Generalized index for spatial data sets as a measure of complete spatial randomness

Emily J. Hackett-Jones,[1] Kale J. Davies,[2] Benjamin J. Binder,[2] and Kerry A. Landman[1,*]
[1]*Department of Mathematics and Statistics, University of Melbourne, Victoria 3010, Australia*
[2]*School of Mathematical Sciences, University of Adelaide, South Australia 5005, Australia*

(Received 21 March 2012; published 8 June 2012)

Spatial data sets, generated from a wide range of physical systems can be analyzed by counting the number of objects in a set of bins. Previous work has been limited to equal-sized bins, which are inappropriate for some domains (e.g., circular). We consider a nonequal size bin configuration whereby overlapping or nonoverlapping bins cover the domain. A generalized index, defined in terms of a variance between bin counts, is developed to indicate whether or not a spatial data set, generated from exclusion or nonexclusion processes, is at the complete spatial randomness (CSR) state. Limiting values of the index are determined. Using examples, we investigate trends in the generalized index as a function of density and compare the results with those using equal size bins. The smallest bin size must be much larger than the mean size of the objects. We can determine whether a spatial data set is at the CSR state or not by comparing the values of a generalized index for different bin configurations—the values will be approximately the same if the data is at the CSR state, while the values will differ if the data set is not at the CSR state. In general, the generalized index is lower than the limiting value of the index, since objects do not have access to the entire region due to blocking by other objects. These methods are applied to two applications: (i) spatial data sets generated from a cellular automata model of cell aggregation in the enteric nervous system and (ii) a known plant data distribution.

DOI: 10.1103/PhysRevE.85.061908 PACS number(s): 87.10.−e, 87.18.Ed, 87.18.Hf

I. INTRODUCTION

The spatial distribution of a set of objects arises throughout physical, biological, and social processes, for example, in fluid mixing [1–5], cell biology [6–9], plant ecology [10–14], and pedestrian and traffic flow [15,16]. They also arise naturally in agent-based models, known as cellular automata (CA) models [17–21]. The objects can be either (i) pointlike objects which only represent the locations of some quantity of interest within the spatial domain (e.g., fluid particles) or (ii) finite-sized objects that exclude volume at locations within the domain (e.g., cells, plants, pedestrians, cars, and CA agents). In the first case, many objects can be colocated at the same point, whereas in the latter, only one object can be located in the same space in the domain. The finite size of the objects is important to many applications (e.g., traffic flow and cellular tumor invasion), giving rise to excluded volume effects—these are known as simple exclusion processes [22].

Measures have been developed which indicate whether or not a spatial data set is at the complete spatial randomness (CSR) state [23,24]—this state occurs when each object is equally likely to lie in any part of the spatial domain. For example, objects dispersed uniformly at random throughout the domain (e.g., as the result of a diffusive process) are at the CSR state. Binder and Landman [25] derived a CSR limiting value for an index [5], when objects exclude volume from the domain. The index was defined by partitioning the domain into *equal size bins* and calculating a scaled variance of the bin counts. The CSR limiting value was an approximation based on the assumption that the bin counts followed a Pólya-Eggenberger (Pólya) distribution [26], with the bin size being much larger than the size of the object [25,27].

Some spatial domains are not easily partitioned into equal size bins. The spatial distribution of fluid particles in a circular batch mixer mixed by a stirring rod [3,4], and the position of bacteria on the surface of a circular Petri dish are two examples where the circular domain is readily partitioned into bins which are concentric circles [see Fig. 1(c)]. With such a bin configuration, the bin counts are easily recorded by simply measuring the object's distance from the center of the domain. Here we generalize the index and CSR limiting value by considering an arrangement of *nonequal size bins* that may either overlap each other (for example, in a nested arrangement) or partition the domain. When bins of equal size are used, this generalized index reduces to the previously discussed index [25]. Using examples with single sized objects and different sized objects, we investigate trends in the generalized index as a function of density and compare the results with those using equal size bins.

For a set of objects known to be placed uniformly at random throughout a domain, we calculate values of the generalized index and compare them to the CSR limiting value. We determine that the CSR limiting value is an excellent predictor of when the CSR state has been attained, provided all the current unoccupied space within the spatial domain is accessible to each object as it is placed in the domain. This is always the case for pointlike objects, as they do not exclude volume and therefore the entire spatial domain is always accessible for every object placement. However, when objects exclude volume, a group of neighboring objects may render unoccupied space between them inaccessible to the subsequent placement of objects. This phenomenon is known as "blocking" [28–34] and leads to calculated values of the generalized index that are lower than those predicted by the CSR limiting value. In this case the CSR limit is still a useful indicator of a spatial data set's proximity to the CSR state. By examining the trend in the calculated values of the generalized

[*] kerryl@unimelb.edu.au

index for two or more bin configurations, we can ascertain whether or not a data set is at the CSR state.

Aggregation patterns, arising from widely different mechanisms, are observed in many biological and ecological applications [35–38]. For example, during the development of the enteric nervous system (ENS), aggregates of neuronal cells (known as ganglia) form behind a fast-moving invasion wave of neuronal precursor cells [18]. Adhesion molecules on the cell surface are responsible for the clustering of the cells into aggregates. CA agent-based modeling has successfully replicated many of the features of the formation of ganglia in the developing gut [18]. We implement the CA model [18] to generate spatial data sets of agent aggregates, from an initially dispersed population of CA agents. The generalized index and CSR limiting value are used to analyze the spatial distribution of these aggregates.

We also use the generalized index to investigate a spatial data set from plant ecology. These applications demonstrate that the generalized index is an easy to use measure, useful to many biological and physical contexts.

II. GENERALIZED INDEX

In this theory section we discuss three-dimensional data sets, bin volumes, and excluded volume of objects. This can be replaced by two-dimensional data sets, bin areas, and excluded area of objects. The examples in later sections are in two dimensions.

Consider a domain of volume A which is populated with a total of n objects, each of volume s. The domain is divided into M bins, each with volume S_j for $j = 1,\ldots,M$. The bins can either overlap each other or are nonoverlapping and partition the domain. If the objects are evenly distributed throughout the domain we expect to observe $\bar{b}_j = nS_j/A$ of them in each bin. This is simply the product of the total number of objects with the jth bin volume fraction. Therefore, we quantify the deviation between each bin count b_j and the evenly distributed state by the statistic

$$\sigma^2 = \frac{1}{M}\sum_{j=1}^{M}(b_j - \bar{b}_j)^2. \tag{1}$$

The statistical measure (1) is scaled by

$$\sigma_0^2 = \frac{n^2}{M}\sum_{j=1}^{M}\frac{S_j}{A}\left(1 - \frac{S_j}{A}\right). \tag{2}$$

The reason for this choice will become clear when we take the CSR limit. Note that for equal-sized bins ($S_j = S$ for $j = 1,\ldots M$) this is the same scaling as in Binder and Landman [25]. This defines a generalized index

$$I = \frac{\sigma^2}{\sigma_0^2}. \tag{3}$$

When the bins are equal in size (with $S_j = S$ and $\bar{b}_j = \bar{b}$ for $j = 1,\ldots,M$), Eqs. (1)–(3) reduce to those of Phelps and Tucker [5]. Therefore our formulation generalizes their index. For an even distribution of objects $\sigma^2 = I = 0$. The generalized index (3) therefore quantifies the deviation of a spatial data set from the evenly distributed state. However, this state is not often realized. A more likely scenario is one where each of the objects is equally likely to lie in any part of the domain, termed the CSR state [2,23,24]. We note in passing that, in contrast to the index for equal size bins, the maximum value of the generalized index, corresponding to a completely segregated state, is no longer unity but is typically greater than unity (Appendix A). Next, we approximate the CSR limiting value for the generalized index (3).

A. CSR limit

The bin counts can be represented by a random variable B_j with observed values b_j and expected value

$$E[B_j] = \bar{b}_j. \tag{4}$$

Taking the expectation of (1) we find

$$E[\sigma^2] = \frac{1}{M}\sum_{j=1}^{M}\mathrm{Var}(B_j). \tag{5}$$

To proceed further we need to determine the distribution of B_j's. Binder and Landman [25] showed that, for equal size bins, the bin counts can be approximated by the Pólya distribution [26], provided the object size is much smaller than the bin size [25,27]. In this more general case, we assume that each random variable B_j follows a Pólya distribution with parameters $n, S_j, A - S_j, -s$, and mean $\bar{b}_j$. Using the formula for the variance of the Pólya distribution and the assumption that $s \ll S_j$ for $j = 1,\ldots,M$, we find

$$\mathrm{Var}(B_j) = \frac{nS_j}{A}\left(1 - \frac{S_j}{A}\right)\left(1 - \frac{ns}{A}\right), \quad j = 1,\ldots,M. \tag{6}$$

Equations (5) and (6) then give

$$\sigma_{\mathrm{CSR}}^2 = E[\sigma^2] = \frac{n}{M}\left(1 - \frac{ns}{A}\right)\sum_{j=1}^{M}\frac{S_j}{A}\left(1 - \frac{S_j}{A}\right). \tag{7}$$

Scaling (7) by σ_0^2 we obtain the CSR limit

$$I_{\mathrm{CSR}} = \frac{\sigma_{\mathrm{CSR}}^2}{\sigma_0^2} = \frac{1}{n}\left(1 - \frac{ns}{A}\right) = \frac{1-d}{n}, \tag{8}$$

where

$$d = \frac{ns}{A} \tag{9}$$

is the volume fraction (density) of the domain that is occupied by objects. Note that the choice of scaling σ_0^2 in (8) gives the same CSR limiting value as determined by Binder and Landman [25] for equal size bins.

Next, for specified values of the density d, we simulate the CSR state and calculate the index for each simulation. We average over N simulations and calculate the average generalized index

$$\langle I\rangle = \frac{1}{N}\sum_{i=1}^{N}I_i, \tag{10}$$

where I_i is the index of the ith realization. The average generalized index is compared to the CSR limiting value (8).

B. Single species simulation of the CSR state

The CSR state is simulated by placing equal size objects uniformly at random onto a two-dimensional domain, as is shown in Figs. 1(a) and 1(c) and Figs. 2(a) and 2(c). Calculated values of the generalized index and CSR limiting values are plotted in Figs. 1(b) and 1(d) and Figs. 2(b) and 2(d).

First consider the results in the first row of Fig. 1. Unit square objects with $s = 1$ are placed uniformly at random on unit square lattice sites. Only one object is allowed to occupy a lattice site—an example of a simple volume exclusion process [22]. The maximum density of these objects occurs when each lattice site is populated with one object, corresponding to $d = 1$. The generalized index is calculated using an overlapping square bin configuration, with the origin being the position of the lower leftmost corner of each bin [Fig. 1(a)]. The CSR limiting value (8) accurately predicts the state of the system [Fig. 1(b)].

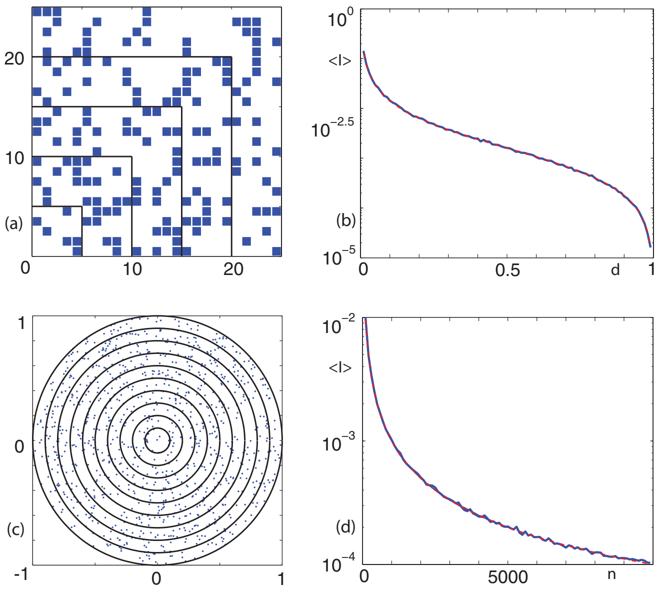


FIG. 1. (Color online) Simulations of the CSR state with overlapping bins. The results show that the CSR limiting value (8) accurately predicts the CSR state. (a) and (b) Overlapping (nested) bins of sizes $S_j = 25j^2$ for $j = 1,\ldots,5$ with object size $s = 1$ and $A = 625$. (a) Typical simulation for density $d = 0.25$. (b) The average generalized index from $N = 10$ simulations [blue (dark gray) curve] plotted as a function of the density d, and the CSR limiting value (dashed red curve). (c) and (d) Overlapping (nested) bins of sizes $S_j = (j/10)^2\pi$ for $j = 1,\ldots,10$, with object size $s = 0$ and $A = \pi$. (c) Typical simulation for $n = 1000$ pointlike objects. (d) The average generalized index from $N = 10$ simulations [blue (dark gray) curve] plotted as a function of the number of objects n, and the CSR limiting value [dashed red (medium gray) curve]. Note that the smallest (or average) bin size in (a) and (b) is equal to the constant size bins for the results shown in (c) and (d). The dashed and solid lines in (b) and (d) are more or less indistinguishable.

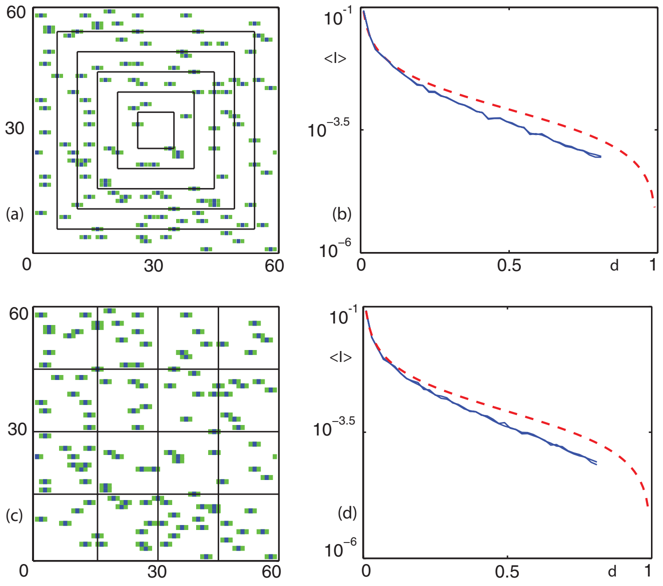


FIG. 2. (Color online) Simulations of the CSR state with objects of size $s = 3$ and $A = 3600$. The average generalized index tends to a limiting value lower than the predicted CSR limiting value (8). (a) and (b) Centered overlapping square bins. (a) Typical arrangement of bins and simulation for density $d = 0.1$. (b) The average generalized index from $N = 200$ simulations, plotted as a function of the density d [two blue (dark gray) curves] and the CSR limiting value (dashed red curve). The two configurations of overlapping square bins are $S_j = (10j)^2$ with $j = 1,\ldots,6$ and $S_j = (6j)^2$ with $j = 1,\ldots,10$. (c) and (d) Nonoverlapping equal size square bins. (c) Typical arrangement of bins and simulation for density $d = 0.2$. (d) The average generalized index from $N = 200$ simulations, plotted as a function of the density d [two blue (dark gray) curves] and the CSR limiting value (dashed red curve). The two configurations of nonoverlapping equal size bins are $S_j = 225$ with $j = 1,\ldots,16$ and $S_j = 900$ with $j = 1,\ldots,4$. The two blue (solid) lines in (b) and (d) are more or less indistinguishable.

The results in the second row of Fig. 1 are for pointlike objects with $s = 0$, placed (off lattice) on a circular domain. Pointlike objects do not exclude volume from the domain ($d = 0$). Consequently it can be populated with any finite number of objects $n < \infty$. The generalized index is calculated using a configuration of concentric circular bins centered at the origin [Fig. 1(c)]. The CSR limiting value accurately predicts the state of the system [Fig. 1(d)].

The average generalized index in Fig. 1 was also calculated for other configurations of (nonequal size) bins, for example bins that partitioned the two domains. The black lines and curves in Figs. 1(a) and 1(c) illustrate the boundaries of the partitioning. In Fig. 1(a) the nonoverlapping bins are one square bin and four L-shaped bins increasing with size as their distance increases from the lower leftmost corner of the domain. In Fig. 1(c) the nonoverlapping bins are one circle and nine annuli that increase in size with distance from the origin. The average generalized indices calculated with these nonoverlapping bins are indistinguishable from those shown in Figs. 1(b) and 1(d). This demonstrates that the results shown in Fig. 1 are independent of the bin configuration we have used.

Next, we consider the placement of rectangular objects with $s = 3$ on unit square lattice sites, as illustrated in Figs. 2(a) and 2(c). The objects are not allowed to overlap—the objects exclude volume from the domain. The central position (on the lattice) of each object is blue (dark gray) with the remaining volume being green (light gray). The object is included into the bin count of the bin containing the position of the blue central part. (Note that if a bin boundary lay precisely in the center of the object, then the object would be assigned randomly to one of the bins on either side. This cannot occur for our bin configurations.) The average generalized index for two configurations of centered overlapping square bins are more or less indistinguishable [Fig. 2(b)], suggesting that the average generalized index is independent of the size and number of bins used in each of these configurations. However, the average generalized index is lower than the CSR limiting value [Fig. 2(b)]. A similar set of results is found using two configurations of nonoverlapping equal size bins [Fig. 2(d)]. Therefore the generalized index gives consistent results irrespective of whether unequal or equal size bins are used for the bin counts. Furthermore, we deduce that the CSR limiting value (8) is overestimating the *true* CSR limiting value for this system, recalling that the spatial data set is known to be at the CSR state. This can be explained as follows.

In the derivation of the CSR limiting value (8), it is assumed that each object placement has access to all the current unoccupied space within the domain. This was indeed true for the results shown in Fig. 1, where square objects size $s = 1$ are placed on a square lattice and point objects size $s = 0$ are placed on a circular domain. Hence for these two cases, the CSR limiting value (8) accurately predicts the CSR state. However, this not true for the simulations with objects of size $s = 3$ shown in Fig. 2. For example, two objects placed with central positions at $(x - 2, y)$ and $(x + 2, y)$ render the volume in between them at (x, y) inaccessible to the subsequent placement of objects. Effectively, an extra unoccupied lattice site as well as the six occupied lattice sites have been removed from the domain. [Note that eight places would be removed if the two objects were placed with central positions at $(x - 3, y)$ and $(x + 2, y)$.] This phenomenon of extra volume removal is well known in the random sequential adsorption literature [28–34] where it is called blocking. The continued placement of objects ultimately leads to the so-called "jamming limit." This is the density at which no further objects can be placed in the domain. The jamming limit (or jamming density) causes the blue curves in Figs. 2(b) and 2(d) to terminate at a density d_{jam} strictly less than unity. This explains the increasing difference between the calculated average generalized index and the CSR limiting value [Figs. 2(b) and 2(d)], where the blocking becomes more prevalent as the jamming density is reached.

When blocking occurs in the spatial system being analyzed, the CSR state can be predicted by examining the trend in the calculated values of the generalized index for different bin sizes and configurations. For some sufficiently small bin sizes S_j, the generalized index can be greater than or equal to the CSR limiting value. [Note, if all the bins are exactly size s, then the index is exactly the CSR value, an artifact of the choice of bin size; see Appendix B.] As the bin sizes S_j increase, the average generalized index falls below the CSR limit. As this process is continued, for all bin configurations where the bin sizes are all sufficiently large, the values of the average generalized index converge to a limiting value. In the examples here, we illustrate this with two bin configurations—the average generalized index is approximately the same for each of these [blue curves, Figs. 2(b) and 2(d)], and the curves are lower than that of the CSR limiting value (8) [dashed red curves, Figs. 2(b) and 2(d)]. Comparisons between the average generalized index calculated from different bin configurations are more or less indistinguishable when $s \ll \min\{S_j, \ \ j = 1, \ldots, M\}$ [the condition used in deriving the CSR limiting value (8)].

Therefore, we must determine and compare the generalized index calculated from at least two bin configurations (satisfying $s \ll \min\{S_j, \ j = 1, \ldots, M\}$); if their values are approximately the same, then we might conclude that the single species spatial data set is at the CSR state. The density of objects in the domain is only needed if we wish to calculate the CSR limiting value, which will be larger than the calculated values of the generalized index due to blocking effects. We now investigate spatial data sets that consist of objects with different sizes; that is, multiple species of objects.

C. Multispecies simulations

The CSR state is simulated by placing different size objects uniformly at random onto a two-dimensional domain, with unit square lattice sites, as is shown in Figs. 3(a) and 3(c).

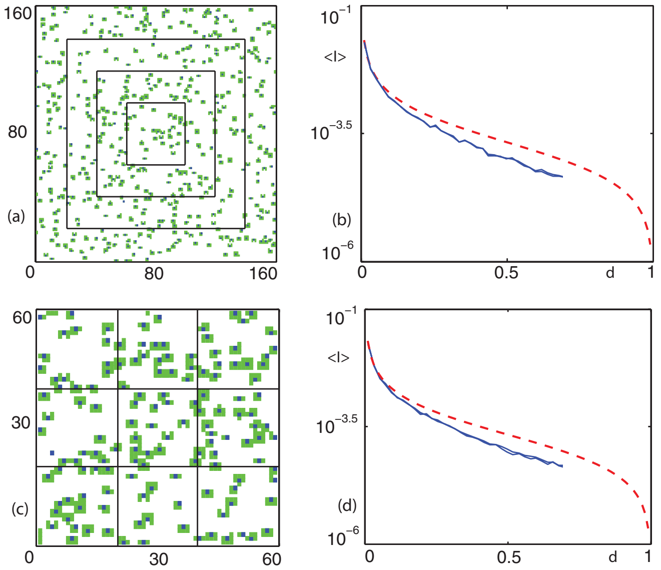


FIG. 3. (Color online) Simulations of the CSR state with a distribution (11) of different size objects. The mean object size is $\alpha = 5$ with $\beta = 0.2$ and $A = 25\,600$. The average generalized index tends to a limiting value lower than predicted CSR limiting value (8). (a) and (b) Centered overlapping (nested) square bins. (a) Typical arrangement of bins and simulation for density $d = 0.1$. (b) The average generalized index from $N = 200$ simulations, plotted as a function of the density d (two blue curves) and the CSR limiting value [dashed red (medium gray) curve]. The two configurations of overlapping square bins are $S_j = (40j)^2$ with $j = 1, \ldots, 4$ and $S_j = (16j)^2$ with $j = 1, \ldots, 10$. (c) and (d) Nonoverlapping equal size square bins. (c) Typical arrangement of bins and simulation for density $d = 0.1$. Note that this is a portion of the domain which more clearly illustrates the spatial distribution of the objects. (d) The average generalized index from $N = 200$ simulations, plotted as a function of the density d [two blue (dark gray) curves] and the CSR limiting value [dashed red (medium gray) curve]. The two configurations of nonoverlapping equal size bins are for $S_j = 1600$ with $j = 1, \ldots, 16$ and $S_j = 6400$ with $j = 1, \ldots, 4$. The two blue (solid) lines in (b) and (d) are more or less indistinguishable.

Calculated values of the average generalized index and CSR limiting values are plotted in Figs. 3(b) and 3(d).

The algorithm for populating the domain is described as follows. (i) A finite number or set of discrete size objects G with varying size $\{s | s = 1, \ldots, 9\}$ is chosen, using the discrete (Gaussian type) probability mass function

$$P(s) = \frac{e^{-\beta(s-\alpha)^2}}{\sum_{j=1}^{2\alpha-1} e^{-\beta(j-\alpha)^2}}. \tag{11}$$

Here, α is the mean object size and β is related to the variance of the distribution of object sizes. (ii) A size s object is selected at random from the set of objects G. (iii) An unoccupied lattice site (x,y) is chosen randomly from the domain. The placement of the size s object onto this site and up to eight of its neighbors in the Moore neighborhood is considered, and if successful, the object is included into the bin count of the bin containing the position (x,y) [blue in Figs. 3(a) and 3(c)]. The shape of the object is determined by selecting at random with equal probability $s - 1$ Moore neighbors [green in Figs. 3(a) and 3(c)] of the lattice site (x,y). The object is placed onto the lattice only if all $s - 1$ neighboring lattice sites are unoccupied; otherwise the process is aborted. (iv) Steps (ii)–(iv) are repeated until the set of objects G is empty.

The multispecies results [Figs. 3(b) and 3(d)] are similar to those found for a single species of objects (Fig. 2). The average generalized index for different configurations of both overlapping unequal size bins and nonoverlapping equal size bins give the same limiting curves as a function of density for sufficiently large bin sizes. These curves are lower than the CSR value, because of the blocking property in this spatial system. Qualitatively similar results are found for changes in the precise rules of the algorithm (i)–(iv), provided the unoccupied lattice site (x,y) is chosen uniformly at random from the domain [stage (ii) of the algorithm].

To summarize, if the generalized indices for at least two bin configurations are approximately the same, then we may conclude that a multispecies spatial data set is at the CSR state. The average density of objects in the domain is only needed if we wish to calculate the CSR limiting value, which will be larger than the calculated values of the generalized index.

So far we have only considered spatial data sets known to be distributed uniformly at random. For these cases, the average generalized index is consistent for different bin arrangements (as long as $s \ll \min\{S_j, j = 1, \ldots, M\}$), but the values are below the predicted CSR limiting value (8). However, we recognize these properties as a measure of the CSR state. To be a useful measure, the generalized index must also be able to detect whether a spatial data set is not distributed randomly.

For spatial data sets known to be distributed nonuniformly at random throughout the domain, we checked that calculated values of the generalized index do not incorrectly predict that the CSR state has been attained. One example is presented where a (Gaussian type) distribution [Eq. (11)] of different size objects is placed nonuniformly at random onto two-dimensional unit square lattice sites (Fig. 4). The algorithm is essentially the same as the one above, but with a change in the rule at stage (ii). Now the unoccupied lattice site (x,y) is no longer chosen uniformly at random, but instead with a higher probability of being placed at the center of the domain. In particular, a Gaussian type spatial distribution was used. Visual examination of a simulation [Fig. 4(a)] makes it clear that the spatial data is not at the CSR state. However, this deduction is not made so easily when considering a smaller central portion of the domain [Fig. 4(b)]. The average generalized index for two bin configurations is considerably higher than the CSR limiting value [Fig. 4(c)]. Other bin configurations give different values of the average index—they do not converge, as in the case of randomly placed objects. These results indicate that the spatial data set is not at the CSR state. Next we compare values of the average generalized index to the CSR limiting value for spatial data sets that are generated from a CA model, which simulates the formation of cell aggregates in the ENS.

III. APPLICATION: CELL AGGREGATION IN THE ENTERIC NERVOUS SYSTEM

Agent-based modeling has an important role to play in the understanding of mechanisms that govern mesoscale spatial

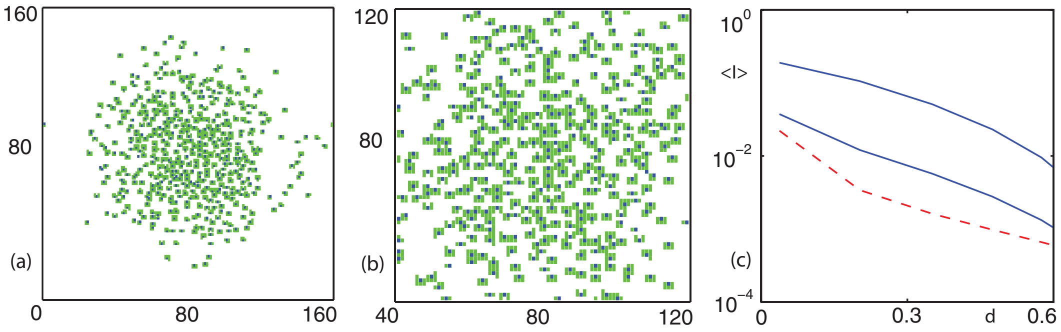


FIG. 4. (Color online) Simulations of a non-CSR state with a distribution (11) of different size objects. The objects were placed using a Gaussian spatial distribution [centered at (80.5,80.5) and variance 1000]. The mean object size is $\alpha = 5$ with $\beta = 0.2$ and $A = 25\,600$. The average generalized index [using centered overlapping (nested) square bins] is above the CSR limiting value (8). (a) Typical simulation with density $d = 0.1$. (b) Typical simulation of the central region of the domain in (a) with (local) density $d = 0.3$. (c) The average generalized index from $N = 200$ simulations of the central region as shown in (b), plotted as a function of the density d [two blue (dark gray) curves] and the CSR limiting value pdashed red (medium gray) curve]. The two configurations of overlapping square bins are $S_j = (10j)^2$ with $j = 1, \ldots, 8$ (upper blue curve) and nonoverlapping equal size bins $S_j = 400$ with $j = 1, \ldots, 16$ (lower blue curve).

patterning and the emergence of ganglionic groups or ganglia in the developing ENS [17,18]. Hackett-Jones *et al.* [18] showed that CA agent-based models can predict the formation of aggregates which resemble ganglia. We analyze spatial data sets of different size agent aggregates (i.e., multispecies) generated by the time evolution of the algorithm used for the ENS application. Only the most important details of the algorithm are provided here.

At time $t = 0$ a two-dimensional domain (area A) consisting of unit square lattice sites is populated randomly with $\mathcal{N}$ (unit square) CA agents, giving a density $d = \mathcal{N}/A$. Consequently, the spatial distribution of CA agents is at the CSR state initially. Each lattice site is either unoccupied or occupied by at most one agent. The agents undergo a biased random walk to one of their four nearest neighbor sites, according to the occupancy of their local neighborhood. During each time step of the algorithm, $\mathcal{N}$ agents are randomly and sequentially selected to move. When an agent at site $\mathbf{v} = (x,y)$ is chosen to move it inspects its immediate Moore neighborhood (comprising the eight adjoining lattice sites), as well as the Moore neighborhood of the four potential new sites, $\mathbf{v}' = \{(x \pm 1, y),(x, y \pm 1)\}$. The scaled local coordination number K at each of the potential new sites is calculated and used to compute a probability of moving and a probability of not moving, in terms of a binding function $f(K)$. Let $P(\mathbf{v}'|\mathbf{v})$ be the conditional transition probability that an agent will move from site $\mathbf{v}$ to a site $\mathbf{v}' \in \mathcal{T}(\mathbf{v})$, its set of unoccupied nearest neighbor sites. An agent also assesses its current site based on the scaled coordination number. Then

$$P(\mathbf{v}'|\mathbf{v}) = \begin{cases} \frac{f(K_{\mathbf{v}'})}{f(K_{\mathbf{v}})+\sum_{\mathbf{v}'\in\mathcal{T}\{\mathbf{v}\}} f(K_{\mathbf{v}''})}, & \mathbf{v}' \in \mathcal{T}\{\mathbf{v}\}, \\ \frac{f(K_{\mathbf{v}})}{f(K_{\mathbf{v}})+\sum_{\mathbf{v}''\in\mathcal{T}\{\mathbf{v}\}} f(K_{\mathbf{v}''})}, & \mathbf{v}' = \mathbf{v}. \end{cases}$$

Boundary conditions must be imposed—both no flux boundary conditions and periodic boundary conditions are implemented.

Here the binding function $f(K) = e^{\gamma K}$ is chosen to reflect whether agents prefer to move to regions of low agent density ($\gamma < 0$) or high agent density ($\gamma > 0$). When $\gamma > 2$, the agents cluster and form aggregates [17]. After a number of time steps of the algorithm a quasisteady state is reached, with agent aggregates (green) dispersed throughout the domain.

Typical aggregate patterns are illustrated at two densities [Figs. 5(a) and 5(c)]. The Hoshen-Kopelman algorithm [39] is used to determine the size of each aggregate and the center of mass of each aggregate. The aggregate is included into the bin count of the bin containing the center of mass. The corresponding distribution of aggregate sizes over 100 realizations is given [Figs. 5(b) and 5(d)].

To analyze the spatial distribution of the CA agent aggregates we calculate the average generalized index and compare it to the CSR limiting value with no flux boundary conditions [Fig. 6(a)] and periodic boundary conditions [Fig. 6(b)]. We begin by examining the no flux boundary condition case.

The average generalized index calculated using an overlapping square bin configuration telescoping from the left hand corner [as in Fig. 1(a)] gives values which are lower than the CSR limiting value for densities $d < 0.3$ [Fig. 6(a), lower blue curve]. This suggests that the spatial domain may be at the CSR state. However, we obtain larger values of the average generalized index [Fig. 6(a), upper blue curve] when using an overlapping square bin configuration setup telescoping from the center [as in Fig. 2(a)]. If the spatial domain is at the CSR state for a specified d, the two calculations of the generalized index using different bin configurations should tend to the same value that is lower than the CSR limiting value.

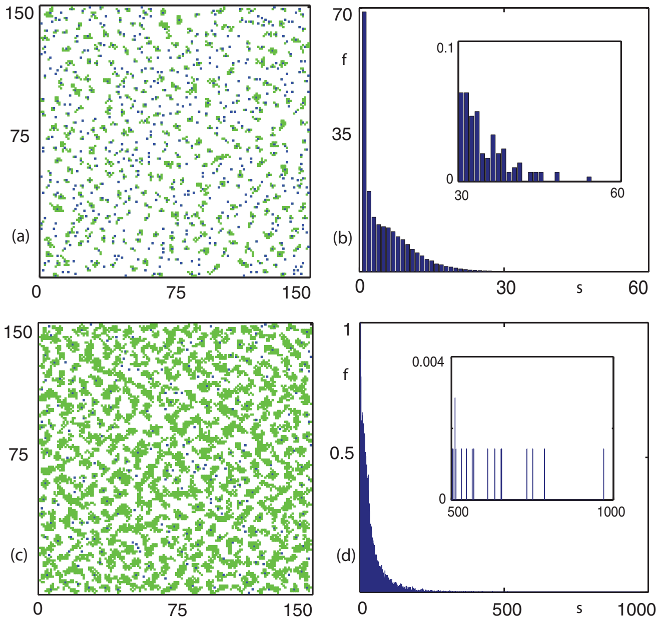


FIG. 5. (Color online) Quasisteady state (at time $t = 20$) of the CA agent aggregation model $\gamma = 8$, $A = 22\,500$ and no flux boundary conditions. The resulting spatial distribution of agent aggregates and variability in agent aggregate sizes are evident. (a) and (c) Single realizations with densities $d = 0.1$ and $d = 0.35$, respectively. The blue (dark gray) markers represent the position of the center of mass of each aggregate; note that some very irregular shaped aggregates have their center of mass lying outside the aggregate. (b) and (d) Frequency distributions of agent aggregate sizes from $N = 100$ simulations with densities $d = 0.1$ and $d = 0.35$, respectively. The frequency axis has been truncated in the main plot of (d).

On closer inspection of the CA agent aggregate data, we find that there is a tendency for aggregates to accumulate more frequently along the boundary of domain. This is due to the no flux boundary conditions implemented in the simulations. This explains why the central telescoping overlapping square bin configuration results are generally above the CSR limiting value in Fig. 6(a), as all the interior bins are underpopulated. This bin configuration correctly predicts that the data set is not at the CSR state. The overlapping bin configuration telescoping from the left corner [lower blue curve in Fig. 6(a)] incorrectly predicts that the spatial data may be at the CSR state. This occurs because each bin contains the same proportion of the domain boundaries relative to its total size. In other words, the bins are not underpopulated with this configuration. This finding underscores the importance of performing at least two bin configurations, as discussed in the previous section.

Next we examine the results from simulations with periodic boundary conditions. Both types of overlapping square bin configurations are considered. The average generalized index is more or less indistinguishable [Fig. 6(b)], suggesting that the values of the generalized index are independent of the bin configuration used. For values of the density $d < 0.3$ the

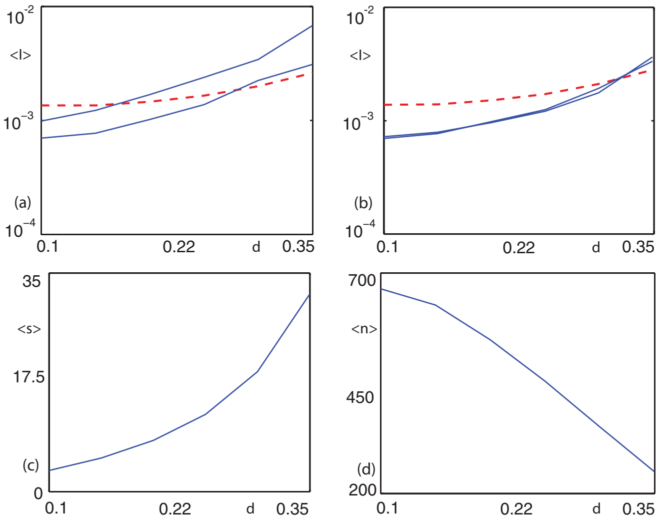


FIG. 6. (Color online) Various quantitative measures for the CA agent aggregation model at time $t = 20$ (quasisteady state) with $\gamma = 8$, $A = 22\,500$, and $N = 100$. (a) and (b) The average generalized index with $S_j = (30j)^2$ for $j = 1,\ldots,5$ [blue (dark gray) curves], compared to the CSR limiting value [dashed red (medium gray) curve]. (a) No flux boundary conditions. The lower blue curve is for an overlapping (nested) bin configuration telescoping from the domain origin [as in Fig. 1(a)]. The upper blue curve is for an overlapping bin configuration, telescoping from the domain center [as in Figs. 2(a) and 3(a)]. (b) Periodic boundary conditions. The blue curves are for the two different bin configurations described in (a). Note that nonoverlapping equal-sized bin ($S = 900$) results match those of the lower blue curve (not shown). (c) and (d) Plots of the average agent aggregate size and average number of agents as a function of density d. The mean agent aggregate size increases and the mean number of agents decreases as the density increases.

blue curves tend to a limiting value that is lower than the CSR limiting value. Consequently, we deduce that the spatial domain is at the CSR state. When periodic boundary conditions are implemented, there is no longer an accumulation of agent aggregates along the boundary of the domain.

We see that for increasing values of the density $d > 0.3$ the blue curves approach the CSR limiting value from below [Fig. 6(b)]. This is due to the increasing number of agent aggregates that are larger than the smallest bin size [Figs. 5(a) and 5(d)] used in the calculations of the generalized index, at these larger values of density. The average aggregate size $\langle s \rangle$ increases with density [Fig. 6(c)] but the average number of aggregates $\langle n \rangle$ decreases with density [Fig. 6(d)]. When the average aggregate size is comparable to the bin size, the CSR limiting value is no longer valid, as the assumption that $\langle s \rangle \ll \min\{S_j,\ j = 1,\ldots,M\}$ made in its derivation is no longer true. For these large densities, our test for determining whether or not a spatial data set is at the CSR state, which compares calculated values of the generalized index to the CSR limiting value, is therefore inconclusive.

IV. DISCUSSION

We have generalized a statistical measure, called a generalized index, and its limiting value. Whether a spatial data set is at the CSR state or not [2,23,24] can be determined by calculating the generalized index in terms of object counts within an arrangement of nonequal size bins that may either overlap each other or partition the domain. This is especially useful for domains which do not easily divide into equal size bins; for example, circular or spherical domains. However, if bins of equal size are used, the generalized index reduces to the previously discussed index [25]. The generalized index and its limiting value are defined for exclusion and nonexclusion processes, which is when the volume of objects cannot overlap and when the objects are pointlike.

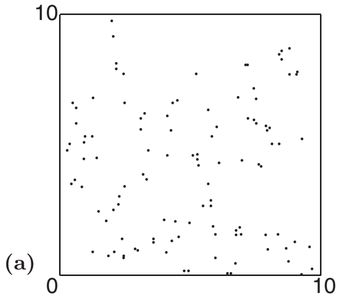
A number of examples of spatial data sets where objects are known to be randomly placed were investigated. Consistent results were obtained if the objects were all the same size or had different sizes. We have shown that the generalized index is a well-defined quantity which is independent of the bin configuration as long as the mean size of the objects is much smaller than the smallest bin size $\langle s \rangle \ll \min\{S_j,\ j = 1,\ldots,M\}$.

For randomly placed objects, the generalized index exactly matches the CSR limiting value for some special cases: (i) when the objects are points and take up no volume and (ii) when the objects have unit volume and they populate a lattice structure with bonds of unit length. In the latter case, any unoccupied space is available to any additional objects placed randomly in the domain. The CSR limiting value may also occur if the objects are deformable or can rearrange on shaking. More generally, when objects exclude volume, the current object placement may render unoccupied space between them inaccessible to the subsequent placement of objects. This blocking phenomena leads to the generalized index being lower than those predicted by the CSR limiting value. The differences are small for low densities, and they increase as the density d increases, as the blocking is enhanced. In this case the CSR limit still proves to be a useful indicator of whether a spatial data set is close to the CSR state. We can distinguish between a spatial data set being at the CSR state or not by comparing the values of generalized index for different bin configurations—the values will be approximately the same if the data is at the CSR state, while the values will differ if the data set is not at the CSR state.

Furthermore, we have demonstrated that the generalized index is able to detect subtle biases in the data. Our example concerned biases due to the boundary conditions used in generating the data, which were undetected by visual inspection. These differences were made apparent using a centered overlapping bin configuration.

Besides constructing data sets, we also applied the generalized index to an agent-based model used to simulate cell aggregation in the ENS. The index will of course be useful in many other applications. We conclude by analyzing a published data set for a plant species. Spatial distributions for plants are studied extensively in the ecological literature [10–14] to provide important information about the system's history, the underlying inter- and intraspecific competition, and the population dynamics of the system. Many statistical methods have been used to analyze such systems.

For example, a range of statistical methods have been applied to the distribution of Mediterranean subshrub *Anthyllis cytisoides* L. [Fig. 7(a)]. Using Ripley's K function (a point-to-point distance method) [23,40,41] with various edge



(b)

Bin Configuration	Generalized Index
(1) $S_j = 6.25, j = 1, \dots M$	0.0138
(2) $S_j = 11.11, j = 1, \dots M$	0.0159
(3) overlapping square bins telescoping from (10,10), $S_j = \frac{100j^2}{36}j, j = 1, \dots 6$	0.0253
(4) overlapping square bins telescoping from (0,0), $S_j = \frac{100j^2}{36}, j = 1, \dots 6$	0.058
CSR limiting value $1/n, n = 111$	0.009

FIG. 7. Distribution pattern of Mediterranean subshrub *Anthyllis cytisoides* L. (a) Spatial arrangement on a 10 × 10 m domain (adapted from [10]). (b) Generalized index for various bin arrangements.

corrections, the natural stand of *Anthyllis cytisoides* is shown to be clumped at distances of up to 0.8 m and again at 3–5 m, and therefore is not at the CSR state. We investigate the data using our generalized index. The index varies for different bin sizes and is above the CSR state (estimated for point-size objects) [Fig. 7(b)]. This trend in the generalized index values matches that of the nonrandom/patchy example discussed in Sec. II C. We therefore conclude that the distribution is not at the CSR state, which agrees with the results of Haase [10] obtained using alternate techniques. The generalized index is an easy to use quantitative measure for establishing whether objects are at their completely spatial random state and will be useful for many biological, physical, and social applications.

APPENDIX A: MAXIMAL INDEX FOR OVERLAPPING BINS

We show that the maximum value of the generalized index (3), corresponding to the completely segregated states, is typically greater than unity. Consider a set of overlapping bins $\{A_j,\ j = 1,\dots,M\}$, where $A_1 \subset A_2 \subset \dots \subset A_{M-1} \subset A_M$, where A_M is the whole domain with volume A and each bin A_j has volume S_j. There are M possible completely segregated states, corresponding to the placement of all n objects in one of $A_1, A_2 - A_1, \dots, A_M - A_{M-1}$. We determine the variance for each of these cases.

If all n objects are placed in A_1, we denote the variance in Eq. (1) by $\sigma_1^2 = n^2\sum_{j=1}^{M-1}(1-\frac{S_j}{A})^2$, while if all objects are placed in the bin $A_2 - A_1$, the variance is $\sigma_2^2 = n^2[\sum_{j=2}^{M-1}(1-\frac{S_j}{A})^2 + (\frac{S_1}{A})^2]$. It is easy to show that $\sigma_1^2 = \sigma_2^2 - 2n^2\frac{S_1}{A} + n^2$. In a similar fashion, if all the objects are in $A_i - A_{i-1}$, for some $i \in \{3,\dots,M\}$, the associated variance is $\sigma_i^2 = n^2[\sum_{j=i}^{M-1}(1-\frac{S_j}{A})^2 + \sum_{j=1}^{i-1}(\frac{S_j}{A})^2]$. In general, we can write $\sigma_{i-1}^2 = \sigma_i^2 - 2n^2\frac{S_{i-1}}{A} + n^2$.

Since $\frac{S_1}{A} < \frac{S_2}{A} < \dots < \frac{S_{M-1}}{A} < 1$, either σ_1^2 or σ_M^2 must give the maximal variance σ^2 and therefore the corresponding maximal index. If all the partial volumes satisfy $\frac{S_i}{A} < \frac{1}{2}$ for all $i = 1,\dots,M-1$, then the variance σ_1^2 is maximal. Therefore, if $\frac{S_{M-1}}{A} < \frac{1}{2}$, the maximal index occurs when all objects are in the most internal bin A_1. Otherwise, if $\frac{S_{M-1}}{A} > \frac{1}{2}$, the maximal index occurs when all objects are in the outer bin $A_M - A_{M-1}$. In both cases, the maximal index is greater than unity.

APPENDIX B: GENERALIZED INDEX WHEN $S_j \approx s$

For simplicity we consider the case of equal-sized bins when $S_j = S$ for all $j = 1,\dots,M$. We show that if small bins are used, with $S \approx s$, the index is close to or above the CSR limiting value, whether or not the distribution is uniformly at random or not, and therefore is an artefact of the bin choice. Suppose small bins of size S are chosen so that at most one object size s can be in each bin. Then $\sigma^2 = \frac{1}{M}[n(1-\frac{n}{M})^2 + (M-n)(\frac{n}{M})^2] = \frac{n}{M}(1-\frac{n}{M})$. Using $A = MS$, $\sigma_0^2 \approx n^2/M$ for $M \gg 1$, so that $I = \frac{1}{n}(1-\frac{n}{M}) = \frac{1}{n}(1-\frac{dS}{s})$. If $S = s$, the index will be precisely at the CSR value, generalizing the known result for $s = 1$ and $S = 1$ [25]. If $S < s$ then the index will lie above the CSR value. However, we note that the condition $s \ll S_j$ for all j is necessary in deriving the CSR limiting value in Sec. II A.

[1] H. Aref and S. W. Jones, Phys. Fluids A **1**, 470 (1989).
[2] S. W. Jones, Phys. Fluids A **3**, 1081 (1991).
[3] B. J. Binder and S. M. Cox, Fluid Dynam. Res. **40**, 34 (2008).
[4] B. J. Binder, Phys. Lett. A **374**, 3483 (2010).
[5] J. H. Phelps and C. L. Tucker, Chem. Eng. Sci. **61**, 6826 (2006).
[6] D. Volfson, S. Cookson, J. Hasty, and L. S. Tsimring, Proc. Natl. Acad. Sci. USA **105**, 15346 (2008).
[7] H. Cho, H. Jönsson, K. Campbell, P. Melke, J. W. Williams, B. Jedynak, A. M. Stevens, A. Groisman, and A. Levchenko, PLoS Biol **5**, 2614 (2007).
[8] H. Jönsson and A. Levchenko, Multiscale Model. Simul. **3**, 346 (2005).
[9] F. Hammad, R. Watling, and D. Moore, Mycol. Res. **97**, 275 (1993).
[10] P. Haase, J. Veg. Sci. **6**, 575 (1995).
[11] D. Malkinson, R. Kadmon, and D. Cohen, J. Veg. Sci. **14**, 213 (2003).
[12] P. Haase, F. I. Pugnaire, S. C. Clark, and I. D. Incoll, J. Veg. Sci. **7**, 527 (1996).
[13] J. Szwagrzyk and M. Czerwczak, J. Veg. Sci. **4**, 469 (1993).
[14] C. P. H. Mulder, E. Bazeley-White, P. G. Dimitrakopoulos, A. Hector, M. Scherer-Lorenzen, and B. Schmid, Oikos **107**, 50 (2004).

[15] W. G. Weng, T. Chen, H. Y. Yuan, and W. C. Fan, Phys. Rev. E **74**, 036102 (2006).

[16] A. Schadschneider, Physica A **313**, 153187 (2002).

[17] M. J. Simpson, K. A. Landman, B. J. Hughes, and A. E. Fernando, Physica A **389**, 1412 (2010).

[18] E. J. Hackett-Jones, K. A. Landman, D. F. Newgreen, and D. C. Zhang, J. Theor. Biol. **287**, 148 (2011).

[19] B. J. Binder, K. A. Landman, M. J. Simpson, M. Mariani, and D. F. Newgreen, Phys. Rev. E **78**, 031912 (2008).

[20] B. J. Binder and K. A. Landman, J. Theor. Biol. **259**, 541 (2009).

[21] D. Zhang, I. M. Brinas, B. J. Binder, K. A. Landman, and D. F. Newgreen, Dev. Biol. **339**, 280 (2010).

[22] D. Chowdhury, A. Schadschneider, and K. Nishinari, Phys. Life. Rev. **2**, 318 (2005).

[23] B. D. Ripley, *Spatial Statistics* (Wiley, New York, 1981).

[24] P. J. Diggle, *Statistical Analysis of Spatial Point Patterns* (Academic, London, 1983).

[25] B. J. Binder and K. A. Landman, Phys. Rev. E **83**, 041914 (2011).

[26] F. Eggenberger and G. Pólya, Z. Angew. Math. Mech. **1**, 279 (1923).

[27] B. J. Binder, E. J. Hackett-Jones, S. J. Tuke, and K. A. Landman, ANZIAM J. (to be published).

[28] V. Privman, J.-S. Wang, and P. Nielaba, Phys. Rev. B. **43**, 3366 (1991).

[29] Lj. Budinski-Petković and U. Kozmidis-Luburić, Physica A **236**, 211 (1997).

[30] S. S. Manna and N. M. Švrakić, J. Phys. A **24**, L671 (1991).

[31] J. W. Evans, Rev. Mod. Phys. **65**, 1281 (1993).

[32] E. Eisenberg and A. Baram, J. Phys. A **33**, 1729 (2000).

[33] M. C. Bartelt and V. Privman, J. Chem. Phys. **93**, 6820 (1990).

[34] M. D. Penrose, Commun. Math. Phys. **218**, 153 (2001).

[35] E. Palsson, J. Theor. Biol. **254**, 1 (2008).

[36] C. P. Beatrici and L. G. Brunnet, Phys. Rev. E **84**, 031927 (2011).

[37] N. J. Armstrong, K. J. Painter, and J. A. Sherratt, J. Theor. Biol. **243**, 98 (2006).

[38] C. M. Topaz, A. L. Bertozzi, and M. A. Lewis, Bull. Math. Biol. **68**, 1601 (2006).

[39] J. Hoshen and R. Kopelman, Phys. Rev. B **14**, 3438 (1976).

[40] B. D. Ripley, J. Appl. Probab. **13**, 255 (1976).

[41] B. D. Ripley, J. R. Stat. Soc. Ser. B (Methodol.) **41**, 368 (1979).