An Efficient Quantile Spatial Scan Statistic for Finding Unusual Regions in Continuous Spatial Data with Covariates

Travis Moore
School of EECS
Oregon State University
Corvallis, OR 97331
moortrav@eecs.oregonstate.edu

Weng-Keen Wong
School of EECS
Oregon State University
Corvallis, OR 97331
wongwe@eecs.oregonstate.edu

Abstract

Domains such as citizen science biodiversity monitoring and real estate sales are producing spatial data with a continuous response and a vector of covariates associated with each spatial data point. A common data analysis task involves finding unusual regions that differ from the surrounding area. Existing techniques compare regions according to the means of their distributions to measure unusualness. Comparing means is not only vulnerable to outliers, but it is also restrictive as an analyst may want to compare other parts of the probability distributions. For instance, an analyst interested in unusual areas for high-end homes would be more interested in the 90th percentile of home sale prices than in the mean. We introduce the Quantile Spatial Scan Statistic (QSSS), which finds unusual regions in spatial data by comparing quantiles of data distributions while accounting for covariates at each data point. We also develop an exact incremental update of the hypothesis test used by the QSSS, which results in a massive speedup over a naive implementation.

1 INTRODUCTION

Analysis of spatial data often involves finding spatial regions that are different from the surrounding area. For example, epidemiologists are interested in finding regions with an unusually high incidence of disease while criminologists are interested in identifying crime hotspots. The spatial scan statistic (SSS) (Kulldorff, 1997) is a widely used technique to discover unusual regions from a Bernoulli or Poisson point process. The SSS searches over a given set of regions, scoring each region according to how a quantity of interest (eg. the disease rate) inside the region differs from outside the region. Finally, the SSS computes the p-value of the highest scoring region using a randomization test.

Many spatial data sets, however, are more complex than point processes, which focus on the spatial locations of the data. Real-world spatial data sets from domains such as citizen science biodiversity monitoring and real estate associate a response value with each point as well as a set of covariates (i.e. features). For example, in a real estate data set, each data point has a location, a sale price, and associated features such as square footage, number of bedrooms, age, etc. Formally, we represent the ith data point of dataset $\boldsymbol{D}$ as a tuple $(Y_i, X_{i,1}, \ldots, X_{i,p}, L_{i,1}, \ldots, L_{i,d})$, where Y_i is a continuous response, $(X_{i,1}, \ldots, X_{i,p})$ are the p covariates and $(L_{i,1}, \ldots, L_{i,d})$ are coordinates in d-dimensions; for simplicity, we assume $d = 2$. In later sections, we will refer to the data as $\boldsymbol{D} = \{\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{L}\}$ to represent the distinct aspects of response, covariates and locations.

We can follow the SSS framework to find unusual regions in this more complex setting. For each region, we fit a model that captures the relationship between the features and the response variable. Then, we use a scoring function to compare the models from the "inside" versus the "outside" regions, by using a hypothesis test that compares the means of the models. While such an approach seems reasonable, there are two shortcomings. First, the approach is not robust as the mean is well known to be vulnerable to outliers and the models for each region can be badly skewed by outliers with extreme values for the response variable (Rousseeuw and Leroy, 1987). Second, many real-world tasks involve comparing spatial regions using other parts of their distributions besides the mean. For instance, a real-estate agent interested in high-end homes may want to compare regions based on the 90th percentile of the sale price distribution. To overcome both of these problems, we develop a novel method for comparing quantiles of spatial regions.

To accomplish our goal of comparing quantiles of spatial regions, we modify the proposed SSS variant by fitting quantile regression models to the "inside" and "outside" regions. Unfortunately, this naive approach is computationally expensive; fitting a quantile regression requires a linear program and this step would be required in the inner loop of the algorithm. To make the algorithm efficient, we replace the likelihood ratio test with the rank test, which is a non-parametric hypothesis test that avoids the need to fit quantile regressions to the "inside" regions. However, performing a rank test from scratch every time we score a new region is also computationally expensive. Instead, we develop an incremental version of the rank test that allows the rank test from a smaller region to be updated when the region is grown to include more spatial data points.

The contributions of our work are as follows. First, we introduce the Quantile Spatial Scan Statistic (QSSS), which discovers unusual regions for continuous spatial data with covariates. The comparison between regions to determine unusualness is based on a comparison of the τ-th quantile of the response variable distributions. To our knowledge, no such version of the SSS currently exists in the literature. This algorithm is also robust to outliers, unlike an analogous algorithm that makes comparisons based on the mean of a region. Second, we show how to make the QSSS over an order of magnitude faster than a naive implementation by introducing an incremental update to the rank test. This update is exact and not an approximation. Finally, we evaluate the QSSS on simulated data and also show interesting results from case studies on three real-world datasets.

2 RELATED WORK AND BACKGROUND

We first discuss related work and then provide some background needed to understand our approach. A large body of work that is seemingly related to our task has focused on producing disease maps that illustrate how disease cases vary across space (eg. (Best et al., 2005)). Researchers have also investigated spatial quantile regression (eg. (Reich et al., 2011; Macmillan, 2013)). These modeling approaches generally produce a probabilistic surface, which results in a useful visualization but does not directly solve our goal of identifying specific unusual regions. Achieving this goal requires a human to inspect the probabilistic surface, manually segment it into unusual regions and rank these regions according to some unusualness criterion. This human intervention is not desirable when the spatial region is large and also if the goal is to create an automated monitoring system. Our QSSS algorithm essentially automates these steps in a computationally efficient manner.

2.1 THE SPATIAL SCAN STATISTIC

The Spatial Scan Statistic, introduced by Kulldorff (1997) is a widely used approach for finding anomalous regions. For the SSS, each spatial data point is represented by a tuple (c_i, b_i) along with its location. The value c_i corresponds to a count at location i (e.g. the number of disease cases) and b_i is the baseline value (e.g. the population) at location i. The value c_i is Poisson distributed with mean qb_i, where q is the probability of an event of interest occurring.

The original SSS used a scanning window in the shape of a circle to discover unusual regions. While in theory the search should be over all circular regions, the search is, in practice, often limited to circles with centers determined by a fixed grid superimposed on the spatial area. Let $\boldsymbol{\mathcal{C}}$ be the set of all circular regions searched by the SSS and let $C \in \boldsymbol{\mathcal{C}}$ be the region under consideration. For a region C under consideration, let $c_{in} = \sum_{i \in C} c_i$, $c_{out} = \sum_{i \notin C} c_i$, $b_{in} = \sum_{i \in C} b_i$, $b_{out} = \sum_{i \notin C} b_i$. Let q_{in} be the event probability inside the region C and let q_{out} be the probability outside the region C.

Under the null hypothesis H_0, the event probability is uniform across the entire area i.e. $q_{in} = q_{out}$. Under the alternate hypothesis $H_1(C)$, $q_{in} > q_{out}$. We estimate q_{in} and q_{out} using maximum likelihood estimation. The SSS uses the likelihood ratio test to score a region C:

$$Score(C) = \frac{P(\boldsymbol{D}|H_1(C))}{P(\boldsymbol{D}|H_0)}$$
$$= \left(\frac{c_{in}}{b_{in}}\right)^{c_{in}} \left(\frac{c_{out}}{b_{out}}\right)^{c_{out}} \left(\frac{c_{in}+c_{out}}{b_{in}+b_{out}}\right)^{-(c_{in}+c_{out})}$$

if $(\frac{c_{in}}{b_{in}}) > (\frac{c_{out}}{b_{out}})$ and 1 otherwise.

The SSS then selects the region with the highest score i.e. $C^* = \underset{C \in \boldsymbol{\mathcal{C}}}{\operatorname{argmax}}\ Score(C)$. Due to the multiple hypothesis testing problem, we cannot interpret the score from the likelihood ratio test as a true p-value. Instead, we estimate the p-value through a randomization test. In each replication of the randomization test, we maintain the same underlying population as the original problem, but generate events assuming a uniform probability. Then, the search for the best scoring region is performed. The process is repeated for R replications to produce an empirical distribution which determines how likely it is to obtain a best score of C^*.

Many researchers have extended the original SSS approach, including using scanning windows that are arbitrarily shaped (Duczmal and Assuncao, 2004) and incorporating mobility patterns (Lan et al., 2014). One variant

goes beyond shifts in means by discovering which subpopulation is most affected by a treatment (McFowland et al., 2018). We point out that performing a quantile-based comparison results in a fundamentally different type of optimization problem and past work on speeding up the SSS (eg. (Neill and Moore, 2004; Neill, 2012)) is not readily applicable. Finally, Moore and Wong (2015) use the SSS to find species rich hotspots but they do not compare quantiles of distributions.

2.2 QUANTILE REGRESSION

Suppose we have a continuous random variable Y with distribution function $F(Y) = P(Y \leq y)$. The τ-th quantile $q(\tau)$, with $0 < \tau < 1$, is defined as $q(\tau) = F^{-1}(\tau) = \inf_y\{F(y) \geq \tau\}$. For example, when $\tau = 0.5$, we get the median. Given a dataset $Y_1, \ldots, Y_n$, the τ-th sample quantile $\hat{q}(\tau)$, can be computed by solving the optimization problem:

$$\hat{q}(\tau) = \underset{q}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_\tau(Y_i - q)$$

where $\rho_\tau(r) = r(\tau - I(r < 0))$.

Quantile regression, introduced by Koenker and Bassett (1978), fits a regression to the conditional τ-th quantile of the response variable. Given a dataset $\boldsymbol{D} = \{(Y_1, \boldsymbol{X}_1), \ldots, (Y_N, \boldsymbol{X}_N)\}$ where Y_i is the response variable and $\boldsymbol{X_i}$ are the covariates, fitting a quantile regression involves solving:

$$\hat{\boldsymbol{\beta}}(\tau) = \underset{\boldsymbol{\beta}}{argmin} \sum_{i=1}^{n} \rho_\tau(Y_i - \boldsymbol{X_i}\boldsymbol{\beta})$$

The solution $\hat{\boldsymbol{\beta}}(\tau)$ produces a conditional quantile function $Q_Y(\tau|\boldsymbol{X} = \boldsymbol{x}) = \boldsymbol{x}'\hat{\boldsymbol{\beta}}(\tau)$, similar to how a standard regression produces the conditional mean when the coefficients are multiplied with the covariate values.

Quantile regression is a useful tool for analyzing specific parts of a distribution. It can model the data extremes by setting τ close to either 1 or 0, or it can reduce the influence of these points by modeling τ close to 0.5.

There are several methods for comparing two quantile regression models. These test include applications of Wald's test, the Likelihood Ratio test, and Rank test (Koenker and Machado, 1999). Mood's median test (Mood, 1950) can also be adapted to perform a comparison at a given quantile. While fast to compute, this version of Mood's test lacks the power of the Wald, Likelihood Ratio, and Rank alternatives. Any of these methods are still usable when the covariate set $\boldsymbol{X}$ is empty by using the quantiles of $\boldsymbol{Y}$. We use the Rank test, as it can be implemented without repeatedly re-estimating the quantile regression coefficients for each data subset, thereby reducing its computation time without sacrificing power. In the following section we explain the Rank test for quantile regression.

2.3 RANK TEST FOR QUANTILE REGRESSION

Let the regression model for the τth quantile have the form $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta_1} + \tilde{\boldsymbol{X}}\boldsymbol{\beta_2}$ where each row $\boldsymbol{X}_i$ corresponds to a data point. For a given data subset $\boldsymbol{C} \subseteq \boldsymbol{D}$, $\tilde{\boldsymbol{X}}_{\boldsymbol{i}} = \boldsymbol{X_i}$ if $\boldsymbol{X_i} \in \boldsymbol{C}$ and $\tilde{\boldsymbol{X}}_{\boldsymbol{i}} = \boldsymbol{0}$ if $\boldsymbol{X_i} \notin \boldsymbol{C}$. This model will simultaneously fit a regression to $\boldsymbol{C}$ and $\boldsymbol{D} \setminus \boldsymbol{C}$. In the spatial scan context $\boldsymbol{C}$ is the region inside our circle and $\boldsymbol{D} \setminus \boldsymbol{C}$ is the region outside. The goal is then to test the null hypothesis $H_0 : \boldsymbol{\beta_2} = \boldsymbol{0}$ against the alternative $H_1 : \boldsymbol{\beta_2} \neq \boldsymbol{0}$ to see if the subset $\boldsymbol{C}$ is sufficiently different from the full distribution of $\boldsymbol{D}$.

The Rank test is an application of the score test, using a score function and ranking process to estimate the data distribution when the true likelihood is unknown. In general terms, the score test statistic is composed of the product of the square of a score vector, an approximation of the derivative using a score function in place of the true likelihood, and the inverse of the Fischer information. The Rank test statistic takes the following form when applied to quantile regression for the null hypothesis above (Gutenbrunner et al., 1993).

$$T = \boldsymbol{S}'\boldsymbol{M}^{-1}\boldsymbol{S}/\Psi^2 \tag{1}$$

We include the following definitions along with the dimensions of each term in braces for clarity.

$$\begin{aligned}
\boldsymbol{S}_{[p\times 1]} &= n^{-1/2}(\tilde{\boldsymbol{X}} - \boldsymbol{H}\tilde{\boldsymbol{X}})'\hat{\boldsymbol{b}} \\
\boldsymbol{H}_{[n\times n]} &= \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' \\
\hat{\boldsymbol{b}}_{[n\times 1]} &= -\int \psi(t)d\hat{\boldsymbol{a}}(t) \\
\boldsymbol{M}_{[p\times p]} &= n^{-1}(\tilde{\boldsymbol{X}} - \boldsymbol{H}\tilde{\boldsymbol{X}})'(\tilde{\boldsymbol{X}} - \boldsymbol{H}\tilde{\boldsymbol{X}})
\end{aligned}$$

We now provide an intuitive explanation for each term. $\boldsymbol{S}$ is the score vector for the test. It represents an approximate derivative of $\boldsymbol{\beta}$ under the null hypothesis. By formulating $\boldsymbol{S}$ with the matrix $\tilde{\boldsymbol{X}} - \boldsymbol{H}\tilde{\boldsymbol{X}}$, the influence of $\boldsymbol{X}$ (and $\boldsymbol{\beta_1}$) is removed, focusing the approximate derivative on $\boldsymbol{\beta_2}$, the parameters of interest. When $\boldsymbol{S}$ is large, it indicates that the null hypothesis is ill-suited to the data.

With the true likelihood unknown, $\boldsymbol{S}$ is calculated using $\hat{\boldsymbol{b}}$, an n vector of scores calculated for each data point. These scores are computed by integrating the score function $\psi(t)$ with respect to the regression rankscores $\hat{\boldsymbol{a}}$ de-

fined by Gutenbrunner and Jureckov (1992). The regression rankscores allow the rank test to be applied to quantile regression by converting the multi-dimensional data into a single-dimensional ranking for the chosen quantile. $\hat{\boldsymbol{a}}$ is equal to the dual solution of the quantile regression under the null hypothesis, and can be calculated using the primal solution $\boldsymbol{\beta_1}$. The value $\hat{\boldsymbol{a}}_i = 1$ if $\boldsymbol{\beta_1 X_i} > 0$, 0 if $\boldsymbol{\beta_1 X_i} < 0$, and between 0 and 1 if $\boldsymbol{\beta_1 X_i} = 0$, satisfying $\boldsymbol{X'}\hat{\boldsymbol{a}} = (1-\tau)\boldsymbol{X'1}$.

$\Psi^2 = \int(\psi(t) - \bar{\psi})^2 dt$ is an additional normalization term for the covariance of the score function. Koenker and Machado (1999) highlight the quantile score function $\psi(t) = \tau - I(t < \tau)$, which focuses the test on a specific quantile. This gives us $\hat{\boldsymbol{b_i}} = \hat{\boldsymbol{a_i}}(\tau) - (1-\tau)$ and $\Psi^2 = \tau(1-\tau)$. With this choice of score function, $\hat{\boldsymbol{b_i}}$ is either τ if $\boldsymbol{\beta_1 X_i} > 0$, $\tau - 1$ if $\boldsymbol{\beta_1 X_i} < 0$, or a value inbetween otherwise.

The test statistic T follows a Chi-squared distribution with p degrees of freedom under the null hypothesis. It has the desirable properties of not depending on the error distribution, and not needing to learn the model under the alternative hypothesis. The values $\hat{\boldsymbol{b}}$ are calculated under the null hypothesis that $\boldsymbol{\beta_2} = \boldsymbol{0}$.

3 METHODOLOGY

We start with a high level overview of our QSSS[1]. Given a dataset $\boldsymbol{D} = \{\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{L}\}$, and a list of starting locations $\boldsymbol{P}$, the QSSS searches over circular areas in $\boldsymbol{L}$, beginning at each starting location in $\boldsymbol{P}$ and growing the regions one data point at a time, starting from some minimum number of points. The regions are grown as circles of increasing radius. Each time the region grows, we calculate its test statistic using our Incremental Rank test (Section 3.1). Once the region cannot be grown any larger, or reaches a maximum size, we move on to the next starting point in $\boldsymbol{P}$. After all starting points have been exhausted, an adjusted p-value is calculated for the region with the highest test statistic using a Gumbel correction (Section 3.2). We chose the Gumbel correction because it is much faster than the traditional randomization test. If the adjusted p-value is significant then the algorithm returns the region, otherwise it says that no significant region was found.

3.1 FASTER RANK TEST FOR QSSS

In the QSSS framework, the Rank test needs to be performed for every circular subset $\boldsymbol{C} \subseteq \boldsymbol{D}$. We can choose a set of starting points (either each data point or a grid formed over $\boldsymbol{L}$) for the regions and grow each one, recalculating our hypothesis test each time the region overlaps a new data point. The inclusion of a new data point i into the region will change the ith row of $\tilde{\boldsymbol{X}}$ from a row of zeros to the ith row of $\boldsymbol{X}$. Under the framework of the Rank test, $\boldsymbol{X}$, $\boldsymbol{H}$, and $\hat{\boldsymbol{b}}$ will be the same for every choice of region $\boldsymbol{C}$. Thus our only task is to update T as $\tilde{\boldsymbol{X}}$ changes.

The primary bottlenecks in updating T are in updating $\boldsymbol{S}$ and recomputing $\boldsymbol{M}^{-1}$. $\boldsymbol{M}^{-1}$ can be updated incrementally using applications of the Sherman-Morrison formula (Sherman and Morrison, 1950), but a more efficient update can be performed by leveraging the special structure of T. Note that we can re-write T as

$$T = \hat{\boldsymbol{b}}\boldsymbol{Z}(\boldsymbol{Z'Z})^{-1}\boldsymbol{Z'}\hat{\boldsymbol{b}}/\Psi^2 \tag{2}$$

where $\boldsymbol{Z} = \tilde{\boldsymbol{X}} - \boldsymbol{H}\tilde{\boldsymbol{X}}$. $\boldsymbol{Z}(\boldsymbol{Z'Z})^{-1}\boldsymbol{Z'}$ is by definition a projection matrix onto the space of $\boldsymbol{Z}$. If we let $\boldsymbol{U}$ be an n x p orthonormal column basis of $\boldsymbol{Z}$, then

$$T = \hat{\boldsymbol{b}}\boldsymbol{U}\boldsymbol{U'}\hat{\boldsymbol{b}}/\Psi^2 \tag{3}$$

The inverse in Equation 2 is a normalization term. Since $\boldsymbol{U}$ is already normalized, the formulation of Equation 3 allows us to forgo the matrix inverse calculation. Thus we can quickly recalculate T by performing incremental updates to our orthonormal basis $\boldsymbol{U}$ as $\tilde{\boldsymbol{X}}$ changes.

3.1.1 Incremental Orthogonalization of Rank Test

Our goal is to take an existing orthonormal basis at iteration t (i.e. $\boldsymbol{U}^t$), and calculate $\boldsymbol{U}^{t+1}$ based on the (small) change in $\tilde{\boldsymbol{X}}$ when a new data point is added to the inside region. To do this efficiently, we leverage the QR decomposibility of $\boldsymbol{Z}^t$, which enables a rank one update. However, we also need to efficiently preserve the QR decomposibility of $\boldsymbol{Z}^{t+1}$, which we do through a series of Givens rotations.

Let $\boldsymbol{K}_{[n\times n]}$ be a row selector matrix, where $K_{[j,j]} = 1$ if point j is in $\boldsymbol{C}$, and all other values are zero. For a current region $\boldsymbol{C}^t$ at iteration t, $\tilde{\boldsymbol{X}} = \boldsymbol{K}^t\boldsymbol{X}$. If we add point i to $\tilde{\boldsymbol{X}}$ during iteration $t+1$, then this is equivalent to changing the ith diagonal of $\boldsymbol{K}^t$ from 0 to 1. We can express this change as a matrix sum $\boldsymbol{K}^{t+1} = \boldsymbol{K}^t + \boldsymbol{K}_i$ where $\boldsymbol{K}_i$ is zero except for the ith diagonal. This allows us to decompose the change in $\boldsymbol{Z}^{t+1}$ as follows:

[1]Matlab code for our experiments and algorithms can be found at https://github.com/moortrav/QSSS

$$\boldsymbol{Z}^{t+1} = (\boldsymbol{K}^t + \boldsymbol{K}_i)\boldsymbol{X} - \boldsymbol{H}(\boldsymbol{K}^t + \boldsymbol{K}_i)\boldsymbol{X} \quad (4)$$

$$= \boldsymbol{Z}^t + \boldsymbol{K}_i\boldsymbol{X} - \boldsymbol{H}\boldsymbol{K}_i\boldsymbol{X} \quad (5)$$

$$= \boldsymbol{Z}^t + (\boldsymbol{e}_i - \boldsymbol{H}_i)'\boldsymbol{X}_i \quad (6)$$

where $\boldsymbol{e}_i$ is the ith unit basis vector of size $1 \times n$. Note that $\boldsymbol{e}_i'\boldsymbol{X}_i$ is an outer product producing a matrix of size $n \times p$. $\boldsymbol{H}$ is a symmetric matrix, so we use the row vector $\boldsymbol{H}_i$ to keep our notation consistent. In Equation 6 we have reduced the update to $\boldsymbol{Z}^t$ to the product of a column and row vector i.e. $(\boldsymbol{e}_i - \boldsymbol{H}_i)'\boldsymbol{X}_i$. This means that the matrix added to $\boldsymbol{Z}^t$ has a rank of one, and the change to each column of $\boldsymbol{Z}^t$ is a multiple of the same column vector $(\boldsymbol{e}_i - \boldsymbol{H}_i)'$. We can use this special update structure in an algorithm to find $\boldsymbol{U}^{t+1}$ efficiently.

If the QR factorization of $\boldsymbol{Z}^t$ is known, where $\boldsymbol{R}$ is an upper triangular matrix and $\boldsymbol{Q} = \boldsymbol{U}^t$ is an orthonormal column basis, then the factorization for $\boldsymbol{Z}^{t+1}$ can be found with the rank one update algorithm detailed in section 12.5.1 of Golub and Loan (2012). This algorithm lets us find the factorization $\boldsymbol{Z}^{t+1} = \boldsymbol{Q}^{t+1}\boldsymbol{R}^{t+1}$ using the previous factorization $\boldsymbol{Z}^t = \boldsymbol{Q}^t\boldsymbol{R}^t$, giving us $\boldsymbol{U}^{t+1} = \boldsymbol{Q}^{t+1}$ for our update to the test statistic T.

Let $\boldsymbol{v} = \boldsymbol{e}_i - \boldsymbol{H}_i$. We start by refactoring the update as

$$\boldsymbol{Z}^{t+1} = \boldsymbol{Q}^t\boldsymbol{R}^t + \boldsymbol{v}'\boldsymbol{X}_i = \boldsymbol{Q}^t(\boldsymbol{R}^t + \boldsymbol{w}'\boldsymbol{X}_i) \quad (7)$$

Where $\boldsymbol{w}' = (\boldsymbol{Q}^t)^{-1}\boldsymbol{v}' = (\boldsymbol{Q}^t)'\boldsymbol{v}'$. Our goal is to turn $\boldsymbol{Z}^{t+1}$ into the product of an orthonormal matrix (which will be $\boldsymbol{Q}^{t+1}$) and an upper triangular matrix to be produced from $(\boldsymbol{R}^t + \boldsymbol{w}'\boldsymbol{X}_i)$ through Givens rotations. The details of the Golub and Loan (2012) algorithm that does this can be found in the supplemental materials.

This algorithm is not ideal in its current form, because creating the upper triangular matrix takes $O(n)$ Givens rotations, a result of $\boldsymbol{Q}$ being $n \times n$. However, the first p columns of $\boldsymbol{Q}$ and p rows of $\boldsymbol{R}$, denoted as $\boldsymbol{Q}_{[\cdot,1:p]}$ and $\boldsymbol{R}_{[1:p,\cdot]}$, are sufficient to reconstruct $\boldsymbol{Z}$, as $\boldsymbol{Z} = \boldsymbol{Q}_{[\cdot,1:p]}\boldsymbol{R}_{[1:p,\cdot]} = \boldsymbol{Q}\boldsymbol{R}$. Working with this reduced factorization would reduce the storage and number of Givens rotations required for the algorithm.

Unfortunately this representation is insufficient to perform the update. If we were to compute the vector $\boldsymbol{w}$ from Equation 7 with $\boldsymbol{Q}_{[\cdot,1:p]}$, then $\boldsymbol{w}' = \boldsymbol{Q}'_{[\cdot,1:p]}\boldsymbol{v}' = \boldsymbol{Q}'_{[\cdot,1:p]}\boldsymbol{e}_i' - \boldsymbol{Q}'_{[\cdot,1:p]}\boldsymbol{H}_i' = \boldsymbol{0}$. To see this, note that $\boldsymbol{Q}'_{[\cdot,1:p]}\boldsymbol{e}_i'$ is zero because the ith row of $\boldsymbol{Z}^t$ and $\boldsymbol{Q}$ is zero, since the ith data point has not been added to the inside region yet. $\boldsymbol{Q}'_{[\cdot,1:p]}\boldsymbol{H}_i'$ is also zero because $\boldsymbol{H}_i$ is perpendicular to $\boldsymbol{Z}^t$ and thus perpendicular to $\boldsymbol{Q}_{[\cdot,1:p]}$. With $\boldsymbol{w}' = \boldsymbol{0}$, Equation 7 becomes $\boldsymbol{Z}^{t+1} = \boldsymbol{Q}^t_{[\cdot,1:p]}\boldsymbol{R}^t_{[1:p,\cdot]}$, which completely ignores the update term. Intuitively speaking, we cannot update the column basis of $\boldsymbol{Z}^t$ by only considering that basis.

Fortunately, there is a way to summarize the influence of the last n-p columns of $\boldsymbol{Q}$, denoted $\boldsymbol{Q}_{[\cdot,(p+1):n]}$, into a single vector. When the Givens rotations zero out element j in $\boldsymbol{w}$, it changes element $j-1$ to $\sqrt{w_{j-1}^2 + w_j^2}$. Consequently, the result of rotations $\boldsymbol{J}'_{p+1} \dots \boldsymbol{J}'_{n-1}$ will set $w_{p+1} = \sqrt{w_{p+1}^2 + \dots + w_n^2} = \sqrt{\sum_{j=p+1}^{n}(\boldsymbol{Q}'_j\boldsymbol{H}_i)^2)} = |\boldsymbol{Q}_{[\cdot,(p+1):n]}\boldsymbol{H}_i|$. Because $\boldsymbol{Q}_{[\cdot,1:p]}$ is perpendicular to $\boldsymbol{H}_i$, the columns $\boldsymbol{Q}_{[\cdot,p+1]} \dots \boldsymbol{Q}_{[\cdot,n]}$ are an orthonormal basis of $\boldsymbol{H}_i$. Projecting $\boldsymbol{H}_i$ onto its own basis will preserve its length, giving us $|\boldsymbol{Q}_{[\cdot,(p+1):n]}\boldsymbol{H}_i| = |\boldsymbol{H}_i|$. Thus, we can summarize all $n-p$ Givens rotations with a single vector $\boldsymbol{q}$ such that $\boldsymbol{q}\boldsymbol{H}_i = |\boldsymbol{H}_i|$, which gives us $\boldsymbol{q} = \boldsymbol{H}_i/|\boldsymbol{H}_i|$. If we append $\boldsymbol{q}$ as a new column of $\boldsymbol{Q}_{[\cdot,1:p]}$ to produce $\tilde{\boldsymbol{Q}}_{[\cdot,1:p]}$ and a zero row to the bottom of $\boldsymbol{R}_{[1:p,\cdot]}$ to produce $\tilde{\boldsymbol{R}}_{[1:p,\cdot]}$ then we can run the algorithm with only $O(p)$ Givens rotations and still produce the same result. Since $\boldsymbol{q}$ is normalized and perpendicular to $\boldsymbol{Q}_{[\cdot,1:p]}$, $\tilde{\boldsymbol{Q}}_{[\cdot,1:p]}$ is still orthonormal.

We can perform the rank one update on $\tilde{\boldsymbol{Q}}^t_{[\cdot,1:p]}$ and $\tilde{\boldsymbol{R}}^t_{[1:p,\cdot]}$ using the algorithm in Golub and Loan (2012). The first p columns of $\tilde{\boldsymbol{Q}}^{t+1}_{[\cdot,1:p]}$ make our new orthonormal column basis $\boldsymbol{U}^{t+1}$ used to calculate our test statistic T.

Algorithm 1 Incremental Rank Test

Inputs: $\boldsymbol{X}, \boldsymbol{H}, \hat{\boldsymbol{b}}, \boldsymbol{Q}, \boldsymbol{R}, \tau, i$
$\boldsymbol{v} = \boldsymbol{e_i} - \boldsymbol{H}_i$
$\boldsymbol{Q}, \boldsymbol{R}$ =qr_update$(\boldsymbol{Q}, \boldsymbol{R}, \boldsymbol{v}, \boldsymbol{X}_i)$
$T = \hat{\boldsymbol{b}}'\boldsymbol{Q}\boldsymbol{Q}'\hat{\boldsymbol{b}}/(\tau(1-\tau))$
Return(T)

Algorithm 1 shows the incremental rank test which calls qr_update. It takes the index i of the datapoint being added to the region, along with the QR factorization for the previous iteration as inputs. The details of qr_update can be found in the supplementary materials. We can run the algorithm with either the full QR factorization, or the abridged form represented by $\tilde{\boldsymbol{Q}}_{[\cdot,1:p]}$ and $\tilde{\boldsymbol{R}}_{[1:p,\cdot]}$

Note that our incremental rank test is not an approximation as it computes the test statistic (Equation 1) exactly.

3.1.2 Update Runtime

With our compact representation for $\tilde{\boldsymbol{Q}}_{[\cdot,1:p]}$ and $\tilde{\boldsymbol{R}}_{[1:p,\cdot]}$, the rank one update to our QR factorization takes $O(np)$ time. Each Givens rotation is an $O(n)$ operation, and we perform $O(p)$ of them in total. Once $\boldsymbol{U}^{t+1}$ is found,

T^{t+1} can be calculated in $O(np)$ time by computing $\hat{\boldsymbol{b}}\boldsymbol{U}^{t+1} = \boldsymbol{u}$, and then finding $T^{t+1} = \boldsymbol{u}\boldsymbol{u}'$. Thus the entire update to T can be performed in $O(np)$ time when a single point is added to $\tilde{\boldsymbol{X}}$.

3.2 MULTIPLE HYPOTHESIS TEST CORRECTION

To account for the multiple hypothesis test problem, we perform a correction using the method in Abrams et al. (2010). We generate 1000 simulations of the data under the null hypothesis. The maximum test statistic from each of these simulations are used to fit the parameters μ, γ of a Gumbel distribution. We calculate the adjusted p-value of a region with test statistic T as $1 - g(T|\mu, \gamma)$, where g is the CDF of the Gumbel distribution. This tells us the rarity of drawing a value at least as large as T from the distribution of maximum test statistics. In all of our applications we report the most significant region found by QSSS, provided that the adjusted p-value of the region is less than 0.05. Otherwise no significantly different region is found.

4 RESULTS

4.1 SYNTHETIC DATA EXPERIMENTS

We begin by demonstrating the speedup from our incremental formulation of the Rank test, followed by a comparison of the Rank test to other possible choices of hypothesis tests. We use synthetic data to evaluate these two criteria, as it is easy to generate in large quantities, and it can contain a verifiable ground truth.

4.1.1 Simulator

The purpose of our simulator is to inject data points in spatial regions where the data distribution is altered at a specific percentile. We start with a default distribution, then modify a specific range of the distribution for a random spatial region. This acts as the target region for the algorithm to identify. A detailed description of our simulator can be found in the supplemental materials.

4.1.2 Incremental Rank Test Timing

Using our simulated data, we compare the runtime of our incremental version of the Rank test to its naive (non-incremental) formulation. For each algorithm we calculated the Rank test statistic T, starting from a base radius, then expanding to include 100 new points. In Figure 1 we show the average time, in milliseconds, that each algorithm took to calculate T when a new point was added. These tests were done for increasing values of n while keeping p constant at 5. The two algorithms start out at similar times when $n = 1000$, but quickly diverge. At $n = 16,000$ our incremental Rank test takes only 2.83 ms to compute each update, while the non-incremental version takes 166.9 ms. The incremental speedup for the Rank test makes it usable within the framework of the QSSS, while the naive calculation would take far too much time to be feasible for large n.

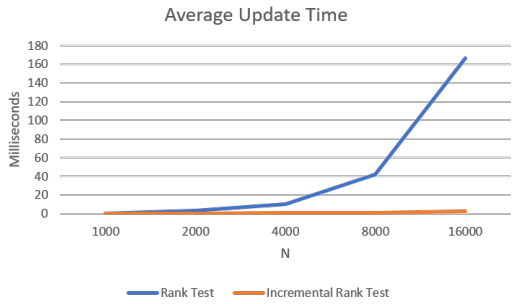


Figure 1: Average time to update the Rank test statistic, for the full and incremental formulations. Times were averaged over 100 updates for randomly generated data with different values of n and constant $p = 5$.

4.1.3 Comparison against Other Baselines

We are unaware of other methods that can solve exactly the same problem as the QSSS. As a result, we develop three baseline algorithms that can be used for comparison. We create the first baseline by adapting the SSS to account for covariates by modeling the response for the inside (and outside) region using least squares regression. This baseline compares means (not quantiles) of distributions by using a likelihood ratio test. We add the Mean test to show the ineffectiveness of a mean-based test statistic in finding regions that differ only in specific quantiles. For the second baseline, we modify the SSS to focus on a specific quantile of the distribution. We do this by fitting a τth quantile regression with coefficients $\boldsymbol{\beta}$ to the entire data set; then, for each test region, this baseline calculates Mood's test at τ, which is a statistic from a 2×2 Chi-Squared table that compares the number of points above and below the $\boldsymbol{\beta}$ plane from both inside and outside of the region. Finally, our third baseline is similar to the second baseline but it replaces Mood's test with the more powerful but computationally expensive likelihood ratio test from Koenker and Machado (1999) (LR) for quantile regression. This LR test forms a Chi-squared statistic from the residuals of the quantile regressions fit to the null and alternative models.

Using our simulator, we produced 30 randomly generated data sets with $n = 1000$. $\boldsymbol{B_2}$ (parameters for the

injected data) is the same as $\boldsymbol{B_1}$ (parameters for the normal data) in these datasets, except between the 70th and 100th percentiles of the distribution. 100 of the points are generated from $f(\boldsymbol{B_2})$ and 900 of the points are generated from $f(\boldsymbol{B_1})$. Our Moods, LR, and Rank test search for regions that differ at the 90th percentile. For each algorithm, we look at the most significant region found for each dataset, provided it has a p-value of at most 0.05 after the Gumbel correction. Otherwise we count the algorithm as finding no significant region for that dataset. Note that this experiment setup is extremely challenging. The ground truth region to detect makes up 10% of the total dataset, but only 30% of the points in the region on average indicate that it has a different distribution. Adding in the random noise term further complicates the detection task.

P=3	**Moods**	**LR**	**Rank**	**Mean**
Precision	0.322	0.499	**0.576**	0.405
Recall	0.353	**0.548**	0.500	0.334
F1	0.337	0.522	**0.535**	0.366
Time (s)	**3.32***	350.73	46.64	31.06
P=5	**Moods**	**LR**	**Rank**	**Mean**
Precision	0.259	0.395	**0.508**	0.216
Recall	0.320	0.416	**0.484**	0.110
F1	0.286	0.405	**0.495**	0.146
Time (s)	**3.02***	310.65	84.36	39.25
P=10	**Moods**	**LR**	**Rank**	**Mean**
Precision	0.286	0.243	**0.676***	0.197
Recall	0.344	0.278	**0.442**	0.169
F1	0.312	0.259	**0.535***	0.182
Time (s)	**3.26***	379.32	83.31	38.25

Table 1: The precision, recall, F1 and running time of QSSS on synthetic data using various algorithms. The * indicates statistical significance (paired t-test, $\alpha = 0.05$).

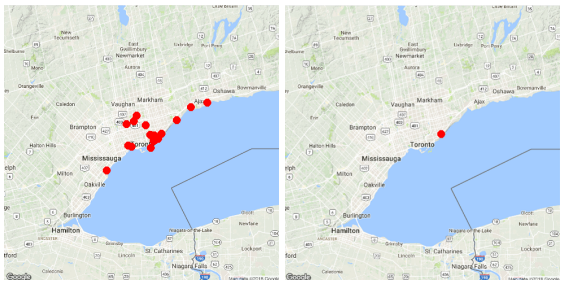
Table 1 shows the results of the simulation experiments. The precision, recall and F1 score of each algorithm is reported in the task of finding the region generated from $\boldsymbol{B_2}$ in each dataset. These values are calculated on a per data point basis for each dataset, then averaged over the 30 datasets. Three experiment runs were performed, with dimensionality $p = 3$, 5 and 10. LR and Rank are the two most accurate tests for $p = 3$ and 5, with Rank being the most accurate for $p = 10$. The poor performance of Mood's and Mean is expected, since Mood's is a low power test and Mean is ill-suited to find such subtle distributional variations.

Table 1 also shows the average total runtime of each spatial scan algorithm on the simulation data. This table illustrates the speed of Mood's test compared to the other hypothesis tests. We can also see that the LR test is significantly slower than the others, a result of needing to fit a quantile regression to the alternative model for every test region. In our implementation of LR, we use warmstarting to increase the speed of the quantile regression algorithm as the regions grow point by point. Even with warmstarting, the LR algorithm still takes at least four times longer to run on the simulation data than our other hypothesis choices.

Comparing the accuracy and timing results, we see that while the Rank and LR test are the most accurate, the Rank test offers the best tradeoff in terms of usability between speed and power. We found it infeasible to use the computationally expensive LR test, even on moderately sized datasets. While the Mood's and Mean tests were faster, neither one was very capable at detecting differences in our simulated data. Due to these result, we primarily use the Rank test in our case studies below; we also include results from the Mean test to illustrate the differences between the two.

4.2 ROBUSTNESS TO OUTLIERS

One of the benefits of quantile based analysis is that it is more robust to data outliers than mean-based methods. We illustrate how this can affect spatial scan analysis using eButterfly data as an example use case.



(a) 50th Percentile (b) Mean

Figure 2: Most significant regions found from eButterfly data, with data points in the region shown as red dots. The figures are zoomed in on the Toronto region for visibility. Figure 2a is from the QSSS algorithm at the 50th percentile, Figure 2b is from the mean regression spatial scan. The region in 2b represents a single location, rather than an area, as all the data from that subset have the same location parameters.

Citizen science biodiversity monitoring programs, such as eButterfly (Prudic et al., 2017), play an important role in ecology as it informs species distribution models and also conservation programs. Citizen scientists participating in these programs submit checklists which record observations of certain types of organisms, such as butterflies in the case of eButterfly, identified by species.

We construct a dataset out of the abundance counts of monarch butterflies (i.e. the number of butterfly individuals observed) in Ontario in 2016. Quantile regression on count data can be addressed using the smoothing method in Machado and Silva (2002), which turns the counts into continuous values by adding uniform noise. This transformation allows us to perform inference with the Rank test as we would with continuous data.

In our analysis, we include the time spent observing for each checklist as the covariate, since there should be a strong correlation between this value and the number of monarchs observed. We ran our QSSS algorithm on several different quantiles and compared the top region for each to the top region found by a mean-based least squares spatial scan.

Figure 2 shows the most significant regions found by the mean regression spatial scan and QSSS at the 50th percentile[2]. Inspection of the data, and verification with domain experts at eButterfly reveal two interesting results identified by the algorithms. Within the data time window there is a single observer who heavily skews the distribution. This observer was involved in a monarch tagging project, and submitted a significant number of very high monarch checklists. The region found by the mean spatial scan only includes the checklists from this observer, all at the same spatial location. When QSSS is run at the 50th percentile, a different trend emerges. The checklists from the observer has much less influence on the model at this level, and the algorithm instead picks up an area of high monarch counts due to migration routes around the great lakes.

If we were limited to only mean-based spatial scans, we would have to filter out the outlier data from the monarch tagging observer to find the desired trends in the dataset. Being able to adjust the percentile of QSSS allows us to reduce the influence of outliers as desired, without explicit removal of outliers from the data.

4.3 QUANTILE BASED REGION DETECTION

We now demonstrate the usefulness of detecting unusual spatial regions based on different quantiles.

4.3.1 Education and Unemployment Data

We combine the county-level education and unemployment datasets from the USDA Economic Research Service web page (Parker, 2017). We use the county-level unemployment rates from 2016 as the response variable, and combine the education percentages from 2012-2016 with median household income (as percentage of state total) values from 2016 as the covariates. The education percentages are the proportion of adults in each county with less than a high school diploma, just a high school diploma, one to three years of college, and four years of college or more. We only use the counties from the continental US.

We ran our QSSS algorithm on the 10th and 90th percentile of the data, along with a mean-based approach using least squares regression. Figure 3 shows the most significant region found by each algorithm. Both the mean and 90th percentile search found the Appalachian region that intersects Kentucky, West Virginia and Virginia, which is well-known to have high unemployment rates with the collapse of the coal industry (Caruthers, 2016). In the 10th percentile region, South Dakota, North Dakota, Nebraska, and Colorado are rated 2,3,4, and 6 in unemployment in the continental US as a whole. This middle region of the country enjoys lower unemployment rates due to the local oil industry and relatively low fallout from the Great Recession (DePillis, 2018). The most significant region discovered at the 10th percentile has a 2 point lower unemployment rate on average, which is abnormally low even when compared to other low unemployment areas.

The unemployment data results highlight the fact that the QSSS, unlike the mean scan, can identify multiple trends in a dataset by changing the modeled quantile.

4.3.2 eBird

The final case study presents the results of applying QSSS to eBird (Sullivan et al., 2014) data. The eBird project collects bird observation checklists from citizen scientists around the world. We compiled two datasets from eBird data collected in 2017 between March and April. These datasets correspond to two different Bird Conservation Regions (BCRs) within the U.S. We divide the data by BCR because they represent cohesive habitats for different bird species. Our choice of March and April is to mitigate the effects of seasonality on the algorithms.

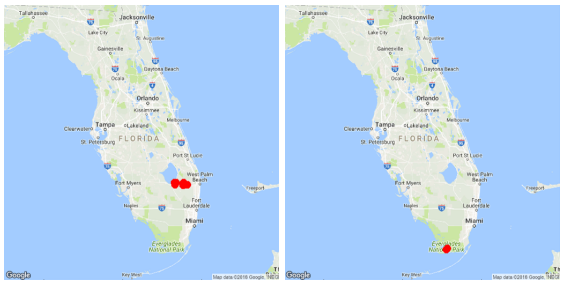
Different from our eButterfly study, we used the total number of species observed from each checklist as our response variable, and the time spent observing as the covariate. Past work has shown that the number of species observed per unit time is highly predictive of the skill level of an observer (Kelling et al., 2015). We use the same count smoothing approach on the eBird data as we did on eButterfly to fit the quantile regression model.

Figure 4 shows the most significant regions found for the mean spatial scan and our QSSS run at the 90th percentile. We corresponded with domain experts from

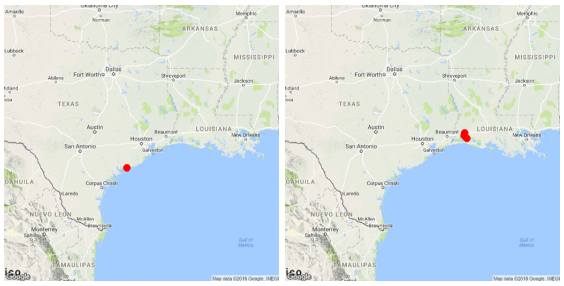
[2]All maps generated using ggmap in R (Kahle and Wickham, 2013)

(a) 10th Percentile (b) 90th Percentile (c) Mean

Figure 3: Most significant region found by the QSSS algorithm for the 10th and 90th percentiles of the Education and Unemployment dataset. Most significant region by the mean spatial scan is included for comparison. Regions are illustrated by the centroids of the counties they contain.



(a) BCR 31 90th Percentile (b) BCR 31 Mean



(c) BCR 37 90th Percentile (d) BCR 37 Mean

Figure 4: The most significant regions found by QSSS at the 90th percentile and by the mean spatial scan on eBird data from BCRs 31 (Florida) and 37 (Gulf coast). Data points within a region are shown as red dots.

eBird, who offered an analysis on the regions detected.

For BCR 31, the QSSS found an unusual birding location – one that is less frequented by beginners. The birders who visit this region are highly skilled and are able to continue observing a high number of bird species as they stay there. In contrast, the mean scan found a popular species-rich hotspot in the Everglades frequented by both experts and novices. This region has many large wading birds which are easy to see and identify initially.

In BCR 37, the QSSS found a hotspot in Matagorda Bay because it has an unusually high number of bird species along the shoreline that can be readily observed as compared to the surrounding area. Our domain expert commented that the area found by the mean scan was an area that was not particularly high in species. Upon inspecting the models for the inside versus outside region, we found that the models indicate that observers appear to find less species initially inside that area than outside that area.

The mean scan and QSSS algorithms both found very different but meaningful regions for the BCRs. We hypothesize that the QSSS is finding unusual areas in terms of the observation process for more skilled observers (as in BCR 31) and we will continue our analysis on other BCRs in future work.

5 FUTURE WORK AND CONCLUSION

The QSSS discovers unusual spatial regions that differ from the surrounding area. The inner loop of the algorithm relies on comparing quantile regressions fit to data from inside and outside a region under consideration. To perform these comparisons efficiently, we developed an incremental rank test, which is over an order of magnitude faster than a naive implementation. Our results on simulated data and on three real-world datasets show that QSSS enables a new type of analysis for spatial data that is different from mean-based methods and that the QSSS is also robust to outliers. For future work, we would like to investigate reporting the top K most unusual regions rather than the top 1 and we would also like to extend our work to find unusual regions in both space and time.

ACKNOWLEDGEMENTS

We thank our collaborators from eButterfly (Katy Prudic, Max Larrivee, Jeffrey Oliver and Jeremy Kerr) and eBird (Daniel Fink, Chris Wood). Moore was supported by NSF grant CCF-1521687. This material is based upon work while Wong was serving at NSF. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NSF.

References

Abrams, A., Kleinman, K., and Kulldorff, M. (2010). Gumbel based p-value approximations for spatial scan statistics. *Int J Health Geogr*, 9(1):61.

Best, N., Richardson, S., and Thomson, A. (2005). A comparison of Bayesian spatial models for disease mapping. *Stat Methods Med Res*, 14(1):3559.

Caruthers, A. (2016). Mapping poverty in the appalachian region. https://www.communitycommons.org/2016/08/mapping-poverty-in-the-appalachian-region/.

DePillis, L. (2018). What america can learn from cities with super-low unemployment. http://money.cnn.com/2018/01/12/news/economy/cities-unemployment/index.html.

Duczmal, L. and Assuncao, R. (2004). A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters. *Comput Stat Data Anal*, 45:269286.

Golub, G. and Loan, C. V. (2012). *Matrix computations*, volume 3. JHU Press.

Gutenbrunner, C. and Jureckov, J. (1992). Regression rank scores and regression quantiles. *The Annals of Statistics*, pages 305–330.

Gutenbrunner, C., Jurekov, J. K. R. S., Koenker, R., and Portnoy, S. (1993). Tests of linear hypotheses based on regression rank scores. *J Nonparametr Stat*, 2(4):307–331.

Kahle, D. and Wickham, H. (2013). ggmap: Spatial visualization with ggplot2. *The R Journ al*, 5(1):144–161.

Kelling, S., Johnston, A., Hochachka, W. M., Iliff, M., Fink, D., Gerbracht, J., Lagoze, C., Sorte, F. A. L., Moore, T., Wiggins, A., Wong, W.-K., Wood, C., and Yu, J. (2015). Can observation skills of citizen scientists be estimated using species accumulation curves? *PLoS ONE*, 10(10):e0139600.

Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica*, 46(1):33–50.

Koenker, R. and Machado, J. A. (1999). Goodness of fit and related inference processes for quantile regression. *J Am Stat Assoc*, 94(448):1296–1310.

Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics: Theory and Methods*, 26(6):1481–1496.

Lan, L., Malbasa, V., and Vucetic, S. (2014). Spatial scan for disease mapping on a mobile population. In *Proceedings of the 28th AAAI Conference*, pages 431–437.

Machado, J. and Silva, J. S. (2002). Quantiles for counts. *J Am Stat Assoc*, 100(472):1226–1237.

Macmillan, D. P. (2013). *Quantile Regression for Spatial Data*. Springer-Verlag, Berlin.

McFowland, III, E., Somanchi, S., and Neill, D. B. (2018). Efficient Discovery of Heterogeneous Treatment Effects in Randomized Experiments via Anomalous Pattern Detection. *ArXiv e-prints*, 1803.09159.

Mood, A. (1950). *Introduction to the Theory of Statistics*. McGraw Hill Book Co.

Moore, T. and Wong, W.-K. (2015). Discovering hotspots and coldspots of species richness in ebird data. In *Proceedings of the AAAI-15 Workshop on Computational Sustainability*.

Neill, D. and Moore, A. W. (2004). Rapid detection of significant spatial clusters. In *Proceedings of the 10th ACM SIGKDD Conference*, pages 256–265.

Neill, D. B. (2012). Fast subset scan for spatial pattern detection. *J R Stat Soc Series B Stat Methodol*, 74(2):337–360.

Parker, T. (2017). Download data. United States Department of Agriculture. https://www.ers.usda.gov/data-products/county-level-data-sets/county-level-data-sets-download-data/.

Prudic, K., McFarland, K., Oliver, J., Hutchinson, R., Long, E., Kerr, J., and Larrive, M. (2017). ebutterfly: leveraging massive online citizen science for butterfly conservation. http://www.mdpi.com/2075-4450/8/2/53/htm.

Reich, B. J., Fuentes, M., and Dunson, D. B. (2011). Bayesian spatial quantile regression. *J Am Stat Assoc*, 106(493):6–20.

Rousseeuw, P. J. and Leroy, A. M. (1987). *Robust Regression and Outlier Detection*. John Wiley and Sons.

Sherman, J. and Morrison, W. J. (1950). Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *The Annals of Mathematical Statistics*, 21(1):124–127.

Sullivan, B. L., Aycrigg, J. L., Barry, J. H., Bonney, R. E., Bruns, N., Cooper, C. B., Damoulas, T., Dhondt, A. A., Dietterich, T., Farnsworth, A., Fink, D., Fitzpatrick, J. W., Fredericks, T., Gerbracht, J., Gomes, C., Hochachka, W. M., Iliff, M. J., Lagoze, C., Sorte, F. L., Merrifield, M., Morris, W., Phillips, T. B., Reynolds, M., Rodewald, A. D., Rosenberg, K. V., Trautmann, N. M., Wiggins, A., Winkler, D. W., Wong, W.-K., Wood, C. L., Yu, J., and Kelling, S. (2014). The ebird enterprise: An integrated approach to development and application of citizen science. *Biological Conservation*, 169:31–40.