

# Distribution-free Comparison of Multiple Spatial Point Patterns

Andrew A. Merton      Jennifer A. Hoeting      Colleen T. Webb\*

August 20, 2008

## Abstract

We develop a procedure to determine whether there are significant differences between two or more realizations where (spatial) location is one of the response measures. We generalize the multi-response permutation procedure (MRPP) to incorporate georeferenced data for both Euclidean and non-Euclidean spaces and designate the resultant procedure as the spatial multi-response permutation procedure (SMRPP). The SMRPP inherits all of the benefits of the MRPP test: the SMRPP does not require distributional assumptions for the observed data, it is invariant with respect to the spatial domain, and the analysis space coincides with the data space. Furthermore the test can be applied to non-Euclidean geometric spaces, e.g., the surface of the globe and stream networks. The utility of the test is illustrated through simulation as well as recovery data for Northern Pintails (*Anas acuta*) across the contiguous United States for the years 2000 through 2002.

KEYWORDS: Spatially distributed observations, permutation testing, Northern Pintails (*Anas acuta*)

## 1 Introduction

In this article we propose a method to compare spatial patterns of one or more response variables. There are a number of advantages of our proposed test over previous methods. First, the proposed method is invariant with respect to the spatial domain and thus not dependent upon some arbitrarily assigned origin. Second, the method is applicable to any spatial domain (e.g., Euclidean or non-Euclidean), so long as the distance measure within the space is a metric. Third, our method does not require any estimation or modeling of

---

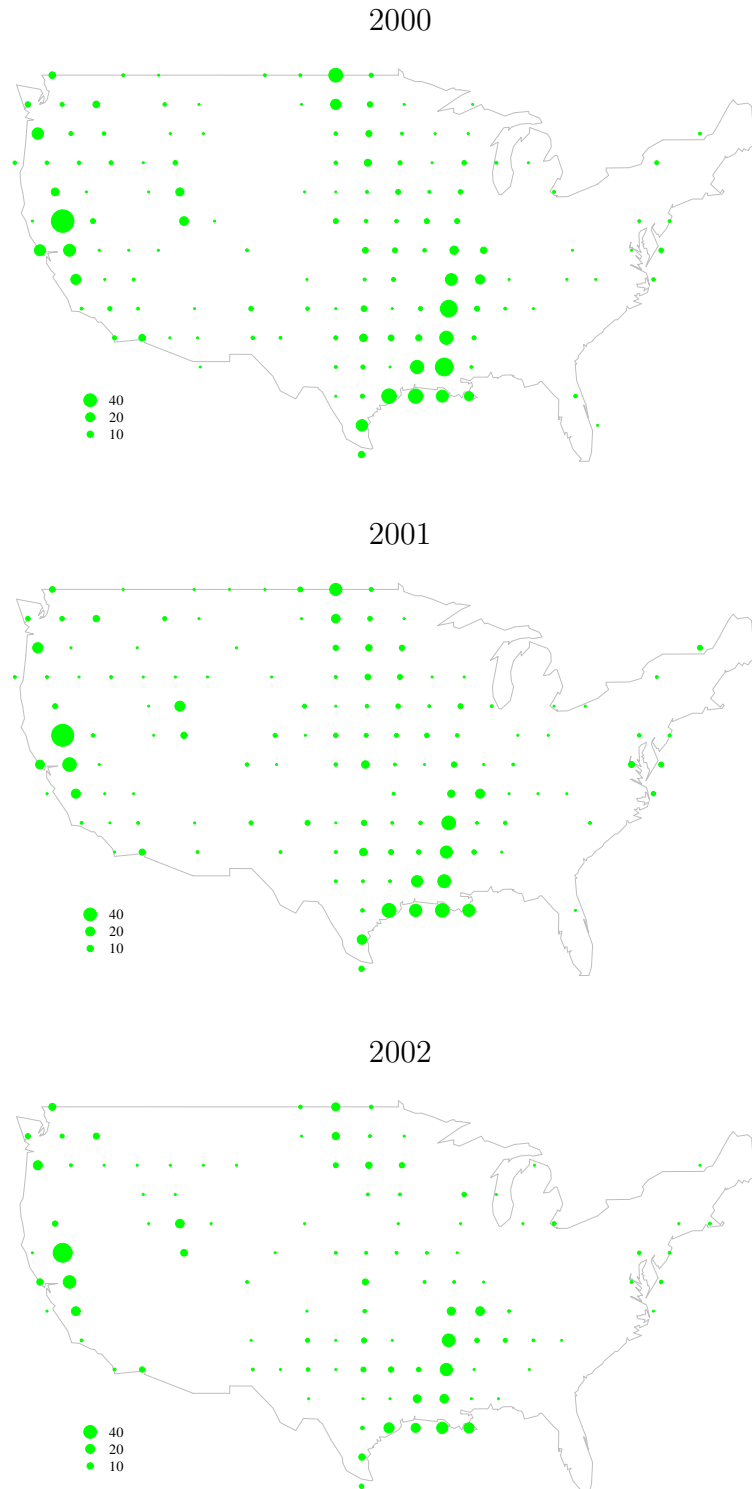
\*A. A. Merton is a Post-doctoral Fellow, Department of Statistics (merton@stat.colostate.edu) J. A. Hoeting is Associate Professor, Department of Statistics (jah@lamar.colostate.edu). C. T. Webb is Assistant Professor, Departments of Biology and Mathematics (ctwebb@lamar.colostate.edu). Colorado State University, Fort Collins, CO 80523.

the spatial correlation. Finally, the method can be applied to multivariate response data. The method generalizes the multi-response permutation procedure (MRPP) as presented by Mielke and Berry (2007) to any non-Euclidean spatial domain (metric space); thus the method is not dependent upon any distributional assumptions.

We motivate the problem by considering the spatial distribution of recovery data for the Northern Pintail (*Anas acuta*), a type of duck, across the contiguous United States. Each year a number of federal and state entities band birds throughout Alaska and other northern latitudes in the spring months after breeding, recording the location, date, sex, and age class. The data are forwarded to the Bird Banding Laboratory (BBL) in Pawtuxent, Maryland where a large, continuously updated catalog of the banded individuals is maintained. Throughout the year as banded individuals are recovered, typically by hunters in the fall and winter months, the recovery location and dates are reported to the BBL. For example, Figure 1 plots the recovery locations for Northern Pintails aggregated onto an approximately regular lattice for the calendar years 2000–2002. Over the past 70 years, the BBL database has grown to include several million records consisting of tens to hundreds of thousands of records for many bird species including the Northern Pintail. This database provides ample opportunity to explore hypotheses regarding waterfowl migration patterns, hunting pressures, etc. Our initial goal is to determine whether the spatial distribution of recoveries has changed across years. Ultimately, we may want to infer why observed patterns differ. For example, there may be fluctuations in weather such as drought or severe cold, changes in hunting pressure due to changes in regulation or participation, changes in the number and location of banding, disease within the waterfowl population, etc. However, prior to any modeling endeavor to investigate any of these factors it is important to identify whether there are significant differences between observed spatial patterns across time.

Spatial data analysis often involves 1) characterizing the observed realization and 2) fitting a model (or series of models). There are several diagnostics for describing the underlying structure of spatially distributed data such as indices of non-randomness for quadrat sampling (e.g., the index of dispersion), Pielou’s index of non-randomness for distance sampling (Pielou (1977)), Moran’s  $I$  and Geary’s  $C$  statistics for spatial point patterns (Schabenberger and Gotway, 2005, p. 22-22), and Ripley’s  $K$ -function for mapped data sets (Schabenberger

Figure 1: Location and number of Northern Pintail band recoveries in 2000–2002 across the contiguous United States (1240, 1040, 716 birds recovered in 2000–2002). The area of the circles is proportional to the number of individuals recovered within an (approximate) 40,000 square km region. Ocean locations are due to lattice configuration.



and Gotway, 2005, p. 101). These indices share the common benchmark of comparing the realized data against the properties of a completely spatially random (CSR) distribution. If the evidence suggests that the data are not CSR, these measures classify the extent to which the data are clustered or regularly distributed and/or estimate the strength of the spatial correlation. Figure 1 is a clear example of the clustering nature associated with the recovery data for Northern Pintails; many more recoveries occur along the southern Mississippi River and central valley of the West Coast. Classification is followed by the selection of one or more appropriate models, e.g., Strauss or Cox processes for spatial point patterns or generating interpolating surfaces (kriging) for continuous response variables, and model parameters are estimated, e.g., using maximum-likelihood.

In contrast, our goal is to directly compare two or more spatial point patterns; we are currently not interested in identifying (or modeling) the underlying process that generated the realizations, but rather in whether or not the observed realizations differ significantly. If the evidence suggests that there is a difference, then one would proceed to attempt to identify how the distributions differ. If the evidence is insufficient to detect a difference across the distributions being compared, then the researcher is free to either i) treat each realization as an independent sample from the underlying process or ii) aggregate (collapse) the data into a single realization. Examples of this type abound in the literature; for example, Lower and Armstrong (2005) examine historical data of abundance for female red king crabs (*Paralithodes camtschaticus*) in Bristol Bay, Alaska. The distributions of broodstock are compared (pairwise) across the years 1975 through 2001 identifying the years where the population changed. Other examples compare juvenile and adult crab spatial distributions (Nielson et al., 2007) and agricultural weed distributions over time (Heijting et al., 2007).

The three studies cited above use a procedure developed by Syrjala (1996) who proposes quantifying the “distance” between the (spatial) empirical distribution functions (EDF). The procedure, which is an extension of the work presented by Zimmerman (1993) for testing for CSR using a modified Cramér-von Mises statistic, uses a Monte Carlo resampling (permutation) technique to generate numerous samples from the original observed realizations against which the original data are compared. In brief, the EDFs are constructed using the original (unpermuted) data and the test statistic, say  $T_0$ , is set equal to the integrated squared

distance between the EDFs over the domain of study. The procedure is repeated  $K$  times using permutations of the data and  $T_k$  computed for  $k = \{1, \dots, K\}$  for each; an “unusual” value of  $T_0$  compared with the collection of  $T_k$ ’s is evidence that the two distributions differ. Syrjala (1996) illustrates the procedure using Pacific cod data off the coast of Alaska, first by comparing the distributions by gender (no significant difference) and second by comparing the distributions of juveniles against adults (significantly different).

A deficiency associated with the test proposed by Syrjala (1996) is that the test is not invariant; the choice of location for the “origin” for computing the EDFs is arbitrary and can change the outcome of the test. Zimmerman (1993) originally proposed an ad hoc solution that averages the test statistic using several different “origins.” For example, if the domain of interest is rectilinear in two-dimensions, one defines four origins, one at each “corner.” This solution, although viable, can become cumbersome for the three-dimensional rectilinear domain. Furthermore, this solution is still susceptible to chance rejections of the null hypothesis (or failure to reject) due to “unlucky” placement of the origin(s). A secondary deficiency is the limitation of the procedure to rectilinear domains (in one-, two-, or three spatial dimensions). This constraint can be problematic simply because the domain of interest may not be well represented by a rectilinear domain; for example the “shape” of the fjords in the study presented by Nielson et al. (2007) are very irregular.

To account for the aforementioned deficiencies we propose a new procedure, the spatial multi-response permutation procedure (SMRPP). We recognize that a MRPP-like test would be invariant with respect to the origin, and we extend the MRPP for use with many types of spatial location data. One of the great strengths of the SMRPP test is that no distributional assumptions are required. The original data are compared to permutations of the data across realizations (e.g., across time periods) with minimal (and sensible) constraints. If the observed data are unusual with respect to the distribution of the permuted realizations, then the null hypothesis is rejected in favor of the alternative, i.e., that the observed realizations differ significantly. The extension to non-Euclidian space is relevant to many ecological scenarios.

Below we describe the MRPP and then generalize the procedure to incorporate geo-referenced data. In Section 3 we illustrate the spatial multiresponse permutation procedure

(SMRPP) through simulation for selected spatially discrete and continuous distributions, evaluating the probability of making a Type I error and approximating power curves as a function of sampling effort. Section 4 illustrates the SMRPP by using the recovery data for Northern Pintails across the contiguous United States for the years 2000 through 2002. We discuss our findings in Section 5.

## 2 Methodology

In Section 2.1, we review the MRPP, closely following the development of the MRPP given in Mielke and Berry (2007, pp. 84-85). In Section 2.2 we propose the spatial multi-response permutation procedure (SMRPP).

### 2.1 Multi-response permutation procedure

The MRPP was originally designed as a non-parametric alternative to a one-way analysis of variance (ANOVA) for univariate data and MANOVA for multivariate data (Mielke and Berry, 1976). The null hypothesis in the MRPP is that the treatment groups are equivalent in some sense. Numerous extensions have been developed (See Cade and Richards, 2005, for a review).

Let  $\Omega = \{\omega_1, \dots, \omega_N\}$  denote a finite collection of  $N$  objects and let  $\mathbf{y}_j = [y_{j1}, \dots, y_{jR}]'$  denote an  $R$ -vector of response measurements for object  $\omega_j$ ,  $j = 1, \dots, N$ . Let  $\Delta : \Omega \times \Omega \rightarrow \mathbb{R}$  be a metric on  $\Omega$  defined to have the properties

1.  $\Delta(\omega_a, \omega_b) \geq 0$  for all  $\omega_a, \omega_b \in \Omega$  and  $\Delta(\omega_a, \omega_b) = 0$  if and only if  $\omega_a = \omega_b$ .
2.  $\Delta(\omega_a, \omega_b) = \Delta(\omega_b, \omega_a)$  for all  $\omega_a, \omega_b \in \Omega$ .
3.  $\Delta(\omega_a, \omega_c) \leq \Delta(\omega_a, \omega_b) + \Delta(\omega_b, \omega_c)$  for all  $\omega_a, \omega_b, \omega_c \in \Omega$ .

For the subsequent presentation we restrict attention to the Minkowski family of distance metrics, i.e.,

$$\Delta_{j,k} = \left( \sum_{r=1}^R |y_{jr} - y_{kr}|^p \right)^{1/p},$$

where  $p \geq 1$ . When  $p = 1$  the metric is referred to as the city-block or Manhattan distance function. The Euclidean distance metric corresponds to  $p = 2$ .

Let  $S_1, \dots, S_I$  denote an exhaustive partitioning of the  $N$  objects of  $\Omega$  into  $I$  disjoint categories where  $n_i \geq 2$  is the number of objects in each category  $S_i$  ( $i = 1, \dots, I$ ) and  $\sum_{i=1}^I n_i = N$ . Let  $\xi_i$  be the mean distance between the response vectors for category  $i \in \{1, \dots, I\}$  where

$$\xi_i = \binom{n_i}{2}^{-1} \sum_{j < k} \Delta_{j,k} \mathbf{1}_i(\omega_j) \mathbf{1}_i(\omega_k), \quad (1)$$

where the sum is over all  $j$  and  $k$  such that  $1 \leq j < k \leq N$ , and  $\mathbf{1}_i(\omega_j)$  is the indicator function equal to 1 when  $\omega_j \in S_i$ . The MRPP statistic is the average within-category difference, weighted by the number of objects  $n_i$  in category  $S_i$  ( $i = 1, \dots, I$ ), given by

$$\delta = \sum_{i=1}^I C_i \xi_i$$

where  $C_i = n_i/N$ ,  $\sum_{i=1}^I C_i = 1$ , and  $i = 1, \dots, I$ .

The null hypothesis ( $H_0$ ) assumes that there are no differences across the  $I$  categories. Testing proceeds by first evaluating the MRPP statistic using the observed data and denoting the value by, say,  $\delta_0$ . Next, the statistic is evaluated for every possible permutation of the category labels. For  $M$  possible permutations the exact p-value equals the number of  $\delta$ 's that are less than or equal to  $\delta_0$ , i.e.,

$$P(\delta \leq \delta_0 | H_0) = \frac{1}{M} \sum_{k=1}^M \mathbf{1}(\delta_k \leq \delta_0),$$

where  $\delta_k$  corresponds to the  $k$ th evaluation of the MRPP statistic for  $k = 1, \dots, M$ . Typically the number of possible permutations  $M$  is prohibitively large,  $\mathcal{O}(I^N)$ . Hence one approximates the p-value by evaluating  $\delta$  for a sufficiently large number of Monte Carlo samples  $K$  where  $K \ll M$ . The p-value is then approximated by

$$\hat{P}(\delta \leq \delta_0 | H_0) = \frac{1}{K+1} \sum_{k=0}^K \mathbf{1}(\delta_k \leq \delta_0).$$

When the data include multiple responses, i.e.,  $R \geq 2$ , the response measurements are typically expressed in different units which must be commensurated (standardized) prior to analysis. Recall that for each object  $\omega_j \in \Omega$  the (non-commensurate) response vector is

denoted  $\mathbf{y}_j = [y_{j1}, \dots, y_{jR}]'$ . Define the commensurate response vector  $\mathbf{x}_j = [x_{j1}, \dots, x_{jR}]'$  where  $x_{jr} = y_{jr}/\phi_r$  and

$$\phi_r = \left[ \sum_{k<l}^N |y_{kr} - y_{lr}|^\nu \right]^{1/\nu} \quad (2)$$

for  $r = 1, \dots, R$ . Note that the commensurate data have the property that

$$\sum_{k<l}^N |x_{kr} - x_{lr}|^\nu = 1$$

for  $r = 1, \dots, R$  and any  $\nu > 0$ . The commensuration is termed Euclidean commensuration when  $\nu = 1$ . Mielke and Berry (2007) describe other commensuration procedures, e.g., Hotelling commensuration, and we defer to that text for further details. First make the data commensurate, then proceed as described above substituting  $\mathbf{y}_j$  with  $\mathbf{x}_j$  within the Minkowski distance function.

## 2.2 Spatial multi-response permutation procedure

We now extend the MRPP statistic to incorporate non-Euclidean spatial location data. Begin by assuming that associated with every object  $\omega_i \in \Omega$  for  $i = 1, \dots, N$  is a (random) geo-referenced location, say  $\mathbf{s}_i$ , such that  $\mathbf{s}_i \in \mathcal{D}$  where  $\mathcal{D}$  is a fixed, finite space. For convenience we set  $y_{i1} = \mathbf{s}_i$ , i.e., set the first entry of the response vector to the spatial location. Note that location is often a vector quantity; thus the response vector  $\mathbf{y}$  must be generalized to be a data structure (or list). Examples of vector locations include  $\mathbf{s}_i = (x_i, y_i, z_i)$  in  $\mathbb{R}^3$  Euclidean space and  $\mathbf{s}_i = (\text{latitude}, \text{longitude})$  for the surface of a sphere. The spatial multi-response permutation procedure (SMRPP) statistic is written  $\Delta_{j,k} = d(\mathbf{s}_j, \mathbf{s}_k)$  when  $R = 1$  and when  $R \geq 2$ ,

$$\Delta_{j,k} = \left( d(\mathbf{s}_j, \mathbf{s}_k)^p + \sum_{r=2}^R |y_{jr} - y_{kr}|^p \right)^{1/p}, \quad (3)$$

where  $d(\cdot, \cdot)$  is a distance measure satisfying the three properties listed in Section 2.1. For  $R > 1$  commensuration of the data proceeds as discussed above (see equation 2) with the notable exception that since  $y_{i1} = \mathbf{s}_i$  for  $i = 1, \dots, N$ , the scaling coefficient  $\phi_1$  is set to the sum of the pairwise distances, i.e.,

$$\phi_1 = \sum_{j<k}^N d(\mathbf{s}_j, \mathbf{s}_k).$$

The SMRPP statistic (3) allows for testing to proceed within non-Euclidean spatial domains. For example, distance along the surface of the Earth is often measured using the great circle distance. By approximating the Earth as a perfect sphere, location is specified using two coordinates, e.g., latitude ( $\delta \in [-90^\circ, +90^\circ]$ ) and longitude ( $\lambda \in [-180^\circ, +180^\circ]$ ). The shortest distance between any two points on a sphere,  $\mathbf{s}_i = (\delta_i, \lambda_i)$  and  $\mathbf{s}_j = (\delta_j, \lambda_j)$ , is a segment of the great circle connecting the points and is computed using the trigonometric relation

$$d(\mathbf{s}_i, \mathbf{s}_j) = a \cos^{-1} [\cos(\delta_i) \cos(\delta_j) \cos(\lambda_i - \lambda_j) + \sin(\delta_i) \sin(\delta_j)], \quad (4)$$

where  $a \approx 6378$  km is the approximate radius of the Earth (Weisstein, 2008). Note that the great circle distance function satisfies the metric properties 1-3 listed above. A second example is distance as measured within a network of connected streams and/or rivers. If “movement” is restricted to the waterway (read “as the fish swims”) then the corresponding spatial domain is of fractional dimension, see for example Tarboton et al. (1988); however, the distance measure satisfies the properties of a metric. For Euclidean spatial domains, the distance measure in (3) can be set to, say, the  $p$ -norm, i.e.,  $d(\mathbf{s}_j, \mathbf{s}_k) = \left( \sum_{i=1}^d |s_{j_i} - s_{k_i}|^p \right)^{1/p}$ .

By treating location as a single entry in the response data structure, the SMRPP test statistic (3) can be applied to a larger class of problems than the MRPP statistic. First, consider the situation where location is the only response. If the spatial domain is Euclidean such that  $\mathcal{D} \subset \mathbb{R}^d$  where  $d \in \{1, 2, \dots\}$  and the data are not commensurate, then the MRPP and the SMRPP test statistics will exactly coincide if  $d(\mathbf{s}_j, \mathbf{s}_k) = \Delta_{j,k}$ ;  $R = 1$  for the SMRPP test and  $R = d$  for the MRPP test. Berry et al. (1983) illustrate precisely this scenario using the MRPP to detect differences in the intra-site patterns of artifact distributions in an archaeological dig. The domain of interest is a  $9 \times 7$  meter rectilinear space in which 226 artifacts each belonging to one of three classes are distributed. The location of each artifact was assigned to the nearest node of a  $9 \times 7$  regular lattice superimposed on the domain and the MRPP performed with  $R = 2$ . We repeated the analysis using the SMRPP with  $R = 1$  and obtained the exact same solution. Now consider the case where the spatial domain is along the surface of a sphere. For the MRPP test, the location vectors are separated into the scalar components  $\delta$  and  $\lambda$ . Since “movement” is constrained to the surface of the sphere, a change of 10 degrees in longitude, i.e.,  $\Delta\lambda = 10^\circ$ , at the equator is much larger than

the same change in longitude at a latitude of  $45^\circ$ . Consequently, the Minkowski family of distance measures is not appropriate for spherical coordinates. In contrast, the SMRPP can accommodate non-Euclidean spatial domains so long as the distance measure is a metric with respect to the domain, i.e., a metric space.

We recommend restricting the distance measure  $\Delta$  in (3) to be a metric; this forces the analysis space to coincide with the data space. Mielke and Berry (2007) illustrate counter-intuitive results when a non-metric measure for  $\Delta$  such as squared distance (which does not satisfy the triangle inequality) is used. Since we are not imposing any distributional assumptions for testing, it is not required that we analyze the data in a transformed space for, say, ease of computations.

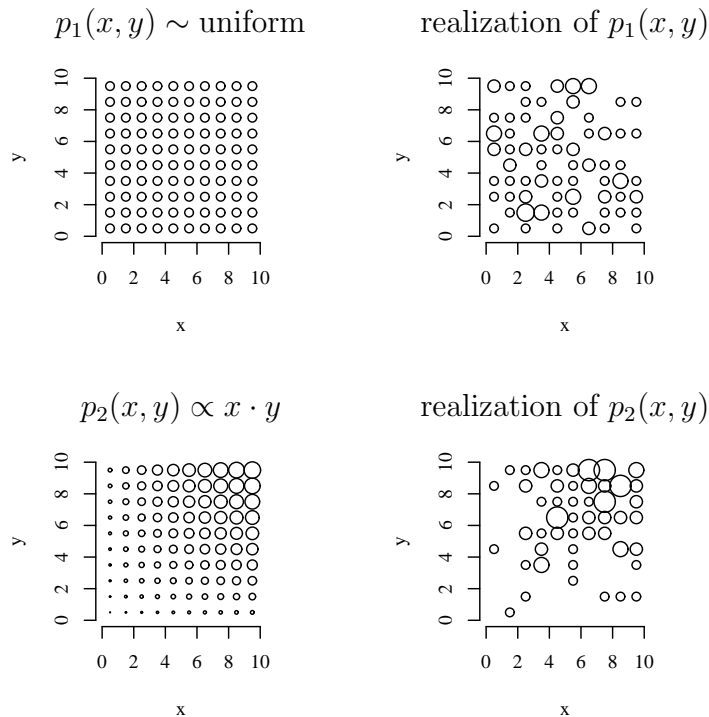
### 3 Simulation Studies

We performed a number of simulation studies to quantify the performance of the SMRPP test statistic. The first series of simulations generated count data at locations on a fixed, regular lattice in  $\mathbb{R}^2$ ; the second series generated count data on a regular lattice superimposed onto a sphere. Both simulation studies have a single response (location), i.e.,  $R = 1$ , such that  $\mathbf{s}_i \in \mathcal{D}$  for  $i = 1, \dots, N$ .

#### 3.1 Spatially distributed count data

For the first set of simulations we generate count data along a two-dimensional regular lattice using three different distributions. We began by constructing a regular  $10 \times 10$  lattice over the domain  $\mathcal{D} = [0, 10] \times [0, 10]$  such that  $\mathbf{s}_i = (x_i, y_i)$  and  $x_i, y_i \in \{0.5, 1.5, \dots, 9.5\}$ . Next we assumed that  $n$  independent observations were obtained such that each observation was geo-referenced to a location  $\mathbf{s}_i \in \mathcal{D}$  for  $i = \{1, 2, \dots, n\}$ . The probability of an observation occurring at location  $\mathbf{s}_i$  is defined by one of two probability mass functions (pmf):  $p_1(x, y) = 100^{-1}$  and  $p_2(x, y) \propto xy$ . The first pmf is a uniform distribution over  $\mathcal{D}$  whereas  $p_2(x, y)$  assigns a larger probability to the upper right-hand corner; Figure 2 shows the pmfs along with a single sample realization for  $n = 100$ . The circle areas for each pmf (left panels) are proportional to the probability of making an observation at that locale (lattice point);

Figure 2: Distributions from which independent count data were simulated; the left panels plot the pmfs and the right panels illustrate a single realization from each distribution for  $n = 100$ .

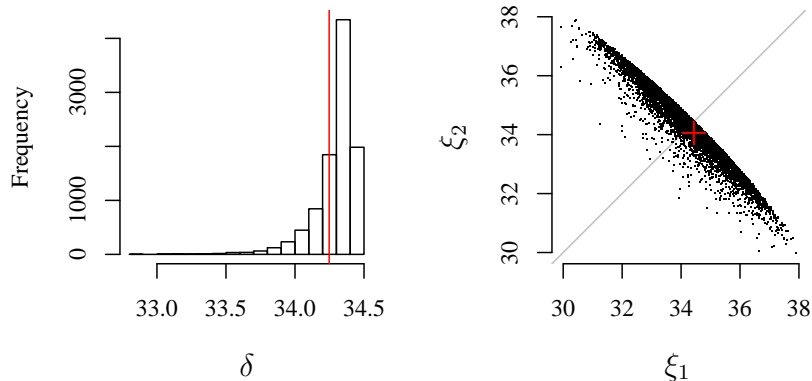


the circles area in the realization plots (right panels) are proportional to the number of observations made at each location.

To evaluate the probability of making a type I error, rejecting the null hypothesis when it is “true”, we performed a set of simulations estimating the p-value for each and recording the (approximate) proportion of time that the null hypothesis was rejected. For each simulation, two realizations of spatial count data were independently drawn from the same distribution e.g.,  $p_1(x, y)$ , and tested using the SMRPP.

Figure 3 plots the results for a single simulation where two independent realizations were independently drawn from  $p_1(x, y)$  with  $n_1 = n_2 = 100$ . The left panel plots the resultant sample distribution of the  $\delta$  statistic using  $K = 9999$  permutations. The estimated p-value for this simulation was 0.2495, i.e.,  $\hat{P}(\delta \leq \delta_0) = 10000^{-1} \sum_{i=0}^{9999} \mathbf{1}(\delta_i \leq \delta_0)$ . The right panel illustrates the joint distribution of the sampling distribution of  $\xi_i = \{\xi_{i_1}, \xi_{i_2}\}$  for the same

Figure 3: Simulation results for count data using the discrete uniform distribution,  $p_1(x, y)$  where  $n_1 = n_2 = 100$ . The left panel plots the distribution of the  $\delta$  statistic for 9999 permutations of the observed data where the red line indicates the value of  $\delta_0$ ; the estimated p-value is 0.2495, i.e.,  $\hat{P}(\delta \leq \delta_0) = 0.2495$ . The right panel plots the joint distribution of  $\xi_i = \{\xi_{1_i}, \xi_{2_i}\}$  for all 9999 permutations where the red “+” indicates  $\xi_0$ .



data where  $\xi_{i_1}$  is the value of equation (1) for realization 1 of the  $i$ th simulation; the black dots plot the sample  $\xi_i$ 's for  $i = \{1, \dots, 9999\}$  and the red “+” plots the observed value  $\xi_0$ . Note that the  $\xi_0$  falls well within the sample data “cloud” indicating that the observed value is not unusual, i.e., the p-value is relatively large. The shape of the data cloud arises from the resampling constraint where  $n_1$  and  $n_2$  must remain constant across permutations. Table 1 summarizes the observed Type I error rate at significance level  $\alpha = 0.05$  for different sample sizes where the total sampling effort was fixed ( $N = n_1 + n_2 = 200$ ).

To evaluate the power of the SMRPP, simulations were performed to measure the probability of rejecting the null hypothesis when comparing realizations generated from different pmfs. Independent realizations were drawn from pmf 1 and 2 with equal sample sizes where  $n = \{5, 10, \dots, 50, 60, \dots, 100\}$  and tested with the SMRPP. The resultant power is given in Figure 4. The SMRPP method has high power to differentiate the two point patterns with fairly small sample sizes, achieving over 90% power with only 30, 35, and 50 observations at a 0.10, 0.05, and 0.01 significance level respectively.

We investigated a number of other scenarios and obtained similar results. For example,

Table 1: Estimated probability of making a Type I error (incorrectly rejecting  $H_0$ ) at significance level  $\alpha = 0.05$  for the two pmfs under study where the total sampling effort was held constant such that  $N = n_1 + n_2 = 200$ . Each simulation was repeated 1000 times with  $K = 999$ .

$(n_1, n_2)$	$p_1$	$p_2$
(50, 150)	0.041	0.052
(75, 125)	0.042	0.036
(100, 100)	0.044	0.056

Figure 4: Empirical power curves for the pairwise comparisons of the simulation pmfs at significance levels 0.10, 0.05, and 0.01.

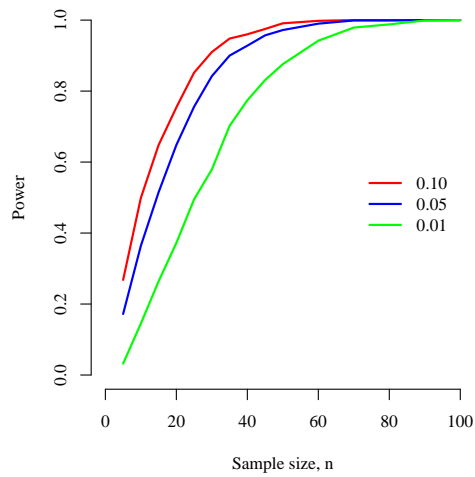
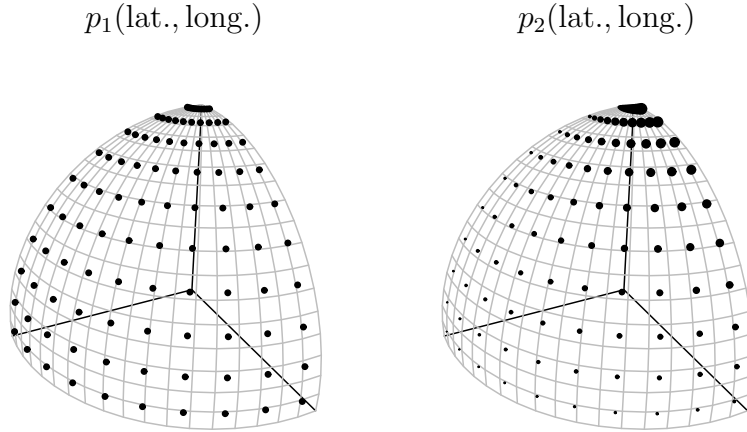


Figure 5: Distributions of the mass along the surface of a unit sphere using the probability mass functions  $p_1$  and  $p_2$ .



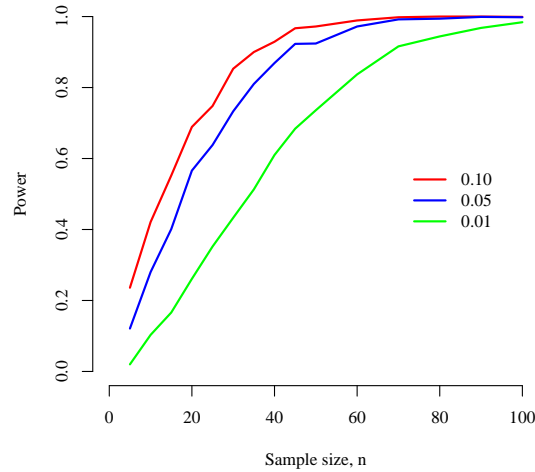
we examined a third pmf  $p_3(x, y) \propto x + y$  and found similar results to Table 1. Since  $p_3$  is more similar to  $p_1$  and  $p_2$ , larger sample sizes were required to differentiate among the pmfs as compared to Figure 4. We also compared point processes for a continuous analog to the results shown above using continuous versions of the discrete densities, e.g.  $f_1(x, y) = 1/100$  and  $f_2(x, y) = (1/2500)xy$ . The results for the continuous versions were very similar to the simulations for the discrete data.

### 3.2 Curvilinear surface

We repeated the simulation study by superimposing the pmfs  $p_1$  and  $p_2$  onto the surface of a unit sphere. Figure 5 illustrates the distribution of the mass for the two pmfs; the area of the circles is proportional to the mass at that location. The point masses are located along the surface such that  $\mathbf{s}_i = (\text{lat.}, \text{long.})$  where  $\text{lat.}, \text{long.} \in \{4.5^\circ, 13.5^\circ, \dots, 85.5^\circ\}$ . The distance between locations  $\mathbf{s}_i$  and  $\mathbf{s}_j$  was measured using the great circle distance (4) which satisfies the properties of a metric. Note that the distance between adjacent locations at a fixed latitude decreases as one moves northward (from the equator,  $\text{lat.} = 0^\circ$ , to the pole,  $\text{lat.} = 90^\circ$ ). Consequently, both  $p_1$  and  $p_2$  assign mass at a higher density (per unit surface area) in the northern latitudes.

The simulation results along the sphere were very similar to the results from the regular

Figure 6: Empirical power curves for the pairwise comparisons of the simulation pmfs superimposed onto the surface of a sphere at significance levels 0.10, 0.05, and 0.01.



lattice presented above. Figure 6 illustrates how power was suppressed for the same sampling effort; to achieve at least 90% power, samples of (approximately) size 35, 45, and 70 observations are required at the 0.10, 0.05, and 0.01 significance levels, respectively. This illustrates the importance of considering the metric space of the data and why the SMRPP is a needed generalization of the MRPP.

## 4 Band Recoveries for Northern Pintails

We applied the SMRPP test to band-recovery data collected by the Bird Band Laboratory (BBL) located in Pawtuxent, MD. Banding and recovery locations are recorded to the nearest 10 minute block in both latitude and longitude. For illustrative purposes we have aggregated the locations to the nearest node of an approximately regular lattice of 200 km per side superimposed across the contiguous United States. Figure 1 shows the (approximate) recovery location and number of banded Northern Pintails for the years 2000 through 2002. Note that although the number of individuals recovered changes across year, the general characteristics of the distributions appear to be invariant. For example, large numbers of recoveries were made in the Central Valley along the West Coast, in the Prairie Pot Hole region (northern North Dakota and Minnesota), and throughout Louisiana along the Mississippi river and Gulf Coast for all three years, whereas few recoveries were made along the Eastern seaboard.

In addition to recovery locations the Northern Pintail data set includes the number of weeks between banding and recovery. Figure 7 plots the marginal distributions of time-to-recovery by year. Note the cyclic nature of the data; recall that a majority of the banding occurs in the spring and summer whereas hunting season typically starts in autumn and continues through February. If a banded individual survives the hunting season, then it is not subject to further hunting pressures until the next season, and hence gaps appear in the distribution for time-to-recovery.

Questions of interest include: are there significant differences in the observed marginal distributions of the recovery locations (“Location”), the time-to-recovery (“Time”), or in the joint distributions? To test for differences, we considered the ensemble of all three years ( $I = 3$ ) and each pairwise combination ( $I = 2$ ). We employed the SMRPP to test differences in the joint distribution and in the marginal distribution of Location. We employed the MRPP to test differences in the marginal distribution of Time. Since the domain of interest, the contiguous United States, is of the continental scale, great circle distance between individuals was approximated by assuming the Earth to be a perfect sphere, i.e.,  $\mathcal{D}$  is curvilinear. The test results are summarized in Table 2: the first column indicates the years being compared, the second and third columns list the p-values associated with the overall tests for both Manhattan ( $p = 1$ ) and Euclidean distance ( $p = 2$ ), the last two columns list the observed p-values for the marginal distributions of recovery locations over time and time-to-recovery respectively.

The main focus is on the results in the top row of the “Location and Time” column which answers the question: are there significant differences in the patterns over location and/or time (Table 2)? The subsequent values are sub-tests that focus on where the differences are observed. The results indicate that there are significant differences over location and time, and that these are mainly driven by differences in time-to-recovery between 2000 and 2002 and somewhat less so by differences in time-to-recovery between 2001 and 2002. Furthermore the results suggest that the joint distributions for location and time-to-recovery are the same for 2000 and 2001. Note that if a Bonferroni correction were used, then we would compare the three single-year comparisons in each column to an  $\alpha$ -level of  $0.05/3 = 0.0167$ . The results indicate that an analysis that is focused only on spatial patterns can be collapsed

Figure 7: Marginal distributions for the total number of weeks elapsed between banding and recovery.

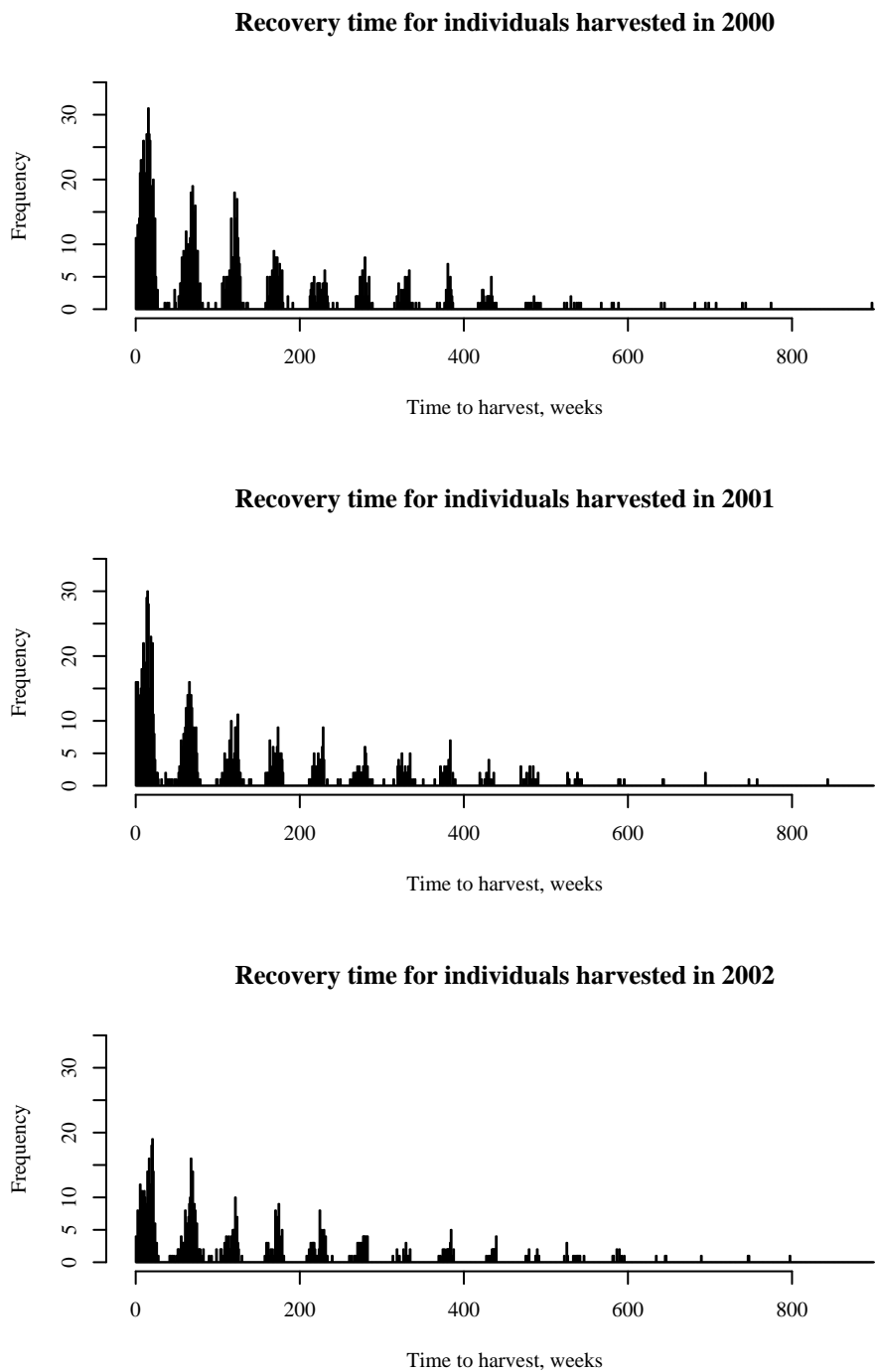


Table 2: Estimated p-value associated with the (S)MRPP testing for the Northern Pintail recovery data. The reported p-values were calculated using  $K = 9999$  permutations.

Years compared	Location and Time		Location	Time
	$p = 1$	$p = 2$		
2000–2002	0.0048	0.0045	0.1674	0.0022
2000, 2001	0.6063	0.6463	0.4295	0.4252
2000, 2002	0.0030	0.0038	0.0745	0.0013
2001, 2002	0.0042	0.0038	0.2053	0.0096
Test	SMRPP		SMRPP	MRPP

over time for these data, making any such analysis simpler. However, no such simplification is appropriate for an analysis of time-to-recovery data. These results also suggest that biologists may want to consider why the time-to-recovery patterns differ over time.

## 5 Discussion

The above examples illustrate the utility and flexibility of the SMRPP test. Treating location as a single response, independent of the space in which location is measured, proves to be a satisfactory method for incorporating location into the MRPP test. The SMRPP test statistic enjoys the same advantages of the MRPP test statistic in that it is distribution-free. Furthermore, the test is invariant; there is no need to define an (arbitrary) “origin” as required by Zimmerman (1993) and Syrjala (1996). The SMRPP allows implementation of the test to non-Euclidean (closed) spaces such as the surface of sphere. This is especially advantageous for large scale ecological, environmental, and climatic studies where data has been collected at the global scale. Lastly, the spatial locations need not be constrained to a regular (or otherwise) lattice.

As with all permutation-based methods, the SMRPP test can be computationally intense and may require careful implementation for very large data sets. However, we implemented SMRPP easily in R (R Development Core Team, 2008) and this has code has been suitable

for all SMRPP analyses we have performed to date.

After a SMRPP test identifies significant differences between realizations, researchers may be interested in determining where the realizations differ. A common method currently being employed to identify such smaller-scale differences is to use scan statistics built upon the work of Kulldorff (1997). The procedure implements a likelihood ratio test by superimposing a moving window on the spatial (or spatial-temporal) distribution to identify the region(s) where the count within the window most greatly exceeds the expected count. Applications of this method include several epidemiologic studies: Christiansen et al. (2006), Jung et al. (2007), Huang et al. (2007), and Cook et al. (2007). The former two articles analyze discrete data identifying spatial clusters of i) *Campylobacter* in broiler flocks across Denmark and ii) prostate cancer grade (and stage) throughout Maryland, respectively. The latter two articles extend the spatial scan statistic to (un)censored survival data, a continuous response measure, for prostate cancer patients in Connecticut and Asthma sufferers in Massachusetts, respectively.

## Acknowledgments

This work is supported by USDA Cooperative Agreement 07-7100-0228. The authors would like to thank the following people for their important contributions to this paper: Paul Mielke for his unending patience in answering our many questions concerning the MRPP; Lance Waller for being a sounding board against which we could test ideas; Paul Doherty for allowing us to tap into his expertise with respect to waterfowl (migration and life cycles); and to Ryan Miller and Matt Farnsworth of the USDA who spent countless hours acquiring, cleaning, and compiling the Northern Pintail data.

## References

Berry, K. J., Kvamme, K. L., and Mielke, Jr., P. W. (1983). Improvements in the permutation test for the spatial analysis of the distribution of artifacts into classes. *American Antiquity*, 48(3):547–553.

- Cade, B. S. and Richards, J. D. (2005). *User Manual for Blossom Statistical Software*. United States Geological Service, Open-File Report 2005-1353.
- Christiansen, L. E., Andersen, J. S., Wegener, H. C., and Madsen, H. (2006). Spatial scan statistics using elliptical methods. *Journal of Agricultural, Biological, and Environmental Statistics*, 11(4):411–424.
- Cook, A. J., Gold, D. R., and Li, Y. (2007). Spatial cluster detection for censored outcome data. *Biometrics*, 63:540–549.
- Heijting, S., van der Werf, W., Stein, A., and Kropff, M. J. (2007). Are weed patches stable in location? Application of an explicitly two-dimensional methodology. *Weed Research*, 47(5):381–395.
- Huang, L., Kulldorff, M., and Gregorio, D. (2007). A spatial scan statistic for survival data. *Biometrics*, 63:109–118.
- Jung, I., Kulldorff, M., and Klassen, A. C. (2007). A spatial scan statistic for ordinal data. *Statistics in Medicine*, 26:1594–1607.
- Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics – Theory and Methodology*, 26(6):1481–1496.
- Lower, T. and Armstrong, D. A. (2005). Historical changes in the abundance and distribution of ovigerous red king crabs (*Paralithodes camtschaticus*) in Bristol Bay (Alaska), and potential relationship with bottom temperature. *Fisheries Oceanography*, 14(4):292–306.
- Mielke, Jr., P. W. and Berry, K. J. (1976). Multi-response permutation procedures for a priori classifications. *Communications in Statistics - Theory and Methods*, 5:1409–1424.
- Mielke, Jr., P. W. and Berry, K. J. (2007). *Permutation Methods: A distance function approach*. Springer, New York, 2nd edition.
- Nielson, J. K., Taggart, S. J., Shirley, T. C., and Mondragon, J. (2007). Spatial distribution of juvenile and adult female tanner crabs (*Chionoecetes bairdi*) in a glacial fjord ecosystem: implications for recruitment processes. *ICES Journal of Marine Science*, 64(9):1772–1784.

- Pielou, E. C. (1977). *Mathematical ecology*. Wiley-Interscience [John Wiley & Sons], New York, second edition.
- R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
- Schabenberger, O. and Gotway, C. A. (2005). *Statistical Methods for Spatial Data Analysis*. Chapman and Hall/CRC Press, Boca Raton, Florida, USA.
- Syrjala, S. E. (1996). A statistical test for a difference between the spatial distributions of two populations. *Ecology*, 77(1):75–80.
- Tarboton, D. G., Bras, R. L., and Rodriguez-Iturbe, I. (1988). The fractal nature of river networks. *Water Resources Research*, 24(8):1317–1322.
- Weisstein, E. W. (2008). Great circle. MathWorld—A Wolfram Web Resource.
- Zimmerman, D. L. (1993). A bivariate Cramer-von Mises type of test for spatial randomness. *Applied Statistics*, 42(1):43–54.