



An empirical study of tests for uniformity in multidimensional data

Adam Petrie^{a,*}, Thomas R. Willemain^b

^a Statistics, Operations, Management Science, University of Tennessee, Knoxville, TN 37996, United States

^b Industrial and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY 12180, United States

HIGHLIGHTS

- A test of uniformity determines if data is uniform or possesses underlying structure.
- Tests of uniformity should be the first step in any pattern or exploratory data analysis.
- We develop new tests of uniformity and compare with many in the literature.
- One test outperforms others when regularity (minimum spacings) exists in the data.
- Our other test is the only option for testing on arbitrary supports and dimensions.

ARTICLE INFO

Article history:

Received 25 April 2012

Received in revised form 17 October 2012

Accepted 8 February 2013

Available online 27 February 2013

Keywords:

Minimum spanning tree

Hamiltonian path

Discrepancies

Spacings

K-function

ABSTRACT

An important problem in high-dimensional data analysis is determining whether sample points are uniformly distributed (i.e., exhibit complete spatial randomness) over some compact support, or rather possess some underlying structure (e.g., clusters or other non-homogeneities). We propose two new graph-theoretic tests of uniformity which utilize the minimum spanning tree and a snake (a short non-branching acyclic path connecting each data point). We compare the powers of statistics based on these graphs with other statistics from the literature on an array of non-uniform alternatives in a variety of supports. For data in a hypercube, we find that test statistics based on the minimum spanning tree have superior power when the data displays regularity (e.g., results from an inhibition process). For arbitrarily shaped or unknown supports, we use run length statistics of the sequence of segment lengths along the snake's path to test uniformity. The snake is particularly useful because no knowledge or estimation of the support is required to compute the test statistic, it can be computed quickly for any dimension, and it shows what kinds of non-uniformities are present. These properties make the snake unique among multivariate tests of uniformity since others only function on specific and known supports, have computational difficulties in high dimension, or have inconsistent type I error rates.

© 2013 Elsevier B.V. All rights reserved.

1. Introduction

We consider a dataset to be 'uniformly distributed' over a d -dimensional compact support S if it is a random sample from a uniform distribution on S . In other words, it is a realization of a spatially homogeneous Poisson process on S . No underlying structures such as clusters, density variations, or patterns (e.g., minimum spacing between points) are present.

Testing for the presence of non-uniformities within a given region or over the entirety of a multidimensional dataset should perhaps be the first step in any type of exploratory multivariate data analysis. For example, clustering algorithms

* Corresponding author. Tel.: +1 865 974 1631.

E-mail addresses: apetrie@tennessee.edu, apetrie@utk.edu (A. Petrie), willet@rpi.edu (T.R. Willemain).

assume the existence of regions of locally high density, so if none of these regions exist then cluster analysis itself can be skipped and the problem of characterizing the data reduces to merely finding the support, or shape, of the datacloud, which may be approximated by finding its convex hull (Barber et al., 1996) or other methods (Cuevas and Rodriguez-Casal, 2004; Berrendero et al., 2012a).

Classical univariate tests for uniformity such as the Kolmogorov–Smirnov or Cramer–von Mises, which check the empirical cumulative distribution function (CDF) of the data for consistency with a uniform distribution, do not extend readily to higher dimensions due to computational difficulties in finding the null distribution of the test statistics; see Zimmerman (1993) and Justel et al. (1997). Rather, tests for uniformity in two or three dimensions have been thoroughly explored in the spatial statistics literature and are considered tests of ‘complete spatial randomness’ (Ripley, 2004) or CSR. Applications abound, for instance, in ecological data where one might consider whether positions of trees, cells, or animal habitats are random. See Baddeley et al. (2006) for many examples or Wiegand and Moloney (2004) and references therein. Tests in three dimensions are less common, but have appeared regarding the positions of lacunae (pits) in skulls (Baddeley et al., 1993), locations of galaxies in the universe (Barrow et al., 1985), and positions of neurons (Jafari-Mamaghani et al., 2010). Standard CSR tests for two or three dimensions may be based on quadrat counts (number of points falling into particular regions), the distribution of nearest neighbors, and ‘second order’ characteristics such as the nearest neighbor function or Ripley’s K -function. See Ripley (2004), Diggle (2003), or Illian et al. (2008) and references therein for a good review of commonly used tests and Ho and Chiu (2007) for a recent summary and comparison.

Tests for uniformity in higher-dimensional data are much rarer. In addition to being the first step in any clustering or segmentation analysis, multivariate uniformity tests are useful for checking the validity of random number generators, e.g. numbers from the infamous RANDU generator form a series of 15 stacked two-dimensional planes when points consisting of three consecutive pseudorandom values are plotted in three dimensions (Marsaglia, 1968). A similar utilization could lie in checking to see if the distribution of high-dimensional cryptographic keys are free from structures which hackers could exploit. A further application is in projection pursuit (Friedman and Tukey, 1974), where one removes all structure found in ‘interesting’ projections of the data until no remaining structure is detectable. Generally, non-interestingness is associated with multivariate normality (Friedman, 1987), but uniformity could be viewed as non-interesting as well. A test could also confirm assumptions of local (or global) uniformity used by some statistical methods, e.g., a maximum likelihood estimator of intrinsic dimension (Levina and Bickel, 2005) or spatial query estimation (Theodoridis et al., 1998).

Standard tests for CSR are easy conceptually to extend to higher dimensions, but their fundamental assumptions are not met by the samples we wish to consider. The standard model in spatial statistics assumes that the point process extends throughout all space (i.e., the support S is \mathbb{R}^d) but is observed only inside a pre-defined region called the sampling window W . For an arbitrary dataset (e.g., fluxes from many different wavelengths of a catalog of astronomical objects), observations (either real or potential) will not fill all of \mathbb{R}^d . Indeed, we desire a test for uniformity that works when the support S is finite but not pre-defined since physical limits on variables are rarely known in advance.

Even if it were appropriate to extend standard tests for CSR to higher dimensions, the computation of their test statistics becomes extremely difficult due to “edge effects”. For example, the nearest neighbor function estimates the probability that the distance between nearest neighbors is less than some specified value r . If a point is closer to the boundary of the sampling window than r , the estimate must be corrected since that point’s nearest neighbor could be outside the sampling window and thus unobserved, but still closer than the distance r . Correction factors have only been derived for two or three dimensions. An alternative to deriving correction factors is to measure distances in terms of the toroidal metric, but this essentially forces the shape of the sampling window to be some hyperrectangle.

Other tests for uniformity in $d \geq 2$ dimensions include Berrendero et al. (2012b, 2006) who use the distribution of distances from points to the boundary of support, Berrendero et al. (2012a) who use multivariate spacings, Tenreiro (2007) who uses the L_2 distance between the kernel density estimate of the underlying probability density and the uniform distribution, Ho and Chiu (2007) and Liang et al. (2001) who use discrepancies (quantities related to the difference between the exact value of an integral and its Monte Carlo approximation using the point set), Jain et al. (2002) who perform a two-sample test on a subsample of points in a high-density region and a subsample in a low-density region of the dataset, Smith and Jain (1984) who perform a two-sample test on a uniform sample generated on the support and the point set, and Hoffman and Jain (1983) who compare the distribution of edge lengths in the dataset’s minimum spanning tree with the corresponding distribution for uniform data in the hypercube. However, none of these tests are completely satisfactory.

Ideally, a general test for uniformity should work on an arbitrary compact support S and the null distribution of the test statistic should be easy to calculate or simulate. Unfortunately, nearly all of these higher-dimensional tests require a specific support (typically a hyperrectangle) and require *a priori* knowledge of the boundary of the support. Only Berrendero et al. (2012a,b), Jain et al. (2002), and Smith and Jain (1984) provide methods which function without foreknowledge of the support (the boundary of the support is estimated), but for $d > 2$ the former two are extremely computationally difficult while the latter two we will show have undesirable statistical properties which render them unusable.

In this work, we introduce two new tests of multivariate uniformity designed to overcome some of the shortcomings of standard methods in spatial statistics and other modern techniques. The first is useful for data with a known support and uses the length (sum of all edge weights) of the dataset’s minimum spanning tree (MST). The MST connects all n points with $n - 1$ edges so that a unique path exists between every two points, and the sum of the edge weights (Euclidean interpoint distances) is a minimum. See Fig. 1 for an illustration. Uniformity is rejected if the length is too long or too short to be consistent with a uniform sample.

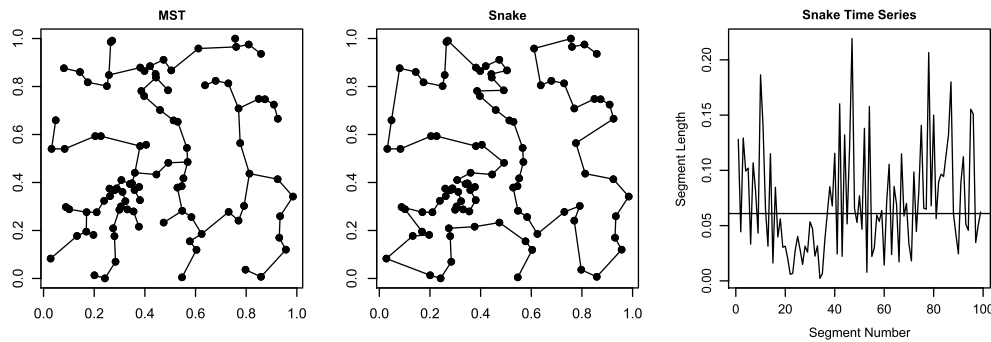


Fig. 1. A dataset with its MST and Snake. The time series is just a plot of segment lengths along the path of the snake. The cluster of points around (0.3, 0.3) appears as a long run of short segments below the median segment length (horizontal line) in the time series.

The second analyzes a ‘snake’, a locally optimal (i.e., short) Hamiltonian path connecting all points. The snake traces the local structure in the datacloud (see Fig. 1 for an illustration) and is effective at detecting variations in the underlying density (Petrie and Willemain, 2010). By plotting the sequence of segment lengths along the path, we effectively transform the multidimensional dataset into a one-dimensional “time series”. When the data is not uniform, a region of high density is manifested as a sequence of short segments (i.e., a long run below the median segment length), while a region of low density is manifested as a sequence of long segments. If the run length statistics from the time series are not consistent with those from uniform samples, uniformity is rejected. Our experiments suggest that the sampling distributions of run length statistics are robust with regards to the shape of the support, so approximate p -values for a test of uniformity can be found without any information about S once the null distribution is simulated for the unit hypercube for the relevant n and d .

We conduct a simulation study to assess the type I error rates and powers of our statistics along with the existing tests of uniformity. When S is known in advance and is the hypercube, the discrepancies of Ho and Chiu (2007) and Liang et al. (2001) typically have the most power for a variety of non-uniform alternatives. However, test statistics based on the MST have superior power against regular alternatives, e.g., data generated by an inhibition process where points are prevented from being within a certain distance of each other or data created by bad pseudo-random number generators where points have a lattice structure.

When S is non-convex or unknown in advance (but may be estimated), the snake has a more consistent type I error rate than the tests in Jain et al. (2002) and Smith and Jain (1984) and generally has superior power to the spacings test in Berrendero et al. (2012a). In the scenarios we have considered, neither the snake nor the distance to boundary method (where S is estimated from the λ -convex hull of the dataset) in Berrendero et al. (2012b) is consistently more powerful than the other, but since the snake can be computed easily in arbitrary dimensions it has the edge as the most suitable test for higher dimensions on arbitrary supports.

We further extend our study to four real-world datasets from the spatial statistics literature and elsewhere. In these examples, the extra utility of the snake is highlighted, since once the time series of segment lengths is plotted it reveals what kind of non-uniformity (e.g., clustering) is present in the data, a property unique to this test.

In Section 2 we introduce our test statistics for uniformity and give a proof of concept for why they should be useful. In Section 3 we report the results of a simulation study that looks at type I error rates and powers of our proposed methods and other current methods for a variety of non-uniform alternatives. In Section 4 we summarize our results and recommend when different tests for uniformity should be used.

2. Proposed new tests for uniformity

Our two tests for uniformity use graph-theoretic constructs built on the data using the data points as nodes and the Euclidean interpoint distances as edge weights. The minimum spanning tree tests require a known support S , while the snake test obtains approximate p -values without explicit knowledge of the support.

2.1. Length of the MST (MST)

The length L_{MST} of the minimum spanning tree (MST) is equal to the sum of its edge weights (Euclidean interpoint distances) and is asymptotically normal, provided the data generating distribution f is absolutely continuous with compact support S and meets a few extra minor requirements (Lee, 1999). The asymptotic expected length of the MST when f is the uniform distribution on S , using the results of Steele (1988) can be shown to be:

$$\lim_{n \rightarrow \infty} \frac{E[L_{MST}]}{n^{(d-1)/d}} = \beta_d \int_S f(x)^{(d-1)/d} dx = \beta_d V^{1/d} \quad (1)$$

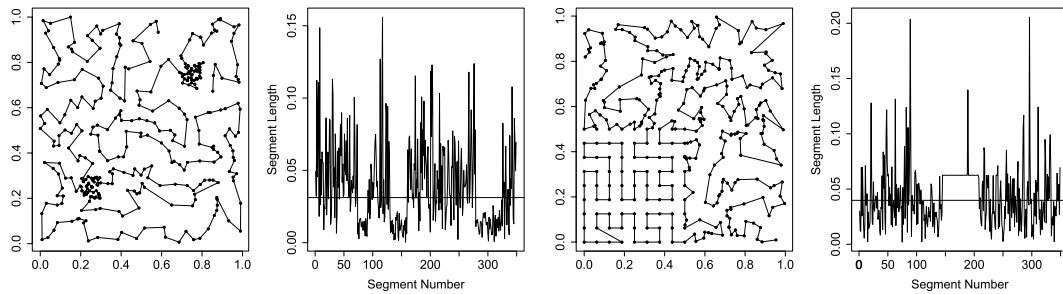


Fig. 2. Clumps (left panels) are manifested as long runs below the median segment length (horizontal line), while a lattice (right panels) is manifested as a long run with nearly equal segment lengths.

where β_d is a constant that depends only on d and not on f or S , and V is the volume of S . Thus, in principal, the length of the MST provides a test statistic that requires only knowledge of the volume of the support (and not its boundary or shape) and is easily calculated (the exact MST can be computed in $O(dn^2) + O(n^2 \log n)$ time; an MST on 10,000 points for $d = 2$ can be found with a greedy algorithm in about 5 s with a 2.16 GHz Intel Core Duo MacBook Pro). Unfortunately, while expressions for β_d exist in theory (Avram and Bertsimas, 1992), they are practically impossible to evaluate for $d > 2$. In principle, expressions for the variance of the asymptotic normal distribution could be derived, but are so complex that in practice the null distribution of MST lengths must be simulated.

The test of uniformity reduces to checking whether the observed MST length falls outside of some central interval of the null distribution obtained via simulations. This test can be made one-sided if it is desired to reject uniformity when clusters or nearly empty regions exist (either will make the length of the MST be smaller than if the data were uniformly distributed) or when regular structures (e.g., lattices, or when there is a minimum allowed distance between points) are present (either will make the length of the MST be larger than if the data were uniformly distributed). We will refer to this method as ‘MST’ in the text and tables.

2.2. Snake run length distribution (Snake)

The snake (Petrie and Willemain, 2010) connects all points in the dataset along a single path and is typically called a Hamiltonian path in the graph-theoretic literature. The shortest possible path naturally traces the local density variations in a datacloud: path segments are long in a region of low density and short in a region of high density. When the density is uniform, there is a predictable amount of variation in the segment lengths. Though finding this shortest path is an NP-hard problem, any reasonably ‘short’ path is typically sufficient for following density variations.

We use the two-opt algorithm (Johnson and McGeoch, 1997) to find a short snake. In this algorithm, an initial path is constructed, either randomly or by starting with a snake composed of a single segment and connecting to it whatever point is closest to the ends of the snake, etc. After an initial path is found, a pair of segments is then deleted and the path is reconnected in the other possible configuration. In other words, one exchanges a pair of segments already in the snake (e.g., perhaps connecting points 3&7 and 9&10) with a pair that is not (e.g., segments connecting points 3&9 and 7&10). If the resulting snake is shorter, the path is updated and the process starts over again; otherwise it attempts to exchange a different pair of segments. A two-opt optimal snake has a length that cannot be shortened by a single exchange of segments and is thus a local optimum. Using efficient data structures, Johnson and McGeoch (1997) shows in simulations that the average running time for planar points is $O(n^{1.2})$, and our own simulations suggest this time is independent of dimension.

After the snake has been constructed, the resulting path may be analyzed by plotting the sequence of segment lengths along the path as a time series (e.g., see Figs. 1 and 2).

We use run length statistics for our test of uniformity. A run is a series of sequential segments whose lengths are all either at least as large or at least as small as the median segment length. Since the snake traces the local density and structure, a cluster or region of high density is manifested as a long run of short segments and a region of low density is manifested as a long run of long segments. We use the median length, rather than the mean, because the distribution of segment lengths for uniform data is skewed right and using the mean segment length inflates the number of long runs, making the detection of non-uniformities more difficult. See Fig. 2 for snakes on datasets with clusters and regular structures.

The maximum run length is a natural test statistic to use to test for uniformity. However, we have found that a more powerful statistic is the ratio of the total number of runs to the maximum run length (Petrie, 2007) (we refer to this ratio statistic as Snake in the text and tables). For example, when a cluster or nearly empty region is present in the dataset there is a long run in the snake and consequently relatively few total runs; thus, the ratio of the two will be small. The intuitive decision criterion is to reject the null hypothesis of uniformity if the ratio is too small. The test can easily be made two sided, though we are unsure as to what type of non-uniformity a large ratio corresponds.

Finding the null distribution of run length statistics is beyond the scope of this paper and may be analytically intractable. However, simulation provides an easy approximation. Fortunately, our experiments show that the null distributions of run length statistics are relatively insensitive to the shape and extent of S , even when it is non-convex. Thus, approximate

Table 1

p -values for the Kolmogorov–Smirnov test comparing the null distribution of Snake when S is the hypercube to when S is another support. Null distributions are found by simulating 1000 uniform datasets. S = random convex region is not tested for $d = 10$ because finding the convex hull is too computationally intensive.

Support	$n = 100$					$n = 500$				
	$d = 2$	$d = 3$	$d = 5$	$d = 7$	$d = 10$	$d = 2$	$d = 3$	$d = 5$	$d = 7$	$d = 10$
Hypercross	0.24	0.43	0.83	0.54	0.08	0.24	0.95	0.76	0.43	0.34
Hypercrescent	0.97	0.98	0.43	0.40	0.76	0.12	0.01	0.57	0.69	0.69
Random convex	0.35	0.18	0.89	10^{-7}	–	0.07	0.22	0.96	10^{-9}	–
Hypersphere surface	0.05	0.18	10^{-4}	10^{-9}	10^{-12}	0.95	0.37	10^{-9}	0	0

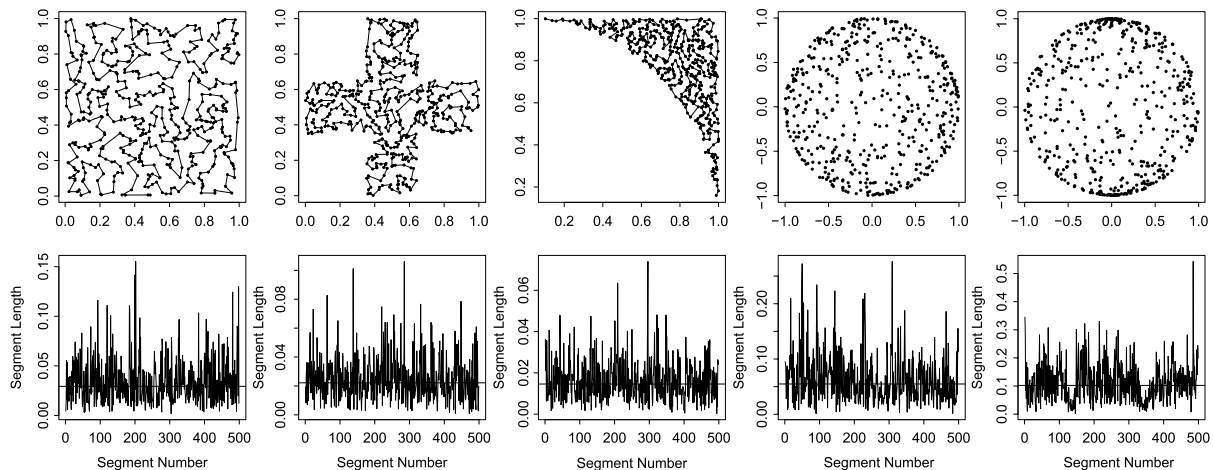


Fig. 3. A snake on uniform data in the unit square (left), inside a cross, inside a crescent, on the surface of a sphere, and on clumpy data on the surface of a sphere (right). The snake behaves in statistically similar ways when the data is uniform regardless of the shape of the domain.

p -values for the test statistic can be found by appealing to its estimated sampling distribution on an easily simulated support, e.g., $U[0, 1]^d$ data, provided the same heuristic (initial path, edge exchange procedure) is used in finding all snakes involved in the computation. The type I and II error rates change very little.

We cannot mathematically prove such a bold assertion, but the proof of concept is illustrated in Fig. 3. The sequence of segment lengths on uniform data in the unit square, inside a two-dimensional cross, inside a crescent (the space outside the unit circle but inside the unit square), and on the surface of a three-dimensional sphere are quite similar in appearance and in run length distribution. In contrast, the snake on data which is clumped on the poles of a sphere looks quite different than on uniform data. Indeed, the snake has two long runs of short segments corresponding to the polar high density regions.

Concrete evidence that the sampling distribution of the ratio is robust to S can be found by comparing it when S is the hypercube to when S is another support using a Kolmogorov–Smirnov test. We simulate the null distribution of the ratio on each support using 1000 trials and present the p -values for the test of equal distributions in Table 1. For $d \leq 10$, the null hypothesis that the sampling distributions are identical is typically not rejected when S is the hypercross, hypercrescent, or a random convex region (defined to be the convex hull of ten randomly chosen points in the hypercube).

One exception is for $d = 7$ for a random convex region where the null distributions are noticeably different. The reason for this is the extreme “pointiness” of the support. Much like the convex hull of a set of three points in two dimensions (a triangle) or a set of four points in three dimensions (a pyramid), the convex hull of seven points in ten dimensions is an extremely marked departure from a hypercube. However, we will eventually see in Table 3 that this difference only results in a slightly liberal test with a type I error rate higher than desired (and the type I error rate is no worse than for other methods which estimate the boundary of the support).

We also find that the sampling distributions in the hypercube and on the surface of a hypersphere are not statistically equivalent, but we should not expect them to be. The surface of a d -dimensional hypersphere has only $d - 1$ dimensions. However, we always assume a full d -dimensional space so that we use no special knowledge regarding the shape of S or its ‘intrinsic dimension’. The difference in sampling distributions makes Snake a conservative test if S happens to be the surface of the hypersphere.

Finally, Section 3.2 reports empirical type I error rates and powers of Snake on supports other than the hypercube and provides additional explicit evidence that the snake is effective at testing uniformity for general datasets when p -values are found by approximating the null distribution of Snake by its sampling distribution in the hypercube.

We must add that since a two-opt snake is a local optimum in terms of length, its structure (and the resulting test) will depend somewhat on the initial path used in construction and the specific sequence of edge exchanges. Through simulation, we find that there can be many local optima and thus many suitable short snakes, though much of the sequence remains

the same from snake to snake. Thus, to conduct the uniformity test it is important to be consistent in the procedure used to construct each snake. If snakes are found by using a random initial path and edge exchanges are attempted starting with the longest edge first when finding the null distribution of run length statistics, then the same construction procedure must be used when analyzing the dataset in question.

3. Simulation study and real data examples

We conduct a simulation study on artificial data and analyze four real datasets in order to quantify the empirical type I error rates ($\hat{\alpha}$) and powers of our new test statistics and the following tests from the spatial statistics and general literature:

1. $G(r)$ – This test estimates the nearest neighbor function $G(r)$, which gives the probability that a point's nearest neighbor is at most r , and compares it with the function for uniform data form via a Kolmogorov–Smirnov type test:

$$T = \sup_{r \leq r_0} |G(r) - \hat{G}(r)|. \quad (2)$$

The power of the test is sensitive to the choice of r_0 . We choose $r_0 = 0.5$ because then $G(r) = 1 - (1 - b_d r^d)^{n-1}$ for n points in the hypercube when distances are measured using the toroidal metric. Here, $b_d = \pi^{d/2} / \Gamma(d/2 + 1)$ is the volume of a d -dimensional unit sphere. The estimate $\hat{G}(r)$ comes from calculating all interpoint distances r with the toroidal metric, selecting all those that are at most $r_0 = 0.5$, then finding the fraction of nearest neighbor distances less than r for each observed value of $r < r_0$.

2. $K(r)$ – This test estimates Ripley's K -function, which gives the expected number of points within a distance r from an arbitrary point divided by the average number of observations per unit hypervolume, and compares it to the function on uniform data with the same type of Kolmogorov–Smirnov for $G(r)$. Again, we choose $r_0 = 0.5$ and use the toroidal metric so that $K(r) = (1 - 1/n)b_d r^d$. The estimate $\hat{K}(r)$ comes from calculating all interpoint distances r with the toroidal metric, selecting all those that are at most $r_0 = 0.5$, then finding the fraction of interpoint distances less than r for each observed value of $r < r_0$.
3. A_n , T_n , UD_2 , and WD_2 – These tests use the discrepancies of the point set (the former two are discussed in [Liang et al., 2001](#) and the latter two are discussed in [Ho and Chiu, 2007](#)). Discrepancies arise in the error analysis of quasi-Monte Carlo methods for evaluating an integral. If I is the value of the integral of some function over the unit hypercube, and Q is the Monte Carlo approximation of the integral using the point set, then the absolute difference between I and Q is no greater than $D \times V$, where D (the discrepancy) measures the quality of the point set and V is a measure of the variation of the function being integrated. Since there are many possible ways to define a discrepancy, we use statistics based only on those discrepancies that have performed well in previous studies. Specifically, we use two statistics A_n and T_n based on the symmetric discrepancy (see [Liang et al., 2001](#) for calculation and discussion), another statistic using the L_2 -unanchored discrepancy (U_2), and a final statistic using the L_2 -wrap-around discrepancy (WD_2). See [Ho and Chiu \(2007\)](#) for calculation and discussion.
4. Spacings – This test uses the maximum multivariate spacing (essentially the size of the largest ball that does not intersect the boundary of the support and that does not contain any of the sample points) as discussed in [Berrendero et al. \(2012a\)](#). When the boundary of the support is unknown it is estimated by the λ -convex hull of the sample. Computation of these spacings for $d \geq 3$ has yet to be worked out, so we present these results only for $d = 2$. Spacings calculated assuming the support is the unit square or using an estimated support are labeled accordingly. We take $\lambda = 1$ unless otherwise specified, and note that one drawback to this method is the correct tuning of λ , whose value can affect the type I error rates and powers. In two dimensions, plotting the λ -convex hull is typically sufficient for choosing λ , but if it becomes feasible to calculate the hull in higher dimensions some heuristic will need to be invented.
5. DBM – This test uses the distance to boundary methods discussed in [Berrendero et al. \(2006, 2012b\)](#), which finds the distribution of each point's minimum distance to its nearest boundary in the support. When the support is a hypercube, hypersphere, or any convex polygon (with deepest point at the origin) circumscribed to a ball, the distribution of these distances is beta with parameters $a = 1$ and $b = d$ so the test of uniformity reduces to a one-dimensional goodness-of-fit test (we use the Kolmogorov–Smirnov test statistic) between the empirical and theoretical distributions of the distances-to-boundary. For the case of unknown planar support ([Berrendero et al., 2012b](#)), the support is estimated using the λ -convex hull of the sample taking $\lambda = 1$ unless otherwise noted. The null distribution of distances-to-boundary is then estimated from the empirical distribution of distances-to-boundary from a uniform distribution with $n = 5000$ generated inside this hull. A two-sample Kolmogorov–Smirnov test compares these distributions and provides a p -value for the test. For the case of estimated support, we refer to the test as DBM_{est} .
6. MSTCDF – This test compares the empirical CDF of the $n - 1$ edges of the MST to its theoretical CDF ([Hoffman and Jain, 1983](#)) on uniform data. To obtain the theoretical CDF we simulate 100,000 MSTs on $U[0, 1]^d$ data, then tabulate the fraction of edges that are at most r , where r is a set of 500 values equally spaced between zero and the maximum edge length observed over all trees simulated at that dimension. The test statistic is the maximum difference between the empirical and theoretical values of the CDF at these 500 distances.
7. Convex – This test performs a two-sample test of equality between the original sample and a uniform sample generated from within the original sample's convex hull ([Smith and Jain, 1984](#)). For moderate dimensions ($d \leq 8$), existing software

such as the quickhull algorithm (Barber et al., 1996) can find the convex hull in reasonable computational time. It is possible to generate a uniform sample within any convex support using a simple rejection method: generate a uniform point in a region larger than the convex hull (e.g., a large hypercube), and reject the point if it falls outside the convex hull. We use the energy test (Szekely and Rizzo, 2004) found in package `energy` in R for the two-sample test.

8. Resampling – This test compares a ‘low-density’ subset to a ‘high-density’ subset generated from the original sample. See Jain et al. (2002) for how these subsets are chosen. Uniformity is rejected if the ‘high-density’ and ‘low-density’ samples fail a two-sample test of equality (we once again use the energy test to compute a p -value), or if the dataset fails the ‘edge-length’ test, i.e., the longest edge in the MST is more than six standard deviations away from the average length. This method has the advantage that no estimation of the support of the data is required, nor is the support required to be convex. However, it is admittedly based more on heuristics and does not constitute a formal test since Jain et al. (2002) do not show that this resampling procedure is valid. Indeed, it performs poorly in our simulations and is not recommended.

In our study, we take the target type I error rate to be $\alpha = 5\%$. In general, to find the p -value for a test statistic, we first find \hat{H} , the empirical distributions of the test statistic under the null distribution of uniformity in the hypercube observed over 10 000 independent samples. The p -value for the statistic on a new dataset is taken to be the fraction of these 10 000 values that are at least as extreme as the observed test statistic. Although expressions for asymptotic or exact p -values exist (e.g., Discrepancies and DBM respectively), we do not use them here so that no approximations are required for any method and each are on an equal footing. The empirical type I error rate $\hat{\alpha}$ (or power, depending on the context) is thus taken to be the fraction of p -values that are smaller than 5%.

The exceptions to this scheme are for Convex, Resampling, Spacings, and DBM_{est} . For Convex, the p -value for the test is the p -value of the two-sample test (found from 5000 permutations of the sample labels) comparing the original sample with a single new uniform sample generated inside the original sample’s convex hull. The p -value for Resampling is either taken to be zero if the initial edge-length test is failed (which happens an insignificant amount of time for truly uniform data), or taken to be the p -value of the two-sample test (again found from 5000 permutations of the sample labels) comparing a single ‘high-density’ and ‘low-density’ sample generated from the original dataset. The p -value for Spacings is output from code provided from Berrendero et al. (2012a). The p -value for DBM_{est} is the p -value of the two-sample Kolmogorov–Smirnov test.

In our study of type I error rates we look at uniform data when S is a hypercube, hypersphere, hypercrescent (area outside a hypersphere but inside the hypercube), hypercross (a d -dimensional ‘plus’ sign, ranging between 0.35 and 0.65 in each dimension), random convex polyhedron (generated within the convex hull of ten randomly chosen points in the hypercube), and the surface of a hypersphere. See Fig. 3 for examples of these supports in two dimensions. To compare powers, we consider data from various non-uniform distributions in the hypercube: beta, meta-uniform, contaminated uniform mixtures, a Neyman–Scott clustering process, uniform data with Poisson cluster contamination, and a Matern inhibition process. We also compare powers for non-uniform data on the surface of a hypersphere, inside a hypercross, and multivariate normal data. We conclude with four real-data examples.

Whenever S is not the hypercube, we compare Snake and Resampling (since these are the methods which can be implemented without prior knowledge or estimation of the support) and Spacings and DBM_{est} (for $d = 2$) since Berrendero et al. (2012a,b) do provide a means of estimating the support and conducting their test in this instance. Results are found from 5000 trials with $n = 50, 100, 250, 500$ and $d = 2 - 10$ unless otherwise noted. In the interest of space only results for selected n and d are reported, though other conclusions may be addressed separately in the text.

3.1. Null case – S is the hypercube

We consider those test statistics whose type I error rates may differ from the target of 5%: Convex (a $U[0, 1]^d$ sample may not be statistically equivalent to a new uniform sample generated from within its convex hull), Resampling (‘high’ and ‘low’ density subsamples generated from the original $U[0, 1]^d$ sample may not be statistically equivalent), Spacings (whether the support is specified or estimated the p -values of the test statistic are approximate), DBM_{est} (the support may not be estimated well) and Snake (the test statistic is discrete so the type I error rates are bounded above by 5%). Since the other statistics’ type I error rates are calculated from uniform data in the hypercube, they are 5% by definition.

For $d > 7$ we do not compute Convex since calculation of the convex hull and the generation of a uniform sample is very computationally intensive. As it turns out, $\hat{\alpha}_{Convex}$ increases dramatically with dimension so it is unlikely to perform well in the omitted dimensions. Table 2 displays the empirical type I error rates found after 10 000 trials. For each test statistic, any value between 3.6% and 6.4% is nominally consistent with 5% at a significance level of 0.05. Values not consistent with 5% are marked with a *.

Convex reproduces the target type I error rate for $d = 2$, but $\hat{\alpha}_{Convex}$ increases dramatically as d increases, becoming over 50% by $d = 5$. The issue is that the uniform sample generated inside the original sample’s convex hull invariably has a higher point density due to it having a smaller hypervolume than that of the original $U[0, 1]^d$ sample, so it typically fails the test of equality. When d increases, the difference between these volumes, and the resulting point density, is exacerbated. Potentially, one could correct for the difference by estimating the support with the λ -convex hull as in Berrendero et al. (2012a,b), but this is not computationally feasible for $d > 3$. Thus, Convex is unsuitable for testing uniformity in higher dimensions so we forgo any further testing of this method, though we do note that this method could have potential if a better estimator of the support of a high-dimensional dataset becomes available.

Table 2Empirical type I error rates ($\hat{\alpha}$) on $U[0, 1]^d$ data based on 10 000 trials.

	$n = 100$				$n = 500$			
	$d = 2$	$d = 3$	$d = 5$	$d = 10$	$d = 2$	$d = 3$	$d = 5$	$d = 10$
Convex	5.06	7.60*	85.62*	–	–	–	–	–
Snake	5.02	5.19	5.05	5.06	5.03	5.12	5.01	5.00
Resampling	2.06*	2.38*	2.00*	0.48*	7.70*	8.30*	9.82*	11.60*
Spacings _{square}	3.12*	–	–	–	4.86	–	–	–
Spacings _{est}	2.26*	–	–	–	4.44	–	–	–
DBM _{est}	4.25	–	–	–	4.85	–	–	–

* indicates inconsistency with 5%.

Table 3

Empirical type I error rates. Results based on 5000 trials.

n		Random convex polygon				Hypercross			
		$d = 2$	$d = 3$	$d = 5$	$d = 7$	$d = 2$	$d = 3$	$d = 5$	$d = 10$
100	Snake	4.24	5.83	4.48	8.21 ^a	6.00	4.54	5.32	4.80
	Resampling	11.58 ^a	3.62	2.90 ^a	1.94 ^a	2.26 ^a	2.20 ^a	2.14 ^a	0.48 ^a
	Spacings _{Est}	2.00 ^a	–	–	–	1.52 ^a	–	–	–
	DBM _{Est}	3.92	–	–	–	5.27	–	–	–
500	Snake	4.38	5.73	5.84	10.9 ^a	6.30	5.06	4.98	4.96
	Resampling	34.72 ^a	13.40 ^a	12.10 ^a	13.52 ^a	8.00 ^a	8.72 ^a	10.46 ^a	15.06 ^a
	Spacings _{Est}	4.46	–	–	–	4.40	–	–	–
	DBM _{Est}	5.64	–	–	–	18.8 ^a	–	–	–
		Hypercrescent				Hypersphere surface			
		$d = 2$	$d = 3$	$d = 5$	$d = 10$	$d = 3$	$d = 4$	$d = 6$	$d = 10$
100	Snake	4.94	5.32	4.86	5.98	1.98 ^a	2.38 ^a	2.50 ^a	1.64 ^a
	Resampling	19.76 ^a	3.26 ^a	1.70 ^a	0.48 ^a	1.44 ^a	2.16 ^a	1.62 ^a	0.26 ^a
	Spacings _{Est}	1.62 ^a	–	–	–	–	–	–	–
	DBM _{Est}	11.0 ^a	–	–	–	–	–	–	–
500	Snake	4.99	5.04	5.38	4.30	1.70 ^a	1.64 ^a	1.62 ^a	0.80 ^a
	Resampling	74.64 ^a	23.28 ^a	11.12 ^a	11.34 ^a	14.06 ^a	6.62 ^a	8.14 ^a	8.66 ^a
	Spacings _{Est}	4.67	–	–	–	–	–	–	–
	DBM _{Est}	13.8 ^a	–	–	–	–	–	–	–

^a not consistent with $\hat{\alpha}$ for the hypercube.

Resampling misses the target of 5% entirely. Specifically, $\hat{\alpha}_{\text{Resampling}}$ is too low for $n = 50, 100$ (and appears to decrease with dimension) and too large for $n = 250, 500, 1000$ (and appears to increase with dimension). The resampling procedure proposed by Jain et al. (2002) is clearly not yielding two statistically equivalent samples when the actual data density is uniform.

Snake and DBM_{est} have $\hat{\alpha}$ consistent with the target value of 5%. The empirical type I error rate of spacings is too low when (correctly) assuming that the point set is the unit square, but achieves the target 5% when the support is estimated. Spacings has an empirical type I error rate too small for $n = 100$, but achieves its target level for $n = 500$.

3.2. Null case – S is not a hypercube

Typically, the support of a dataset is unknown in advance so only the Snake and Resampling statistics can immediately be calculated unless the support can be estimated, in which case $\text{Spacings}_{\text{set}}$ and DBM_{set} can be found for $d = 2$. Although the null distribution of Snake is affected by the shape of the support, our simulations show that the desired type I error rates are typically achieved if we simply approximate its p -value by using the null distribution in the hypercube. Section 2.2 shows that in general, the differences between the sampling distributions on various supports are not statistically significant.

Table 3 further confirms that this approximation is typically good ($\hat{\alpha}$ consistent with 5%) when S is the convex hull of ten randomly chosen points in the hypercube (we take $d \leq 7$ since computation in higher dimensions takes too long), when S is a hypercross, and when S is a hypercrescent. The only exceptions are for $n = 500$ and $d = 7$ with the random convex region where the empirical type I error rates are near 10% (recall this is due to the support being very “pointy” and quite unlike a hypercube). Further, the snake yields a conservative test ($\hat{\alpha}$ typically between 1% and 2%) when S is the surface of the hypersphere, but this is because the “wrong” sampling distribution is being used: the surface of a d -dimensional hypersphere has only $d - 1$ dimensions but we always assume a full d -dimensional space. In principle we could use the appropriate d for the sampling distributions, but we wish to apply our method without any special knowledge regarding the shape of S or its ‘intrinsic dimension’.

Resampling once again misses the target type I error rate, though in an inconsistent manner. For example, it is too conservative when S is the surface of a hypersphere and $n = 100$ ($\hat{\alpha} = 1.4\%$), but liberal when $n = 500$ ($\hat{\alpha} = 14.1\%$). The type I error rates can be unacceptably high, e.g., $\hat{\alpha} = 35\%$ for $n = 500$, $d = 2$ when S is a random convex polygon, and $\hat{\alpha} = 75\%$ for $n = 500$, $d = 2$ when S is a hypercrescent. Because Resampling is not reliable in regards to its type I error rate (and we cannot think of an obvious way to make corrections), we drop this test from further discussions.

Further, Table 3 also shows that $\text{Spacings}_{\text{set}}$ remains a conservative test when $n = 100$ on two-dimensional non-hypercubical supports (computation of the test statistic for $d \geq 3$ is not yet possible), with type I error rates typically being less than 5%. For $n = 500$ however, it does achieve the target significance level (perhaps reflecting that the support is much better estimated). For each case, $\text{Spacings}_{\text{set}}$ is estimating the support using the λ -convex hull with $\lambda = 1.0$, except for the cross where $\lambda = 0.2$ (too large a λ consistently places the largest spacing outside the cross itself).

Finally, we see that also DBM_{est} fails to achieve the desired type I error rate in half the cases ($n = 500$ for the hypercross and both sample sizes for the hypercrescent): $\hat{\alpha}$ is 10% or larger. The poor performance may be due to the λ -convex hull (with $\lambda = 1$ for the hypercrescent or $\lambda = 0.2$ for the hypercross) being a poor estimate of S , though visually there does not look to be issues. Rather, it may be due to the fact that DBM_{est} underestimates the proportion of points near the boundary of S combined with these supports having a high perimeter to area ratio.

To summarize, we find that the snake is the most successful at achieving the target type I error rate on non-hypercubical supports, even though the sampling distribution of the test statistic is approximated by its sampling distribution in the hypercube. Resampling appears to have significant difficulty achieving the desired type I error rate on any support considered, even the hypercube, and is thus useless in this general setting. $\text{Spacings}_{\text{set}}$ and DBM_{est} often achieve the desired type I error rate, though the former is often too conservative and the latter may be too liberal in certain settings. For the critical dimension of $d = 2$, DBM_{est} especially has difficulty achieving the target error rate on non-convex supports.

3.3. Alternative cases of theoretical interest

Previous studies such as Liang et al. (2001) and Berrendero et al. (2006, 2012a) have examined their tests on a variety of non-uniform alternatives in the hypercube. To compare with these previous works, we recreate their scenarios.

The first alternative has each dimension being an independent beta distribution with both parameters equal to 0.8 (giving each marginal distribution a U-shape peaking at 0 and 1). The second alternative is a ‘measure-contaminated uniform’ (discussed in Berrendero et al., 2006) where $U[0, 1]^d$ data has been mixed with some percentage of uniformly distributed points within a smaller hypercube of volume 0.5. Let ϵ be the fraction of contamination in the mixture, then the distribution is $(1 - \epsilon)U[0, 1]^d + \epsilon U[x, 1 - x]^d$, where x solves the equation $(1 - 2x)^d = 1/2$. The length of an edge of the hypercube containing the contamination grows with d (by $d = 10$, each edge is between $[0.03, 0.97]$), so the contamination looks more and more distributed as $U[0, 1]^d$. We take $\epsilon = 0.2$.

The third alternative is a correlated ‘meta-uniform’ distribution obtained by taking some correlated multivariate distribution and transforming it into the unit hypercube using the (marginal) inverse CDF transform on each dimension separately. The marginals of a meta-uniform distribution are uniform, but the joint is not. We take the baseline distribution to be a multivariate normal centered at the origin with unit variances and each covariance equal to 0.5.

The powers of the statistics estimated from 5000 trials can be found in Table 4 (for $n = 100$). We consider the results of tests which must estimate or do not use the support (Snake, DBM_{est} , $\text{Spacings}_{\text{Est}}$) separately in the text, but incorporate the powers in the table to see how much is “given up” by these more flexible methods.

For tests that explicitly use the fact that the support in these simulations is a hypercube, the general conclusion is that no test statistic is always the most powerful, e.g., with UD_2 being the best for the beta alternative, DBM being best for the measure-contaminated alternative, and T_n being the best for the meta-uniform alternative. The powers of the test statistics increase with n , though the overall rankings of the statistics remain the same.

Not surprisingly, tests that estimate or use no knowledge of the support (Snake, $\text{Spacings}_{\text{est}}$, and DBM_{est}) typically have much lower power than tests that exploit the support. We compare these tests against each other for a larger sample of $n = 500$, see Table 5. Snake is comparable to or outperforms Spacings, though DBM_{est} is clearly more powerful for these larger sample sizes. With such large sample sizes with these types of alternatives, the λ -convex hull is an excellent approximation to the unit square and there is little difference between DBM_{est} and DBM itself.

3.4. Alternative cases of practical interest

In data analysis, the more interesting alternative to uniformity is the presence of some type of clustering. In spatial statistics, another common alternative is that the points display regularity, e.g., there is some imposed minimum distance between points. Fig. 4 shows examples of alternatives we will consider.

In a Neyman–Scott clustering processes, ‘parent’ points are realizations from a spatial Poisson process around which ‘offspring’ are clustered. For our first alternative, we consider $n_{\text{parent}} = 0.1n$ parent points randomly generated in the hypercube. Offspring are assigned at random to the parents, and then are placed randomly and uniformly within a hypersphere with radius 0.5 centered on their parents. An offspring is rejected if it falls outside the hypercube (and more offspring are generated for that parent until one falls inside the hypercube).

Table 4
Power for alternatives of theoretical interest for $n = 100$. Boxed powers represent the best for a particular alternative. Results based on 5000 trials.

	Beta with $a, b = 0.8$				Measure contaminated $U[0, 1]^d \epsilon = 0.2$				Meta-uniform (normal)			
	$d = 2$	$d = 3$	$d = 5$	$d = 10$	$d = 2$	$d = 3$	$d = 5$	$d = 10$	$d = 2$	$d = 3$	$d = 5$	$d = 10$
A_n	62	81	95	100	37	33	26	15	9	32	93	100
T_n	52	71	91	100	24	21	18	12	65	99	100	100
UD_2	67	83	96	100	33	31	21	14	6	11	14	31
WD_2	27	34	45	73	12	8	5	5	4	5	9	60
$G(r)$	5	6	7	8	5	5	5	5	12	33	71	99
$K(r)$	28	32	34	11	15	9	6	5	24	47	95	100
DBM	56	72	89	99	44	41	43	43	10	23	57	94
DBM _{Est}	20	–	–	–	27	–	–	–	5	–	–	–
MSTCDF	8	19	53	91	8	13	13	9	36	87	100	100
MST	11	24	59	95	9	14	15	10	49	94	100	100
Snake	4	4	4	5	8	7	6	6	12	22	31	41
Spacings _{Square}	12	–	–	–	1	–	–	–	7	–	–	–
Spacings _{Est}	6	–	–	–	1	–	–	–	3	–	–	–

Table 5
Power for alternatives of theoretical interest for $n = 500, d = 2$. Results based on 5000 trials.

	Beta with $a, b = 0.8$	Measure contaminated $U[0, 1]^d \epsilon = 0.2$	Meta-uniform (normal)
Snake	10	15	47
Spacings _{Est}	14	4	32
DBM _{Est}	95	97	30

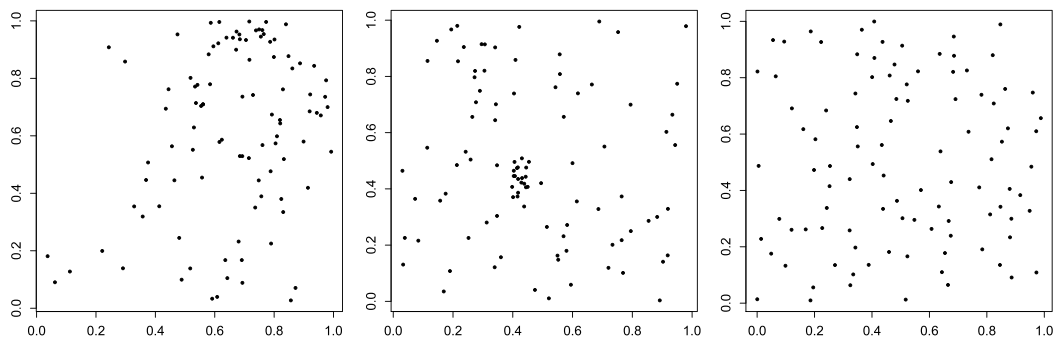


Fig. 4. Left: realization of Neyman–Scott clustering processes with 10 parents (with offspring uniformly in a risk of radius 0.25); middle: realization of $U[0, 1]^2$ data with 20% contamination from a cluster with one parent and 19 offspring; right: realization of Matern inhibition process.

For the second alternative, we consider $U[0, 1]^d$ data with 20% contamination from a cluster. The contamination consists of a single parent whose location is randomly chosen to fall between 0.4 and 0.6 (uniformly and independently chosen in each dimension), and generate offspring from an uncorrelated multivariate normal distribution whose standard deviation is equal to $0.8\bar{s}/\sqrt{d}$, where \bar{s} is the average segment length of a snake on $U[0, 1]^d$ for that value of n . An offspring is rejected if it lies outside the hypercube. The reasoning behind this rather unconventional generating procedure is that we want the interpoint distances inside the cluster to be smaller than the interpoint distances in the surrounding noise so that, as the snake travels through this clump, a ‘dip’ should occur in the resulting time series of segment lengths.

The third alternative has regularity. We consider a Matern simple inhibition process. Two points are prevented from being within a distance Δ of each other, as if the points themselves were disks of size Δ that cannot overlap (Diggle, 2003). For example, trees in a forest must be a minimum distance from each other as they compete for light. Although an analogous situation in higher dimensions may not be readily apparent (although bad random number generators will often have lattice-like structures in higher dimensions), we examine this anti-clustered alternative for completeness. In our simulations, we take Δ to be the 20th percentile of the edge length distribution of MSTs on $U[0, 1]^d$ data for that particular value of n (the CDF of MST edge lengths has been found in simulations previously since our test statistic MSTCDF requires its values).

Table 6Power for alternatives of practical interest for $n = 50$. Results based on 5000 trials.

	Neyman–Scott clustering $n_{parent} = 5$				$U[0, 1]^d$ with cluster contamination				Matern inhibition process			
	$d = 2$	$d = 3$	$d = 5$	$d = 10$	$d = 2$	$d = 3$	$d = 5$	$d = 10$	$d = 2$	$d = 3$	$d = 5$	$d = 10$
A_n	61	92	100	100	66	79	86	95	2	3	6	17
T_n	86	100	100	100	51	66	78	89	0	1	3	15
UD_2	41	47	57	76	62	71	77	84	2	3	5	10
WD_2	76	97	100	100	50	45	35	22	18	15	9	7
$G(r)$	55	99	100	100	12	12	10	16	100	98	76	11
$K(r)$	82	99	100	100	57	53	49	19	0	1	5	10
DBM	32	37	40	42	65	63	52	37	1	3	6	7
DBM_{est}	15	–	–	–	68	–	–	–	1	–	–	–
MSTCDF	80	100	100	100	29	29	38	63	100	100	81	59
MST	87	100	100	100	20	29	44	73	50	62	61	64
Snake	23	34	50	52	50	39	23	13	5	4	4	4
Spacings _{square}	18	–	–	–	16	–	–	–	0	–	–	–
Spacings _{est}	5	–	–	–	4	–	–	–	0	–	–	–

Table 6 shows the estimated powers over 5000 trials for these alternatives for $n = 50$ (for $n = 100, 500$ the powers of many of the methods are nearly 100%, making them hard to compare). Again, there is no consistently most powerful statistic.

For the Neyman–Scott clustering alternative, the test statistics based on the MST do very well, with the length of the MST outperforming the generalization of the technique by Hoffman and Jain (1983) which uses the CDF of the edge lengths for the typical case of $d = 2$ in spatial statistics. In fact, MST has comparable power to the best performing discrepancy (T_n) and is at least as powerful as the summary functions. For $d > 3$ all but Snake, DBM, Spacings, and UD_2 have essentially 100% power.

For uniformly distributed data with cluster contamination, A_n is the clear winner, though DBM performs well for $d = 2$. In fact, for this one instance DBM_{est} has comparable power to DBM itself. We believe this is because the non-uniformity is located near the center of the support, so accurate estimation of the boundary is much less critical. The power of DBM (and summary functions) decreases or stays constant with d while the powers of the discrepancies and MST based statistics increase with d .

For the regular (Matern inhibition) alternative, MSTCDF is the clear winner, though MST becomes comparable at $d = 10$ and $G(r)$ performs comparably for the critical dimensions of $d = 2, 3$. The discrepancies, DBM, Spacings, and Snake have little to no power against this alternative. It is interesting to note that both DBM_{est} and Spacings_{est} have powers much less than 5%. Indeed, for most dimensions all of these except Snake have powers that are significantly below 5%.

Of the methods that require no knowledge of or that estimate the support (Snake, DBM_{est} , and Spacings_{est}), Snake has the most power for $n = 50, 100$, but by $n = 500$ we find that DBM_{est} is the most powerful (with Snake and Spacings_{est} having comparable power), indicating again that once the support is well-estimated (and nearly hypercubical), the distance to boundary method is preferred.

In summary, when S is the hypercube there is no clear most powerful test among those which exploit foreknowledge of S . The MST, DBM, summary functions, and discrepancies all excel in specific alternatives. We do feel that without a specific alternative in mind, T_n is the safest bet because while it is not always the most powerful, it has the least noticeable failures (only the inhibition process) compared to the other discrepancies (all of which perform poorly for the inhibition process and in at least one other scenario, e.g., A_n , UD_2 , and WD_2 fail for meta-uniform normal data in low dimensions), the summary functions (both of which typically perform poorly when the alternatives are not clustering or regularity), DBM (which performs poorly on meta-uniform, Neyman–Scott, and regular alternatives), and the MST (which performs poorly in low-dimensional cases where the alternative is not clustering or regularity). Before recommending a test when the support is unknown, we consider a few more scenarios.

3.5. Alternative – non-uniform data when S is not the hypercube

The discrepancies, summary functions, and MST-based statistics all require explicit (direct or indirect) knowledge of the support and it becomes non-trivial to apply these techniques to arbitrary supports. One may approximate the support \hat{S} as the intersection of boundaries of hyperspheres as in Cuevas and Rodriguez-Casal (2004) or by its λ -convex hull as in Berrendero et al. (2012a). However, we find in simulations that \hat{S} from Cuevas and Rodriguez-Casal (2004) does not yield accurate sampling distributions under uniformity, and the λ -convex hull is currently too difficult to calculate except for $d = 2$.

Table 7

Power vs. non-uniform distributions on non-hypercubical supports. Results based on 5000 trials.

	Clumpy hypersphere surface ($n = 50$)				Clumpy hypercross ($n = 100$)			
	$d = 3$	$d = 4$	$d = 6$	$d = 10$	$d = 2$	$d = 4$	$d = 6$	$d = 8$
Snake	21	96	97	95	40	53	54	40
Spacings	–	–	–	–	1	–	–	–
DBM_{est}	–	–	–	–	93	–	–	–
Gine	85	100	100	100	–	–	–	–
	Multivariate normal ($n = 50$)				Multivariate normal ($n = 250$)			
	$d = 2$	$d = 3$	$d = 5$	$d = 10$	$d = 2$	$d = 3$	$d = 5$	$d = 10$
Snake	27	62	45	22	84	90	85	55
Spacings	1	–	–	–	69	–	–	–
DBM_{est}	80	–	–	–	100	–	–	–

Let us consider three additional scenarios for those tests that do not require explicit knowledge of the boundary of the support (Snake, Spacings_{est}, and DBM_{est}). The first is non-uniform data on the surface of a d -dimensional hypersphere, which can easily be made by uniformly generating angles $\theta_1, \dots, \theta_{d-1}$ between 0 and 2π and using the hyperspherical transform:

$$x_1 = \cos \theta_1 \quad x_2 = \sin \theta_1 \cos \theta_2 \quad x_{d-1} = \sin \theta_1 \cdots \sin \theta_{d-2} \cos \theta_{d-1} \quad x_d = \sin \theta_1 \cdots \sin \theta_{d-2} \sin \theta_{d-1}.$$

This transformation results in data that is clumped at the poles of the hypersphere when $d \geq 3$ (see Fig. 3). There are a large number of tests specifically designed to test uniformity on the surface of hyperspheres, e.g., Gine's test (see Ch. 10 of Mardia and Jupp (2000) for a discussion). Although this is not a fair comparison because Gine's test exploits knowledge of S and is specifically designed for this scenario, the differences in power between it and Snake are illuminating. DBM cannot be used because the support has no boundaries, and currently there is no algorithm to find the largest multivariate spacing on the surface of a sphere.

The second alternative is non-uniform data in a hypercross, which we make by generating uniform data within a $1.5d$ -dimensional hypercross and then projecting it down to d dimensions by deleting the latter $0.5d$ descriptors (this results in the center of the cross being denser than the arms). The final alternative is uncorrelated multivariate normal data, whose support is technically unbounded. Table 7 shows the powers of Snake for various d and for Spacings_{set} and DBM_{set} for $d = 2$.

Unsurprisingly, compared to Gine's specialized test for the clumpy hypersphere, Snake has less power. Spacings lacks power compared to Snake when the alternative is a clumpy hypercross or multivariate normal, even for larger sample sizes. Thus, overall we believe that the snake is more useful than Spacings, especially since its calculation is possible for any dimension. DBM_{set} , however, is typically much more powerful than the snake (with the exception being meta-uniform normal and Neyman–Scott clustering for smaller dimensions).

Further, we note that the value of λ used in Spacings_{set} and DBM_{set} appears to become much more important as n increases. Using $\lambda = 1$ when $n = 500$ for bivariate normal data results in the largest sphere containing no points to be located at the very edge of the datacloud extending outward. Indeed, if λ is beyond some threshold, the size of this sphere is essentially set by the value of λ and not by the location of any particular data points. We have chosen $\lambda = 0.2$ when $n = 500$, and note that this tuning of λ is a drawback of the Spacings method in general settings. Additionally, we do have some lingering doubts as to the appropriateness of DBM_{est} , namely its type I error rates on non-convex supports seem to be much higher than desired, it cannot be implemented effectively $d > 2$, and it cannot be implemented if S has no boundary, like the surface of a sphere or torus.

To summarize, Snake appears to always be comparable and actually is typically much more powerful than Spacings_{set}, so it is to be preferred especially since its calculation is possible for any dimension. DBM_{set} is a great choice for $d = 2$ when S looks like a square or circle, but for more general supports we recommend using Snake.

3.6. Examples with real data

To complement the simulation study, we offer four examples using real data that further compare the tests for uniformity while also showing the utility of the new MST and Snake tests. Our first three examples consider datasets in spatial statistics available in the package spatstat in R. The final example is a 93-dimensional dataset containing the fluxes from many different wavelength bands of 531 astronomical objects taken from the UCI data repository (Low Resolution Spectrometer Dataset). In all cases, we are interested in whether the data is uniformly distributed over its support, and if not, what kind of non-homogeneities exist.

We begin by examining the Lansing dataset previously considered in Berrendero et al. (2012a). This dataset contains the locations of different types of trees in a square plot near Lansing, Michigan. We consider the distribution of the 929 oak trees only since all test statistics considered in this paper readily reject uniformity for hickories and maples. Upon visual inspection (see Fig. 5), we see that most of the plot is roughly consistent with a uniform random sample with the exception of some potential clusters near the edges of the square (we disagree with the assessment of Berrendero et al. (2012a) p. 268 who claim there is “no obvious structure”). Table 8 gives the p -values for each test statistic we have considered. The consensus is that the distribution of trees is not uniform, with uniformity being retained only by Spacings and $G(r)$.

Table 8
 p -values for the test of uniformity on the landings oaks.

A_n	T_n	UD_2	WD_2	$G(r)$	$K(r)$	DBM
0.020	< 0.001	0.003	0.004	0.699	0.008	0.002
DBM_{est}	MSTCDF	MST	Snake	$Spacings_{square}$	$Spacings_{est}$	
0.003	0.043	0.016	< 0.001	0.350	0.313	

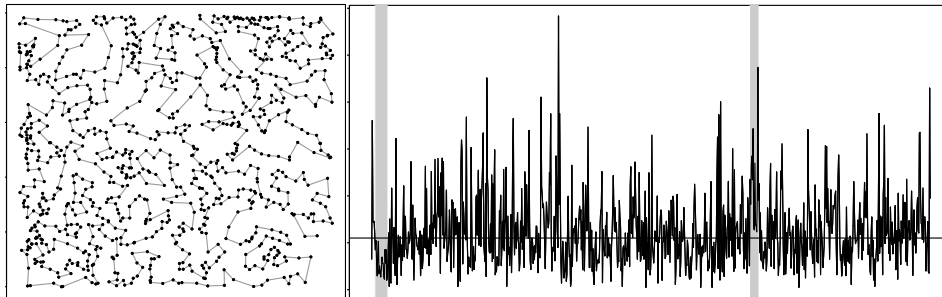


Fig. 5. Left: Positions of 929 oak trees in the landings with the snake superimposed. Right: Snake highlighting two unusually long runs. The run of length 22 at the beginning of the snake corresponds to a cluster on the upper right edge of the support, while the run of length 16 near the end of the snake corresponds to the sequence of unusually long segments found on the bottom right perimeter and indicates a region of relatively low density. Although most test statistics reject uniformity, Snake is the only one that reveals the types of non-uniformities in the data.

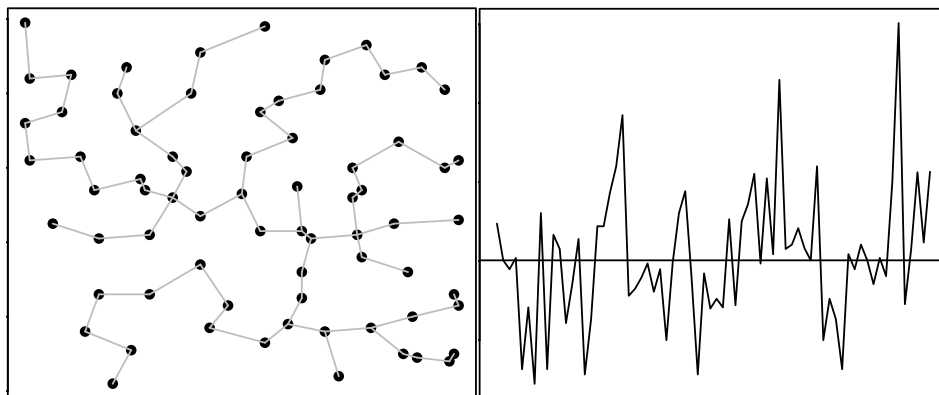


Fig. 6. Left: Positions of 71 trees from swedishpines with the MST superimposed. Right: Snake. The locations of the trees appear to result from a type of inhibition process. The discrepancies, spacings, and distance to boundary method are unable to detect the regularity in the places, and the snake shows no obvious long runs indicating clustering or sparsity. Indeed, only the MST (which is “too long” to be consistent with uniform data) and the nearest neighbor function reject uniformity.

The snake offers an additional glimpse into why uniformity was rejected and what types of non-uniformities are present in the dataset. Through simulation, we find that the probability of observing a run of length 16 or more on uniform data in the unit square with $n = 929$ is 3.2%. Thus, any run of length 16 or longer indicates some type of nonuniformity. The snake has a run of 22 short segments which matches up with the apparent region of high density along the upper right perimeter (see Fig. 5), and it has a run of 16 long segments which trace the lower right perimeter, indicating that this region has unusually low density.

We now consider the swedishpines dataset, which contains the locations of 71 trees in a 10 by 10 meter square. See Fig. 6. Previous analyses, e.g. Baddeley and Turner (2000) fit the data to a Strauss process (a type of inhibition process). Thus, the MST and nearest neighbor function should reject uniformity while the other methods may be unable to detect the nonuniformity. As anticipated, only the $G(r)$, MST, and MSTCDF reject the null hypothesis of uniformity (with p -values less than 0.001). The length of the MST is much too long to be consistent with uniform data. As in our simulations, the discrepancies, spacings, distances-to-boundary, and Snake are unable to detect global regularities in the data.

Next, we consider the urkiola dataset, which contains the locations of 886 birch trees in Urkiola Natural Park in Spain and is the only example with real data in spatstat with an irregular polygonal shape where non-uniformity is not extremely obvious upon visual inspection. The support is known, but for illustration let us incorporate no foreknowledge of the boundary so that we can meaningfully compare Snake, DBM_{est} , and $Spacings_{est}$. The boundary is estimated using the

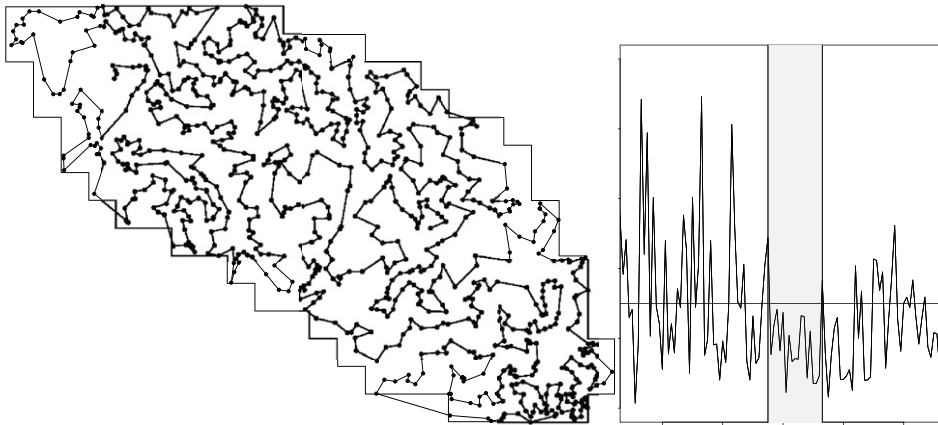


Fig. 7. Left: Positions of 866 birch trees in the Urkiola dataset along with the polygonal boundary of the support; superimposed is the Snake. Right: the snake for 100 segments centered on the apparent non-uniformity in the lower right of the support. The highlighted run of short segments is unusually long for uniform data and contributes greatly to the snake's rejection of the null hypothesis of uniformity. Spacings and DBM_{est} both retain uniformity.

λ -convex hull (taking $\lambda = 1$) after scaling each dimension down by a factor of 100 so that values range typically between 0 and 2). Examining Fig. 7, we see that the density of points is roughly uniform with the exception of a region with seemingly high density near the lower right of the support. Indeed, the snake has a run of 17 short segments in this region, which is highly unusual for truly uniform data (the probability of finding a maximum run length of 17 or larger in uniform data is approximately 2%). The p -value for the test of uniformity is 0.0054 for the snake, leading us to reject uniformity in favor of some type of clustering. The p -values for DBM_{est} and $Spacings_{est}$ are 0.0054, 0.16, and 0.14, respectively, so these tests fail to pick on the localized region of high density.

The Low Resolution Spectrometer (LRS) dataset provides an excellent illustration of the snake as a test for uniformity. Since the data has 93 variables (the flux of astronomical objects measured at 93 different wavelength bands), it is impractical to estimate the support, so the Snake is the only viable means to test uniformity. The data contains four classes of astronomical objects. It is interesting to test whether observations within a given class are uniform or show some additional type of structure (e.g., subclasses are present). Consider classes 3 and 4. Fig. 8 shows the snake on these datasets along with a “typical” Snake on $U[0, 1]^{93}$ data. The p -values for the tests of uniformity on classes 3 and 4 are 0.027 and 0.541, respectively (the p -values for the other two classes are both above 5%), indicating that while class 4 has run length statistics consistent with that of a 93-dimensional uniform sample, class 3 does not. Indeed, the snake on class 3 has one very long run of large segments at the beginning (indicating an area of low density), then a long run of short segments (indicating a region of high density).

4. Summary and conclusions

We have compared many existing tests of multivariate uniformity based on the nearest neighbor function $G(r)$, Ripley's K function $K(r)$, discrepancies of the point set (A_n, T_n, UD_2, WD_2), the distance to boundary method (DBM), comparisons of uniform samples in the data's convex hull (Convex) or of ‘high’-density and ‘low’-density samples (Resampling), and multivariate spacings (Spacings), and have introduced two new tests which use the length of the minimum spanning tree (MST) and run length statistics of a snake (Snake). We have estimated the empirical type I error rates and powers in the hypercube and in other convex and non-convex supports for each method when possible and have shown that Convex and Resampling do not provide acceptable tests for $d > 2$. Further, no one test statistic is consistently the most powerful for every alternative distribution or every support.

The effectiveness of each method depends highly on the shape of the support and the dimensionality of the data. While summary functions of the point set and discrepancies can be easily computed in any dimension, $G(r)$ and $K(r)$ essentially require the support to be a hyper-rectangle while the discrepancies essentially require the support to be a hypercube. Spacings requires $d = 2$ since the procedure for their computation has not been worked out for $d > 2$. Further, the distance to boundary and MST-based methods require a known support. More general methods such as Snake or versions of Spacings and DBM that estimate the support are typically much less powerful than the aforementioned test statistics; the price one must pay for increased flexibility.

The hypercube is an important support because of its connection to general multivariate goodness of fit tests and applicability to many datasets in spatial statistics. For simulated data, we find the discrepancies A_n and T_n to typically be the most powerful statistics against general non-uniform alternatives (beta, meta-uniform, contaminated uniforms, clustering). However, tests based on the minimum spanning tree are the most effective for anti-clustered or regular alternatives. In most instances, the powers of each of these statistics increase with both n and d .

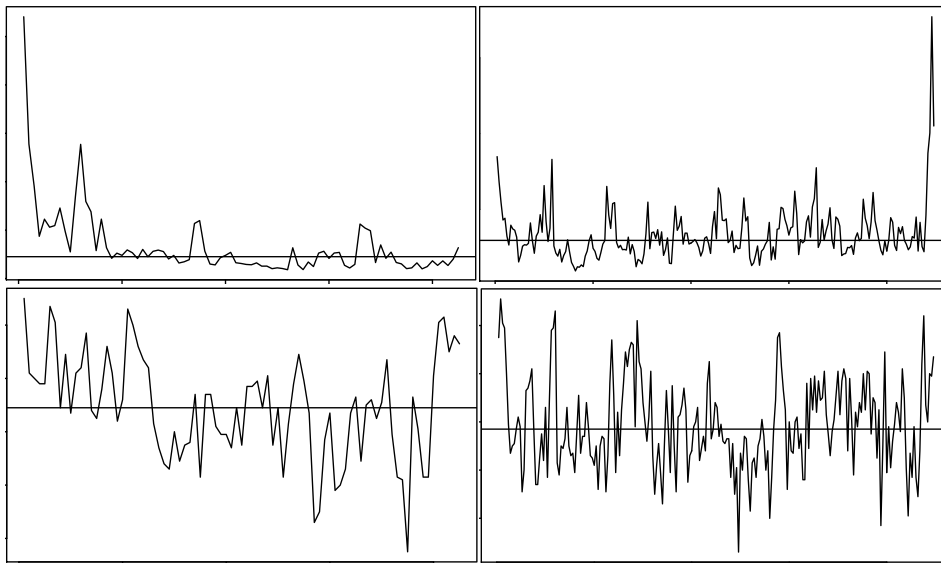


Fig. 8. Left: Snake for class 3 in the LRS dataset (top) along with an example snake on truly uniform 93-dimensional data (bottom). Right: Snake for class 4 in the LRS dataset (top) along with an example snake on truly uniform 93-dimensional data (bottom). The null hypothesis of uniformity is rejected for class 3, and the snake reveals that this is because of the long run of large segments at the beginning (region of low density) and due to a long run of short segments (cluster) in the middle. The null hypothesis of uniformity is retained for class 4, and except for the long segment at the end of the snake the appearance looks very much like what one finds on uniform data.

A more interesting case is when the support of the data is arbitrary and not known in advance. Here, Snake and the generalized Spacings and DBM methods are the only applicable tests. For Snake, we approximate the sample distributions of the run length statistics on the unknown support by their null distributions in the hypercube. Our simulations suggest that the sampling distributions are quite robust to changes in support and the type I error rates on non-hypercubical supports change very little when using this approximation. The generalized Spacings and DBM use the λ -convex hull of the point set to approximate the support, though currently this calculation can only be performed for $d = 2$. When S is a square or circle and n is large enough so that the support is well-estimated by the λ -convex hull, the DBM method has higher power than Snake or Spacings. However, the utility of DBM declines for supports that differ substantially from the square: when S is a cross or crescent, the type I error rate for DBM is nearly twice that desired. We recommend using the generalized distance to boundary method if the λ -convex hull resembles a square or circle, otherwise we recommend Snake since it is consistently more powerful than Spacings.

On real data from the *spatstat* library, we find the tests introduced in this paper are quite useful. The MST detects the anti-clustering present in the *swedishspines* dataset where the rest of the tests (except the nearest neighbor function) retain the null hypothesis of uniformity. Snake easily detects nonuniformities in the *lansing oaks* dataset. Although all tests but Spacings reject uniformity, Snake adds the additional information of why uniformity was failed: a prominent cluster of points and an extended region of unusually low data density. For the *urkiola* dataset, whose support is non-convex and is treated as unknown, only Snake detects that the distribution of trees is non-uniform and indicates that there is a specific cluster of trees that is highly unlikely to have been produced by uniform data. Finally, on the Low Resolution Spectroscopy dataset (which has unknown support with $d = 93$), Snake reveals that collection of fluxes from three of the four types of objects look indistinguishable from uniform data in some high-dimensional space, while one of the objects has clear subregions of high and low density.

Indeed, one property of the snake that is unique among all these tests is its ability to gauge what kinds of nonuniformities are present when the null hypothesis is rejected. By plotting the sequence of segment lengths as a time series, a cluster is represented as a long run of short (relative to the median) segment, and a region of low density is represented as a long run of long segments.

In conclusion, for data in the hypercube we recommend that Discrepancies should be used unless one wants to test for a specific alternative, e.g., anticlustering (use the MST) or special types of contamination (use the DBM). For datasets with unknown support, Snake should be used unless it is believed that it looks like the support resembles the square or circle and is well estimated by the λ -convex hull, in which case we recommend using the distances-to-boundary technique. Snake is the only general method that is appropriate for $d \geq 3$.

Since the potential applicability of Snake has been shown empirically, an obvious future direction would be to work out if any theory can be fleshed out. However, details about shortest paths through random points are sparse and difficult from an analytical point of view (see Beardwood et al. (1959) for a discussion on the asymptotic length of a traveling salesman tour), so we are uncertain if an attempt would be successful. Further, since the short paths that we use are not necessarily optimal paths, a theory that could be worked out may not be directly applicable.

Acknowledgments

The authors would like to thank the three anonymous reviewers for their valuable comments regarding the inclusion of real data and the overall framework and tone of the paper.

References

- Avram, F., Bertsimas, D., 1992. The minimum spanning tree constant in geometrical probability and under the independent model: a unified approach. *The Annals of Applied Probability* 2, 113–130.
- Baddeley, A., Gregori, P., Mateu, J., Stoica, R., Stoyan, D., 2006. *Case Studies in Spatial Point Process Modeling*. In: *Lecture Notes in Statistics*, vol. 185. Springer, New York.
- Baddeley, A., Moyeed, R., Howard, C., Boyde, A., 1993. Analysis of a three-dimensional point pattern with replication. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 42 (4), 641–668.
- Baddeley, A., Turner, R., 2000. Practical maximum pseudolikelihood for spatial point patterns. *Australian and New Zealand Journal of Statistics* 42, 283–322.
- Barber, C., Dobkin, D., Huhdanpaa, H., 1996. The Quickhull algorithm for convex hulls. *ACM Transactions on Mathematical Software* 22 (4), 469–483.
- Barrow, J., Bhavsar, S., Sonoda, D., 1985. Minimal spanning trees, filaments and galaxy clustering. *Royal Astronomical Society, Monthly Notices* 216, 17–35.
- Beardwood, J., Halton, J., Hammersley, J., 1959. The shortest path through many points. *Proceedings of the Cambridge Philosophical Society* 55, 299–327.
- Berrendero, J., Cuevas, A., Pateiro-Lopez, B., 2012a. A multivariate uniformity test for the case of unknown support. *Statistics and Computing* 22, 259–271.
- Berrendero, J., Cuevas, A., Pateiro-Lopez, B., 2012b. Testing uniformity for the case of a planar unknown support. *The Canadian Journal of Statistics* 40, 378–395.
- Berrendero, J., Cuevas, A., Vasquez, F., 2006. Testing multivariate uniformity: the distance-to-boundary method. *Canadian Journal of Statistics* 34, 693–707.
- Cuevas, A., Rodriguez-Casal, A., 2004. On boundary estimation. *Advances in Applied Probability* 36 (2), 340–354.
- Diggle, P., 2003. *Statistical Analysis of Spatial Point Patterns*, second ed. Arnold/Hodder Headline Group, London.
- Friedman, J., 1987. Exploratory projection pursuit. *Journal of the American Statistical Association* 397, 249–266.
- Friedman, J., Tukey, J., 1974. A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers* C23 (9), 881–890.
- Ho, L., Chiu, S., 2007. Testing uniformity of a spatial point pattern. *Journal of Computational and Graphical Statistics* 16, 378–398.
- Hoffman, R., Jain, A., 1983. A test of randomness based on the minimal spanning tree. *Pattern Recognition Letters* 175–180.
- Illian, J., Penttinen, A., Stoyan, H., Stoyan, D., 2008. *Statistical Analysis and Modelling of Spatial Point Patterns*. Wiley, Chichester.
- Jafari-Mamaghani, M., Andersson, M., Krieger, P., 2010. Spatial point pattern analysis of neurons using Ripley's k -function in 3d. *Frontiers in Neuroinformatics* 4 (9), 1–10.
- Jain, A., Xu, X., Ho, T., Xiao, F., 2002. Uniformity testing using minimal spanning tree. In: *Proceedings of the 16th International Conference on Pattern Recognition*, vol. 4, pp. 281–284.
- Johnson, D., McGeoch, L., 1997. The traveling salesman problem: a case study in local optimization. In: Aarts, E., Lenstra, J. (Eds.), *Local Search in Combinatorial Optimisation*. John Wiley and Sons Ltd., pp. 215–310.
- Justel, A., Pena, D., Zamar, R., 1997. A multivariate Kolmogorov–Smirnov test of goodness of fit. *Statistics & Probability Letters* 35 (3), 251–259.
- Lee, S., 1999. The central limit theorem for euclidean minimal spanning trees ii. *Advances in Applied Probability* 31, 969–984.
- Levina, E., Bickel, P., 2005. Maximum likelihood estimation of intrinsic dimension. In: *Advances in Neural Information Processing Systems*, vol. 17. MIT Press, pp. 777–784.
- Liang, J.-J., Fang, K.-T., H. F., Li, R., 2001. Testing multivariate uniformity and its applications. *Mathematics of Computation* 70, 337–355.
- Mardia, K., Jupp, P., 2000. *Directional Statistics*. Wiley, Chichester.
- Marsaglia, G., 1968. Random numbers fall mainly in the planes. *Proceedings of the National Academy of Sciences* 61 (1), 25–28.
- Petrie, A., 2007. *Spanning trees as tools for data analysis*. Ph.D. Thesis, Rensselaer Polytechnic Institute.
- Petrie, A., Willemain, T., 2010. The snake for visualizing and clustering multivariate data. *Statistical Analysis and Data Mining* 3 (4), 236–252.
- Ripley, B., 2004. *Spatial Statistics*. Wiley-Interscience.
- Smith, S., Jain, A., 1984. Testing for uniformity in multidimensional data. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6, 73–81.
- Steele, J., 1988. Growth rates of Euclidean minimal spanning trees with power-weighted edges. *The Annals of Probability* 16, 1767–1787.
- Szekely, G., Rizzo, M., 2004. Testing for equal distributions in high dimension. *Interstat* 5. URL <http://interstat.statjournals.net/YEAR/2004/abstracts/0411005.php>.
- Tenreiro, C., 2007. On the finite sample behavior of fixed bandwidth Bickel–Rosenblatt test for univariate and multivariate uniformity. *Communications in Statistics Simulation and Computation* 36, 827–846.
- Theodoridis, Y., Stefanakis, E., Sellis, T., 1998. Cost models for join queries in spatial databases. In: *Proceedings of IEEE ICDE Conference*, pp. 476–483.
- Wiegand, T., Moloney, K., 2004. Rings, circles, and null-models for point pattern analysis in ecology. *OIKOS* 104 (2), 209–229.
- Zimmerman, D., 1993. A bivariate Cramer–von Mises type of test for spatial randomness. *Applied Statistics* 42 (1), 43–54.