Applied Spatial Statistics for Public Health Data

LANCE A. WALLER

Emory University
Department of Biostatistics
Atlanta, Georgia

CAROL A. GOTWAY

National Center for Environmental Health
Centers for Disease Control and Prevention
Atlanta, Georgia

WILEY-
INTERSCIENCE

A JOHN WILEY & SONS, INC., PUBLICATION

process under consideration [see Cressie (1993, pp. 663–664) for a more formal argument]. We note that the results establish an equivalent Cox process for any Neyman–Scott process (with Poisson numbers of children), but the reverse does not hold, as the class of Cox processes is much larger than those equivalent to Neyman–Scott processes. For example, consider the simple Cox process based on CSR with a variable intensity λ. No realization of this process involves clustering consistent with a Neyman–Scott process (unless one considers a degenerate uniform child-dispersal distribution). Finally, although the conceptual description of Bartlett's equivalence above might suggest that any Poisson cluster process yields an equivalent Cox processes, formalizing the argument mathematically requires precise definitions of valid probability structures, and Cressie (1993, p. 664) points out that the general result remains unproven.

5.5 ADDITIONAL TOPICS AND FURTHER READING

In this chapter we provide only a brief introduction to spatial point processes and their first- and second-order properties. The methods outlined above provide the probabilistic tools to develop the analytic methods in Chapters 6 and 7 for investigating spatial patterns of disease. Many of the applications in Chapters 6 and 7 build from heterogeneous Poisson processes, and our discussion here tends to focus accordingly, resulting in limited treatment of some concepts covered in more detail in more general texts on spatial statistics. In particular, we only provide the barest details regarding tests of CSR. Cressie (1993, Chapter 8) provides a more thorough presentation of such methods.

Due to our focus on heterogeneous Poisson processes, we ignore the sizable literature regarding nearest-neighbor distance distributions. The literature refers to the *F function* and the *G function* to represent cumulative distribution functions of the distances between either a randomly chosen point in the study area or a randomly chosen event to the nearest-neighboring event, respectively. See Diggle (1983, Section 2.3) and Cressie (1993, Sections 8.2.6 and 8.4.2) for further details. In addition, van Lieshout and Baddeley (1996) consider the ratio of the F and G functions (termed the *J function*) as a measure of spatial interaction in a spatial point processes. Van Lieshout and Baddeley (1999) provide an analog to the J function for multivariate point processes (e.g., point processes with more than one type of event, as in the grave site example).

Finally, there are also a wide variety of point process models in addition to the Poisson processes outlined above. We refer the reader to Diggle (1983), Ripley (1988), Chapter 8 of Cressie (1993), Stoyan et al. (1995), and Lawson (2001) for further details and examples.

5.6 EXERCISES

5.1 Suppose that we have a realization of a spatial point process consisting of N event locations $\{s_1, \dots, s_N\}$. Let W_i denote the distance between the ith event and its nearest-neighboring event. The literature refers to the cumulative distribution function of W (the nearest event–event distance) as the *G function*. What is the G function under complete spatial randomness; that is, what is $\Pr[W \leq w]$? (*Hint:* Consider the probability of observing no events within a circle of radius w.)

5.2 Simulate 100 realizations of complete spatial randomness in the unit square with 30 events in each realization. For each realization, calculate the distance between each event and its nearest-neighboring event, denoted W_i for the ith event in the realization. Calculate $2\pi\lambda\sum_{i=1}^{30} W_i^2$ (Skellam 1952) for each realization and compare the distribution of values to a χ^2_{2N} distribution where $N = 30$ denotes the number of events, λ the intensity function (30 for this application), and π is the familiar mathematical constant.

5.3 Repeat Exercise 5.2, letting the number of events in each realization follow a Poisson distribution with mean 30. What changes in the two settings? Under what assumptions is Skellam's chi-square distribution appropriate?

5.4 Simulate 100 realizations of a Poisson cluster process and calculate Skellam's statistic for each realization. Compare the histogram of values to that obtained in Exercise 5.2. How does the distribution of the statistic change compared to its distribution under complete spatial randomness?

5.5 For each of $\lambda = 10, 20$, and 100, generate six realizations of CSR on the unit square. For each realization, construct a kernel estimate of $\lambda(s)$ (supposing you did not know that the data represented realizations of CSR). How does each set of six estimates of $\lambda(s)$ compare to the known constant values of λ? What precautions does this exercise suggest with regard to interpreting estimates of intensity from a single realization (data set)?

5.6 The following algorithm outlines a straightforward acceptance–rejection approach to simulating realizations of N events from a heterogeneous Poisson process with intensity $\lambda(s)$. First, suppose that we can calculate $\lambda(s)$ for any point s in the study area A, and that we know (or can calculate) a bounding value λ^* such that $\lambda(s) \leq \lambda^*$ for all $s \in A$.

Step 1. Generate a "candidate" event at location s_0 under CSR in area A.

Step 2. Generate a uniform random number, say w, in the interval [0, 1].

Step 3. If $w \leq \left(\lambda(s)/\lambda^*\right)$ [i.e., with probability $\left(\lambda(s)/\lambda^*\right)$], keep the candidate event as part of the simulated realization; otherwise, "reject" the candidate and omit it from the realization.

Step 4. Return to step 1 until the collection of accepted events numbers N.

In this algorithm, events have a higher probability of being retained in the realization in locations where the ratio $\left(\lambda(s)/\lambda^*\right)$ is higher. The closer the value λ^* is to $\max_{s \in A} \lambda(s)$, the more efficient the algorithm will be (as fewer candidates will be rejected overall). [See Lewis and Shedler (1979), Ogata (1981), and Stoyan et al. (1995, Section 2.6.2) for more detailed discussions of this and similar algorithms.]

For a heterogeneous intensity $\lambda(s)$ and study area A of your choice, generate six realizations with 30 events each from the same underlying heterogeneous Poisson process. For each realization, estimate $\lambda(s)$ via kernel estimation. Plot the realizations with respect to your known intensity $\lambda(s)$. Provide separate plots of each kernel density estimate and compare to the true intensity function $\lambda(s)$.

On a separate plot, indicate the location of the mode (maximal value) of each kernel estimate of $\lambda(s)$. How do these six values compare to the true mode of $\lambda(s)$? What (if any) implications do your results suggest with respect to identifying modes of disease incidence based on intensity estimated from a single data realization (e.g., a set of incident cases for a single year)?

5.7 The medieval grave site data set introduced in Section 5.2.5 appear in Table 5.2. Estimate the intensity functions for affected and nonaffected sites for a variety of bandwidths. For what bandwidths do the two intensities appear similar? For what bandwidths do they appear different?

Table 5.2 Medieval Grave Site Data[a]

u	v	Aff	u	v	Aff	u	v	Aff
8072	8970	0	9004	7953	0	8614	8528	0
9139	8337	1	8876	8641	1	8996	8039	0
7898	8892	0	8320	9010	0	9052	8923	0
9130	7438	0	9194	6474	0	9338	5737	0
8102	7636	0	9334	6740	0	9183	6073	0
8889	7272	0	8639	6916	0	9110	6393	0
8167	5609	0	9272	7095	0	8341	6903	0
8546	6218	0	9419	4177	1	9215	4570	0
8400	4117	0	9110	5067	0	9310	5450	1
9361	7166	1	8303	4935	0	8536	4226	0
9435	4473	1	8189	5720	0	8797	4787	0
8326	9300	0	8457	4785	0	8326	9541	1
5100	6466	0	8373	9379	0	7042	8761	0
4714	6455	0	4492	7463	0	7212	8262	0
7209	7467	0	7468	7789	1	7768	7972	1
7796	7657	0	7639	7009	1	7237	7299	0
7620	6039	0	6934	6918	1	9149	3588	0
7708	5776	0	7119	7784	0	7042	7264	1
7039	6234	1	8778	3844	0	9485	3319	1
5305	6065	1	5306	6065	1	5456	6353	0
5717	6023	0	5597	7725	0	5231	7472	0
6092	7671	1	4862	5969	0	6252	8271	0

(continued overleaf)

Table 5.2 (*continued*)

u	v	Aff	u	v	Aff	u	v	Aff
6720	7164	1	6569	7391	0	6258	7127	0
9558	9403	0	9208	9114	1	9352	7957	0
9473	8826	0	7505	6024	0	7974	6332	0
7634	6229	0	8126	7269	0	9756	9257	0
9752	6468	0	10073	8273	0	9405	10431	0
10147	8141	0	10100	3085	1	9262	10068	0
9990	3824	0	9305	9661	0	8831	9393	0
9570	9059	0	9656	8356	0	9547	7690	0
9416	9223	0	9502	8846	0	8937	9611	0
10263	4790	0	10324	4389	0	10232	7271	0
9497	7564	0	9412	7463	0	9722	7065	0
9757	5276	1	9879	6309	0	10061	5937	0
9716	6713	0	9699	7240	0	9665	5554	1
10156	5225	1	10143	6317	0	10373	3208	1
8575	8840	0	9072	8894	0	8846	7633	0
9131	6958	1	9230	7068	0	8217	5835	0
8458	5106	0	8685	4497	0	8175	4862	0
8598	5377	0	8789	5006	0	5101	7115	0
4716	6733	0	5109	6590	0	7507	8280	0
7459	6591	0	8861	3882	0	7068	6341	0
5683	7046	0	4612	6147	0	5385	7052	0
6720	7541	0	5952	6278	1	7759	6222	1
7628	6730	0	10070	6739	0	9770	3469	0
9850	3656	1	9667	9541	0	9702	4581	1
10030	4274	0	10292	7562	0	9953	4673	0
10192	5291	0	10148	5222	1			

[a] u and v denote the coordinates of each location and "Aff" indicates whether the grave site included missing or reduced wisdom teeth (Aff = 1) or did not (Aff = 0). See text for details.

5.8 Simulate 10 realizations from the Baddeley and Silverman (1984) process defined in Section 5.3.4 on a 100×100 unit grid. Plot the K function for each realization and plot the average of the 10 K functions at each distance value. Does the average of the K-function estimates appear consistent with the K function for complete spatial randomness? Does each of the 10 estimated K functions appear consistent with the K function from complete spatial randomness?