Proc. Indian Acad. Sci. (Math. Sci.) Vol. 121, No. 2, May 2011, pp. 229–244.

Bias expansion of spatial statistics and approximation of differenced lattice point counts

DANIEL J NORDMAN[1] and SOUMENDRA N LAHIRI[2]

[1]Department of Statistics, Iowa State University, Ames, IA 50011, USA
[2]Department of Statistics, Texas A&M University, College Station, TX 77843, USA
E-mail: dnordman@iastate.edu; snlahiri@stat.tamu.edu

MS received 14 May 2010; revised 2 January 2011

Abstract. Investigations of spatial statistics, computed from lattice data in the plane, can lead to a special lattice point counting problem. The statistical goal is to expand the asymptotic expectation or large-sample bias of certain spatial covariance estimators, where this bias typically depends on the shape of a spatial sampling region. In particular, such bias expansions often require approximating a difference between two lattice point counts, where the counts correspond to a set of increasing domain (i.e., the sampling region) and an intersection of this set with a vector translate of itself. Non-trivially, the approximation error needs to be of smaller order than the spatial region's perimeter length. For all convex regions in 2-dimensional Euclidean space and certain unions of convex sets, we show that a difference in areas can approximate a difference in lattice point counts to this required accuracy, even though area can poorly measure the lattice point count of any single set involved in the difference. When investigating large-sample properties of spatial estimators, this approximation result facilitates direct calculation of limiting bias, because, unlike counts, differences in areas are often tractable to compute even with non-rectangular regions. We illustrate the counting approximations with two statistical examples.

Keywords. Area approximations; convex sets; increasing domain; infill sampling; lattice data; spatial processes; spatial covariance.

1. Introduction

Statistics computed from $\mathbb{R}^2$-spatial data often have biases which depend heavily on the geometry of a spatial sampling region. To investigate the bias and other expectation properties of such spatial statistics in large samples, a common framework supposes that points on the integer lattice $\mathbb{Z}^2$ indicate the locations of observations within a $\mathbb{R}^2$-sampling region which grows in size to allow increasingly more observations to be sampled. In this setting, asymptotic expressions for bias can require computing *differenced* lattice point counts (or, in fact, limits of these) over regions increasing in domain. A complicating factor is that the number of integers inside an expanding $\mathbb{R}^2$-region is generally difficult to determine. On the other hand, a region's area is often much more tractable to compute directly. This leads to the question of how adequately areas can approximate lattice point counts when studying the asymptotic bias of spatial statistics. Nordman and Lahiri [20] illustrated such approximations in calculating the bias of a spatial subsampling method for variance estimation, though no formal justification for these 'area-for-count' approximations was given. Our goal here is to describe and illustrate the general counting problem of interest for spatial lattice data in $\mathbb{R}^2$ (collected on expanding regions) and to rigorously justify the approximation of lattice point counts with areas in a manner useful for explicitly determining the limiting bias of certain spatial statistics, such as covariance estimators.

To formulate the lattice point counting problem, we consider a common sampling framework in spatial statistics [20,26]. Suppose D is a subset of $\mathbb{R}^2$ with non-empty interior serving as a 'template', $b > 1$ is a positive enlargement factor, and a region $D_b = bD$ is obtained by 'inflating' the template D by b. We treat $D_b \subset \mathbb{R}^2$ as a spatial sampling region and suppose that spatial observations are positioned at integer $\mathbb{Z}^2$ points inside D_b. Then, the region's area $|D_b|$ and spatial sample size $\#\mathbb{Z}^2 \cap D_b$ increase as $b \to \infty$ in a so-called 'increasing domain' asymptotic framework for studying spatial statistics [4], where we let $\#B$ denote the cardinality of a finite set B and $|B|$ denote the Lebesgue area of an uncountable (measurable) set B. As will be more carefully illustrated in §2, bias expansions for sample covariances on $\mathbb{Z}^2 \cap D_b$ often require subtracted counts

$$\#\mathbb{Z}^2 \cap D_b \; - \; \#\mathbb{Z}^2 \cap D_b \cap (D_b + \mathbf{k}), \quad \mathbf{k} = (k_1, k_2)' \in \mathbb{Z}^2, \tag{1}$$

corresponding to an expanding region and an intersection of this region with an integer translate of itself. Specifically, a quantity of interest for studying large-sample bias is the scaled limiting difference

$$\mathcal{L}(\mathbf{k}) = \lim_{b \to \infty} \frac{\#\mathbb{Z}^2 \cap D_b - \#\mathbb{Z}^2 \cap D_b \cap (D_b + \mathbf{k})}{b}, \quad \mathbf{k} \in \mathbb{Z}^2, \tag{2}$$

provided this limit exists. The divisor b above corresponds to the typical order of the perimeter length of D_b. The challenge is determining (2) explicitly for all possible values of $\mathbf{k} \in \mathbb{Z}^2$ for use in large-sample bias expansions.

As a region's area is often easier to determine than its lattice point count, we use differenced areas $|D_b| - |D_b \cap (D_b + \mathbf{k})|$ to approximate the corresponding differenced counts (1) required in (2), but caution is needed to justify this approach. To determine the limit (2), the complication is that the error in approximating (1) with areas must be of smaller order than the usual perimeter length $O(b)$ of the sampling region D_b or number of lattice points near the boundary of D_b. That is, we require a side-condition that

$$\text{the error in approximating (1) with } |D_b| - |D_b \cap (D_b + \mathbf{k})| \text{ satisfies } o(b), \tag{3}$$

for any fixed $\mathbf{k} \in \mathbb{Z}^2$ as $b \to \infty$. This makes the approximation problem non-trivial. When considering an inflated region D_b, the discrepancy between the region's lattice point count $\#\mathbb{Z}^2 \cap D_b$ and area $|D_b|$ is often *exactly* of order $O(b)$ due to contributions from straight line boundary components (Lemma 2.1.1, Theorem 2.3.3 of [13]; Theorem 1.7 of [15]). For example, Pick's formula [21] states that if D_b is based on a convex polygonal template $D \subset \mathbb{R}^2$ with vertices defined by integer coordinates, then the discrepancy is

$$\#\mathbb{Z}^2 \cap D_b - |D_b| = 1 + \frac{b}{2} \cdot \#\mathbb{Z}^2 \cap \partial D,$$

where ∂D denotes the boundary of D, so that the error in approximating a count $\#\mathbb{Z}^2 \cap D_b$ with area is exactly $O(b)$ in this case. However, even for these regions, the results to follow show that (3) still holds when approximating *differenced* counts in bias expansions.

The rest of the manuscript is organized as follows. We conclude this section with a brief survey of mathematical literature in approximating lattice point counts. Section 2 provides some simple examples to explain how such differenced counts, like (1) and (2), may arise in large-sample bias expansions of spatial statistics. Section 3 describes the main 'area-for-count' approximation results, which validate the approximation of differenced counts with differenced areas for *any* regions $D_b = bD \subset \mathbb{R}^2$ defined by a convex template $D \subset \mathbb{R}^2$. These approximation results show, for instance, that $\mathcal{L}(\mathbf{k}) = \|\mathbf{k}\|$ in (2) if the template D is a unit circle and $\mathcal{L}(\mathbf{k}) = \max\{|k_1|, |k_2|\}$ if D has a diamond shape (see figure 1 in §2). The results of §3 are then applied in §4 to determine limiting expectations in two statistical examples, which provide concrete motivation and application of the bias expansions here and also aim to illustrate how the geometry of a spatial sampling region D_b influences large-sample bias when D_b is *non-rectangular*. In contrast, related asymptotic studies of spatial statistics for lattice data have often focused on rectangular sampling regions $D_b = bD$ for $D = (0, \ell_1] \times (0, \ell_2]$, integer $b, \ell_1, \ell_2 \geq 1$; see [10] for spectral mean estimation, [22] for spatial subsampling, and [8] for variograms. (With such rectangular regions, the lattice point counts required in (1) are rather trivial to obtain because $\#\mathbb{Z}^2 \cap D_b - \#\mathbb{Z}^2 \cap D_b \cap \big(D_b + \mathbf{k}\big) = |D_b| - |D_b \cap \big(D_b + \mathbf{k}\big)|$ holds exactly along with

$$|D_b| - |D_b \cap \big(D_b+\mathbf{k}\big)| = \begin{cases} b\ell_1|k_2|+b\ell_2|k_1|-\ell_1\ell_2|k_1k_2|, & \text{if } |k_i| \leq b\ell_i,\ i=1,2 \\ b^2\ell_1\ell_2, & \text{otherwise} \end{cases}$$

so that $\mathcal{L}(\mathbf{k}) = \ell_1|k_2|+\ell_2|k_1|$, $\mathbf{k} = (k_1, k_2)' \in \mathbb{Z}^2$ is simple to determine in (2).) Section 5 provides some concluding remarks and describes potential extensions of the approximation problem to higher sampling dimensions $\mathbb{R}^d$, $d \geq 3$. Proofs of the main results are given in §6.

A great amount of literature exists in number theory and enumerative combinatorics concerning lattice point counting and volume approximations; for example, see historical references in [2,13]. Traditionally, lattice point counting in the plane $\mathbb{R}^2$ has considered the size $\Delta(b) = \#\mathbb{Z}^2 \cap D_b - |D_b|$ of the discrepancy, when the template set D is prescribed by the inside of a curve. In this case, the search for the best possible estimates of the remainder $\Delta(b)$ is known as the 'O-problem' in number theory [15]. Usually the curves considered are sufficiently smoothly winding and estimation tools for exponential sums are applied (see Chapter 2 of [13]) where better approximations of exponential sums often lead to sharper 'O-estimates'. For example, for convex sets with a smooth border, van der Corput's [29] bound on the remainder $\Delta(b)$ is $O(b^{46/69+\epsilon})$, while an improved bound, based on more recent analytic methods for exponential sums is $O(b^{46/73+\epsilon})$ for appropriately smooth curves [12]. For inflated polytopes $D_b \subset \mathbb{R}^d$, $d \geq 1$ (usually with vertices defined by rational coordinates), there are formulae for directly computing lattice point counts, based on Ehrhart quasi-polynomials [7,18] and generating functions [28] for example, as well as algorithmic theory and methods for determining these counts (cf. [1,2,6,5,17] and references therein). In considering the present approximation problem (1) with general convex sets of increasing domain and self-intersections with translates, none of the work mentioned above is directly or generally applicable, which led us to consider results in §3 for directly determining (2) from areas.

2. Initial examples of differenced counts/bias expansions

Before presenting the approximation results of §3 or the statistical examples of §4, it is helpful to illustrate how such counting expansions arise with some simple statistics of spatial lattice data. In particular, many spatial measures of association involve statistics computed from averages of observations at different spatial lags (or cross-products of these), such as sample covariances, sample variograms, black–white join-count statistics and unnormalized forms of Moran's I statistic (cf. Chapter 2 of [3]). The examples below and in §4 fit this basic mold.

Consider a sampling region $D_b = bD \subset \mathbb{R}^2$ from §1 along with a collection $\{Y_\mathbf{s} : \mathbf{s} \in \mathbb{Z}^2\}$ of stationary, mean-zero real-valued random variables on the lattice with covariances $c(\mathbf{k}) = \mathrm{Cov}(Y_\mathbf{s}, Y_{\mathbf{s}+\mathbf{k}}) = \mathrm{E}(Y_\mathbf{s} Y_{\mathbf{s}+\mathbf{k}})$, $\mathbf{k}, \mathbf{s} \in \mathbb{Z}^2$, where E denotes the expected value. The available data are $\{Y_\mathbf{s} : \mathbf{s} \in \mathbb{Z}^2 \cap D_b\}$ (sampled at integer points inside D_b) and we let $N_b = \#\mathbb{Z}^2 \cap D_b$ denote the sample size. Fix a lag $\mathbf{k} \in \mathbb{Z}^2$ and consider estimating the covariance $c(\mathbf{k})$ with a sample covariance

$$\hat{c}(\mathbf{k}) = \frac{1}{N_b} \sum_{\mathbf{s} \in \mathbb{Z}^2 \cap D_b \cap (D_b - \mathbf{k})} Y_\mathbf{s} Y_{\mathbf{s}+\mathbf{k}},$$

where the sum is taken over locations $\mathbf{s} \in \mathbb{Z}^2 \cap D_b$ whose translate by $\mathbf{k}$ lies in D_b (implying $\mathbf{s} \in D_b - \mathbf{k}$). Then, the bias in estimating $c(\mathbf{k})$ is given by the expected difference

$$\mathrm{E}\big[\hat{c}(\mathbf{k})\big] - c(\mathbf{k}) = -\frac{c(\mathbf{k})}{N_b}\Big[\#\mathbb{Z}^2 \cap D_b - \#\mathbb{Z}^2 \cap D_b \cap (D_b + \mathbf{k})\Big], \tag{4}$$

using $\#\mathbb{Z}^2 \cap D_b \cap (D_b + \mathbf{k}) = \#\mathbb{Z}^2 \cap D_b \cap (D_b - \mathbf{k})$ and $N_b = \#\mathbb{Z}^2 \cap D_b$. If the template D is convex (or a finite union of convex sets), then the number of integer cubes $\mathbf{i} + [0,1]^2$, $\mathbf{i} \in \mathbb{Z}^2$, intersecting the boundary of D_b is of smaller order $O(b)$ than the totality of observations N_b in D_b so that $N_b/|D_b| = N_b/[b^2|D|] \to 1$ holds as $b \to \infty$. Substituting $b^2|D|$ for N_b in (4) and if the limit (2) can be found, we can obtain a very explicit expansion for the main bias component of the sample covariance as $b \to \infty$ (i.e., in large-samples),

$$\mathrm{E}\big[\hat{c}(\mathbf{k})\big] - c(\mathbf{k}) = -c(\mathbf{k})\frac{\mathcal{L}(\mathbf{k})}{|D|}\frac{1}{b} + o\Big(b^{-1}\Big), \tag{5}$$

where, as mentioned in §1, the geometry of the template D determines $\mathcal{L}(\mathbf{k})$ from (2). Figure 1 demonstrates values of $\mathcal{L}(\mathbf{k})$ for several template D shapes, which were computed by substituting areas in (2) as justified by results in §3. While it is simplistic, this example shows that the count difference (1) and scaled limiting difference (2) enter into the expression for bias (4) and the limiting approximation (5). Since (1) and (2) are

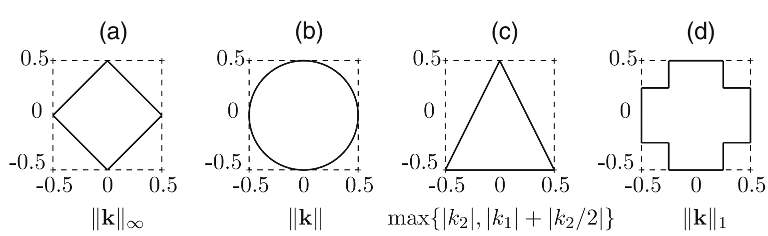

Figure 1. Examples of $\mathbb{R}^2$-templates D outlined by solid lines and associated bias constants $\mathcal{L}(\mathbf{k})$ from (2). For $\mathbf{k} = (k_1, k_2)' \in \mathbb{Z}^2$, vector norms are $\|\mathbf{k}\|_\infty = \max\{|k_1|, |k_2|\}$, $\|\mathbf{k}\| = (|k_1|^2 + |k_2|^2)^{1/2}$ and $\|\mathbf{k}\|_1 = |k_1| + |k_2|$.

difficult to obtain, approximating these quantities by differences in area (3) is of practical importance.

Another more ambitious bias expansion is as follows. Considering the same spatial sampling framework with convex template D and mean-zero stationary process, suppose the covariances satisfy a weak dependence condition $\sum_{\mathbf{k}\in\mathbb{Z}^2}|c(\mathbf{k})| < \infty$. Let $\bar{Y}_b = \sum_{\mathbf{s}\in\mathbb{Z}^2\cap D_b} Y_\mathbf{s}/N_b$ denote the spatial sample mean with $N_b = \#\mathbb{Z}^2 \cap D_b$, which has limiting variance given by $\sigma^2_\infty \equiv \lim_{b\to\infty} N_b \mathrm{E}(\bar{Y}_b^2) = \sum_{\mathbf{k}\in\mathbb{Z}^2} c(\mathbf{k})$. Then, the expected difference between $N_b\bar{Y}_b^2$ and σ^2_∞ (i.e., difference between finite sample and limiting variances) is given by

$$\begin{aligned} N_b\mathrm{E}(\bar{Y}_b^2) - \sigma^2_\infty &= \frac{1}{N_b}\sum_{\mathbf{k}\in\mathbb{Z}^2}\{\#\mathbb{Z}^2\cap D_b\cap(D_b+\mathbf{k})\}c(\mathbf{k}) - \sigma^2_\infty \\ &= -\sum_{\mathbf{k}\in\mathbb{Z}^2}\frac{\{\#\mathbb{Z}^2\cap D_b - \#\mathbb{Z}^2\cap D_b\cap(D_b+\mathbf{k})c(\mathbf{k})\}}{N_b} \\ &= -\frac{1}{b}\sum_{\mathbf{k}\in\mathbb{Z}^2}\frac{\mathcal{L}(\mathbf{k})c(\mathbf{k})}{|D|} + o(b^{-1}); \end{aligned} \tag{6}$$

the last step follows from the limits (2) and $N_b/[b^2|D|] \to 1$ as $b\to\infty$ along with the dominated convergence theorem (DCT) to justify the change to pointwise limits along the entire summation. (The DCT uses covariance summability and the fact that, for convex regions, the count difference (1) is bounded by Cb for some $C > 0$ independent of $\mathbf{k}\in\mathbb{Z}^2$ and $b > 1$.) 'Area-for-count' approximations then allow the main coefficient $\mathcal{L} \equiv -\sum_{\mathbf{k}\in\mathbb{Z}^2}\mathcal{L}(\mathbf{k})c(\mathbf{k})/|D|$ in the large-sample bias expression (6) to be explicitly computed (i.e., determine $\mathcal{L}(\mathbf{k})$ in (2) by substituting areas), where the sampling region's geometry is again reflected in the coefficients $\mathcal{L}(\mathbf{k})$ (e.g. figure 1). We also may view (6) as the spatial extension of a well-known result from time series (cf. Corollary 6.1.1.2 of [9]), which expands the variance of the sample mean $\bar{X}_b = \sum_{i=1}^b X_i/b$ of a stationary, real-valued time stretch $X_1,\ldots,X_b$ (integer $b\ge 1$) as

$$\begin{aligned} b\,\mathrm{Var}(\bar{X}_b) - \sum_{k\in\mathbb{Z}} c_X(k) &= b^{-1}\sum_{k\in\mathbb{Z}}\{(b-|k|)-b\}c_X(k) \\ &\quad + O\left(b^{-1}\sum_{|k|\ge b}|k||c_X(k)|\right) \\ &= -b^{-1}\sum_{k\in\mathbb{Z}}|k|c_X(k) + o\Big(b^{-1}\Big), \end{aligned}$$

assuming covariances $c_X(k) = \mathrm{Cov}(X_0, X_k)$, $k\in\mathbb{Z}$ satisfy $\sum_{k=1}^\infty k|c_X(k)| < \infty$; this derivation for time series requires subtracting two simple lattice point counts $\#\mathbb{Z}\cap(0,b] = b$ and $\#\mathbb{Z}\cap(0,b]\cap((0,b]+k) = b-|k|$ for $|k|\le b$, over an interval $(0,b]$ (i.e., the time series counterpart of the spatial sampling region $D_b\subset\mathbb{R}^2$). Because the spatial expansion in (6) involved considering differenced lattice point counts (and their limits upon scaling by b^{-1}) at many lags $\mathbf{k}\in\mathbb{Z}^2$ simultaneously, this provides a second prototypical example of how such counts can be encountered in limiting biases of spatial statistics and where approximations would again be useful. Section 4 presents further examples of bias/expectation expansions to motivate their use in more concrete statistical settings, with one example illustrating the use of the main coefficient in (6).

3. Main results on approximating differenced counts in $\mathbb{R}^2$

For $\mathbf{k} \in \mathbb{Z}^2$ and a region $D_b = bD \subset \mathbb{R}^2$ with $b > 0$, define the (signed) error

$$\begin{aligned}\Delta(D_b, \mathbf{k}) \equiv\ & \{\#\mathbb{Z}^2 \cap D_b - \#\mathbb{Z}^2 \cap D_b \cap (D_b + \mathbf{k})\} \\ & - \{|D_b| - |D_b \cap (D_b + \mathbf{k})|\}\end{aligned}$$

from approximating the difference of lattice point counts (1) with a difference in areas. Recall from (3) that, for large-sample bias expansions, we require the error $\Delta(D_b, \mathbf{k})$ to be $o(b)$ as $b \to \infty$ for any fixed $\mathbf{k} \in \mathbb{Z}^2$. Theorem 3.1 below shows that this indeed holds for the class of all convex regions $D_b \subset \mathbb{R}^2$. Interestingly, the individual approximation errors with either '$\#\mathbb{Z}^2 \cap D_b - |D_b|$' or '$\#\mathbb{Z}^2 \cap D_b \cap (D_b + \mathbf{k}) - |D_b \cap (D_b + \mathbf{k})|$' can be of exact order $O(b)$ for many convex regions D_b but, upon subtraction, these will cancel to the extent that $\Delta(D_b, \mathbf{k}) = o(b)$ holds.

Theorem 3.1. *Let $\mathbf{k} \in \mathbb{Z}^2$ and suppose that $D \subset \mathbb{R}^2$ is a closed convex domain with non-empty interior and $D_b = bD$ for $b > 0$. Then, the error in approximating* (1) *with $|D_b| - |D_b \cap (D_b + \mathbf{k})|$ is bounded as*

$$\left|\Delta(D_b, \mathbf{k})\right| \leq C\|\mathbf{k}\|, \tag{7}$$

for a constant $C \in (1, \infty)$ depending on the template D, but not on $b > 0$ or $\mathbf{k} \in \mathbb{Z}^2$.

Remark 1. Theorem 3.1 also remains valid if points on the boundary of D are arbitrarily excluded; statistically, this allows for missing observations at the boundaries.

As a small numerical illustration, consider a region $D_b = bD$ formed by the unit circle $D = \{\mathbf{x} \in \mathbb{R}^2 : \|\mathbf{x}\| \leq 1\}$ as a template. For a given $\mathbf{k} \in \mathbb{Z}^2$, we approximate the scaled difference in counts $b^{-1}[\#\mathbb{Z}^2 \cap D_b - \#\mathbb{Z}^2 \cap D_b \cap (D_b + \mathbf{k})]$ (cf. (2), (4), or (6)) with areas determined as

$$\frac{|D_b| - |D_b \cap (D_b + \mathbf{k})|}{b} = b\pi - 2b \arccos\left(\frac{\|\mathbf{k}\|}{2b}\right) + b \sin\left(\frac{\|\mathbf{k}\|}{b}\right) \tag{8}$$

for the circle, which produces an error $\Delta(D_b, \mathbf{k})/b$. As $b \to \infty$ with fixed $\mathbf{k} \in \mathbb{R}$, it can be checked that (8) has a limit $\|\mathbf{k}\|$ while Theorem 3.1 implies $\Delta(D_b, \mathbf{k})/b = o(1)$, so that these together provide an exact form for the limit $\mathcal{L}(\mathbf{k}) = \lim_{b\to\infty} b^{-1}[\#\mathbb{Z}^2 \cap D_b - \#\mathbb{Z}^2 \cap D_b \cap (D_b + \mathbf{k})] = \|\mathbf{k}\|$ as desired in (2). Theorem 3.1 was similarly applied to determine similar limits $\mathcal{L}(\mathbf{k})$ from (2) with other templates D in figures 1(a)–(c). While limits are our main interest here, figure 2 displays a box plot of relative errors $\{\Delta(D_b, \mathbf{k})/[b\mathcal{L}(\mathbf{k})] : \mathbf{k} \in \mathbb{Z}^2, \|\mathbf{k}\| \leq 10\}$, with one box plot for each circular region D_b, $b = 10, \ldots, 100$, to show the finite-sample errors from the approximation (8) at all lags $\mathbf{k} \in \mathbb{Z}^2$, $\|\mathbf{k}\| \leq 10$ (i.e., the errors $\Delta(D_b, \mathbf{k})/b$ relative to the limit $\mathcal{L}(\mathbf{k}) = \|\mathbf{k}\|$); these errors, in accordance with Theorem 3.1, converge to zero as b grows.

As shown in Theorem 3.2 below, similar approximation results are also possible for non-convex regions $D_b = bD$, with templates D representable as a finite union of convex sets. Specifically, we consider $D = S_1 \cup S_2$ formed by the union of two convex sets S_1, S_2 with boundaries that are finitely intersecting. Area approximations are analogously possible for larger unions of convex sets which have finitely many border points in common,

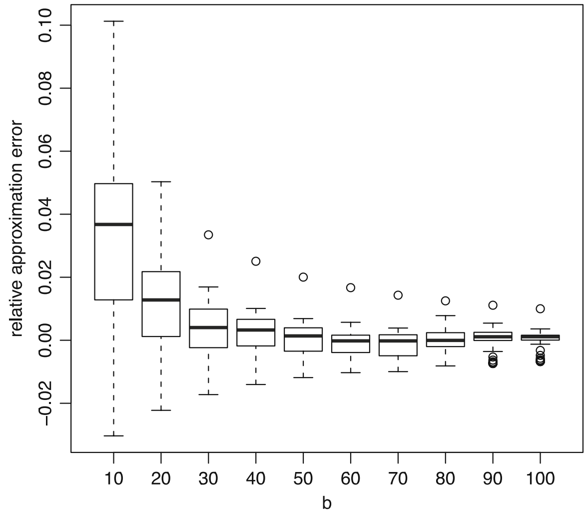

Figure 2. For each region $D_b = bD$, with D as the unit circle and $b = 10, \ldots, 100$, a box plot displays the collection of relative approximation errors $\Delta(D_b, \mathbf{k})/(b\|\mathbf{k}\|)$ for all lags $\mathbf{k} \in \mathbb{Z}^2$, $\|\mathbf{k}\| \leq 10$ (i.e., there is one box plot per region defined by b).

but considering the union of two sets keeps the proof of Theorem 3.2 relatively simple. Write ∂B to denote the boundary of a set $B \subset \mathbb{R}^2$.

Theorem 3.2. *Suppose that* $D_b = bD$ *for* $D = S_1 \cup S_2$; $S_i \subset \mathbb{R}^2$ *is a closed convex domain*, $i = 1, 2$; $S_3 = S_1 \cap S_2$ *has a nonempty interior*; *and* $\partial S_1 \cap \partial S_2$ *is finite. For some* $C > 0$ (*depending on D but not on* $b > 0$ *or* $\mathbf{k} \in \mathbb{Z}^2$), *the error in approximating* (1) *with* $|D_b| - |D_b \cap (D_b + \mathbf{k})|$ *is bounded by* $|\Delta(D_b, \mathbf{k})| \leq C\|\mathbf{k}\|^2$ *and*

$$\left|\left(|D_b| - |D_b \cap (D_b + \mathbf{k})|\right) - \left(A_{1,b}(\mathbf{k}) + A_{2,b}(\mathbf{k}) - A_{3,b}(\mathbf{k})\right)\right| \leq C\|\mathbf{k}\|^2,$$

where $A_{j,b}(\mathbf{k}) = |bS_j| - |bS_j \cap (bS_j + \mathbf{k})|$, $j = 1, 2, 3$.

Remark 2. In place of $|D_b| - |D_b \cap (D_b + \mathbf{k})|$, Theorem 3.2 also gives a computationally simpler estimate $A_{1,b}(\mathbf{k}) + A_{2,b}(\mathbf{k}) - A_{3,b}(\mathbf{k})$ of the count difference (1) involving three separately computed area differences $A_{j,b}(\mathbf{k})$ over convex sets bS_j, $j = 1, 2, 3$. This was applied to determine the quantity (2) in figure 1(d) in a simple fashion, as Theorem 3.2 shows that (3) holds for any fixed $\mathbf{k} \in \mathbb{Z}^2$. Given Theorem 3.1 and the fact that S_1, S_2 and S_3 are convex, the key to establishing Theorem 3.2 is bounding the area and lattice point count of $b(S_j \setminus S_{3-j}) \cap (b(S_{3-j} \setminus S_j) + \mathbf{k})$, $j = 1, 2$ by $C\|\mathbf{k}\|^2$; these are $\mathbb{R}^2$-points, not common to either set, which move from bS_j to bS_{3-j} upon translation by $\mathbf{k}$ for $j = 1, 2$. This bound follows from the assumption that S_1, S_2 share finitely many common border points.

Under stronger conditions on the template D, tighter bounds on the approximation errors in Theorems 3.1 and 3.2 may also be possible. If, for example, the boundary of D is a smooth curve with finite radius of curvature, then certain boundary contributions (e.g., B_1 and B_3 from the proof in §6) to the discrepancy $\Delta(D_b, \mathbf{k})$ could possibly be estimated with exponential sum techniques to improve the bound in Theorem 3.1 (cf. Ch. 8 of [13]). The improvement of the bounds presents a more complicated problem and is not of direct interest for the asymptotic bias expansions which require validating (3).

4. Examples and applications

Here we illustrate two asymptotic investigations of spatial statistics, which involve differenced lattice point counts and the approximation results from §3. Section 4.1 addresses spatial subsampling variance estimators with spatial lattice data collected in an 'increasing domain asymptotic' structure (i.e., our standard formulation of sampling regions in this paper). Section 4.2 considers a problem with variogram estimators when sampling spatial lattice data in a 'fixed-domain infilling' structure (cf. [4]), which can be treated with our approximation results though the sampling framework is different.

4.1 *Spatial subsampling variance estimators*

Subsampling, as outlined by [24], is a general method for nonparametric estimation which loosely involves carving 'mini' data-sets (called subsamples) out of a larger available set of data. Subsamples provide several replications of the original data, though on a smaller scale, which can be combined for inference. Several authors have proposed subsampling for estimating the variance of spatial statistics computed from spatial lattice data [22,25–27]. We briefly describe a subsampling variance estimator for the spatial sample mean (other statistics are handled similarly) and expand its large-sample bias, which can be determined by approximating differenced lattice point counts with areas as in (3).

As in §2, let $\{Y_{\mathbf{s}} : \mathbf{s} \in \mathbb{Z}^2\}$ denote a stationary, real-valued spatial process with mean $\mathrm{E}(Y_{\mathbf{s}}) = \mu \in \mathbb{R}$ and covariances $c(\mathbf{k}) = \mathrm{Cov}(Y_{\mathbf{s}}, Y_{\mathbf{s}+\mathbf{k}})$, $\mathbf{k} \in \mathbb{Z}^2$. Suppose the available data $\{Y_{\mathbf{s}} : \mathbf{s} \in \mathbb{Z}^2 \cap R_n\}$ are those observations located in a spatial sampling region $R_n = \lambda_n R_0 \subset \mathbb{R}^2$, defined by a template $R_0 \subset \mathbb{R}^2$ and a sequence $\{\lambda_n\}$ of scaling factors with $\lambda_n \to \infty$ as $n \to \infty$. More observations become available as $n \to \infty$ and the region R_n grows in size. If $\bar{Y}_n = \sum_{\mathbf{s}\in\mathbb{Z}^2\cap R_n} Y_{\mathbf{s}}/N_n$ denotes the sample mean of observations in R_n, where $N_n = \#\mathbb{Z}^2 \cap R_n$ is the sample size, then subsampling allows nonparametric estimation of the variance $\sigma_n^2 \equiv N_n \mathrm{Var}(\bar{Y}_n) = N_n \mathrm{E}(\bar{Y}_n - \mu)^2$, as described next.

Let $D \subset \mathbb{R}^2$ be a convex template and define a 'subregion' $D_b = bD$ with a sequence of scaling factors $b \equiv b_n > 0$ such that $b_n \to \infty$, $b_n^{3/2}/\lambda_n \to 0$ as $n \to \infty$ (i.e., as $n \to \infty$, subregions are intended to grow in size but remain 'small' relative to the original sampling region $R_n = \lambda_n R_0$). Now create several subsample regions as translates of D_b within R_n by $D_{\mathbf{i},b} = D_b + \mathbf{i}$, $\mathbf{i} \in \mathcal{I}_n$, where $\mathcal{I}_n \equiv \{\mathbf{i} \in \mathbb{Z}^2 : D_{\mathbf{i},b} \subset R_n\}$ denotes the index set of all (overlapping) subsample regions contained completely within R_n. These subregions $D_{\mathbf{i},b}$, $\mathbf{i} \in \mathcal{I}_n$ provide repeated copies of spatial data and, for studying the sample mean $\bar{Y}_n$ on R_n, we compute the sample average $\bar{Y}_{\mathbf{i},b} = \sum_{\mathbf{s}\in\mathbb{Z}^2\cap D_{\mathbf{i},b}} Y_s/N_b$ of each subsample, where $N_b = \#\mathbb{Z}^2 \cap D_b$ denotes the number of observations in a subsample. Then, the subsampling variance estimator of $\sigma_n^2 \equiv N_n \mathrm{Var}(\bar{Y}_n)$ is given by

$$\hat{\sigma}_n^2(b) \equiv \frac{1}{\#\mathcal{I}_n} \sum_{\mathbf{i}\in\mathcal{I}_n} N_b \left(\bar{Y}_{\mathbf{i},b} - \hat{\mu}_n\right)^2, \qquad \hat{\mu}_n \equiv \frac{1}{\#\mathcal{I}_n} \sum_{\mathbf{i}\in\mathcal{I}_n} \bar{Y}_{\mathbf{i},b},$$

which is simply a scaled sample variance of subsample means. Note that $\hat{\sigma}_n^2(b)$ depends on the subsample scaling b and template D.

As mentioned in §1, Nordman and Lahiri [20] provided an investigation of the large-sample properties of spatial subsampling variance estimators, which partly motivated the development of count approximations here. Specifically, under some mixing conditions on the process, the bias of $\hat{\sigma}_n^2(b)$ in large samples can be expanded as

$$\begin{aligned} \mathrm{E}[\hat{\sigma}_n^2(b)] - \sigma_n^2 &= N_b \mathrm{E}(\bar{Y}_b^2) - \sigma_\infty^2 + o(b^{-1}) \\ &= \frac{\mathcal{L}}{b} + o(b^{-1}), \end{aligned}$$

where we define $\sigma_\infty^2 \equiv \sum_{\mathbf{k}\in\mathbb{Z}^2} c(\mathbf{k})$, $\bar{Y}_b \equiv \sum_{\mathbf{s}\in\mathbb{Z}^2\cap D_b}(Y_s - \mu)/N_b$ and $\mathcal{L} \equiv -\sum_{\mathbf{k}\in\mathbb{Z}^2} \mathcal{L}(\mathbf{k})c(\mathbf{k})/|D|$. While our bias derivation is compact, the second line above follows from the first by the same arguments as in (6) for approximating lattice point counts, where the limits $\mathcal{L}(\mathbf{k})$, $\mathbf{k} \in \mathbb{Z}^2$ in (2) may be computed from area approximations (as in §3). Although the main bias of the subsampling estimator is $O(1/b)$, the proportionality constant $\mathcal{L}$ is heavily influenced by a combination of the underlying covariances $c(\mathbf{k})$ and the geometry of the subsample template D (i.e., impacting $\mathcal{L}(\mathbf{k})$ terms). Using this bias expansion, the mean squared error of the subsampling estimator can also be determined in large samples as

$$\begin{aligned} \mathrm{MSE}_n(b) &\equiv \mathrm{E}(\hat{\sigma}_n^2(b) - \sigma_n^2)^2 = (\mathrm{E}[\hat{\sigma}_n^2(b)] - \sigma_n^2)^2 + \mathrm{Var}(\hat{\sigma}_n^2(b)) \\ &\approx \left(\frac{\mathcal{L}}{b}\right)^2 + V_D \frac{b^2|D|}{|R_n|}, \end{aligned} \tag{9}$$

where $V_D b^2|D|/|R_n|$ represents an expansion of the variance $\mathrm{Var}(\hat{\sigma}_n^2(b))$ of the subsampling estimator and $V_D > 0$ is a constant that depends on a subsample's shape D, not size (e.g., $V_D = 2/5$ for any triangle D or $1 - 16/(3\pi^2)$ for any ellipse D in $\mathbb{R}^2$); see [20] for more details. Minimization of (9) for b gives a form of the (large-sample) optimal subsample scaling (proportional to $|R_n|^{1/2}$), and this theoretical scaling can be used to find a data-driven ('plug-in') estimator of subsample scale b in practice (cf. [23]). In the time series setting, similar consideration has been given to determine bias and mean squared error expansions to derive block sizes for block resampling methods [11,16], so that the above results are spatial analogs. Additionally, most authors [22,26,27] have traditionally considered spatial subregions $D_b = bD$ to have the same shape as the original region $R_n = \lambda_n R_0$ (i.e., a template $D = R_0$). Using (9) with count approximations to determine $\mathcal{L}$, we can also examine how subsampling bias and mean squared error are influenced by different choices of the subsample shape D as well.

4.2 *Variogram estimation with infill sampling structure*

The variogram plays a key role in describing spatial dependence in geostatistical analyses [4]. Fuentes [8] examined a variogram estimator for spatial lattice data collected from a Gaussian process, where the main goal was to determine the variance of this estimator (when evaluated at shrinking spatial lags) under an infill sampling design. Fuentes considered a rectangular sampling region and established an order for the estimator's variance to verify consistency properties. Using counting approximations here, we consider the same essential problem but provide a very explicit expression for the variogram estimator's variance while treating sampling regions which need not be rectangular.

To outline a sampling and estimation framework analogous to [8], suppose that $R_0 \subset \mathbb{R}^2$ is a fixed, convex sampling region (not necessarily rectangular) and that $\{Y_\mathbf{s} : \mathbf{s} \in \mathbb{R}^2\}$ denotes a stationary, Gaussian process with variogram $\gamma(\mathbf{h}) \equiv \mathrm{Var}(Y_\mathbf{s} - Y_{\mathbf{s}+\mathbf{h}}) = \mathrm{E}(Y_\mathbf{s} - Y_{\mathbf{s}+\mathbf{h}})^2$, $\mathbf{h} \in \mathbb{R}^2$; it is assumed that $\gamma : \mathbb{R}^2 \to [0, \infty)$ has continuous and bounded third-order partial derivatives. For a positive real sequence $\{\eta_n\}$ such that $\eta_n \to 0$ as $n \to \infty$, assume that a corresponding sequence of data $\{Y_{\mathbf{s}} : \mathbf{s} \in \eta_n \mathbb{Z}^2 \cap R_0\}$ are available, sampled at positions in R_0 on the scaled integer grid $\eta_n \mathbb{Z}^2$. That is, as $n \to \infty$, the process is observed at an increasing finer scale within the region R_0, which describes a fixed-domain infill sampling structure. Similar to [8], we consider the sample variogram for estimating the process variogram $\gamma(\mathbf{v}_n)$ at a shrinking lag $\mathbf{v}_n = \beta_n \eta_n \mathbf{v}_0$, defined with an arbitrary, but fixed, $\mathbf{v}_0 \in \mathbb{Z}^2$ and positive integer sequence $\{\beta_n\}$ where $\eta_n \beta_n \to 0$. Expressed in terms of $Z_{n,\mathbf{s}} \equiv Y_{\mathbf{s}} - Y_{\mathbf{s}+\mathbf{v}_n}$, $\mathbf{s} \in \mathbb{R}^2$, Matheron's sample variogram [19] is

$$\hat{\gamma}_n(\mathbf{v}_n) = \frac{1}{\#\mathcal{J}_n} \sum_{\mathbf{s} \in \mathcal{J}_n} Z_{n,\mathbf{s}}^2,$$

where $\mathcal{J}_n \equiv \{\mathbf{s} \in \eta_n \mathbb{Z}^2 \cap R_0 : \mathbf{s} + \mathbf{v}_n \in \eta_n \mathbb{Z}^2 \cap R_0\}$ denotes the index set of all available observations $Z_{n,\mathbf{s}}$ from sampling sites $\eta_n \mathbb{Z}^2$ in R_0. This estimator is the unbiased $\mathrm{E}[\hat{\gamma}_n(\mathbf{v}_n)] = \hat{\gamma}_n(\mathbf{v}_n)$ and will be mean-squared error consistent if its variance decays to zero as $n \to \infty$.

To expand the variance of $\hat{\gamma}_n(\mathbf{v}_n)$, let $\mathbf{k}_n = \eta_n \mathbf{k}$ and define counts $N_n(\mathbf{k}) \equiv \#\{\mathbf{s} \in \mathcal{J}_n : \mathbf{s} + \mathbf{k}_n \in \mathcal{J}_n\}$ for $\mathbf{k} \in \mathbb{Z}^2$; note $N_n(\mathbf{0}) = \#\mathcal{J}_n$ for the zero vector $\mathbf{0} \in \mathbb{Z}^2$. Then, by stationarity, we have $\beta_n^{-4}[N_n(\mathbf{0})]^2 \operatorname{Var}(\hat{\gamma}_n(\mathbf{v}_n)) = \beta_n^{-4} \sum_{\mathbf{k} \in \mathbb{Z}^2} N_n(\mathbf{k}) c_n(\mathbf{k})$ with $c_n(\mathbf{k}) \equiv \operatorname{Cov}(Z_{n,\mathbf{0}}^2, Z_{n,\mathbf{k}_n}^2)$, $\mathbf{k} \in \mathbb{Z}^2$. Since $\{Y_{\mathbf{s}} : \mathbf{s} \in \mathbb{R}^2\}$ is Gaussian, it holds that $c_n(\mathbf{k}) = 2[\operatorname{Cov}(Z_{n,\mathbf{0}}, Z_{n,\mathbf{k}_n})]^2$, where $\operatorname{Cov}(Z_{n,\mathbf{0}}, Z_{n,\mathbf{k}_n}) = [\gamma(\mathbf{k}_n + \mathbf{v}_n) - \gamma(\mathbf{k}_n) + \gamma(\mathbf{k}_n - \mathbf{v}_n) - \gamma(\mathbf{k}_n)]/2$, and the smoothness properties of the variogram then imply $|c_n(\mathbf{k}) - 2^{-1}[\mathbf{v}_n' \Psi(\mathbf{k}_n) \mathbf{v}_n]^2| \leq C \|\mathbf{v}_n\|^5$ for a constant $C > 0$ (not depending on n or $\mathbf{k} \in \mathbb{Z}^2$), where $\Psi(\mathbf{s})$ denotes the 2×2 matrix of second-order partial derivatives of $\gamma(\cdot)$ at $\mathbf{s} \in \mathbb{R}^2$. Recalling $\mathbf{v}_n = \beta_n \eta_n \mathbf{v}_0$, we may then write

$$\beta_n^{-4}[N_n(\mathbf{0})]^2 \operatorname{Var}(\hat{\gamma}_n(\mathbf{v}_n)) = 2^{-1} \eta_n^4 \sum_{\mathbf{k} \in \mathbb{Z}^2} N_n(\mathbf{k}) [\mathbf{v}_0' \Psi(\mathbf{k}_n) \mathbf{v}_0]^2 + O(\beta_n \eta_n), \tag{10}$$

with the order of the remainder found from $N_n(\mathbf{k})$, $\#\{\mathbf{k} \in \mathbb{Z}^2 : N_n(\mathbf{k}) > 0\} \leq N_n(\mathbf{0})$ and (as shown below) $N_n(\mathbf{0}) = O(\eta_n^{-2})$.

We now need to handle lattice point counts $N_n(\mathbf{k})$ in (10), which can be re-written as

$$N_n(\mathbf{k}) = \#\mathbb{Z}^2 \cap D_b \cap (D_b + \mathbf{k}), \quad \mathbf{k} \in \mathbb{Z}^2,$$

to fit into the approximation framework of §§1–3, involving an increasing domain $D_b = bD$ with template $D = R_0 \cap (R_0 - \mathbf{v}_n)$ and scaling factors $b \equiv b_n = \eta_n^{-1} \to \infty$ as $n \to \infty$. (This template D is technically not 'fixed' but the count approximations of Theorem 3.1 may still be verified; intuitively, since $\mathbf{v}_n \to 0$, the template D is asymptotically equivalent to R_0.) When $N_n(\mathbf{k}) > 0$ in (10), write $N_n(\mathbf{k}) = N_n(\mathbf{0}) - [N_n(\mathbf{0}) - N_n(\mathbf{k})]$ and approximate/replace the count difference $N_n(\mathbf{0}) - N_n(\mathbf{k})$ with corresponding areas $A_n(\mathbf{k}) \equiv |D_b| - |D_b \cap (D_b + \mathbf{k})|$, so that (10) becomes

$$\begin{aligned} &\beta_n^{-4}[N_n(\mathbf{0})]^2 \operatorname{Var}(\hat{\gamma}_n(\mathbf{v}_n)) \\ &= 2^{-1} \eta_n^2 \sum_{\mathbf{k} \in \mathbb{Z}^2, N_n(\mathbf{k}) > 0} \eta_n^2 [N_n(\mathbf{0}) - A_n(\mathbf{k})][\mathbf{v}_0' \Psi(\mathbf{k}_n) \mathbf{v}_0]^2 + E_n + O(\beta_n \eta_n) \qquad (11) \\ &= 2^{-1} \int_{\mathbb{R}^2} f_n(\mathbf{s}) d\mathbf{s} + O(\eta_n) + O(\beta_n \eta_n); \qquad (12) \end{aligned}$$

in (11), E_n denotes the error introduced from applying area approximations in (10), which Theorem 3.1 bounds as

$$|E_n| \leq C\eta_n^2 \sum_{\mathbf{k}\in\mathbb{Z}^2, N_n(\mathbf{k})>0} \eta_n^2 \|\mathbf{k}\| [\mathbf{v}_0' \Psi(\mathbf{k}_n)\mathbf{v}_0]^2 = O(\eta_n)$$

using that $\#\{\mathbf{k} \in \mathbb{Z}^2 : N_n(\mathbf{k}) > 0\} = O(\eta_n^{-2})$, that $\mathbf{v}_0'\Psi(\cdot)\mathbf{v}_0$ is bounded, and that $\max\{\eta_n\|\mathbf{k}\| : \mathbf{k} \in \mathbb{Z}^2, N_n(\mathbf{k}) > 0\} \leq \sup\{\|\mathbf{x}-\mathbf{y}\| : \mathbf{x},\mathbf{y} \in R_0\} < \infty$. Equation (12) then re-writes the sum in (11) as an integral (with respect to the $\mathbb{R}^2$-Lebesgue measure) of a step-function

$$f_n(\mathbf{s}) \equiv \begin{cases} \eta_n^2(N_n(\mathbf{0}) - A_n(\lfloor \eta_n^{-1}\mathbf{s}\rfloor))[\mathbf{v}_0'\Psi(\eta_n\lfloor\eta_n^{-1}\mathbf{s}\rfloor)\mathbf{v}_0]^2, & \text{if } N_n(\lfloor\eta_n^{-1}\mathbf{s}\rfloor) > 0,\\ 0, & \text{otherwise,}\end{cases}$$

$\mathbf{s} = (s_1, s_2)' \in \mathbb{R}^2$ with $\lfloor\eta_n^{-1}\mathbf{s}\rfloor \equiv (\lfloor\eta_n^{-1}s_1\rfloor, \lfloor\eta_n^{-1}s_2\rfloor)'$ using the floor function $\lfloor\cdot\rfloor$. Because $|R_0|/|D| \to 1$ and $N_n(\mathbf{0})/|D_b| = N_n(\mathbf{0})/[\eta_n^{-2}|D|] \to 1$ as $n \to \infty$ (by convexity as in §2) and $\eta_n^2 A_n(\lfloor\eta_n^{-1}\mathbf{s}\rfloor) = |D| - |D \cap (D + \eta_n\lfloor\eta_n^{-1}\mathbf{s}\rfloor)|$, we have that the step-function converges pointwise (almost everywhere) as $\lim_{n\to\infty} f_n(\mathbf{s}) = |R_0 \cap (R_0 + \mathbf{s})|[\mathbf{v}_0'\Psi(\mathbf{s})\mathbf{v}_0]^2$, $\mathbf{s} \in \mathbb{R}^2$, and so the DCT applied to (12) gives

$$\lim_{n\to\infty} \beta_n^{-4}[N_n(\mathbf{0})]^2 \operatorname{Var}(\hat{\gamma}_n(\mathbf{v}_n)) = 2^{-1}\int_{\mathbb{R}^2} |R_0 \cap (R_0+\mathbf{s})|[\mathbf{v}_0'\Psi(\mathbf{s})\mathbf{v}_0]^2 d\mathbf{s}.$$

Beyond an order $\beta_n^4[N_n(\mathbf{0})]^{-2} = O((\beta_n\eta_n)^4) = O(\|\mathbf{v}_n\|^4)$ for $\operatorname{Var}(\hat{\gamma}_n(\mathbf{v}_n))$, driven by the lag $\mathbf{v}_n$, this result shows how the geometry of the sampling region $R_0 \subset \mathbb{R}^2$ asymptotically influences the sample variogram's variance. Note that, under the infill sampling design, the derivation above did not require limits of 'scaled' differenced lattice point counts as in (2) (i.e., scaled by b^{-1}), but area-for-count approximations facilitated the arguments.

5. Concluding remarks and extensions

We have motivated how bias expansions for certain statistics, related to covariance estimators with spatial lattice data in $\mathbb{R}^2$, require approximating a difference between two lattice point counts. This counting problem differs from the usual 'O-problem' considered in number theory. For all convex regions in $\mathbb{R}^2$, approximation results here show that differenced areas can adequately approximate differenced lattice point counts, in the sense of (3), for purposes of explicitly computing limiting bias and other expectation expressions.

While spatial data collected in the plane are most common, sampling in higher spatial dimensions $\mathbb{R}^d$, $d \geq 3$, leads to analogous problems with bias expansions. In this case, we consider a spatial region $D_b = bD$ defined by inflating a given template $D \subset \mathbb{R}^d$. Sample covariances of §2, for example, can be generalized to any sampling dimension $d \geq 1$ as $\hat{c}(\mathbf{k}) = N_b^{-1}\sum_{\mathbf{s}\in\mathbb{Z}^d\cap D_b\cap(D_b-\mathbf{k})} Y_\mathbf{s}Y_{\mathbf{s}+\mathbf{k}}$ with $N_b = \#\mathbb{Z}^d \cap D_b$ and $\mathbf{k} = (k_1,\ldots,k_d)'$ and the bias expansion in (4), (5) also generalizes naturally as

$$\begin{aligned}\mathrm{E}[\hat{c}(\mathbf{k})] - c(\mathbf{k}) &= -\frac{c(\mathbf{k})}{N_b}[\#\mathbb{Z}^d \cap D_b - \#\mathbb{Z}^d \cap D_b \cap (D_b + \mathbf{k})]\\ &= -\frac{c(\mathbf{k})\mathcal{L}(\mathbf{k})}{|D|}\frac{1}{b} + o(b^{-1}),\end{aligned}$$

assuming $N_b/|D_b| = N_b/[b^d|D|] \to 1$ as $b \to \infty$. We define the extension of (2) to be

$$\mathcal{L}(\mathbf{k}) = \lim_{b\to\infty} \frac{1}{b^{d-1}}\{\#\mathbb{Z}^d \cap D_b - \#\mathbb{Z}^d \cap D_b \cap (D_b + \mathbf{k})\}, \quad \mathbf{k} \in \mathbb{Z}^d,$$

where $O(b^{d-1})$ is the typical order of D_b's surface area in $\mathbb{R}^d$. The problem in $\mathbb{R}^d$ identifies conditions under which the counts $\#\mathbb{Z}^d \cap D_b - \#\mathbb{Z}^d \cap D_b \cap (D_b + \mathbf{k})$ can be approximated by the areas $|D_b| - |D_b \cap (D_b + \mathbf{k})|$ with an order $o(b^{d-1})$ error (as opposed to $o(b)$ in (3) when $d = 2$). More work is required to characterize the sampling regions, whether convex or not, for which this approximation approach is valid.

6. Proofs

Let PQ and $\overline{PQ}$ denote open and closed line segments between points P and $Q \in \mathbb{R}^2$; $\mathrm{Ln}(PQ)$ denotes the infinite line through P, Q; and $\mathrm{Ry}(PQ)$ denotes the open ray from P (not included) through Q. Fix $\mathbf{k} \in \mathbb{Z}^2$ in the following. If $\mathbf{k}$ has gradient $k_1/k_2 = a/q$ (a, q are integers, $q \geq 1$) with $gcd(a, q) = 1$, then $\mathbf{k} = \ell(a, q)'$ and $\|\mathbf{k}\| = \ell\sqrt{a^2 + q^2}$ for some positive integer ℓ. There also exist integers i_a, i_q, where $ai_a - qi_q = 1$, whereby the $\mathbb{Z}^2$-integers may be written as

$$\mathbb{Z}^2 = \{m(i_a, i_q)' + \ell^{-1}n\mathbf{k} : n, m \in \mathbb{Z}\}. \qquad (13)$$

Any line in $\mathbb{R}^2$ parallel to $\mathbf{k}$ may be expressed as $l_c \equiv \{\mathbf{x} \in \mathbb{R}^2 : \langle \mathbf{x}, (a, -q)' \rangle = c\}$ for some $c \in \mathbb{R}$, using the vector inner product $\langle \cdot, \cdot \rangle$, and the perpendicular distance between two lines l_{c_1} and l_{c_2} is $|c_1 - c_2|/\sqrt{a^2 + q^2}$ for $c_1, c_2 \in \mathbb{R}$ (p. 99 of [14]). In the following, define the (signed) discrepancy between the area and lattice point count of a subset $S \subset \mathbb{R}^2$ as $\Delta(S) = \#\mathbb{Z}^2 \cap S - |S|$.

Proof of Theorem 3.1. For a fixed $\mathbf{k} \in \mathbb{Z}^2$, the points of the region D_b, not in $D_b + \mathbf{k}$, form the set $D_b \backslash (D_b + \mathbf{k})$. The signed error in $\Delta(D_b, \mathbf{k})$ in Theorem 3.1 is equivalent to $\Delta(D_b \backslash (D_b + \mathbf{k}))$, which we seek to bound. For some $\epsilon \in (0, 1)$ and $\mathbf{x} \in D$, a closed ball of radius ϵ around $\mathbf{x}$ (including $\mathbf{x} \pm \epsilon\mathbf{k}/\|\mathbf{k}\|$) is interior to D. Assume that $\alpha < 1$ holds for $\alpha \equiv \|\mathbf{k}\|/(2\epsilon b)$, implying that $D_b \cap (D_b + \mathbf{k})$ is nonempty. The case $\alpha \geq 1$ will be treated lastly.

Consider intersections of the closed, convex set $D_b = bD$ by parallel lines l_c to $\mathbf{k}$. Over some interval $c \in [c_1, c_2]$, the intersections of lines l_c with D_b are non-empty and each intersection yields a closed line segment of length, say, $\mathrm{len}(c)$. By the convexity of D_b, $\mathrm{len} : [c_1, c_2] \to [0, \infty)$ will be a continuous, concave function with a maximum $M_b > 0$. Also, there will be a subinterval $[m_1, m_2] \subseteq [c_1, c_2]$, for which $\mathrm{len}(c) = M_b$, $c \in [m_1, m_2]$ holds and, by convexity, $\mathrm{len}(c)$ must be strictly increasing on $[c_1, m_1]$ if $c_1 < m_1$ and strictly decreasing on $[m_2, c_2]$ if $m_2 < c_2$ (i.e, D_b would not be convex if $\mathrm{len}(\cdot)$ had two local maxima). There are two possible cases:

Case 1. $\mathrm{len}(c_1) = \mathrm{len}(c_2) = 0$ so that the first and the last $\mathbf{k}$-parallel lines intersect D_b at single points as in figure 3(a).

Case 2. $\mathrm{len}(c_1) + \mathrm{len}(c_2) > 0$ (i.e., either the first or the last line intersection with D_b produces a closed line segment of non-zero length), as in figure 3(b).

We consider Case 1 first. Because $M_b > \|\mathbf{k}\|$ holds when $\alpha < 1$, there are two unique intersections of D_b having length $\|\mathbf{k}\|$, say $\overline{PQ}$ and $\overline{P'Q'}$, where $P = Q + \mathbf{k}$, $Q' = P' + \mathbf{k}$.

In addition, $\partial D_b \cap (\partial D_b + \mathbf{k})$ has exactly two elements by the convexity of D_b, corresponding to the endpoints Q, Q'. The boundary B of the set $D_b \backslash (D_b + \mathbf{k})$ consists of four parts: B_1, the boundary of D_b from Q to P; B_2, the boundary of D_b from P to P'; B_3, the boundary of D_b from P' to Q'; and B_4, the boundary of $D_b + \mathbf{k}$ from Q' to Q, which is congruent to B_2 but described in the opposite direction; see figure 3(a).

We divide the set $D_b \backslash (D_b + \mathbf{k})$ correspondingly into three regions: U_1, U_2 and U_3. U_1 is bounded by the line PQ and the part B_1 of B. U_1 includes all points on PQ and also the point P. U_2 is bounded by the lines PQ and $P'Q'$ and the parts B_2 and B_4 of B. All points of the line PQ are excluded from U_2, while U_2 includes all points on the line $P'Q'$ as well as the point P'. U_3 is bounded by the line $P'Q'$ and the part B_3 of B. All points of $P'Q'$ are excluded from U_3. Points Q, Q' do not belong to $D_b \backslash (D_b + \mathbf{k}) = \cup_{i=1}^{3} U_i$. For each of these three parts we wish to find the (signed) discrepancy:

$$\Delta(D_b \backslash (D_b + \mathbf{k})) = \Delta(U_1) + \Delta(U_2) + \Delta(U_3).$$

The discrepancy $\Delta(U_2)$ of U_2 is the same as the discrepancy of the parallelogram U_4, $PP'Q'Q$, with the convention that P' and all points of PP' and $P'Q'$ are included in U_4, while all points of $Q'Q$ and PQ as well as points P, Q', Q are excluded from U_4. Note that if the perpendicular distance between PQ and $P'Q'$ is $w/\sqrt{a^2+q^2}$ for some integer w, then the parallelogram U_4 $(PP'Q'Q)$ has area ℓw and, by (13), it contains exactly ℓw integer points, in which case $\Delta(U_2) = \Delta(U_4) = 0$. If w is not an integer, then we expand U_4 to a larger parallelogram U_5 of width $\lceil w \rceil/\sqrt{a^2+q^2}$, using the ceiling function $\lceil \cdot \rceil$. Then, the change in area from U_4 to U_5 is at most $\|\mathbf{k}\|/\sqrt{a^2+q^2} = \ell$ and if there are extra lattice points in U_5, then they lie on a single line parallel to $\mathbf{k}$ and there are ℓ of them by (13). Since $\Delta(U_5) = 0$, we have $|\Delta(U_2)| = |\Delta(U_4)| \le 2\ell + |\Delta(U_5)| = 2\ell$.

For $\Delta(U_1)$, a counting squares argument (Lemma 2.1.1 of [13]) shows that its contribution can be estimated trivially by the U_1-perimeter length: $|\Delta(U_1)| \le 4[\|\mathbf{k}\| + \text{length } B_1] + 8$. To bound the length of B_1, let A represent the point formed by the intersection of the first l_{c_1} or last l_{c_2} line with the boundary B_1 of U_1; see figure 3a. By convexity, the length of B_1 must be less than the length $\|\mathbf{k}\|$ of PQ added to twice the perpendicular distance from A to PQ. Since $\alpha = \|\mathbf{k}\|/(2\epsilon b) < 1$ holds, the points $Q_1 = A(1-\alpha) + \alpha(b\mathbf{x} + b\epsilon\mathbf{k}/\|\mathbf{k}\|)$ and $P_1 = A(1-\alpha) + \alpha(b\mathbf{x} - b\epsilon\mathbf{k}/\|\mathbf{k}\|)$ must be interior to D_b as well as the $\mathbf{k}$-parallel segment $\overline{P_1Q_1}$ by convexity. The $\mathbf{k}$-parallel line passing through P_1 and Q_1 has an intersection with D_b of length exceeding

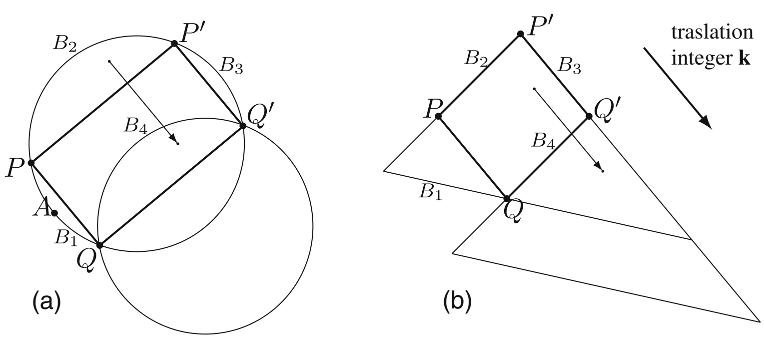

Figure 3. Two examples of planar regions $D_b = bD$ and their resulting translations by an integer vector $\mathbf{k} \in \mathbb{Z}^2$.

length $\overline{P_1Q_1} = \|\mathbf{k}\|$; hence, this line cannot pass between A and PQ since lengths of line intersections with D_b increase from 0 to $\|\mathbf{k}\|$ between A and PQ. That is, the perpendicular distance from A to PQ is less than the perpendicular distance $|d|$ from A to P_1Q_1, where $d = \alpha\langle A - b\mathbf{x}, (a, -q)'\rangle/\sqrt{a^2+q^2} = \ell\langle b^{-1}A - \mathbf{x}, (a, -q)'\rangle/(2\epsilon)$ for bounded points $b^{-1}A, \mathbf{x} \in D$. Finally, $|\Delta(U_1)| \le 8[\|\mathbf{k}\| + |d|] + 8 \le C\|\mathbf{k}\|$ for some $C > 0$ (not depending on $\mathbf{k} \in \mathbb{Z}^2$ or b); a similar argument holds for U_3. This establishes Case 1.

In Case 2, the first or the last intersection of bD by a $\mathbf{k}$-parallel line (l_{c_1} or l_{c_2}) is a closed line segment with length, say, $\text{len}(c_1) > 0$. In this case, the same essential arguments in Case 1 apply (e.g., handling U_4) except that, when $\text{len}(c_1) > \|\mathbf{k}\|$, the boundaries B_1 or B_3 from Case 1 may collapse to closed segments $\overline{PQ}$ or $\overline{P'Q'}$ of length $\|\mathbf{k}\|$; see figure 3b. If this happens, at most ℓ integers lie on $\overline{PQ}$ (or $\overline{P'Q'}$) by (13) and the same bound on the discrepancy $\Delta(D_b\backslash(D_b+\mathbf{k}))$ from Case 1 applies.

Finally, if $\alpha \equiv \|\mathbf{k}\|/(2\epsilon b) \ge 1$, a counting squares argument (Lemma 2.1.1 of [13]) shows that $\Delta(D_b\backslash(D_b+\mathbf{k}))$ is bounded by 4[perimeter length of D_b] + 8. Since D is convex, the perimeter length of D_b is bounded by $C_1 b$ for some $C_1 > 1$. It then follows that $\Delta(D_b\backslash(D_b+\mathbf{k})) \le C\|\mathbf{k}\|$ for $C = 10\epsilon^{-1}C_1$ independent of b and $\mathbf{k} \in \mathbb{Z}^2$. □

Proof of Theorem 3.2. For each $j = 1, 2, 3$, let $\Omega_j(b, \mathbf{k}) = bS_j\backslash(bS_j+\mathbf{k})$ and define sets $\tilde{\Omega}_j(b, \mathbf{k}) = b(S_{3-j}\backslash S_j) \cap (b(S_j\backslash S_{3-j}) + \mathbf{k})$ for $j = 1, 2$. With algebra,

$$f(D_b\backslash(D_b+\mathbf{k})) = \sum_{j=1}^{2}\left[f(\Omega_j(b,\mathbf{k})) - f(\tilde{\Omega}_j(b,\mathbf{k}))\right] - f(\Omega_3(b,\mathbf{k}))$$

holds where the set operation $f(\cdot)$ may denote either area $|\cdot|$, lattice point count $\#\mathbb{Z}^2\cap(\cdot)$, or discrepancy $\Delta(\cdot)$. Because $|\Delta(\Omega_j(b,\mathbf{k}))| \le C\|\mathbf{k}\|$ holds by Theorem 3.1 for $j = 1, 2, 3$, it suffices to show that $|f(\tilde{\Omega}_j(b,\mathbf{k}))| \le C\|\mathbf{k}\|^2$ for $j = 1, 2$ for some $C > 0$ (not depending on $\mathbf{k}$ or b), which bounds the area (or count) of those points (or integer points) in $b(S_1\backslash S_2)$ which may move into $b(S_2\backslash S_1)$ upon translation by $\mathbf{k}$. We focus on $\tilde{\Omega}_1(b,\mathbf{k})$ as similar arguments apply to $\tilde{\Omega}_2(b,\mathbf{k})$.

Since $S_3 = S_1 \cap S_2$ has a non-empty interior, the intersection of the boundary curves is $\partial bS_1 \cap \partial bS_2 = \{P_1, \ldots, P_m\}$ for some $m \ge 2$ (i.e., m cannot be 1). Assume first that $m = 2$. For some $\mathbf{x} \in S_3$, suppose a closed ball radius $\epsilon \in (0, 1)$ lies inside S_3 and that $\|\mathbf{k}\|/(2\epsilon b) < 1$, implying $bS_3 \cap (bS_3+\mathbf{k})$ is nonempty. (As in the proof of Theorem 3.1, we can trivially bound $\Delta(D_b\backslash(D_b+\mathbf{k}))$ with $C\|\mathbf{k}\|$ by counting squares argument when $\|\mathbf{k}\|/(2\epsilon b) \ge 1$.) The intersection of the line $\text{Ln}(P_1P_2)$ with D_b is $\overline{P_1P_2}$ where the open segment P_1P_2 is interior to bS_3 since the ray from any interior point of convex body intersects the boundary exactly once [14]. Hence, $b(S_1\backslash S_2)$ and $b(S_2\backslash S_1)$ must lie on opposite open half-spaces defined by $\text{Ln}(P_1P_2)$; see figure 4.

If $\text{Ln}(P_1P_2)$ is parallel to $\mathbf{k}$, then $f(\tilde{\Omega}_1(b,\mathbf{k})) = 0$ as the set $bS_1\backslash bS_2$, upon translation by $\mathbf{k}$, slides parallel to $bS_2\backslash bS_1$ on the complementary half-spaces of $\text{Ln}(P_1P_2)$. If $\text{Ln}(P_1P_2)$ is not parallel to $\mathbf{k}$, we take the points $\mathbf{x} \pm \epsilon\mathbf{k}/\|\mathbf{k}\|$ interior to S_3 by construction and scale these by b to obtain $Q, Q' \in bS_3$, where the segment $\overline{QQ'}$ is interior to bS_3 and parallel to $\mathbf{k}$. By taking the ray from any point on $\overline{QQ'}$ to P_1, a V-shaped cone is formed with vertex P_1, where all points of the cone (except P_1) belong to the complement D_b^c of D_b and all points of the closed triangle $QQ'P_1$ (except P_1) belong to the interior of bS_3 as in figure 4; analogous constructions with rays hold with respect to P_2. There exist two points $A_1', B_1 \in \text{Ry}(Q'P_1)$ such that $A_1 = A_1' - \mathbf{k}$, $B_1' = B_1 + \mathbf{k} \in \text{Ry}(QP_1)$

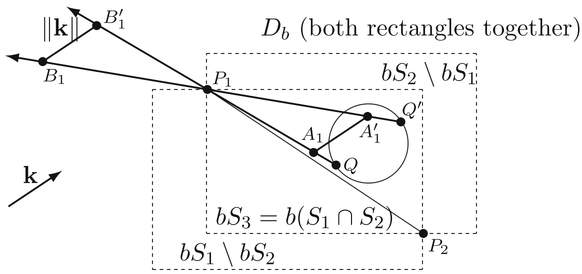

Figure 4. An example of a non-convex (untranslated) region $D_b = bD = bS_1 \cup bS_2$ formed by a union of two rectangles with a rectangular intersection bS_3. An integer vector $\mathbf{k} \in \mathbb{Z}^2$ for translation of D_b is also indicated.

and $\overline{A_1A_1'}$ belongs to the triangle $QQ'P_1$ while $\overline{B_1B_1'}$ belongs to the V-shaped cone in D_b^c with vertex P_1; see figure 4. Analogous points A_2, A_2' and B_2, B_2' exist using rays through P_2. For a point of $bS_1 \backslash bS_2$ to move into $bS_2 \backslash bS_1$ upon translation by $\mathbf{k}$, the point must slide through both rays $\mathrm{Ry}(QP_i)$ and $\mathrm{Ry}(Q'P_i)$ along a $\mathbf{k}$-parallel line lying between $\mathrm{Ln}(A_iA_i')$ and $\mathrm{Ln}(B_iB_i')$ for some $i = 1, 2$; no $\mathbf{k}$-translated point of $bS_1 \backslash bS_2$ may move to $bS_2 \backslash bS_1$ through the strip of bS_3 between $\mathrm{Ln}(A_1A_1')$ and $\mathrm{Ln}(A_2A_2')$ as the closed parallelogram $A_1A_1'A_2'A_2$ is interior to bS_3 with sides A_iA_i', $i = 1, 2$ of length $\|\mathbf{k}\|$, as $A_i' = A_i + \mathbf{k}$. The perpendicular distance between A_iA_i' and B_iB_i' is bounded by $C\|\mathbf{k}\|$, $i = 1, 2$ (cf. distance $|d|$ from the proof of Theorem 3.1) and multiplying this by $\|\mathbf{k}\|$ gives $|f(\tilde{\Omega}_1(b, \mathbf{k}))| \le C\|\mathbf{k}\|^2$, using (13) for counts.

If $m > 2$, then we may assume that the points $P_1, P_2, \ldots, P_m$ fall consecutively in a direction along the curve ∂bS_3 so that the closure/interior of the polygon $P_1P_2 \cdots P_m$ belongs to the closure/interior of bS_3. If $\mathbf{x} \in \tilde{\Omega}_1(b, \mathbf{k})$, then $\mathbf{x}$ and $\mathbf{x} - \mathbf{k}$ must lie in complementary half-spaces defined by a line between some consecutive vertices P_i, P_{i+1}. Similar arguments may be applied to show $|f(\tilde{\Omega}_1(b, \mathbf{k}))| \le C\|\mathbf{k}\|^2$ as above.

Acknowledgements

The authors are grateful to the reviewer for providing helpful and insightful comments which significantly improved the manuscript. This research was partially supported by NSF grants DMS-0707139 and DMS-0906588.

References

[1] Barvinok A I, Polynomial time algorithm for counting integral points in polyhedra when the dimension is fixed, *Math. Oper. Res.* **19** (1994) 769–779

[2] Barvinok A I and Pommersheim J, An algorithmic theory of lattice points in polyhedra, in: New Perspectives in Algebraic Combinatorics (Berkeley, CA, 1996–1997) (Math. Sci. Res. Inst. Publ., vol. 38. Cambridge Univ. Press, Cambridge) (1999) pp. 91–147

[3] Cliff A D, and Ord J K, Spatial processes, models and application (London: Pion Limited) (1981)

[4] Cressie N, Statistics for spatial data, 2nd edition (New York: Wiley) (1993)

[5] De Loerab J A, Hemmeckeb R, Tauzera J and Yoshidab R, Effective lattice point counting in rational convex polytopes, *J. Symbolic Computat.* **38** (2004) 1273–1302

[6] Dyer M, and Kannan R, On Barvinoks algorithm for counting lattice points in fixed dimension, *Math. Oper. Res.* **22** (1997) 545–549

[7] Ehrhart E, Polynômes arithmétiques et methode des polyédres en combinatoire, International Series of Numerical Mathematics, 35 (Basel: Birkhäuser) (1977)

[8] Fuentes M, Fixed-domain asymptotics for variograms using subsampling, *Math. Geol.* **33** (2001) 679–691

[9] Fuller W, Introduction to statistical time series, 2nd edition (New York: Wiley) (1996)

[10] Guyon X, Random fields on a network (Springer, New York) (1995)

[11] Hall P, Horowitz J L and Jing B-Y, On blocking rules for the bootstrap with dependent data, *Biometrika* **82** (1995) 561–574

[12] Huxley M N, Exponential sums and lattice points II, *Proc. London Math. Soc.* **66** (1993) 279–301

[13] Huxley M N, Area, lattice points, and exponential sums (New York: Oxford University Press) (1996)

[14] Kelly P J, and Weiss M L, Geometry and Convexity (New York: Wiley) (1979)

[15] Krätzel E, Lattice points (Berlin: Deutscher Verlag Wiss.) (1988)

[16] Künsch H R, The jackknife and the bootstrap for general stationary observations, *Ann. Statist.* **17** (1989) 1217–1261

[17] Lenstra H W, Integer programming with a fixed number of variable, *Math. Oper. Res.* **8** (1983) 538–548

[18] McAllister T B and Woods K M, The minimum period of the Ehrhart quasi-polynomial of a rational polytope, *J. Combin. Theory* **A109** (2005) 345–352

[19] Matheron G, The theory of regionalized random variables and its applications, Les Cahiers du Centre de Morphologie Mathematique, Fasc. 5, Centre de Geostatistique, Fontainebleau (1971)

[20] Nordman D J and Lahiri S N, On optimal spatial subsample size for variance estimation, *Ann. Statist.* **32** (2004) 1981–2027

[21] Pick G, Geometrishes zur Zahlentheorie, *Sitzenber. Lotos* **19** (1899) 311–319

[22] Politis D N and Romano J P, Large sample confidence regions based on subsamples under minimal assumptions, *Ann. Statist.* **22** (1994) 2031–2050

[23] Politis D N and White H, Automatic block-length selection for the dependent bootstrap, *Economet. Revi.* **23** (2004) 53–70

[24] Politis D N, Romano J P and Wolf M, Subsampling (New York: Springer) (1999)

[25] Possolo A, Subsampling a random field spatial statistics and imaging (ed.) A Possolo, IMS Lecture Notes Monograph Series 20 (CA: Institute of Mathematical Statistics, Hayward) (1991) pp. 286–294

[26] Sherman M, Variance estimation for statistics computed from spatial lattice data, *J. R. Stat. Soc.* **B58** (1996) 509–523

[27] Sherman M and Carlstein E, Nonparametric estimation of the moments of a general statistic computed from spatial data, *J. Amer. Statist. Assoc.* **89** (1994) 496–500

[28] Stanley R P, Enumerative combinatorics, vol I (California: Wadsworth, Belmont) (1986)

[29] van der Corput J C, Über Gitterpunkte in der Ebene, *Math. Ann.* **81** (1920) 1–20