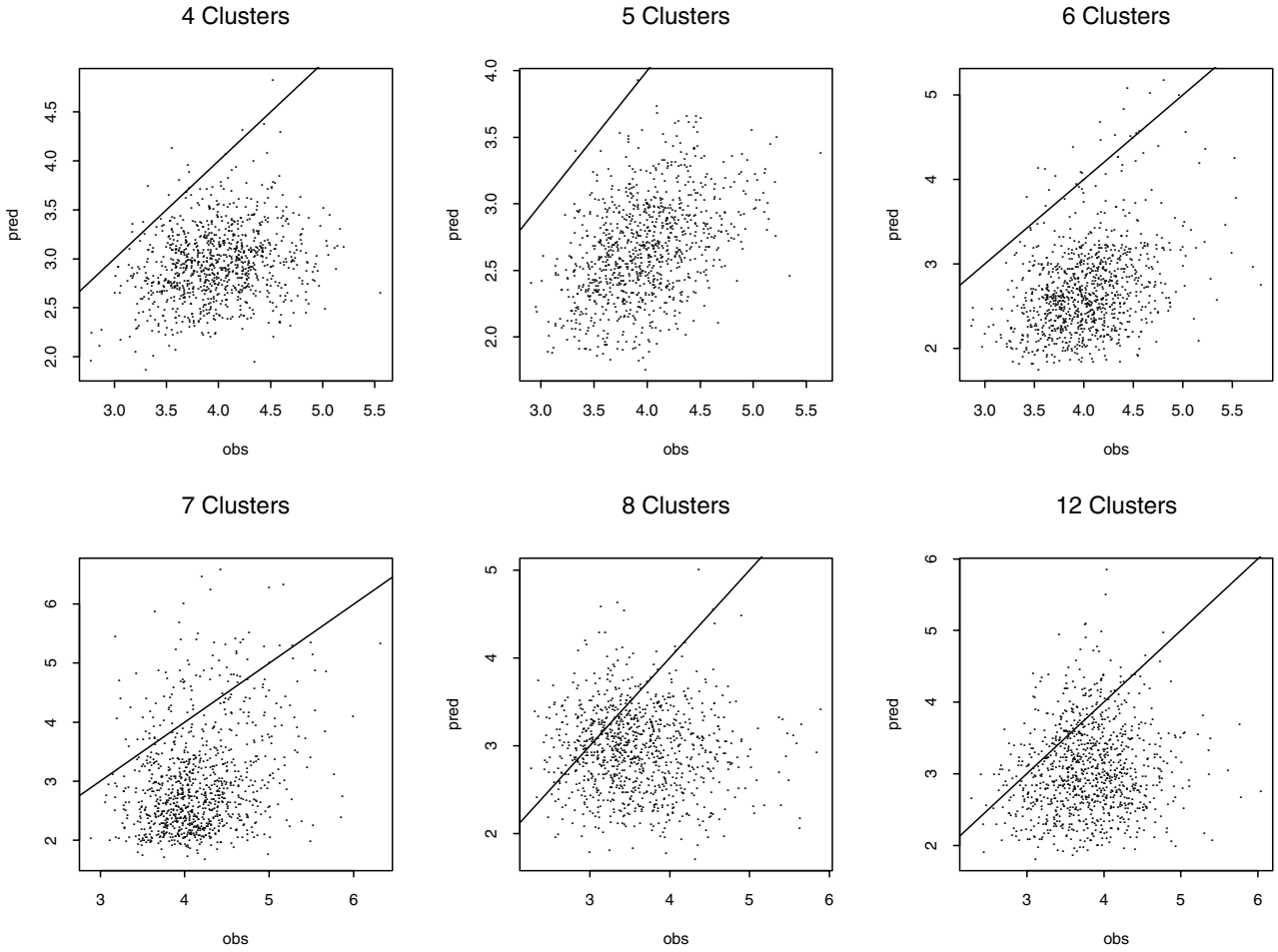

Figure 8. Comparison of the Maximum of the Absolute Value of the Standardized Residuals Under the Posterior Distribution to the Posterior Predictive Distribution of this Statistic for Six Different Numbers of Clusters.

the number of iterations) are necessary to get the Gelman–Rubin $\sqrt{\widehat{R}}$ diagnostic statistics below 1.2. (See Gelman and Rubin 1992 for diagnosing convergence of Markov chains.) For K in the range of 5–10, we used 8 chains and found that typically 100,000 iterations were sufficient to obtain convergence. For models with more than this number of clusters, even more iterations were necessary to obtain convergence (e.g., with $K = 12$, we used 4 chains and 200,000 iterations to get $\sqrt{\widehat{R}} \leq 1.2$ for all of the parameters). Although this may seem unbearable to those accustomed to having a substantial matrix inversion to carry out for each iteration, the calculations are not excessively time-consuming, because the posterior is easy to evaluate.

Given samples from $p(\theta|s)$, it is relatively straightforward to find the normalized version of $H(d)$, but we must attend a slight detail. As mentioned previously, when we standardize, we must calculate what we think $H(d)$ would be had we a homogeneous Poisson process. For the simple method, we simply used the area of the annuli and the total number of points per unit area, but there is the possibility that these annuli overlap and intersect the boundary. Therefore, for each draw from the posterior, we must determine the area covered by the annuli that we set out around the cluster centers. Because these annuli may overlap in arbitrarily complicated ways, finding this area is not so straightforward. Our solution is to use Monte Carlo integration. That is, we define a function over our sampled region that is 1 if the point is in some annulus for a given distance and 0 otherwise, then draw many points (we used 100,000) uniformly over the sampled region and determine what percentage of these result in our function returning 1. Note that this area depends on the locations of the cluster centers and the distance at which we are estimating H; hence we must repeat this calculation for each parameter draw and distance (making the normalization somewhat computationally intensive). One could certainly add this refinement to the construction of the H^* function, but this would detract from its simplicity.

7. DISCUSSION

Although the K function is useful for determining whether a sample realization is from a homogeneous Poisson process, if the process is of a different form, then we should use other functions of the data as summaries. Whereas the $H(d)$ function presented here is useful for the current application, we see this not as an alternative to the standard methodology, but rather as a supplement. We expect that different applications will suggest different summaries for the exploration of point process data. Although the Bayesian version of the H function

is more defensible (because it integrates over all of the uncertainty in the model specification), the simplified version is useful because it is much easier to compute. For routine analyses, the simplified estimate is most likely adequate.

Although the emphasis in this article is on the H function, we also think that modeling point processes with inhomogeneous Poisson processes with intensities that are linear combinations of Gaussian densities is a useful general strategy for analyzing point process data. Just how general such a parameterization is remains an open question. For some applications, more parsimonious parameterizations may be achieved by using functions other than Gaussian densities. We used the Gaussian density for two reasons: the assumed spherical symmetry of the spread of infection and the desire to cluster points based on the intensity. Clearly, use of bicubic splines (as in Ogata and Katsura 1988) would not suit either of these requirements as naturally. Another strength of the current method is the Bayesian formulation. This framework saves us from having to work with models that are stationary (or confining our analysis to datasets that have replicate observations on the point process), and it frees us from having to be concerned with deriving consistency and asymptotic normality results under the stationarity condition.

Although it has been suspected, based on images generated by in situ hybridization, that infection spreads locally (at least during some stages of infection), this has never been demonstrated until now. The model of transmission put forward by Grossman et al. (1998) seemed logical at the time, but some aspects of that model have since been shown to be incorrect, and hence that theory needs some rethinking. In their explanation of PAT, Grossman et al. maintained that activation is a crucial component of the model, because it was thought at that time (based on in vitro studies) that T cells could not be infected unless they were activated. Grossman et al. maintained that stochastic activation events driven by pathogens unrelated to HIV provide the basis for sporadic local expansion of the infection via the PAT mechanism, and these sporadic expansions provide the means for the long-term survival of latent reservoirs of infected cells when no virus can be detected in the blood. Zhang et al. (1999) unexpectedly found that in vivo SIV (and HIV) replicates in nonactivated T cells at low levels (which is not possible in vitro). Hence the activation aspect of the PAT model may be superfluous even though the local spread aspect of the theory is defensible, at least in the early stages of infection. The surprising result of Zhang et al. was interpreted as evidence for a model of propagation of infection in vivo in which the type of cell infected initially is determined by the representation of the cell types in the population and the spread of infection is to the nearest cells in the vicinity of a productively infected cell. In the earliest stages of infection, HIV and SIV must be able to use resting T cells, because these cells are most available to the virus in lymphatic tissues. The spatial pattern of infection predicted by this model, which we have demonstrated here, is that productively infected cells will be clustered about an infectious "parent" during the early stages of infection. Thus we have shown that statistical approaches effectively complement traditional investigations of fundamental problems in viral pathogenesis in the complex in vivo setting by testing the spatial predictions of hypothesized mechanisms.

[Received —- 2001. Revised July 2002.]

REFERENCES

Baddeley, A., and Silverman, B. (1984), "A Cautionary Example for the Use of Second Order Methods for Analyzing Point Patterns," *Biometrics,* 40, 1089–1094.

Banfield, J., and Raftery, A. (1993), "Model-Based Gaussian and Non-Gaussian Clustering," *Biometrics,* 49, 803–821.

Bartlett, M. S. (1964), "The Spectral Analysis of Two-Dimensional Point Processes," *Biometrika,* 51, 299–311.

Belin, T., and Rubin, D. (1995), "A Method for Calibrating False-Match Rates in Record Linkage," *Journal of the American Statistical Association*, 90, 694–707.

Besag, J., and Newell, J. (1991), "The Detection of Clusters in Rare Diseases," *Journal of the Royal Statistical Society,* Ser. A, 154, 143–155.

Box, G., and Tiao, G. (1968), "A Bayesian Approach to Some Outlier Problems," *Biometrika,* 55, 119–129.

Cheynier, R., Hanrichwark, S., Hadida, F., Pelletier, E., Oksenhendler, E., Autran, B., and Wain-Hobson, S. (1994), "HIV and T Cell Expansion in Splenic White Pulps is Accompanied by Infiltration of HIV-1 Specific Cytotoxic T Lymphocytes," *Cell,* 78, 2705–2714.

Cressie, N. (1993), *Statistics for Spatial Data,* New York: Wiley.

Daley, D., and Vere-Jones, D. (1988), *Introduction to the Theory of Point Processes,* New York: Springer.

Gelman, A., Carlin, J., Stern, H., and Rubin, D. (1995), *Bayesian Data Analysis,* London: Chapman and Hall.

Gelman, A., and King, G. (1990), "Estimating the Electoral Consequences of Legislative Redistricting," *Journal of the American Statistical Association,* 85, 274–282.

Gelman, A., and Rubin, D. (1992), "Inference From Iterative Simulation Using Multiple Sequences" (with discussion), *Statistical Science,* 7, 457–511.

Grossman, Z., Feinberg, M., and Paul, W. (1998), "Multiple Modes of Cellular Activation and Virus Transmission in HIV Infection: A Role for Chronically and Latently Infected Cells in Sustaining Viral Replication," *Proceedings of the National Academy of Sciences USA*, 95, 6314–6319.

Haase, A. (1999), "Population Biology of HIV-1 Infection: Viral and $CD4^+$ T Cell Demographics and Dynamics in Lymphatic Tissues," *Annual Review of Immunology,* 17, 625–656.

Kooijman, S. (1979), "The Description of Point Patterns," in *Spatial and Temporal Analysis in Ecology,* eds. R. Cormack and J. Ord, Fairland, MD: International Co-operative Publishing House, pp. 305–331.

Lawson, A. (1988), "On Tests for Spatial Trend in a Nonhomogeneous Poisson Process," *Journal of Applied Statistics,* 15, 225–234.

MacQueen, J. B. (1967), "Some Methods for Classification and Analysis of Multivariate Observations," in *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability,* 1, Berkeley, CA: University of California Press, pp. 281–297.

Ogata, Y., and Katsura, K. (1988), "Likelihood Analysis of Spatial Inhomogeneity for Marked Point Patterns," *Annals of the Institute of Statistical Mathematics*, 40, 29–39.

Perelson, A., Neumann, A., Markowitz, M., Leonard, J., and Ho D. (1996), "HIV-1 Dynamics In Vivo: Virion Clearance Rate, Infected Cell Life-Span, and Viral Generation Time," *Science,* 271, 1582.

Reinhart, T., Rogan, M., Amedee, A., Murphey-Corb, M., Rausch, D., Eiden, L., and Hasse, A. T. (1998), "Tracking Members of the Simian Immunodeficiency Virus Delta B670 Quasispecies Population In Vivo at Single-Cell Resolution," *Journal of Virology,* 72, 113–120.

Ripley, B. (1976), "The Second-Order Analysis of Stationary Point Processes," *Journal of Applied Probability,* 13, 255–266.

——— (1988), *Statistical Inference for Spatial Processes,* New York: Cambridge University Press.

Stoyan, D., Kendall, W., and Mecke, J. (1995), *Stochastic Geometry and its Applications,* New York: Wiley.

Tenner-Racz, K., Racz, P., Schmidt, H., Dietrich, M., Kern, P., Louie, A., Gartner, S., and Popovic, M. (1988), "Immunohistochemical, Electron Microscopic and In Situ Hybridization Evidence for the Involvement of Lymphatics in the Spread of HIV-1," *AIDS,* 2, 299–309.

Titterington, D., Smith, A., and Makov, U. (1985), *Statistical Analysis of Finite Mixture Distributions,* New York: Wiley.

Zhang, Z., Schuler, T., Zupancic, M., Wietgrefe, S., Staskys, K., Reimann, K., Reinhart, T., Rogan, M., Cavert, W., Miller, C., Veazey, R., Notermans, D., Little, S., Danner, S., Richman, D., Havlir, D., Wong, J., Jordan, H., Schacker, T., Racz, P., Tenner-Racz, K., Letvin, N., Wolinsky, S., and Hasse, A. T. (1999), "Sexual Transmission and Propagation of SIV and HIV in Resting and Activated $CD4^+$ T Cells," *Science,* 286, 1353–1357.