NIH Public Access
Author Manuscript
*J R Stat Soc Series B Stat Methodol*. Author manuscript; available in PMC 2016 January 01.

Published in final edited form as:
*J R Stat Soc Series B Stat Methodol*. 2015 January 1; 77(1): 59–83. doi:10.1111/rssb.12064.

# False Discovery Control in Large-Scale Spatial Multiple Testing

**Wenguang Sun**,
University of Southern California, Los Angeles, USA

**Brian J. Reich**,
North Carolina State University, Raleigh, USA

**T. Tony Cai**,
University of Pennsylvania, Philadelphia, USA

**Michele Guindani**, and
UT MD Anderson Cancer Center, Houston, USA

**Armin Schwartzman**
Harvard University, Boston, USA

## Summary

This article develops a unified theoretical and computational framework for false discovery control in multiple testing of spatial signals. We consider both point-wise and cluster-wise spatial analyses, and derive oracle procedures which optimally control the false discovery rate, false discovery exceedance and false cluster rate, respectively. A data-driven finite approximation strategy is developed to mimic the oracle procedures on a continuous spatial domain. Our multiple testing procedures are asymptotically valid and can be effectively implemented using Bayesian computational algorithms for analysis of large spatial data sets. Numerical results show that the proposed procedures lead to more accurate error control and better power performance than conventional methods. We demonstrate our methods for analyzing the time trends in tropospheric ozone in eastern US.

### Keywords

Compound decision theory; false cluster rate; false discovery exceedance; false discovery rate; large-scale multiple testing; spatial dependency

## 1. Introduction

Let $\boldsymbol{X} = \{X(s) : s \in S\}$ be a random field on a spatial domain $S$:

$$X(s)=\mu(s)+\varepsilon(s), \quad (1.1)$$

where $\mu(s)$ is the unobserved random process and $\varepsilon(s)$ is the noise process. Assume that there is an underlying state $\theta(s)$ associated with each location $s$ with one state being dominant ("background"). In applications, an important goal is to identify locations that exhibit significant deviations from background. This involves conducting a large number of spatially correlated tests simultaneously. It is desirable to maintain good power for detecting

true signals while guarding against too many false positive findings. The false discovery rate (FDR, Benjamini and Hochberg, 1995) approach is particularly useful as an exploratory tool to achieve these two goals and has received much attention in the literature. In a spatial setting, the multiple comparison issue has been raised in a wide range of problems such as brain imaging (Genovese et al., 2002; Heller et al., 2006; Schwartzman et al., 2008), disease mapping (Green and Richardson, 2002), public health surveillance (Caldas de Castro and Singer, 2006), network analysis of genome-wide association studies (Wei and Li, 2007; Chen et al., 2011), and astronomical surveys (Miller et al., 2007; Meinshausen et al., 2009).

Consider the following example for analyzing time trends in tropospheric ozone in the Eastern US. Ozone is one of the six criteria pollutants regulated by the US EPA under the Clean Air Act and has been linked with several adverse health effects. The EPA has established a network of monitors for regulation of ozone, as shown in Figure 1a. We are interested in identifying locations with abrupt changing ozone levels using the ozone concentration data collected at monitoring stations. In particular, we wish to study the ozone process for predefined sub-regions, such as counties or states, to identify interesting sub-regions. Similar problems may arise from disease mapping problems in epidemiology, where the goal is to identify geographical area with elevated disease incidence rates. It is also desirable to take into account region specific variables, such as the population in or the area of a county, to reflect the relative importance of each sub-region.

Spatial multiple testing poses new challenges which are not present in conventional multiple testing problems. Firstly, one only observes data points at a discrete subset of the locations but often needs to make inference everywhere in the spatial domain. It is thus necessary to develop a testing procedure which effectively exploits the spatial correlation and pools information from nearby locations. Secondly, a finite approximation strategy is needed for inference in a continuous spatial domain – otherwise an uncountable number of tests needs to be conducted, which is impossible in practice. Thirdly, it is challenging to address the strong dependency in a two or higher dimensional random field. Finally, in many important applications, it is desirable to aggregate information from nearby locations to make cluster-wise inference, and to incorporate important spatial variables in the decision-making process. The goal of the present paper is to develop a unified theoretical and computational framework to address these challenges.

The impact of dependence has been extensively studied in the multiple testing literature. Efron (2007) and Schwartzman and Lin (2011) show that correlation usually degrades statistical accuracy, affecting both estimation and testing. High correlation also results in high variability of testing results and hence the irreproducibility of scientific findings; see Owen (2005), Finner et al. (2007) and Heller (2010) for related discussions. Meanwhile, it has been shown that the classical Benjamini-Hochberg (BH) procedure is valid for controlling the false discovery rate (FDR, Benjamini and Hochberg, 1995) under different dependency assumptions, indicating that it is safe to apply conventional methods as if the tests were independent (see Benjamini and Yekutieli, 2001; Sarkar, 2002; Wu, 2008; Clarke and Hall, 2009; among others). Another important research direction in multiple testing is the optimality issue under dependency. Sun and Cai (2009) introduced an asymptotically optimal FDR procedure for testing hypotheses arising from a hidden Markov model (HMM)

and showed that the HMM dependency can be exploited to improve the existing *p*-value based procedures. This demonstrates that informative dependence structure promises to increase the precision of inference. For example, in genome-wide association studies, signals from individual markers are weak, hence a number of approaches have been developed to increase statistical power by aggregating multiple markers and exploiting the high correlation among adjacent loci (e.g., see Peng et al., 2009; Wei et al., 2009; Chen et al., 2011). When the intensities of signals have a spatial pattern, it is expected that incorporating the underlying dependency structure can significantly improve the power and accuracy of conventional methods. This intuition is supported both theoretically and numerically in our work.

In this article, we develop a compound decision theoretic framework for spatial multiple testing and propose a class of asymptotically optimal data-driven procedures that control the FDR, false discovery exceedance (FDX) and false cluster rate (FCR), respectively. The widely used Bayesian modeling framework and computational algorithms are adopted to effectively extract information from large spatial datasets. We discuss how to summarize the fitted spatial models using posterior sampling to address related multiple testing problems. The control of the FDX and FCR is quite challenging from the classical perspective. We show that the FDR, FDX and FCR controlling problems can be solved in a unified theoretical and computational framework. A finite approximation strategy for inference on a continuous spatial domain is developed and it is shown that a continuous decision process can be described, within a small margin of error, by a finite number of decisions on a grid of pixels. This overcomes the limitation of conventional methods which can only test hypotheses on a discrete set of locations where observations are available. Simulation studies are carried out to investigate the numerical properties of the proposed methods. The results show that by exploiting the spatial dependency, the data-driven procedures lead to better rankings of hypotheses, more accurate error control and enhanced power.

The proposed methods are developed in a frequentist framework and aims to control the frequentist FDR. The Bayesian computational framework, which involves hierarchical modeling and MCMC computing, provides a powerful tool to implement the data-driven procedures. When the goal is to control the FDR and tests are independent, our procedure coincides with the Bayesian FDR approach originally proposed by Newton et al. (2004). Müller et al. (2004) and Müller et al. (2007) showed that controlling the Bayesian FDR implies FDR control. However, those type of results do not immediately extend to correlated tests (see Remark 4 in Pacifico et al. (2004) and Guindani et al. (2009)). In addition, existing literature on Bayesian FDR analysis (Müller et al., 2004, 2007 and Bogdan et al., 2008) has focused on the point-wise FDR control only, and the issues related to FDX and FCR have not been discussed. In contrast, we develop a unified theoretical framework and propose testing procedures for controlling different error rates. The methods are attractive by providing effective control of the widely used frequentist FDR.

The article is organized as follows. Section 2 introduces appropriate false discovery measures in a spatial setting. Section 3 presents a decision theoretic framework to characterize the optimal decision rule. In Section 4, we propose data-driven procedures and discuss the computational algorithms for implementation. Sections 5 and 6 investigate the

numerical properties of the proposed procedures using both simulated and real data. The proofs and technical details in computation are given in the Appendix.

## 2. False Discovery Measures for Spatial Multiple Testing

In this section we introduce some notation and important false discovery measures in a random field, following the works of Pacifico et al. (2004) and Benjamini and Heller (2007). Both point-wise analysis and cluster-wise analysis will be considered.

### 2.1. Point-wise inference

Suppose for each location $s$, we are interested in testing the hypothesis

$$H_0(s):\mu(s) \in A \quad \text{versus} \quad H_1(s):\mu(s) \in A^c, \quad (2.1)$$

where $A$ is the indifference region, e.g. $A = \{\mu : \mu \le \mu_0\}$ for a one-sided test and $A = \{\mu : |\mu| \le \mu_0\}$ for a two-sided test. Let $\theta(s) \in \{0, 1\}$ be an indicator such that $\theta(s) = 1$ if $\mu(s) \in A^c$ and $\theta(s) = 0$ otherwise. Define $S_0 = \{s \in S : \theta(s) = 0\}$ and $S_1 = \{s \in S : \theta(s) = 1\}$ as the null and non-null areas, respectively. In a point-wise analysis, a decision $\delta(s)$ is made for each location $s$. Let $\delta(s) = 1$ if $H_0(s)$ is rejected and $\delta(s) = 0$ otherwise. The decision rule for the whole spatial domain $S$ is denoted by $\boldsymbol{\delta} = \{\delta(s) : s \in S\}$. Then $R = \{s \in S : \delta(s) = 1\}$ is the rejection area, and $S_{\text{FP}} = \{s \in S : \theta(s) = 0, \delta(s) = 1\}$ and $S_{\text{FN}} = \{s \in S : \delta(s) = 1, \delta(s) = 0\}$ are the false positive and false negative areas, respectively. Let $\nu(\cdot)$ denote a measure on $S$, where $\nu(\cdot)$ is the Lebesgue measure if $S$ is continuous and a counting measure if $S$ is discrete. When the interest is to test hypotheses at individual locations, it is natural to control the false discovery rate (FDR, Benjamini and Hochberg, 1995), a powerful and widely used error measure in large-scale testing problems. Let $c_0$ be a small positive value. In practice if the rejection area is too small, then we can proceed as if no rejection is made. Define the false discovery proportion as

$$\text{FDP} = \frac{\nu(S_{\text{FP}})}{\nu(R)} I\{\nu(R) > c_0\}. \quad (2.2)$$

The FDR is the expected value of the FDP: FDR = $E$(FDP). Alternative measures to the FDR include the marginal FDR, mFDR = $E\{\nu(S_{\text{FP}})\}/E\{\nu(R)\}$(Genovese and Wasserman, 2002) and positive FDR (pFDR, Storey, 2002).

The FDP is highly variable under strong dependence (Finner and Roters, 2002; Finner et al., 2007; Heller, 2010). The false discovery exceedance (FDX), discussed in Pacifico et al. (2004), Lehmann and Romano (2005), and Genovese and Wasserman (2006), is a useful alternative to the FDR. FDX control takes into account the variability of the FDP, and is desirable in a spatial setting where the tests are highly correlated. Let $0 \le t \le 1$ be a pre-specified *tolerance level*, the FDX at level $\tau$ is $\text{FDX}_\tau = P(\text{FDP} > t)$, the tail probability that the FDP exceeds a given bound.

To evaluate the power of a multiple testing procedure, we use the missed discovery rate MDR = $E\{\nu(S_{FN})\}$. Other power measures include the false non-discovery rate and average power; our result can be extended to these measures without essential difficulty. A multiple testing procedure is said to be *valid* if the FDR can be controlled at the nominal level and *optimal* if it has the smallest MDR among all valid testing procedures.

### 2.2. Cluster-wise inference

When the interest is on the behavior of a process over sub-regions, the testing units become spatial clusters instead of individual locations. Combining hypotheses over a set of locations naturally reduces multiplicity and correlation. In addition, set-wise analysis improves statistical power as data in a set may show an increased signal to noise ratio (Benjamini and Heller, 2007). The idea of set-wise/cluster-wise inference has been successfully applied in many scientific fields including large epidemiological surveys (Zaykin et al., 2002), meta-analysis of microarray experiments (Pyne et al., 2006), gene set enrichment analysis (Subramanian et al., 2005) and brain imaging studies (Heller et al., 2006).

The definition of a cluster is often application specific. Two existing methods for obtaining spatial clusters include: (i) to aggregate locations into regions according to available prior information (Heller et al., 2006; Benjamini and Heller, 2007); (ii) to conduct a *preliminary* point-wise analysis and define the clusters after inspection of the results (Pacifico et al., 2004). Let $\mathcal{C} = \{C_1, \cdots, C_K\}$ denote the set of (known) clusters of interest. We can form for each cluster $C_k$ a *partial conjunction null hypothesis* (Benjamini and Heller, 2008), $H_0(C_k) : \pi_k \leq \gamma$ versus $H_1(C_k) : \pi_k > \gamma$, where $\pi_k = \nu(\{s \in C_k : \theta(s) = 1\})/\nu(C_k)$ is the proportion of non-null locations in $C_k$ and $0 \leq \gamma \leq 1$ is a pre-specified tolerance level. The null hypothesis could also be defined in terms of the average activation amplitude $\bar{\mu}(C_k) = \nu(C_k)^{-1} \int_{C_k} \mu(s)ds$, that is, $H_0(C_k) : \bar{\mu}(C_k) \leq \bar{\mu}_0$ versus $H_1(C_k) : \bar{\mu}(C_k) > \bar{\mu}_0$, for some pre-specified $\bar{\mu}_0$. Each cluster $C_k$ is associated with an unknown state $\vartheta_k \in \{0, 1\}$, indicating whether the cluster shows a signal or not. Let $S_0 = \cup_{k:\vartheta_k=0} C_k$ and $S_1 = \cup_{k:\vartheta_k=1} C_k$ denote the corresponding null and non-null areas, respectively. In cluster-wise analysis, a universal decision rule is taken for all locations in the cluster, i.e. $\delta(s) = \Delta_k$, for all $s \in C_k$. The decision rule is $\boldsymbol{\Delta} = (\Delta_1, \cdots, \Delta_K)$. Then, the rejection area is $R = \cup_{k:\Delta_k=1} C_k$.

In many applications it is desirable to incorporate the cluster size or other spatial variables in the error measure. We consider the weighted multiple testing framework, first proposed by Benjamini and Hochberg (1997) and further developed by Benjamini and Heller (2007) in a spatial setting, to reflect the relative importance of various clusters in the decision process. The general strategy involves the modifications of either the error rate to be controlled, or the power function to be maximized, or both. Define the false cluster rate

$$\mathrm{FCR} = E\left\{\frac{\sum_k w_k(1-\vartheta_k)\Delta_k}{(\sum_k w_k\Delta_k) \vee 1}\right\}, \quad (2.3)$$

where $w_k$ are cluster specific weights which are often pre-specified in practice. For example, one can take $w_k = \nu(C_k)$, the size of a cluster, to indicate that a false positive cluster with

larger size would account for a larger error. Similarly, we define the marginal FCR as mFCR = $E\{\Sigma_k w_k(1 - \vartheta_k)\Delta k\}/E(\Sigma_k w_k\Delta_k)$.

We can see that in the definition of the FCR, a large false positive cluster is penalized by a larger weight. At the same time, correctly identifying a large cluster that contains signal may correspond to a greater gain; hence the power function should be weighted as well. For example, in epidemic disease surveillance, it is critical to identify aberrations in areas with larger populations where interventions should be first put into place. To reflect that some areas are more crucial, we give higher penalty in the loss function if an important cluster is missed. The same weights $w_k$ are used as a reflective of proportional error and gain. Define the missed cluster rate MCR = $E\{\Sigma_k w_k\vartheta_k (1 - \Delta k)\}$. In cluster-wise analysis the goal is to control the FCR while minimizing the MCR.

## 3. Compound Decision Theory for Spatial Multiple Testing

In this section we formulate a compound decision theoretic framework for spatial multiple testing problems and derive a class of oracle procedures for controlling the FDR, FDX and FCR, respectively. Section 4 develops data-driven procedures to mimic the oracle procedures and discusses their implementations in a Bayesian computational framework.

### 3.1. Oracle procedures for point-wise analysis

Let $X_1, \cdots, X_n$ be observations at locations $S^*=\{s_1^*, \cdots, s_n^*\}$. In point-wise analysis, $S^*$ is often a subset of $S$, and we need to make decisions at locations where no observation is available; therefore the problem is different from conventional multiple testing problems where each hypothesis has its own observed data. It is therefore necessary to exploit the spatial dependency and pool information from nearby observations. In this section, we discuss optimal results on point-wise FDR analysis from a theoretical perspective.

The optimal testing rule is derived in two steps: first the hypotheses are ranked optimally and then a cutoff is chosen along the rankings to control the FDR precisely. The optimal result on ranking is obtained by connecting the multiple testing problem to a weighted classification problem. Consider a general decision rule $\boldsymbol{\delta} = \{\delta(s) : s \in S\}$ of the form:

$$\delta(s)=I(T(s)<t), \quad (3.1)$$

where $T(s) = T_s(X^n)$ is a test statistic, $T_s(\cdot)$ is a function which maps $X^n$ to a real value, and $t$ is a universal threshold for all $T(s)$, $s \in S$. To separate a signal ($\theta(s) = 1$) from noise ($\theta(s) = 0$), consider the loss function

$$L(\boldsymbol{\theta}, \boldsymbol{\delta})=\lambda\nu(S_{\mathrm{FP}})+\nu(S_{\mathrm{FN}}), \quad (3.2)$$

where $\lambda$ is the penalty for false positives, and $S_{\mathrm{FP}}$ and $S_{\mathrm{FN}}$ are false positive and false negative areas defined in Section 2. The goal of a weighted classification problem is to find a decision rule $\boldsymbol{\delta}$ to minimize the classification risk $R = E\{L(\boldsymbol{\theta}, \boldsymbol{\delta})\}$. It turns out that the optimal solution to the weighted classification problem is also optimal for mFDR control when a monotone ratio condition (MRC) is fulfilled. Specifically, define $G_j(t) = \int_S P(T(s) <$

$t$, $\theta(s) = j)d\nu(s)$, $j = 0, 1$. $G_0(t)$ can be viewed as the overall "Type I error" function at all locations in $S$ where the null is true, and $G_1(t)$ can be viewed as the overall "power" function at all locations in $S$ where the alternative is true. In Section XXX of the supplementary material, we show that it is reasonable to assume that $G_0$ and $G_1$ are differentiable when $X(s)$ are continuous random variables on $S$. Denote by $g_0(t)$ and $g_1(t)$ their derivatives. The MRC can be stated as

$$g_1(t)/g_0(t) \text{ is monotonically decreasing in } t. \quad (3.3)$$

The MRC is a reasonable and mild regularity condition in multiple testing which ensures that the mFDR increases in $t$ and the MDR decreases in $t$. Therefore in order to minimize the MDR, we choose the *largest* threshold subject to mFDR ≤ $a$. The MRC reduces to the monotone likelihood ratio condition (MLRC, Sun and Cai, 2007) when the tests are independent. The MLRC is satisfied by the $p$-value when the $p$-value distribution is concave (Genovese and Wasserman, 2002). In a hidden Markov model, the MRC is satisfied by the local index of significance (Sun and Cai, 2009).

Let $\boldsymbol{X}^n = \{X_1, \cdots, X_n\}$. Consider a class of decision rules $\mathcal{D}$ of the form $\boldsymbol{\delta} = \{I\{T(s) < t\} : s \in S\}$, where $\boldsymbol{T} = \{T(s) : s \in S\}$ satisfies the MRC (3.3). The next theorem derives the optimal classification statistic and gives the optimal multiple testing rule for mFDR control.

**Theorem 1—**Let $\Psi$ be the collection of all parameters in random field (1.1) and we assume that $\Psi$ is known. Define the oracle statistic

$$T_{OR}(s) = P_\Psi\{\theta(s) = 0 | \boldsymbol{X}^n\} \quad (3.4)$$

and assume that $G_j(t)$ are differentiable, $j = 0, 1$. Then

**a.** The classification risk is minimized by $\boldsymbol{\delta} = \{\delta(s) : s \in S\}$, where

$$\delta(s) = I\left\{T_{OR}(s) < (1+\lambda)^{-1}\right\} \text{ and.} \quad (3.5)$$

**b.** Let $\boldsymbol{T}_{OR} = \{\boldsymbol{T}_{OR}(s) : s \in S\}$. Then $\boldsymbol{T}_{OR}$ satisfies the MRC (3.3).

**c.** There exists an oracle threshold

$$t_{OR}(\alpha) = \sup\{t : mFDR(t) \le \alpha\} \quad (3.6)$$

such that the oracle testing procedure

$$\boldsymbol{\delta}_{OR} = \{I[T_{OR}(s) < t_{OR}(\alpha)] : s \in S\} \quad (3.7)$$

has the smallest MDR among all $\alpha$-level mFDR procedures in $\mathcal{D}$.

**Remark 1—**Theorem 1 implies that, under the MRC (3.3), the optimal solution to a multiple testing problem (for mFDR control at level $\alpha$) is the solution to an equivalent

weighted classification problem with the loss function (3.2) and penalty $\lambda(a) = \{1 - t_{OR}(a)\}/t_{OR}(a)$. The procedure is called an "oracle" procedure because it relies on the knowledge of the true distributional information and the optimal threshold $t_{OR}(a)$, which are typically unknown in practice.

**Remark 2—**The result in Theorem 1(c) can be used to develop an FDX-controlling procedure. First the hypotheses are ranked according to the values of $T_{OR}(s)$. Since the MDR decreases in $t$, we choose the largest $t$ subject to the constraint on the FDX. The oracle FDX procedure is then given by

$$\boldsymbol{\delta}_{OR,FDX}=\{I(T_{OR}(s)<t_{OR,FDX}):s \in S\}, \quad (3.8)$$

where $t_{OR,FDX} = \arg\max_t\{\mathrm{FDX}_t(t) \le a\}$ is the oracle FDX threshold.

### 3.2. Oracle procedure for cluster-wise analysis

Let $\mathcal{H}_1, \cdots, \mathcal{H}_K$ be the hypotheses on the $K$ clusters $\mathcal{C} = \{C_1, \cdots, C_K\}$. The true states of nature (e.g. defined by partial conjunction nulls) can be represented by a binary vector $\boldsymbol{\vartheta} = \{\vartheta_k : k = 1, \cdots, K\} \in \{0, 1\}^K$. The decisions based on $\boldsymbol{X}^n = \{X_1, \cdots, X_n\}$ are denoted by $\boldsymbol{\Delta} = (\Delta_1, \cdots, \Delta_K) \in \{0, 1\}^K$. The goal is to find $\boldsymbol{\Delta}$ to minimize the MCR subject to FCR $\le a$. It is natural to consider the loss function

$$L_w(\boldsymbol{\vartheta}, \boldsymbol{\Delta})=\sum_{k=1}^{K}\{\lambda w_k(1-\vartheta_k)\Delta_k+\omega_k\vartheta_k(1-\Delta_k)\}, \quad (3.9)$$

where $\lambda$ is the penalty for false positives. As one would expect from Remark 1, the FCR control problem can be solved by connecting it to a weighted classification problem with a suitably chosen $\lambda$. In practice $\lambda$ is an unknown function of the FCR level $a$ and needs to be estimated. In contrast, the weights $w_k$ are pre-specified. Let $T_k$ be a cluster-wise test statistic. Define $p_k = P(\vartheta_k = 1)$, $G_{jk}(t) = P(T_k < t|\vartheta_k = j)$ and $g_{jk}(t) = (d/dt)G_{jk}(t)$, $j = 0, 1$. Consider the generalized monotone ratio condition (GMRC):

$$\frac{\sum_{k=1}^{K} w_k p_k g_{1k}(t)}{\sum_{k=1}^{K} w_k(1-p_k) g_{0k}(t)} \text{is decreasing in } t. \quad (3.10)$$

The GMRC guarantees that the MCR is decreasing in the FCR. Let $\mathcal{D}_c$ be the class of decision rules of the form $\boldsymbol{\Delta} = \{I(T_k < t) : k = 1, \cdots, K\}$, where $\boldsymbol{T} = (T_1, \cdots, T_k)$ satisfies the GMRC (3.10). We have the following results.

**Theorem 2—**Let $\Psi$ be the collection of all parameters in random field (1.1). Assume that $\Psi$ is known. Define the oracle test statistic

$$T_{OR}(C_k)=P_{\Psi}(\vartheta_k=0|\boldsymbol{X}^n) \quad (3.11)$$

and assume that $G_{jk}(t)$ are differentiable, $k = 1, \cdots, K$, $j = 0, 1$. Then

**a.** the classification risk with loss (3.9) is minimized by $\boldsymbol{\Delta} = \{\Delta_k : k = 1, \cdots, K\}$, where

$$\Delta_k = I\{T_{OR}(C_k) < (1+\lambda)^{-1}\}. \quad (3.12)$$

**b.** $\boldsymbol{T}_{OR} = \{T_{OR}(C_k) : k = 1, \cdots, K\}$ satisfies the GMRC (3.10).

**c.** Define the oracle mFCR procedure

$$\boldsymbol{\Delta}_{OR} = \{\Delta_{OR}^k : k = 1, \cdots, K\} = \{I(T_{OR}(C_k) < t_{OR}^c(\alpha)) : k = 1, \cdots, K\}, \quad (3.13)$$

where $t_{OR}^c(\alpha) = \sup\{t : mFCR(t) \leq \alpha\}$ is the oracle threshold. Then the oracle mFCR procedure (3.13) has the smallest MCR among all $\alpha$-level mFCR procedures in $\mathcal{D}_c$.

In Section 4 we develop data-driven procedures to mimic the above oracle procedures.

## 4. False Discovery Controlling Procedures and Computational Algorithms

The oracle procedures are difficult to implement because (i) it is impossible to make an uncountable number of decisions when $S$ is continuous, and (ii) the optimal threshold $t_{OR}$ and the oracle test statistics are essentially unknown in practice. This section develops data-driven procedures for FDR, FDX and FCR analyses to overcome these difficulties. We first describe how a continuous decision process can be approximated, within a small margin of error, by a finite number of decisions on a grid of pixels, then discuss how to calculate the test statistics.

### 4.1. FDR and FDX procedures for point-wise inference

To avoid making inference at every point, our strategy is to divide a continuous $S$ into $m$ "pixels," pick one point in each pixel, and use the decision at that point to represent all decisions in the pixel. We show that as the partition becomes finer, the representation leads to an asymptotically equivalent version of the oracle procedure.

Let $\cup_{i=1}^m S_i$ be a partition of $S$. A good partition in practice entails dividing $S$ into roughly homogeneous pixels, within which $\mu(s)$ varies at most a small constant. This condition is stated precisely as Condition 2 when we study the asymptotic validity of the proposed method. Next take a point $s_i$ from each $S_i$. In practice it is natural to use the center point of $S_i$ but we shall see that the choice of $s_i$ is nonessential as long as Condition 2 is fulfilled. Let $T_{OR}^{(1)} \leq T_{OR}^{(2)} \leq \cdots \leq T_{OR}^{(m)}$ denote the ordered oracle statistics defined by (3.4) and $S_{(i)}$ the region corresponding to $T_{OR}^{(i)}$. The following testing procedure is proposed for FDR control.

**Procedure 1—**(FDR control). Define $R_j = \cup_{i=1}^j S_{(i)}$ and

$$r=\max\left\{j:\nu(R_j)^{-1}\sum_{i=1}^{j}T_{OR}^{(i)}\nu S_{(i)}\le\alpha\right\}. \quad (4.1)$$

The rejection area is given by $R=\cup_{i=1}^{r}S_{(i)}$.

Next we propose an FDX procedure at level ($\gamma$, $\alpha$) based on the same ranking and partition schemes. Let $R_j^m=\{s_1,\cdots,s_m\}\cap R_j$ be the set of rejected representation points. The main idea of the following procedure is to first obtain a discrete version of the $\mathrm{FDX}_\tau$ based on a finite approximation, then estimate the actual FDX level for different cutoffs, and finally choose the largest cutoff which controls the FDX.

**Procedure 2—**(FDX control). Pick a small $\varepsilon_0 > 0$. Define $R_j=\cup_{i=1}^{j}S_{(i)}$ and

$$\mathrm{FDX}_{\tau,j}^m=P_\Psi\left(\nu(R_j)^{-1}\sum_{s_i\in R_j^m}\{1-\theta(s_i)\}\nu(S_i)>\tau-\varepsilon_0|\boldsymbol{X}^n\right), \quad (4.2)$$

where $\theta(s_i)$ is a binary variable indicating the true state at location $s_i$. Let $r=\max\{j:\mathrm{FDX}_{\tau,j}^m\le\alpha\}$, then the rejection region is given by $R=\cup_{i=1}^{r}S_{(i)}$.

Now we study the theoretical properties of Procedures 1 and 2. The first requirement is that $\mu(s)$ is a smooth process that does not degenerate at the boundaries of the indifference region $A = [A_l, A_u]$. To see why such a requirement is needed, define

$$\mu^m(s)=\sum_{i=1}^{m}\mu(s_i)I(s\in S_i),\theta(s)=I\{\mu(s)\in A^c\}\text{ and }\theta^m(s)=I\{\mu^m(s)\in A^c\}.$$

For a particular realization of $\mu(s)$, $\mu^m(s)$ is a *simple function* which takes a finite number of values according to the partition $S = \cup_i S_i$, and converges to $\mu(s)$ point-wise as the partition becomes finer. Note that at locations close to the boundaries, a small difference between $\mu^m(s)$ and $\mu(s)$ can lead to different $\theta(s)$ and $\theta^m(s)$. The following condition, which states that $\mu(s)$ does not degenerate at the boundaries, guarantees that $\theta(s) \neq \theta^m(s)$ only occurs with a small chance when $|\mu^m(s) - \mu(s)|$ is small. The condition holds when $\mu(s)$ is a *continuous* random variable.

**Condition 1:** Let $A = [A_l, A_u]$ be the indifference region and $\varepsilon$ a small positive constant. Then $\int_S P(A_* - \varepsilon < \mu(s) < A_* + \varepsilon)d\nu(s) \to 0$ as $\varepsilon \to 0$, for $A_* = Al$ or $Au$.

To achieve asymptotic validity, the partition $S = \cup_i S_i$ should yield roughly homogeneous pixels so that the decision at point $s_i$ is a good representation of the decision process on pixel $S_i$. Consider the event that the variation of $\mu(s)$ on a pixel exceeds a small constant. The next condition guarantees that the event only occurs with a vanishingly small chance. The

condition holds for the Gaussian and Matérn models that are used in our simulation study and real data analysis.

**Condition 2:** There exists a sequence of partitions $\{S=\cup_{i=1}^{m} S_i : m=1,2,\cdots\}$ such that for any given $\varepsilon > 0$, $\lim_{m\to\infty} \int_S P\{|\mu(s) - \mu^m(s)| \geq \varepsilon\} d\nu(s) = 0$.

Conditions 1 and 2 together guarantee that $\theta(s) = \theta^m(s)$ would occur with overwhelming probability when the partition becomes finer. See Lemma 2 in Section 7.

The next theorem shows that Procedures 1 and 2 are *asymptotically* valid for FDR and FDX control, respectively. We first state the main result for a continuous $S$.

**Theorem 3—**Consider $T_{OR}(s)$ and $FDX_{\tau,j}^{m}$ defined in (3.4) and (4.2), respectively. Let $\{\cup_{i=1}^{m} S_i : m=1,2,\cdots\}$ be a sequence of partitions of $S$ satisfying Conditions 1–2. Then

- **a.** the *FDR* level of Procedure 1 satisfies $FDR \leq \alpha + o(1)$ when $m \to \infty$;
- **b.** the *FDX* level of Procedure 2 satisfies $FDX_{\tau} \leq \alpha + o(1)$ when $m \to \infty$.

When $S$ is discrete, the FDR/FDX control is *exact*; this (stronger) result follows directly from the proof of Theorem 3.

**Corollary 1—**When $S$ is discrete, a natural partition is $S=\cup_{i=1}^{m}\{s_i\}$. Then

- **a.** the *FDR* level of Procedure 1 satisfies $FDR \leq \alpha$;
- **b.** the *FDX* level of Procedure 2 satisfies $FDX_{\tau} \leq \alpha$.

### 4.2. FCR procedure for cluster-wise inference

Now we turn to the cluster-wise analysis. Let $C_1, \cdots, C_K$ be the clusters and $\mathcal{H}_1, \cdots, \mathcal{H}_K$ the corresponding hypotheses. We have shown that $T_{OR}(C_k) = P_{\Psi}(\vartheta_k = 0|\boldsymbol{X}^n)$ is the optimal statistic for cluster-wise inference.

**Procedure 3—**(FCR control). Let $T_{(1)}^{c} \leq \cdots \leq T_{(K)}^{c}$ be the ordered $T_{OR}(C_k)$ values, and $\mathcal{H}_{(1)}, \cdots, \mathcal{H}_{(K)}$ and $w_{(1)}, \cdots, w_{(K)}$ the corresponding hypotheses and weights, respectively. Let

$$r = \max\left\{ j : \frac{\sum_{k=1}^{j} w_k T_{(k)}^{c}}{\sum_{k=1}^{j} w_{(k)}} \leq \alpha \right\}.$$

Then reject $\mathcal{H}_{(1)}, \cdots, \mathcal{H}_{(r)}$.

The next theorem shows that Procedure 3 is valid for FCR control.

**Theorem 4—**Consider $T_{OR}(C_k)$ defined in (3.11). Then the FCR of Procedure 3 is controlled at the level $\alpha$.

It is not straightforward to implement Procedures 1–3 because $T_{OR}(s_i)$, $\mathrm{FDX}^m_{\tau,j}$ and $T_{OR}(C_k)$ are unknown in practice. Next section develops computational algorithms to estimate these quantities based on Bayesian spatial models.

### 4.3. Data-driven procedures and computational Algorithms

An important special case of (1.1) is the Gaussian random field (GRF), where the signals and errors are generated as Gaussian processes with means $\mu$ and 0, and covariance matrices $\Sigma_1$ and $\Sigma_2$, respectively. Let $\Psi$ be the collection of all hyperparameters in random field (1.1).

Consider a general random field model (1.1) defined on $S$. Let $\hat{\Psi}$ be the estimate of $\Psi$. Denote by $\boldsymbol{X}^n = (X_1, \cdots, X_n)$ the collection of random variables associated with locations $s_1^*, \cdots, s_n^*$. Further let $f(\boldsymbol{\mu}|\boldsymbol{X}^n, \hat{\Psi}) \propto \pi(\boldsymbol{\mu}) f(\boldsymbol{X}^n|\boldsymbol{\mu}, \hat{\Psi})$ be the posterior density function of $\boldsymbol{\mu}$ given $\boldsymbol{X}^n$ and $\hat{\Psi}$. The numerical methods for model fitting and parameter estimation in spatial models have been extensively studied (see Gelfand et al. (2010) and the references therein). We provide in the web appendix the technical details in a Gaussian random field model (GRFM), which is used in both the simulation study and real data example. The focus of discussion is on how the MCMC samples, generated from the posterior distribution, can be used to carry out the proposed multiple testing procedures.

We start with a point-wise testing problem with $H_0(s) : \mu(s) \in A$ versus $H_1(s) : \mu(s) \notin A$, $s \in S$. Let $S^m = (s_1, \cdots, s_m)$ denote the collection of the representative points based on partition $S = \cup_{i=1}^m S_i$. We only discuss the result for a continuous $S$ (the result extends to a discrete $S$ by simply taking $S^m = S$). Suppose the MCMC samples are $\{\hat{\boldsymbol{\mu}}_b^m : b = 1, \cdots, B\}$, where $\hat{\boldsymbol{\mu}}_b^m = (\hat{\mu}_b^{m,1}, \cdots, \hat{\mu}_b^{m,m})$ is a $m$-dimensional posterior sample indicating the magnitudes of the signals at locations $s_1, \cdots, s_m$ in replication $b$. Let $\hat{\theta}_b^{m,i} = I(\hat{\mu}_b^{m,i} \notin A)$ denote the estimated state of location $s_i$ in replication $b$. To implement Procedure 1 for FDR analysis, we need to compute

$$T_{OR}(s_i) = P_\Psi\{\theta(s_i) = 0|\boldsymbol{X}^n\} = \int I\{\mu(s_i) \in A\} f_{\boldsymbol{\mu}|\boldsymbol{X}^n}(\boldsymbol{\mu}|\boldsymbol{X}^n, \Psi) d\boldsymbol{\mu}.$$

It is easy to see that $T_{OR}(s_i)$ can be estimated by

$$\hat{T}_{OR}(s_i) = \frac{1}{B}\sum_{b=1}^{B} I(\hat{\mu}_b^{m,i} \in A) = \frac{1}{B}\sum_{b=1}^{B}(1 - \hat{\theta}_b^{m,i}). \quad (4.3)$$

To implement Procedure 2, note that the FDX defined in (4.2) can be written as

$$\mathrm{FDX}^m_{\tau,j} = \int I\left[\nu(R_j)^{-1} \sum_{s_i \in R_j^m} \{1 - \theta(s_i)\}\nu(S_i) > \tau - \varepsilon_0\right] f_{\boldsymbol{\mu}|\boldsymbol{X}^n}(\boldsymbol{\mu}|\boldsymbol{X}^n, \Psi) d\boldsymbol{\mu},$$

where $j$ is the number of points in $\boldsymbol{s}^m$ which are rejected, $R_j=\cup_{i=1}^{j} S_{(i)}$ is the rejection region and $R_j^m=S^m \cap R_j$ is subset of points in $S^m$ which are rejected. Given the MCMC samples $\{\hat{\boldsymbol{\mu}}_b^m: b=1,\cdots,B\}$, the $\mathrm{FDX}_{\tau,j}^m$ can be estimated as

$$\hat{\mathrm{FDX}}_{\tau,j}^m=\frac{1}{B}\sum_{i=1}^{B} I\left\{\nu(R_j)^{-1}\sum_{s_i\in R_j^m}(1-\hat{\theta}_b^{m,i})\nu(S_i)>\tau-\varepsilon_0\right\}. \quad (4.4)$$

Therefore Procedures 1 and 2 can be implemented by replacing $T_{OR}(s_i)$ and $\mathrm{FDX}_{\tau,j}^m$ by their estimates given in (4.3) and (4.4).

Next we turn to cluster-wise testing problems. Let $\cup_{i=1}^{m_k} S_i^k$ be a partition of $C_k$. Take a point $s_i^k$ from each $S_i^k$. Let $\boldsymbol{s}^{m_k}=(s^{m_k,1},\cdots,s^{m_k,m_k})$ be the collection of sampled points in cluster $C_k$, $m=\sum_{k=1}^{K} m_k$ be the count of points sampled in $S$ and $\boldsymbol{s}^m=(\boldsymbol{s}^{m_1},\cdots,\boldsymbol{s}^{m_K})$. If we are interested in testing partial conjunction of nulls $H_0(C_k): \pi_k \le \gamma$ versus $H_1(C_k): \pi_k > \gamma$, where $\pi_k = \nu(\{s\in C_k : \theta(s)=1\})/\nu(C_k)$, then we can define $\vartheta_k^m=I\{\sum_{i=1}^{m_k}\theta(s_i^k)\nu(S_i^k)>\gamma\nu(C_k)\}$ as an approximation to $\vartheta_k=I(\pi_k>\gamma)$. If the goal is to test average activation amplitude, i.e., $H_0(C_k): \bar{\mu}(C_k) \le \bar{\mu}_0$ versus $H_1(C_k): \bar{\mu}(C_k) > \bar{\mu}_0$, then we can define $\vartheta_k^m=I\{\sum_{i=1}^{m_k}\mu(s_i^k)\nu(S_i^k)>\bar{\mu}_0\nu(C_k)\}$. Let $T_{OR}^m(C_k)=P(\vartheta_k^m=0|\boldsymbol{X}^n)$.

To implement Procedure 3, we need to compute $T_{OR}^m(C_k)$. Suppose we are interested in testing partial conjunction of nulls, then

$$T_{OR}^m(C_k)=\int I\left\{\sum_{i=1}^{m_k}\theta(s_i^k)\nu(S_i^k)<\gamma\nu(C_k)\right\} f_{\boldsymbol{\mu}|\boldsymbol{X}^n}(\boldsymbol{\mu}|\boldsymbol{X}^n)d\boldsymbol{\mu}$$

Denote by $\hat{\boldsymbol{\mu}}_b^{m_k}=(\hat{\mu}_b^{m_k,1},\cdots,\hat{\mu}_b^{m_k,m_k})$ be the MCMC samples for cluster $C_k$ at points $\boldsymbol{s}^{m_k}$ in replication $b$, $b=1,\cdots,B$. Further let $\hat{\theta}_b^{m_k,i}=I(\hat{\mu}_b^{m_k,i}\notin A)$. Then $\int_{C_k}\theta(s)ds$ in a particular replication $b$ can be approximated by ${m_k}^{-1}\sum_{i=1}^{m_k}\hat{\theta}_b^{m_k,i}\nu(S_i^k)$ and the oracle statistic $T_{OR}^m(C_k)$ can be estimated by $\hat{T}_{OR}(C_k)=\frac{1}{B}\sum_{b=1}^{B} I\left\{\frac{1}{m_k}\sum_{i=1}^{m_k}\hat{\theta}_b^{m_k,i}\nu(S_i^k)<\gamma\nu(C_k)\right\}$. If the goal is to test average activation amplitude, $T_{OR}^m(C_k)$ can be estimated as $\hat{T}_{OR}(C_k)=\frac{1}{B}\sum_{b=1}^{B} I\left\{\nu(C_k)^{-1}\sum_{i=1}^{m_k}\hat{\mu}_b^{m_k,i}\nu(S_i^k)<\bar{\mu}_0\right\}$.

## 5. Simulation

We conduct simulation studies to investigate the numerical properties of the proposed methods. A significant advantage of our method over conventional methods is that the procedure is capable of carrying out analysis on a continuous spatial domain. However, to permit comparisons with other methods, we first limit the analysis to a Gaussian model for

testing hypotheses at the $n$ locations where the data points are observed. Therefore we have $m = n$. Then we conduct simulations to investigate, without comparison, the performance of our methods for a Matérn model to test hypotheses on a continuous domain based on a discrete set of data points. The R code for implementing our procedures is available at: http://www-bcf.usc.edu/~wenguans/Spatial-FDR-Software.

### 5.1. Gaussian model with observed data at all testing units

We generate data according to model (1.1) with both the signals and errors being Gaussian processes. Let $\|\cdot\|$ denote the Euclidean distance. The signal process $\mu$ has mean $\bar{\mu}$ and powered exponential covariance $\mathrm{Cov}[\mu(s), \mu(s')] = \sigma_\mu^2 \exp[-(\|s - s'\|/\rho_\mu)^k]$, while the error process $\varepsilon$ has mean zero and covariance $\mathrm{Cov}[\varepsilon(s), \varepsilon(s')] = (1-r)I(s = s') + r \exp[-(\|s - s'\|/\rho_\varepsilon)^k]$ so that $r \in [0, 1]$ controls the proportion of the error variance with spatial correlation. For each simulated data set, the process is observed at $n$ data locations generated as $s_1, \ldots, s_n \overset{i.i.d.}{\sim} \mathrm{Uniform}([0, 1]^2)$. For all simulations, we choose $n = 1000$, $r = 0.9$, $\bar{\mu} = -1$, and $\sigma_\mu = 2$; under this setting the expected proportion of positive observations is 33%. We generate data with $k = 1$ (exponential correlation) and $k = 2$ (Gaussian correlation), and for several values of the spatial ranges $\rho_\mu$ and $\rho_\varepsilon$. We only present the results for $k = 1$. The conclusions from simulations for $k = 2$ are similar in the sense that our methods control the FDR more precisely and are more powerful than competitive methods. For each combination of spatial covariance parameters, we generate 200 datasets. For simulations studying the effects of varying $\rho_\mu$ we fix $\rho_\varepsilon = 0.05$, and for simulations studying the effects of varying $\rho_\varepsilon$ we fix $\rho_\mu = 0.05$.

**5.1.1. Point-wise analysis—**For each of the $n$ locations, we test the hypotheses $H_0(s) : \mu(s) \le 0$ versus $H_1(s) : \mu(s) > 0$. We implement Procedure 1 (assuming the parameters are known, denoted by Oracle FDR) and the proposed method (4.3) using MCMC samples (denoted by MC FDR), and compare our methods with three popular approaches: the step-up $p$-value procedure (Benjamini and Hochberg, 1995), the adaptive $p$-value procedure (Benjamini and Hochberg, 2000; Genovese and Wasserman, 2002), and the FDR procedure proposed by Pacifico et al. (2004), which are denoted by BH, AP and PGVW FDR, respectively. We then implement Procedure 2 (assuming the parameters are known, denoted by Oracle FDX) and its MCMC version (MC FDX) based on (4.4), and compare the methods with the procedure proposed by Pacifico et al. (2004) (denoted by PGVW FDX).

We generate the MCMC samples using a Bayes model, where we assume that $k$ is known, and select uninformative priors: $\bar{\mu} \sim \mathrm{N}(0, 100^2)$, $\sigma_\mu^{-2} \sim \mathrm{Gamma}(0.1, 0.1)$, and $r$, $\rho_\mu$, $\rho_\varepsilon \sim$ Uniform(0, 1). The Oracle FDR/FDX procedure fixes these five hyperparameters at their true values to determine the effect of their uncertainty on the results. For each method and each data set we take $\alpha = \tau = 0.1$. Figures 2 plots the averages of the FDPs and MDPs over the 200 datasets.

We can see that the Oracle FDR procedure controls the FDR nearly perfectly. The MC FDR procedure, with uninformative priors on the unknown spatial correlation parameters, also has good FDR control, between 10–12%. As expected, the Oracle and MC FDX methods

tuned to control FDX are more conservative than the FDR methods, with observed FDR between 5–8%. The FDX methods become increasingly conservative as the spatial correlation of the signal increases to appropriately adjust for higher correlation between tests. In contrast, the BH, GW and PGVW procedures are very conservative, with much higher MDR levels. The distribution of FDP is shown in Figures 2c and 2d. In some cases, the upper tail of the FDP distribution approaches 0.2 for the MC FDR procedure. In contrast, the Oracle FDX method has FDP under 0.1 with very high probability for all correlation models. The MC FDX procedure also effectively controls FDX in most cases. The $95^{th}$ percentile of FDP is 0.15 for the smallest spatial range in Figure 2c, and less than 0.12 in all other cases.

**5.1.2. Cluster-wise analysis—**We use the same data-generating schemes and MCMC sampling methods as in the site-wise simulation in the previous section. The whole spatial domain is partitioned into a regular 7 ×7 grid, giving 49 clusters. We consider partial conjunction tests, where a cluster is rejected if more than 20% of the locations in the cluster contain true positive signal ($\mu(s) > 0$). We implement Procedure 3 (assuming parameters are known, denoted by Oracle FCR) and the corresponding MCMC method with non-informative priors (denoted by MC FCR). We compare our methods with the combined *p*-value approach proposed by Benjamini and Heller (2007). To make the methods comparable, we restrict the analysis to the $n = 1000$ data locations. We assume $\alpha = 0.1$ and an exponential correlation with $k = 1$. The simulation results are summarized in Figure 3. We can see that the Oracle FCR procedure control the FCR nearly perfectly. The MC FCR procedure has FCR slightly above the nominal level (less than 0.13 in all settings). In contrast the combined *p*-value method is very conservative, with FCR less than 0.02. Both Oracle FCR and MC FCR procedures have much lower missed cluster rate (MCR, the proportion of missed clusters which contain true signal in more than 20% of the locations).

### 5.2. Matérn model with missing data on the testing units

We use the model $z(s) = \mu(s) + \varepsilon(s)$, but generate the signals $\mu(s)$ and errors $\varepsilon(s)$ as Gaussian processes with Matérn covariance functions. The signal process $\{\mu(s) : s \in S\}$ has mean $\bar{\mu}$ and covariance $\mathrm{Cov}[\mu(s), \mu(t)] = \sigma_\mu^2 M(\|s-t\|; \rho_\mu, \kappa_\mu)$, where the Matérn correlation function, $M$, is determined by the spatial range parameter $\rho_\mu > 0$ and smoothness parameter $\kappa_\mu$. The error process $\{\varepsilon(s) : s \in S\}$ has mean zero and covariance $\mathrm{Cov}[\varepsilon(s), \varepsilon(t)] = (1 - r)I(s = t) + rM(\|s-t\|; \rho_\varepsilon, \kappa_\varepsilon)$ so that $r \in [0, 1]$ controls the proportion of the error variance with spatial correlation.

For each simulated data set, data are generated at $n$ spatial locations $s_i \overset{i.i.d.}{\sim} \mathrm{Uniform}(\mathcal{D})$, where $\mathcal{D}$ is the unit square $\mathcal{D} = [0, 1]^2$. Predictions are made and tests of $\mathcal{H}_0 : \mu(s) \le \mu_0$ versus $\mathcal{H}_1 : \mu(s) > \mu_0$ are conducted at the $m^2$ locations forming the $m \times m$ square grid covering $\mathcal{D}$. For all simulations, we choose $n = 200$, $m = 25$, $r = 0.9$, $\bar{\mu} = 0$, $\mu_0 = 6.41$, and $\sigma_\mu = 5$; under this setting the expected proportion of locations with $\mu(s) > \mu_0$ is 0.1. We generate data with two correlation functions: the first is exponential correlation with $\kappa_\mu = \kappa_\varepsilon = 0.5$ and $\rho_\mu = \rho_\varepsilon = 0.2$; the second has $\kappa_\mu = \kappa_\varepsilon = 2.5$ and $\rho_\mu = \rho_\varepsilon = 0.1$ which gives a smoother spatial process than the exponential but with roughly the same effective range (the distance at which

correlation is 0.05). For both correlation functions we generate 200 datasets, and fit the model with Matérn correlation function and priors $\mu \sim N(0, 1000^2)$, $\sigma_\mu^{-2} \sim \text{Gamma}(0.01, 0.01)$, $r \sim \text{Unif}(0, 1)$, and $\kappa_\mu, \kappa_\varepsilon, \rho_\mu, \rho_\varepsilon \overset{i.i.d.}{\sim} N(-1, 1)$. For comparison we also fit the oracle model with hyperparameters $\mu$, $\sigma_\mu$, $r$, $\kappa_\mu$, $\kappa_\varepsilon$, $\rho_\mu$, and $\rho_\varepsilon$ fixed at their true values.

The results are summarized in Figure 4. For data simulated with exponential correlation, both the data-driven procedure and oracle procedure with FDR-thresholding maintain proper FDR (0.09 for the data-driven procedure and 0.07 for the oracle procedure). The 0.9 quantile of FDP for the data-driven procedure with FDR control is over 0.20. In contrast, the 0.9 quantile for the data-driven procedure with FDX threshold is slightly below 0.1, indicating proper FDX control. The results for the Matérn data are similar, except that all models have lower missed discovery rate because with a smoother spatial surface the predictions are more precise.

We also evaluate the cluster FDR and FDX performance using this simulation design. Data were generated and the models were fit as for the point-wise simulation. We define the spatial cluster regions by first creating a 10×10 regular partition of $\mathcal{D}$, and then combing the final two columns and final two rows to give unequal cluster sizes. This gives 81 clusters and between 4 and 25 prediction locations per spatial cluster. We define a cluster as non-null if $\mu(s) > \mu_0$ for at least 20% of its locations. The FDR and FDX are controlled in all cases, and the power is much higher for the smoother Matérn data. The FDR and FDX of the data-driven procedures are comparable to the oracle procedure with these parameters fixed at their true values, suggesting the proposed testing procedure is efficient even in this difficult setting.

## 6. Ozone data analysis

To illustrate the proposed method, we analyze daily surface-level 8-hour average ozone for the eastern US. The data are obtained from the US EPA's Air Explorer Data Base (http://www.epa.gov/airexplorer/index.htm). Ozone regulation is based on the fourth highest daily value of the year. Therefore, for each of the 631 stations and each year from 1997–2005, we compute the fourth highest daily value of 8-hour average ozone. Our objective is to identify locations with a decreasing time trend in this yearly value.

The precision of our testing procedure shows some sensitivity to model misspecification; hence we must be careful to conduct exploratory analysis to ensure that the spatial model fits the data reasonably well. See the web appendix for a more detailed discussion. After some exploratory analysis, we fit the model $\hat{\beta}(s) = \beta(s) + w(s)\varepsilon(s)$, where $\hat{\beta}(s)$ and $w(s)$ are the estimated slope and its standard error, respectively, from the first stage simple linear regression analysis with predictor year, conducted separately at each site. After projecting the spatial coordinates to the unit square using a Mercator projection, the model for $\beta$ and $\varepsilon$ and the priors for all hyperparameters are the same to those in the simulation study in Section 5. The estimated slopes and corresponding $z$-values are plotted in Figure 1. We can see that the estimated slope is generally negative, implying that ozone concentrations are declining through the vast majority of the spatial domain. Thus we choose to test whether

the decline in ozone is more than 1 ppb per decade, that is, $H_0$:$\beta(s) \geq -0.1$ versus $H_1$:$\beta(s) < -0.1$.

We choose $k = 1$ (exponential correlation) and generate MCMC samples based on the posterior distribution of $\beta$ on a rectangular 100 × 100 grid of points covering the spatial domain (including areas outside the US), and test the hypotheses at each grid cell in the US. Comparing Figures 5a and 1a, we see considerable smoothing of the estimated slopes. The posterior mean is negative throughout most of the domain, but there are areas with a positive slope, including western Pennsylvania and Chesapeake Bay. The estimated decrease is the largest in Wisconsin, Illinois, Georgia, and Florida. The estimates of $1 - \hat{T_{OR}}(s_i)$ are plotted in Figure 5b. The estimated FDR ($\alpha = 0.1$) and FDX ($\alpha = \tau = 0.1$) thresholds for $\hat{T_{OR}}$ are 0.30 and 0.16, respectively. Figures 5c and 5d show that the null hypothesis is rejected using both thresholding rules for the western part of the domain, Georgia and Florida, and much of New England. As expected, the FDX threshold is more conservative; the null is rejected for much of North Carolina and Virginia using FDR, but not FDX.

We also conduct a cluster-wise analysis using states as clusters. Although these clusters are fairly large, spatial correlation persists after clustering. For example, denote $\bar{\beta_j}$ as the average of $\beta(s)$ at the grid locations described above for state $j$. The posterior correlation between $\bar{\beta_j}$ for Florida and other states is 0.51 for Georgia, 0.36 for Alabama, and 0.33 for North Carolina. Table 6 summarizes the cluster-wise analysis. We define the state to have a significant change in ozone if at least 80% of the state has slope less than −0.1ppb. Using this criteria gives $\hat{T_{OR}}(C_k)$ a threshold of 0.27 for an FCR analysis at level $\alpha = .1$, and 10 of the 26 states have a statistically significant trend in ozone. An alternative way to perform cluster-wise analysis is to define a cluster as active if its mean $\bar{\beta_j} < -0.1$. Table 6 gives the posterior probabilities that $\bar{\beta_j} < -0.1$ for each state. All 26 states have a statistically significant trend in ozone using an FCR analysis at level 0.1.

## 7. Proofs

Here we prove Theorems 1 and 3. The proofs of Theorems 2 and 4 and the lemmas are provided in the web appendix.

### 7.1. Proof of Theorem 1

We first state a lemma, which is proved in the web appendix.

**Lemma 1—**Consider a decision rule $\boldsymbol{\delta} = [I\{T(s) < t\}: s \in S]$. If $\boldsymbol{T} = \{T(s): s \in S\}$ satisfies the MRC (3.3), then the mFDR level of $\boldsymbol{\delta}$ monotonically increases in $t$.

**a.** Let $\boldsymbol{\theta} = \{\theta(s): s \in S\}$ and $\boldsymbol{\delta} = (\delta(s): s \in S)$ denote the unknown states and decisions, respectively. The loss function (3.2) can be written as

$$L(\boldsymbol{\theta}, \boldsymbol{\delta}) = \lambda\nu(S_{FP}) + \nu(S_{FN}) = \int_S \lambda\{1-\theta(s)\}\delta(s)d\nu(s) + \int_S \theta(s)\{1-\delta(s)\}d\nu(s).$$

The posterior classification risk is

$$E_{\boldsymbol{\theta}|\boldsymbol{X}^n}\{L(\boldsymbol{\theta},\boldsymbol{\delta})\}=\int_S[\delta(s)\lambda P\{\theta(s)=0|\boldsymbol{X}^n\}+\{1-\delta(s)\}P\{\theta(s)=1|\boldsymbol{X}^n\}]\,d\nu(s)$$
$$=\int_S\delta(s)\,[\lambda P\{\theta(s)=0|\boldsymbol{X}^n\}-P\{\theta(s)=1|\boldsymbol{X}^n\}]\,d\nu(s)+\int_S P(\theta(s)=1|\boldsymbol{X}^n)d\nu(s)$$

Therefore, the optimal decision rule which minimizes the posterior classification risk (also the classification risk) is given by $\boldsymbol{\delta}_{OR} = \{\delta_{OR}(s):s \in S\}$, where

$$\delta_{OR}(s)=I\,[\lambda P\{\theta(s)=0|\boldsymbol{X}^n\}-P\{\theta(s)=1|\boldsymbol{X}^n\}<0]=I[T_{OR}(s)<(1+\lambda)^{-1}].$$

**b.** We have assumed that $G_0(t) = \int_S P\{\theta(s) = 0, T_{OR}(s) < t\}d\nu(s)$ and $G_1(t) = \int_S P\{\theta(s) = 1, T_{OR}(s) < t\}d\nu(s)$ are differentiable. Let $g_1(t)$ and $g_0(t)$ be the derivatives. The goal is to show that $g_1(t)/g_0(t)$ decreases in $t$ for $t \in (0, 1)$. Consider a weighted classification problem with loss function

$$L(\boldsymbol{\theta},\boldsymbol{\delta})=\frac{1-t}{t}\nu(S_{FP})+\nu(S_{FN}).$$

Suppose $\boldsymbol{T}_{OR} = \{T_{OR}(s):s \in S\}$ is used in the weighted classification problem and the threshold is $c$. By Fubini's Theorem the classification risk is

$$E\left\{\tfrac{1-t}{t}\nu(S_{FP})+\nu(S_{FN})\right\}=\tfrac{1-t}{t}\int_S P\{\theta(s)=0,T_{OR}(s)<c\}d\nu(s)+\int_S P\{\theta(s)=1,T_{OR}(s)>c\}d\nu(s)$$
$$=\tfrac{1-t}{t}G_0(c)+\int_S P\{\theta(s)=1\}d\nu(s)-G_1(c)$$

The threshold $c = t^*$ which minimizes the classification risk satisfies: $t^{-1}(1 - t)g_0(t^*) = g_1(t^*)$. By part (a), the optimal threshold $t^* = 1 + t^{-1}(1 - t)^{-1} = t$. Therefore we have

$$\frac{g_1(t)}{g_0(t)}=\frac{1-t}{t},\ \text{for all}\ 0<t<1,$$

and the result follows.

**c.** Let $\boldsymbol{T}$ be a test statistic that satisfies the MRC (3.3). Lemma 1 indicates that for a given $\alpha \in (0, \alpha^*)$ ($\alpha^*$ is the largest mFDR level when the threshold $t = 1$), there exists a threshold $t(\alpha)$ such that the mFDR level of $\boldsymbol{\delta} = [I\{T(s) < t(\alpha)\}:s \in S]$ is $\alpha$, which completes the first part of the proof.

Let ERA($\boldsymbol{T}$, $t(\alpha)$), ETPA($\boldsymbol{T}$, $t(\alpha)$) and EFPA($\boldsymbol{T}$, $t(\alpha)$) be the expected rejection area, expected true positive area and expected false positive area of the decision rule $\boldsymbol{\delta} = [I\{T(s) < t(\alpha)\}:s \in S]$, respectively. Then we have

$$\mathrm{ERA}(\boldsymbol{T},t(\alpha))=E\left[\int_S I\{T(s)<t(\alpha)\}d\nu(s)\right]=\int_S P\{T(s)<t(\alpha)\}d\nu(s).$$

By definition, ERA($\boldsymbol{T}$, $t(a)$) = ETPA($\boldsymbol{T}$, $t(a)$) + EFPA($\boldsymbol{T}$, $t(a)$). Also note that the mFDR level is exactly $a$. We conclude that ETPA($\boldsymbol{T}$, $t(a)$) = $a\int_S P\{T(s) < t(a)\}d\nu(s)$, and EFPA($\boldsymbol{T}$, $t(a)$) = $(1 - a)\int_S P\{T(s) < t(a)\}d\nu(s)$.

Now consider the oracle test statistic $\boldsymbol{T}_{OR}$ defined in (3.5). Part (b) of Theorem 1 shows that $\boldsymbol{T}_{OR}$ satisfies the MRC (3.3). Hence, from the first part of the proof of (c), there exists $t_{OR}(a)$ such that $\boldsymbol{\delta}_{OR} = [I\{T_{OR}(s) < t_{OR}(a)\}:s \in S]$ controls the mFDR at level $a$ exactly. Consider a weighted classification problem with the following loss function

$$L(\boldsymbol{\theta}, \boldsymbol{\delta}) = \frac{1-t_{OR}(\alpha)}{t_{OR}(\alpha)}\nu(S_{FP}) + \nu(S_{FN}). \quad (7.1)$$

Part (a) shows that the optimal solution to the weighted classification problem is $\boldsymbol{\delta}_{OR} = [I\{T_{OR}(s) < t_{OR}(a)\}:s \in S]$. The classification risk of $\boldsymbol{\delta}_{OR}$ is

$$\begin{aligned} E\{L(\boldsymbol{\theta}, \boldsymbol{\delta}_{OR})\} &= \tfrac{1-t_{OR}(\alpha)}{t_{OR}(\alpha)} E\left[\int_S \{1-\theta(s)\}\delta_{OR}(s)d\nu(s)\right] + E\left[\int_S \theta(s)\{1-\delta_{OR}(s)\}d\nu(s)\right] \\ &= \tfrac{1-t_{OR}(\alpha)}{t_{OR}(\alpha)}\mathrm{EFPA}(\boldsymbol{T}_{OR}, t_{OR}(\alpha)) + \int_S P\{\theta(s)=1\}d\nu(s) - \mathrm{ETPA}(\boldsymbol{T}_{OR}, t_{OR}(\alpha)) \\ &= \left\{\tfrac{\alpha-t_{OR}(\alpha)}{t_{OR}(\alpha)}\right\}\mathrm{ERA}(\boldsymbol{T}_{OR}, t_{OR}(\alpha)) + \int_S P\{\theta(s)=1\}d\nu(s) \end{aligned}$$

The last equation is due to the facts that ETPA($\boldsymbol{T}$, $t(a)$) = $a\int_S P\{T(s) < t(a)\}d\nu(s)$, EFPA($\boldsymbol{T}$, $t(a)$) = $(1 - a)\int_S P\{T(s) < t(a)\}d\nu(s)$ and ETPA($\boldsymbol{T}$, $t(a)$) = $a$ERA($\boldsymbol{T}$, $t(a)$).

According to a Markov type inequality, double expectation theorem, and the fact that ETPA($\boldsymbol{T}$, $t(a)$) = $a$ERA($\boldsymbol{T}$, $t(a)$), we conclude that

$$\begin{aligned} t_{OR}(\alpha)\int_S E[I\{T_{OR}(s)<t_{OR}(\alpha)\}]d\nu(s) &> \int_S E\left[I\{T_{OR}(s)<t_{OR}(\alpha)\}T_{OR}(s)\right]d\nu(s) \\ &= \int_S E\left[T_{OR}(s)<t_{OR}, \theta(s)=0\right]d\nu(s) \\ &= \alpha\int_S E\left[I\{T_{OR}(s)<t_{OR}(\alpha)\}\right]d\nu(s) \end{aligned}$$

Hence we always have $t_{OR}(a) - a > 0$.

Next we claim that for any decision rules $\boldsymbol{\delta} = [I\{T(s) < t(a)\}:s \in S]$ in $\mathcal{D}$, the following result holds: ERA($\boldsymbol{T}$, $t(a)$) ≤ ERA($\boldsymbol{T}_{OR}$, $t_{OR}(a)$). We argue by contradiction. If there exists $\boldsymbol{\delta}^* = [I\{T^*(s) < t^*(a)\}:s \in S]$ such that

$$\mathrm{ERA}(\boldsymbol{T}^*, t^*(\alpha)) > \mathrm{ERA}(\boldsymbol{T}_{OR}, t_{OR}(\alpha)). \quad (7.2)$$

Then when $\boldsymbol{\delta}^*$ is used in the weighted classification problem with loss function (7.1), the classification risk of $\boldsymbol{\delta}^*$ is

$$E\{L(\boldsymbol{\theta},\boldsymbol{\delta}^*)\}=\left\{\frac{\alpha-t_{OR}(\theta)}{t_{OR}(\alpha)}\right\}\mathrm{ERA}(\boldsymbol{T}^*,t^*(\alpha))+\int_S P\{\theta(s)=1\}d\nu(s)$$
$$<\left\{\frac{\alpha-t_{OR}(\alpha)}{t_{OR}(\alpha)}\right\}\mathrm{ERA}(\boldsymbol{T}_{OR},t_{OR}(\alpha))+\int_S P\{\theta(s)=1\}d\nu(s)$$
$$=E\{L(\boldsymbol{\theta},\boldsymbol{\delta}_{OR})\},$$

The first equation holds because $\delta(\boldsymbol{T}^*, t^*(\alpha))$ is also an $\alpha$ level mFDR procedure. This contradicts the result in Theorem 1, which claims that $\boldsymbol{\delta}_{OR}$ minimizes the classification risk with loss function (7.1).

Therefore we claim that $\boldsymbol{\delta}_{OR}$ has the largest ERA, hence the largest ETPA (note we always have ETPA=$\alpha$ERA) and the smallest missed discovery region MDR among all mFDR procedures at level $\alpha$ in $\mathcal{D}$.

### 7.2. Proof of Theorem 3

We first state and prove a lemma. Define $\theta(s) = I\{\mu(s) \in A^c\}$ and $\theta^m(s) = I\{\mu^m(s) \in A^c\}$, where $A = [A_l, A_u]$ is the indifference region.

**Lemma 2—**Consider the discrete approximation based on a sequence of partitions of the spatial domain $\{S=\cup_{i=1}^m S_i : m=1,2,\cdots\}$. Then under the conditions of the theorem, we have $\int_S P\{\theta(s) \neq \theta^m(s)\}d\nu(s) \to 0$ as $m \to \infty$.

**Proof of Theorem 3:** (a). Suppose $T_{OR}(s) = P_\Psi\{\theta(s) = 0|\boldsymbol{X}^n\}$ is used for testing. Then Procedure 1 corresponds to the decision rule $\boldsymbol{\delta}^m = \{\delta^m(s): s \in S\}$, where $\delta^m(s)=\sum_{i=1}^m I\{T_{OR}(s_i)<t\}I(s \in S_i)$. We assume that $r$ pixels are rejected and let $R_r$ be the rejected area. The FDR level of $\boldsymbol{\delta}^m$ is

$$\mathrm{FDR} \le E\left\{\frac{\int_S\{1-\theta(s)\}\delta^m(s)d\nu(s)}{\nu(R_r)\vee c_0}\right\}$$
$$=E\left(\frac{1}{\nu(R_r)\vee c_0}\left[\sum_{i=1}^m \delta(s_i)\int_{S_i} E\{1-\theta(s)|\boldsymbol{X}^n\}d\nu(s)\right]\right)$$
$$=E\left(\frac{1}{\nu(R_r)\vee c_0}\left[\sum_{i=1}^m \delta(s_i)T_{OR}(s_i)\nu(S_i)+\sum_{i=1}^m \delta(s_i)\int_{S_i} E\{\theta(s_i)-\theta(s)|\boldsymbol{X}^n\}d\nu(s)\right]\right)$$
$$\le E\left\{\frac{1}{\nu(R_r)\vee c_0}\sum_{i=1}^r T_{OR}^{(i)}\nu S_{(i)}\right\}+Z_m,$$

where $Z_m = E\{\nu(R_r) \vee c_0\}^{-1}\int_S E\{\theta(s) - \theta^m(s)|\boldsymbol{X}^n\}\delta^m(s)d\nu(s)$. The second equality follows from double expectation theorem. The third equality can be verified by first adding and subtracting $\theta(s_i)$, expanding the sum, and then simplifying.

Next note that an upper bound for the random quantity $\{\nu(R_r) \vee c_0\}^{-1}$ is given by $c_0^{-1}$. Applying Lemma 2,

$$Z_m \le \tfrac{1}{c_0}\int_S E[\delta^{(m)}(s)E\{\theta(s)-\theta^m(s)|\boldsymbol{X}^n\}]d\nu(s) \le \tfrac{1}{c_0}\int_S P\{\theta(s)\neq\theta^m(s)\}d\nu(s)\to 0.$$

Since the operation of procedure $\mathcal{S}^m$ guarantees that

$$\frac{1}{\nu(R_r)\vee c_0}\sum_{i=1}^{r} T_{OR}^{(i)}\nu(S_{(i)}) \le \alpha$$

for all realizations of $\boldsymbol{X}^n$, the FDR is controlled at level $\alpha$ asymptotically.

(b). Suppose that $r$ pixels are rejected by Procedure 2. Consider $\delta^m(s)$ defined in part (a). Then the FDX at tolerance level $\tau$ is

$$\begin{aligned}\mathrm{FDX}_\tau &\le P\left[\{\nu(R_r)\vee c_0\}^{-1}\int_S \delta^m(s)\{1-\theta(s)\}d\nu(s)>\tau\right]\\ &= P\left[\{\nu(R_r)\vee c_0\}^{-1}\sum_{i=1}^m \delta(s_i)\int_{S_i}\{1-\theta(s)\}d\nu(s)>\tau\right]\\ = P\left[\{\nu(R_r)\vee c_0\}^{-1}\sum_{i=1}^m \delta(s_i)\{1-\theta(s_i)\}\nu(S_i)+\{\nu(R_r)\vee c_0\}^{-1}\int_S \delta^m(s)\{\theta^m(s)-\theta(s)\}d\nu(s)>\tau\right]\\ &\equiv P(A+B>\tau).\end{aligned}$$

where $A$ and $B$ are the corresponding terms on the left side of the inequality. Let $\varepsilon_0 \in (0, \tau)$ be the small positive number defined in Procedure 2. Then $A+B > \tau$ implies that $A > \tau-\varepsilon_0$ or $B > \varepsilon_0$. It follows that

$$P\{A+B>\tau\} \le P\{A>\tau-\varepsilon_0 \text{ or } B>\varepsilon_0\} \le P(A>\tau-\varepsilon_0)+P(B>\varepsilon_0).$$

Let $I$ denote an indicator function. Apply the double expectation theorem to the first term $P(A > \tau - \varepsilon_0)$, we have

$$P(A>\tau-\varepsilon_0)=E[I\{A>\tau-\varepsilon_0\}]=E\{P(A>\tau-\varepsilon_0|X^n)\}.$$

Replacing $A$ and $B$ by their original expressions, we have

$$\begin{aligned}\mathrm{FDX}_\tau \le E\left(P\left[\{\nu(R_r)\vee c_0\}^{-1}\sum_{i=1}^m \delta(s_i)\{1-\theta(s_i)\}\nu(S_i)>\tau-\varepsilon_0\middle|\boldsymbol{X}^n\right]\right)\\ +P\left[\{\nu(R_r)\vee c_0\}^{-1}\int_S \delta^m(s)\{\theta^m(s)-\theta(s)\}d\nu(s)\ge\varepsilon_0\right]\end{aligned}$$

It is easy to see that

$$\mathrm{FDX}_{\tau,r}^{m} \geq P\left[\{\nu(R_r)\vee c_0\}^{-1}\sum_{i=1}^{m}\delta(s_i)\{1-\theta(s_i)\}\nu(S_i)>\tau-\varepsilon_0\middle|\boldsymbol{X}^n\right].$$

The operation property of Procedure 2 guarantees that $\mathrm{FDX}_{\tau,r}^{m} \leq \alpha$ for all realizations of $\boldsymbol{X}^n$. Therefore the first term of in the expression of $\mathrm{FDX}_{\tau}$ is less than $\alpha$. The second term in the upper bound of $\mathrm{FDX}_{\tau}$ satisfies

$$P\left[\{\nu(R_r)\vee c_0\}^{-1}\int_S \delta^m(s)\{\theta^m(s)-\theta(s)\}d\nu(s)\geq\varepsilon_0\right] \leq (\varepsilon_0 c_0)^{-1}E\left[\int_S \delta^m(s)|\theta^m(s)-\theta(s)|d\nu(s)\right]$$
$$\leq (\varepsilon_0 c_0)^{-1}\int_S P\{\theta(s)\neq\theta^m(s)\}d\nu(s)\to 0$$

and the desired result follows.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgments

Sun's research was supported in part by NSF grants DMS-CAREER 1255406 and DMS-1244556. Reich's research was supported by the US Environmental Protection Agency (R835228), National Science Foundation (1107046), and National Institutes of Health (5R01ES014843-02). Cai's research was supported in part by NSF FRG Grant DMS-0854973, NSF Grant DMS-1208982, and NIH Grant R01 CA 127334. Guindani's research is supported in part by the NIH/NCI grant P30CA016672. We thank the Associate Editor and two referees for detailed and constructive comments which lead to a much improved article.

## References

Benjamini Y, Heller R. False discovery rates for spatial signals. J Amer Statist Assoc. 2007; 102:1272–1281.

Benjamini Y, Heller R. Screening for partial conjunction hypotheses. Biometrics. 2008; 64:1215–1222. [PubMed: 18261164]

Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Statist Soc B. 1995; 57:289–300.

Benjamini Y, Hochberg Y. Multiple hypotheses testing with weights. Scandinavian Journal of Statistics. 1997; 24:407–418.

Benjamini Y, Hochberg Y. On the adaptive control of the false discovery rate in multiple testing with independent statistics. Journal of Educational and Behavioral Statistics. 2000; 25:60–83.

Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. Ann Statist. 2001; 29(4):1165–1188.

Bogdan, M.; Gosh, J.; Tokdar, S. A comparison of the benjamini-hochberg procedure with some Bayesian rules for multile testing. In: Balakrishnan, N.; Peña, E.; Silvapulle, M., editors. Beyond Parametrics in Interdisciplinary Research: Festshcrift in Honor of Professor Pranab K. Sen. IMS Collections; Beachwood, Ohio, USA: Institute of Mathematical Statistics; 2008. p. 211-230.

Caldas de Castro M, Singer B. Controlling the false discovery rate: a new application to account for multiple and dependent tests in local statistics of spatial association. Geographical Analysis. 2006; 38(2):180–208.

Chen M, Cho J, Zhao H. Incorporating biological pathways via a markov random field model in genome-wide association studies. PLoS Genet. 2011; 7(4):e1001353. [PubMed: 21490723]

Clarke S, Hall P. Robustness of multiple testing procedures against dependence. Ann Statist. 2009; 37(1):332–358.

Efron B. Correlation and large-scale simultaneous significance testing. J Amer Statist Assoc. 2007; 102:93–103.

Finner H, Dickhaus T, Roters M. Dependency and false discovery rate: asymptotics. Ann Statist. 2007; 35(4):1432–1455.

Finner H, Roters M. Multiple hypotheses testing and expected number of type i errors. The Annals of Statistics. 2002; 30(1):220–238.

Gelfand, AE.; Diggle, PJ.; Fuentes, M.; Guttorp, P. Handbook of Spatial Statistics. New York: Chapman & Hall/CRC; 2010.

Genovese C, Wasserman L. Operating characteristics and extensions of the false discovery rate procedure. J R Stat Soc B. 2002; 64:499–517.

Genovese CR, Lazar NA, Nichols T. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. Neuroimage. 2002; 15(4):870–878. [PubMed: 11906227]

Genovese CR, Wasserman L. Exceedance control of the false discovery proportion. Journal of the American Statistical Association. 2006; 101:1408–1417.

Green P, Richardson S. Hidden markov models and disease mapping. Journal of the American statistical Association. 2002; 97(460):1055–1070.

Guindani M, Müller P, Zhang S. A bayesian discovery procedure. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2009; 71(5):905–925.

Heller R. Comment: Correlated z-values and the accuracy of large-scale statistical estimates. Journal of the American Statistical Association. 2010; 105(491):1057–1059.

Heller R, Stanley D, Yekutieli D, Rubin N, Benjamini Y. Cluster-based analysis of fmri data. Neuroimage. 2006; 33:599–608. [PubMed: 16952467]

Lehmann, EL.; Romano, JP. Springer Texts in Statistics. 3. New York: Springer; 2005. Testing statistical hypotheses.

Meinshausen N, Bickel P, Rice J. Efficient blind search: Optimal power of detection under computational cost constraints. The Annals of Applied Statistics. 2009; 3(1):38–60.

Miller C, Genovese C, Nichol R, Wasserman L, Connolly A, Reichart D, Hopkins A, Schneider J, Moore A. Controlling the false-discovery rate in astrophysical data analysis. The Astronomical Journal. 2007; 122(6):3492–3505.

Müller, P.; Parmigiani, G.; Rice, K. Fdr and bayesian multiple comparisons rules. In: Bernardo, J.; Bayarri, M.; Berger, J.; Dawid, A.; Heckerman, AD.; Smith; West, M., editors. Bayesian Statistics. Vol. 8. Oxford, UK: Oxford University Press; 2007.

Müller P, Parmigiani G, Robert CP, Rousseau J. Optimal sample size for multiple testing: the case of gene expression microarrays. J Am Statist Assoc. 2004; 99:990–1001.

Newton MA, Noueiry A, Sarkar D, Ahlquist P. Detecting differential gene expression with a semiparametric hierarchical mixture method. Biostatistics. 2004; 5(2):155–176. [PubMed: 15054023]

Owen AB. Variance of the number of false discoveries. J R Stat Soc, Ser B. 2005; 67(3):411–426.

Pacifico MP, Genovese C, Verdinelli I, Wasserman L. False discovery control for random fields. Journal of the American Statistical Association. 2004; 99:1002–1014.

Peng G, Luo L, Siu H, Zhu Y, Hu P, Hong S, Zhao J, Zhou X, Reveille JD, Jin L, Amos CI, Xiong M. Gene and pathway-based second-wave analysis of genome-wide association studies. European Journal of Human Genetics. 2009 Jul; 18(1):111–117. [PubMed: 19584899]

Pyne S, Futcher B, Skiena S. Meta-analysis based on control of false discovery rate: combining yeast chip-chip datasets. Bioinformatics. 2006; 22:2516–2522. [PubMed: 16908499]

Sarkar SK. Some results on false discovery rate in stepwise multiple testing procedures. Ann Statist. 2002; 30:239–257.

Schwartzman A, Dougherty RF, Taylor JE. False discovery rate analysis of brain diffusion direction maps. Ann Appl Stat. 2008; 2(1):153–175.

Schwartzman A, Lin X. The effect of correlation in false discovery rate estimation. Biometrika. 2011; 98(1):199–214. [PubMed: 23049127]

Storey JD. A direct approach to false discovery rates. J R Stat Soc B. 2002; 64:479–498.

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America. 2005; 102(43):15545–15550. [PubMed: 16199517]

Sun W, Cai TT. Oracle and adaptive compound decision rules for false discovery rate control. J Amer Statist Assoc. 2007; 102:901–912.

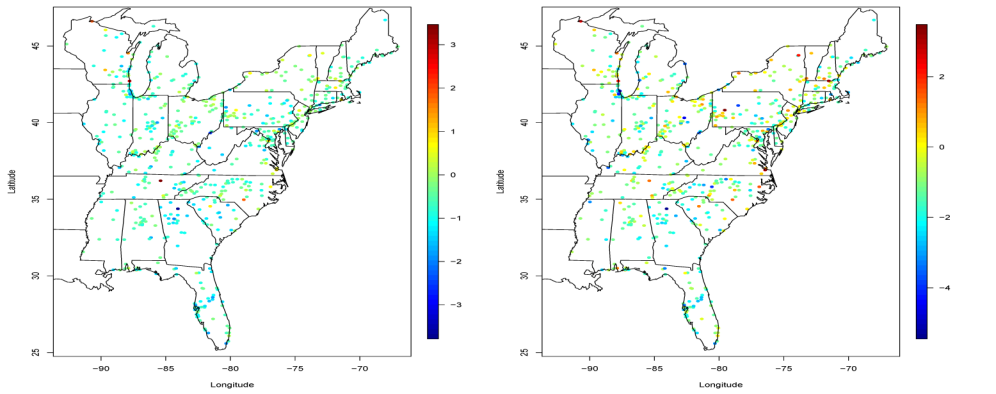
Sun W, Cai TT. Large-scale multiple testing under dependence. J R Stat Soc B. 2009; 71:393–424.

Wei Z, Li H. A markov random field model for network-based analysis of genomic data. Bioinformatics. 2007; 23(12):1537–1544. [PubMed: 17483504]

Wei Z, Sun W, Wang K, Hakonarson H. Multiple testing in genome-wide association studies via hidden markov models. Bioinformatics. 2009; 25:2802–2808. [PubMed: 19654115]

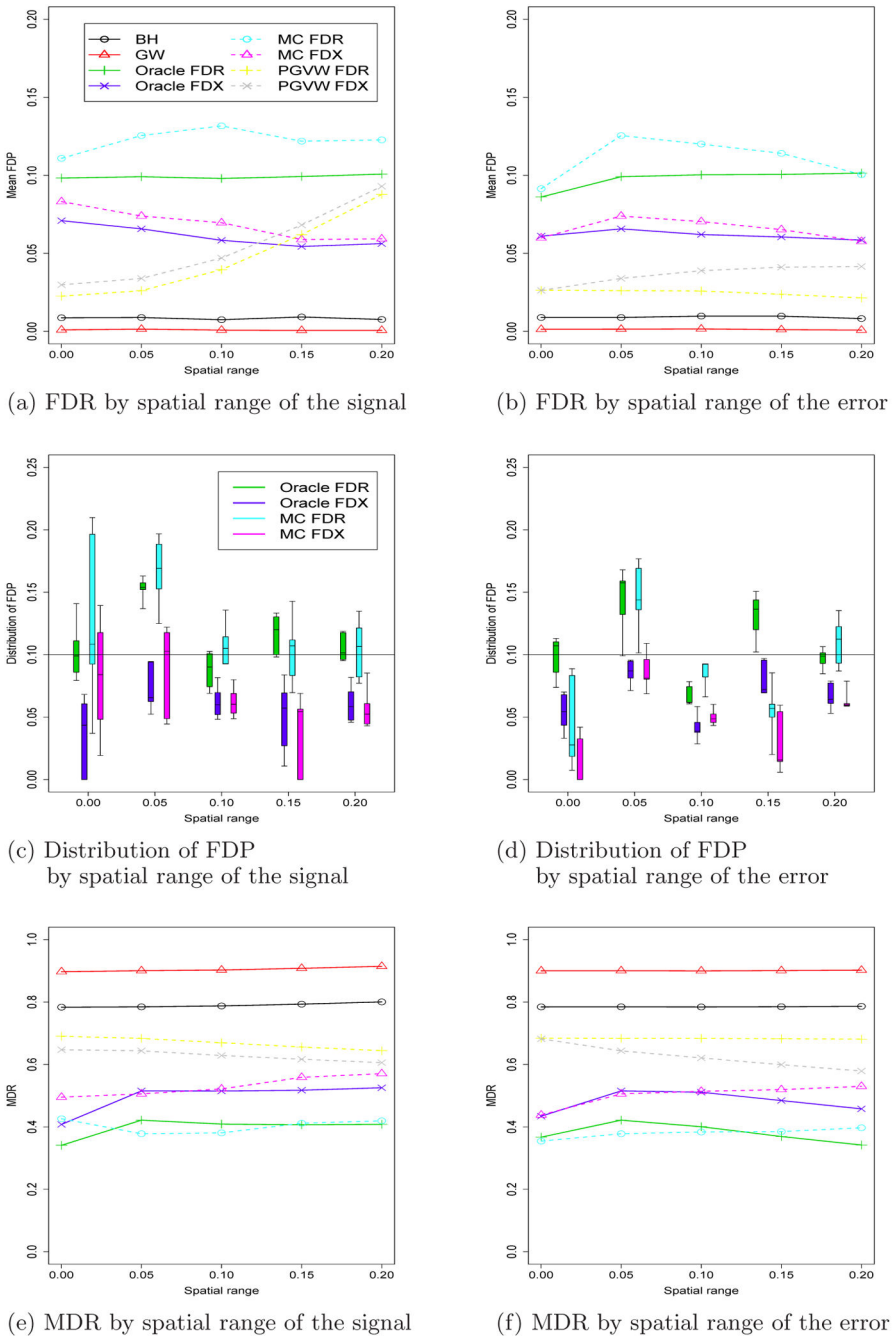
Wu WB. On false discovery control under dependence. Ann Statist. 2008; 36(1):364–380.

Zaykin DV, Zhivotovsky LA, Westfall PH, Weir BS. Truncated product method for combining p-values. Genet Epidemiol. 2002; 22:170–185. [PubMed: 11788962]




(a) First stage estimates, $\hat{\beta}(\mathbf{s})$ (b) First stage z-scores, $z(\mathbf{s}) = \hat{\beta}(\mathbf{s})/w(\mathbf{s})$

**Fig. 1.**
OLS analysis of ozone data, conducted separately at each site.




(a) FDR by spatial range of the signal

(b) FDR by spatial range of the error

(c) Distribution of FDP by spatial range of the signal

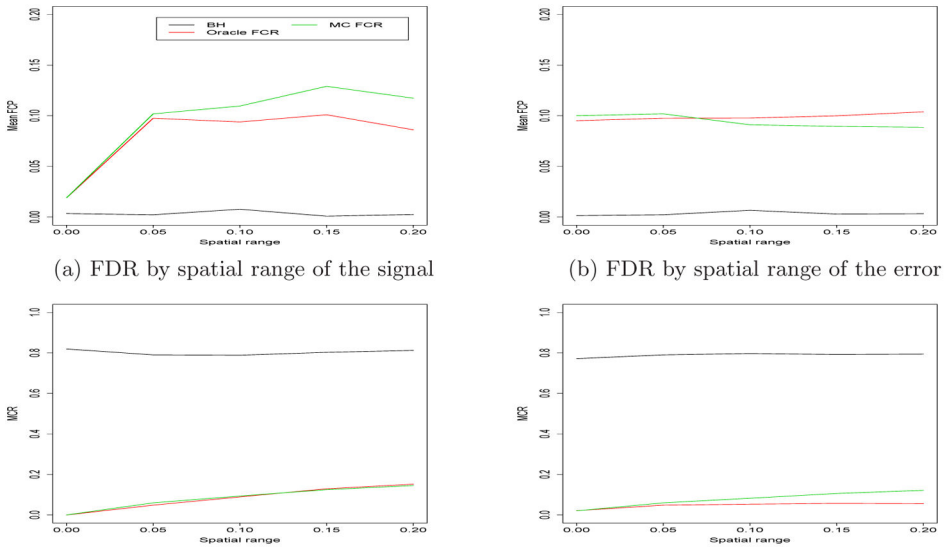
(d) Distribution of FDP by spatial range of the error

(e) MDR by spatial range of the signal

(f) MDR by spatial range of the error

**Fig. 2.**

Summary of the site-wise simulation study with exponential correlation. The horizontal lines in the boxplots in Panels (c) and (d) are the 0.1, 0.25, 0.50, 0.75, and 0.9 quantiles of FDP.




(a) FDR by spatial range of the signal

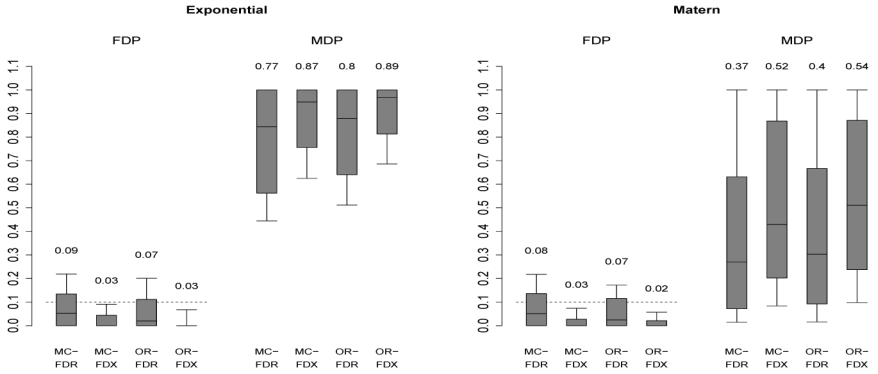
(b) FDR by spatial range of the error

(c) MDR by spatial range of the signal

(d) MDR by spatial range of the error

**Fig. 3.**
Summary of the cluster simulation study.

(a) Point-wise analysis




(b) Cluster analysis

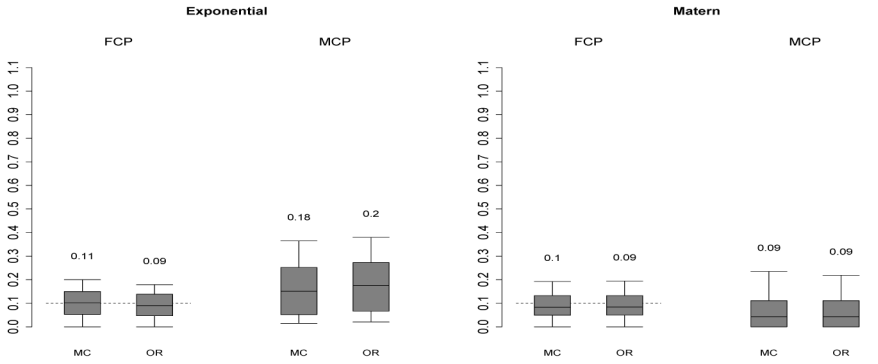
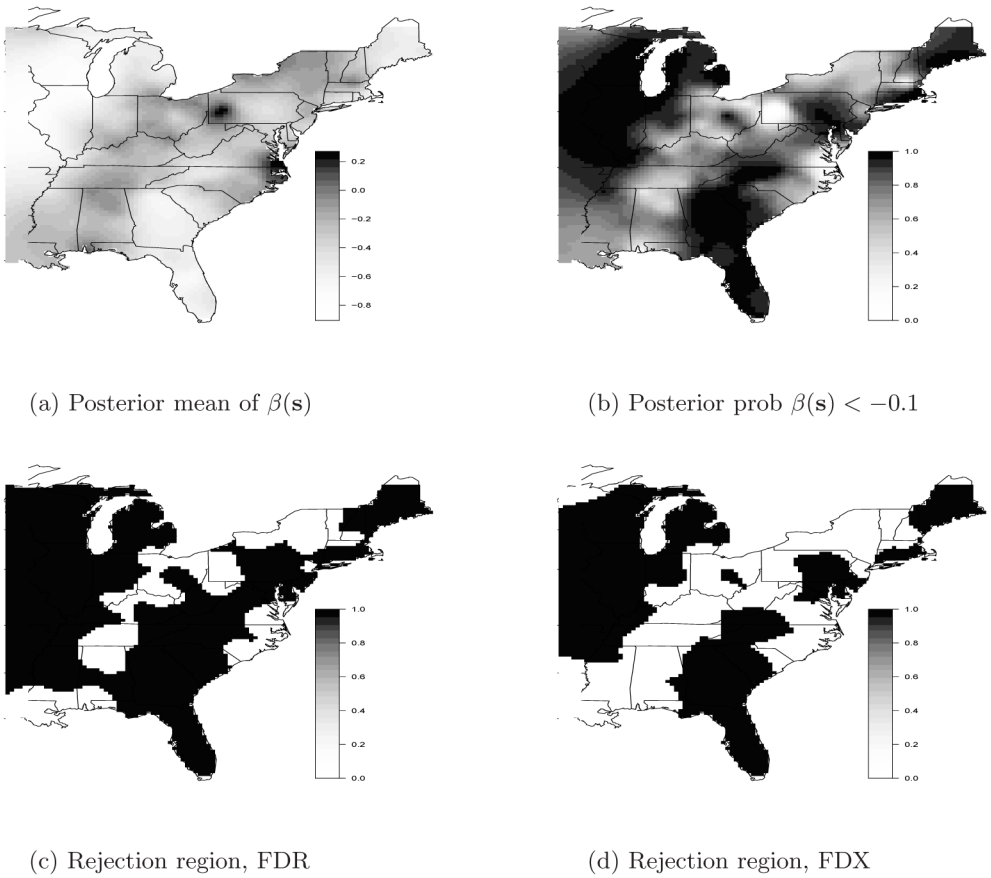



**Fig. 4.**

Simulation results with $n = 200$ with data generated with exponential and Matérn spatial correlation. Plotted are the FDP and MDP for the 200 simulated datasets. The boxplots' horizontal lines are the 0.10, 0.25, 0.50, 0.75, and 0.90 quantiles of FDP and MDP, and the numbers of above the boxplots are the means of FDP (FDR) and MDP (MDR).




(a) Posterior mean of $\beta(\mathbf{s})$

(b) Posterior prob $\beta(\mathbf{s}) < -0.1$

(c) Rejection region, FDR

(d) Rejection region, FDX

**Fig. 5.**

Summary of the ozone data analysis. Panels (a) and (b) give the posterior mean of $\beta(S)$ and the posterior probability that $\beta(S) < -0.1$. Panels (c) and (d) plot the rejection region using FDR and FDX (rejection plotted as a one, and vice versa).

**Table 1**

Cluster analysis for the ozone data. "State ave trend" is the posterior mean of the average of the $\beta(S)$ at the grid cells in the state.

| State | Number of Monitors | Number of grid points | State ave trend | Prob state ave < −0.1 | Proportion non-null | Post prob active |
|---|---|---|---|---|---|---|
| Alabama | 25 | 234 | −0.19 | 0.78 | 0.65 | 0.25 |
| Connecticut | 9 | 28 | −0.38 | 0.97 | 0.92 | 0.86* |
| Delaware | 6 | 8 | −0.36 | 0.95 | 0.91 | 0.81* |
| Florida | 43 | 235 | −0.53 | 1.00 | 0.93 | 0.93* |
| Georgia | 23 | 271 | −0.54 | 1.00 | 0.96 | 0.97* |
| Illinois | 35 | 277 | −0.66 | 1.00 | 0.98 | 0.99* |
| Indiana | 41 | 185 | −0.32 | 0.98 | 0.84 | 0.68 |
| Kentucky | 32 | 195 | −0.25 | 0.91 | 0.75 | 0.45 |
| Maine | 8 | 159 | −0.51 | 0.99 | 0.94 | 0.90* |
| Maryland | 19 | 55 | −0.30 | 0.96 | 0.83 | 0.64 |
| Massachusetts | 14 | 36 | −0.26 | 0.91 | 0.76 | 0.51 |
| Michigan | 26 | 296 | −0.50 | 1.00 | 0.92 | 0.92* |
| Mississippi | 8 | 220 | −0.27 | 0.87 | 0.76 | 0.52 |
| New Hampshire | 13 | 46 | −0.23 | 0.85 | 0.73 | 0.46 |
| New Jersey | 13 | 39 | −0.27 | 0.92 | 0.81 | 0.59 |
| New York | 33 | 262 | −0.15 | 0.65 | 0.59 | 0.18 |
| North Carolina | 41 | 227 | −0.23 | 0.88 | 0.71 | 0.33 |
| Ohio | 48 | 202 | −0.16 | 0.77 | 0.62 | 0.15 |
| Pennsylvania | 46 | 219 | −0.23 | 0.90 | 0.70 | 0.20 |
| Rhode Island | 3 | 8 | −0.47 | 0.99 | 0.98 | 0.96* |
| South Carolina | 20 | 144 | −0.42 | 0.98 | 0.89 | 0.81* |
| Tennessee | 25 | 185 | −0.25 | 0.89 | 0.73 | 0.41 |
| Vermont | 2 | 55 | −0.18 | 0.69 | 0.63 | 0.34 |
| Virginia | 23 | 188 | −0.25 | 0.88 | 0.73 | 0.40 |
| West Virginia | 9 | 115 | −0.24 | 0.83 | 0.72 | 0.45 |
| Wisconsin | 31 | 292 | −0.64 | 0.98 | 0.92 | 0.86* |

States which is significant at $\alpha = 0.1$ are denoted "*".