ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistical Methodology
Series B

J. R. Statist. Soc. B (2015)
77, *Part* 1, *pp.* 59–83

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

University of Texas M. D. Anderson Cancer Center, Houston, USA

and Armin Schwartzman

North Carolina State University, Raleigh, USA

[Received January 2013. Revised November 2013]

Summary. The paper develops a unified theoretical and computational framework for false discovery control in multiple testing of spatial signals. We consider both pointwise and clusterwise spatial analyses, and derive oracle procedures which optimally control the false discovery rate, false discovery exceedance and false cluster rate. 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 procedures proposed lead to more accurate error control and better power performance than conventional methods. We demonstrate our methods for analysing the time trends in tropospheric ozone in eastern USA.

Keywords: Compound decision theory; False cluster rate; False discovery exceedance; False discovery rate; Large-scale multiple testing; Spatial dependence

1. Introduction

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

$$X(s) = \mu(s) + \varepsilon(s), \tag{1.1}$$

where $\mu(s)$ is the unobserved random process and $\varepsilon(s)$ is the noise process. Assume that there is an underlying state $\theta(s)$ that is associated with each location s with one state being dominant ('background'). In applications, an important goal is to identify locations that exhibit significant

Address for correspondence: Wenguang Sun, Department of Information and Operation Management, University of Southern California, 505 Hoffman Hall, Los Angeles, CA 91103, USA.
E-mail: wenguang@marshall.usc.edu

 1369–7412/15/77059

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 genomewide 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 analysing time trends in tropospheric ozone in the eastern USA. Ozone is one of the six criteria pollutants that are regulated by the US Environmental Protection Agency under the Clean Air Act and has been linked with several adverse health effects. The Environmental Protection Agency has established a network of monitors for regulation of ozone, as shown in Fig. 1(a). We are interested in identifying locations with abrupt changing ozone levels by using the ozone concentration data that are collected at monitoring stations. In particular, we wish to study the ozone process for predefined subregions, such as counties or states, to identify interesting subregions. Similar problems may arise from disease mapping problems in epidemiology, where the goal is to identify geographical areas with elevated incidence of disease 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 subregion.

Spatial multiple testing poses new challenges which are not present in conventional multiple-testing problems. Firstly, one observes data points only 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 dependence in a two- or higher dimensional random field. Finally, in many important applications, it is desirable to aggregate information from nearby locations to make clusterwise 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 effect of dependence has been extensively studied in the multiple-testing literature. Efron (2007) and Schwartzman and Lin (2011) showed 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 procedure is valid for controlling the false discovery rate FDR (Benjamini and Hochberg, 1995) under various dependence 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) and Clarke and Hall (2009), among others). Another important research direction in multiple testing is the optimality issue under dependence. Sun and Cai (2009) introduced an asymptotically optimal FDR procedure for testing hypotheses arising from a hidden Markov model and showed that the hidden Markov model dependence 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 genomewide association studies, signals from individual markers are weak; hence several approaches have been developed to increase statistical power by aggregating multiple markers

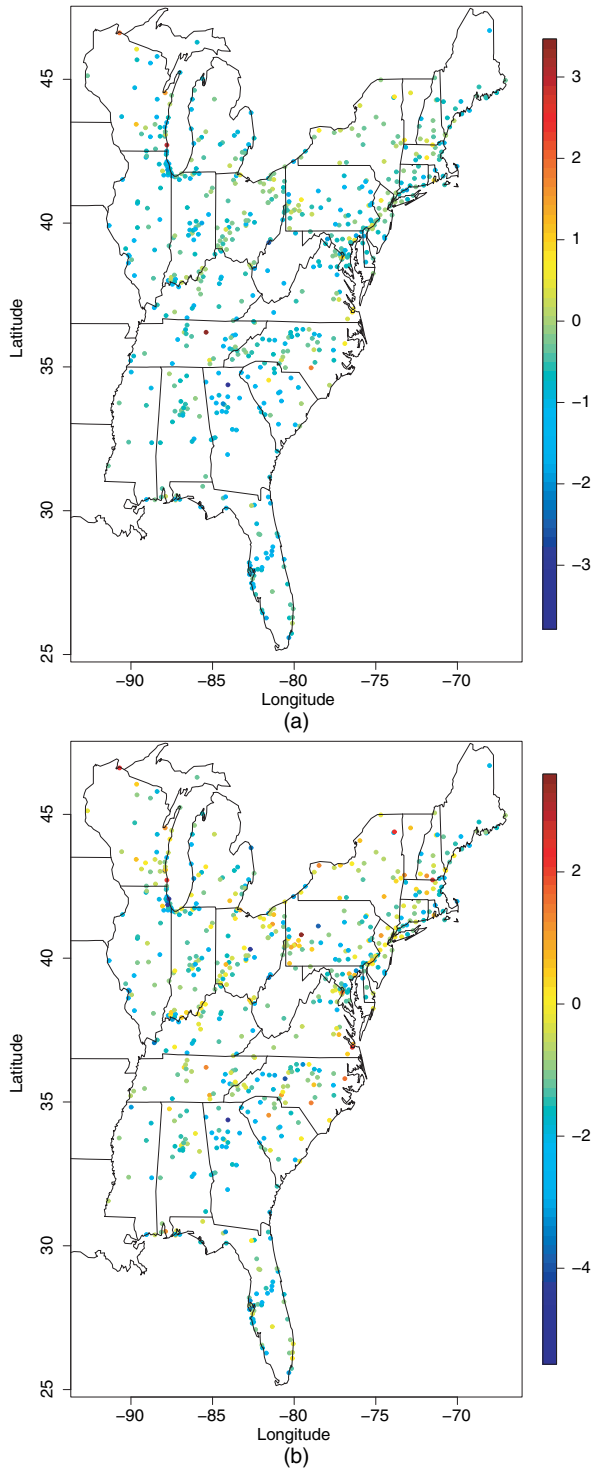

Fig. 1. Ordinary least squares analysis of the ozone data, conducted separately at each site: (a) first-stage analysis, $\hat{\beta}(\mathbf{s})$; (b) first-stage z-scores, $z(\mathbf{s}) = \hat{\beta}(\mathbf{s})/w(\mathbf{s})$

and exploiting the high correlation between adjacent loci (for example, see Peng *et al.* (2009), Wei *et al.* (2009) and Chen *et al.* (2011)). When the intensities of signals have a spatial pattern, it is expected that incorporating the underlying dependence structure can significantly improve the power and accuracy of conventional methods. This intuition is supported both theoretically and numerically in our work.

In this paper, we develop a compound decision theoretic framework for spatial multiple testing and propose a class of asymptotically optimal data-driven procedures that control FDR, the false discovery exceedance FDX and false cluster rate FCR. Widely used Bayesian modelling frameworks and computational algorithms are adopted to extract information effectively from large spatial data sets. We discuss how to summarize the fitted spatial models by using posterior sampling to address related multiple-testing problems. The control of 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 methods proposed. The results show that, by exploiting the spatial dependence, the data-driven procedures lead to better rankings of hypotheses, more accurate error control and enhanced power.

The methods proposed are developed in a frequentist framework and aim to control the frequentist FDR. The Bayesian computational framework, which involves hierarchical modelling and Markov chain Monte Carlo (MCMC) computing, provides a powerful tool to implement the data-driven procedures. When the goal is to control FDR and tests are independent, our procedure coincides with the Bayesian FDR approach that was originally proposed by Newton *et al.* (2004). Müller *et al.* (2004, 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; Bogdan *et al.*, 2008) has focused on pointwise 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 paper 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 Appendix A.

The programs that were used to analyse the data can be obtained from

`http://wileyonlinelibrary.com/journal/rss-datasets`

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 pointwise analysis and clusterwise analysis will be considered.

2.1. Pointwise inference

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

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

where A is the *indifference region*, e.g. $A = \{\mu : \mu \leqslant \mu_0\}$ for a one-sided test and $A = \{\mu : |\mu| \leqslant \mu_0\}$ 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 pointwise 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 $S_{\mathrm{FP}} = \{s \in S : \theta(s) = 0, \delta(s) = 1\}$ and $S_{\mathrm{FN}} = \{s \in S : \theta(s) = 1, \delta(s) = 0\}$ 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), which is 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

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

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

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

To evaluate the power of a multiple-testing procedure, we use the missed discovery rate $\mathrm{MDR} = E\{\nu(S_{\mathrm{FN}})\}$. 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. Clusterwise inference

When the interest is on the behaviour of a process over subregions, the testing units become spatial clusters instead of individual locations. Combining hypotheses over a set of locations naturally reduces multiplicity and correlation. In addition, setwise analysis improves statistical power as data in a set may show an increased signal-to-noise ratio (Benjamini and Heller, 2007). The idea of setwise or clusterwise 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

(a) to aggregate locations into regions according to available prior information (Heller *et al.*, 2006; Benjamini and Heller, 2007) and

(b) to conduct a *preliminary* pointwise analysis and to define the clusters after inspection of the results (Pacifico *et al.*, 2004).

Let $\mathcal{C} = \{C_1, \ldots, C_K\}$ 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), $H_0(C_k) : \pi_k \leqslant \gamma$ *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 $0 \leqslant \gamma \leqslant 1$ is a prespecified tolerance level. The null hypothesis could also be defined in terms of the average activation amplitude $\bar{\mu}(C_k) = \nu(C_k)^{-1} \int_{C_k} \mu(s)\,\mathrm{d}s$, i.e. $H_0(C_k) : \bar{\mu}(C_k) \leqslant \bar{\mu}_0$ *versus* $H_1(C_k) : \bar{\mu}(C_k) > \bar{\mu}_0$, for some prespecified $\bar{\mu}_0$. 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 clusterwise 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 $\boldsymbol{\Delta} = (\Delta_1, \ldots, \Delta_K)$. 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, which was 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

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

where w_k are cluster-specific weights which are often prespecified 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

$$\mathrm{mFCR} = \frac{E\left\{\sum_k w_k(1-\vartheta_k)\Delta_k\right\}}{E\left(\sum_k w_k \Delta_k\right)}.$$

We can see that, in the definition of 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 a higher penalty in the loss function if an important cluster is missed. The same weights w_k are used as reflective of proportional error and gain. Define the missed cluster rate $\mathrm{MCR} = E\{\Sigma_k w_k \vartheta_k (1 - \Delta_k)\}$. In clusterwise analysis the goal is to control FCR while minimizing 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 FDR, FDX and FCR. 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 pointwise analysis

Let $X_1, \ldots, X_n$ be observations at locations $S^* = \{s_1^*, \ldots, s_n^*\}$. In pointwise 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 dependence and to pool information from nearby observations. In this section, we discuss optimal results on pointwise FDR analysis from a theoretical perspective.

The optimal testing rule is derived in two steps: first the hypotheses are ranked optimally and then a cut-off is chosen along the rankings to control 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

$$\delta(s) = I\{T(s) < t\}, \tag{3.1}$$

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

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

where λ is the penalty for false positive results, and S_{FP} and S_{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 $G_j(t) = \int_S P\{T(s) < t, \theta(s) = j\}\, \mathrm{d}\nu(s)$, $j = 0, 1$. $G_0(t)$ can be viewed as the overall 'type I error' function at all locations in S where the null hypothesis 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 on-line 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

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

The MRC is a reasonable and mild regularity condition in multiple testing which ensures that mFDR increases in t and MDR decreases in t. Therefore to minimize MDR, we choose the *largest* threshold subject to $\mathrm{mFDR} \leqslant \alpha$. The MRC reduces to the monotone likelihood ratio condition (Sun and Cai, 2007) when the tests are independent. The monotone likelihood ratio condition 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 $\mathbf{X}^n = \{X_1, \ldots, X_n\}$. Consider a class of decision rules $\mathcal{D}$ of the form $\boldsymbol{\delta} = \{I\{T(s) < t\} : s \in S\}$, where $\mathbf{T} = \{T(s) : s \in S\}$ satisfies the MRC (3.3). The following theorem derives the optimal classification statistic and gives the optimal multiple-testing rule for mFDR control.

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

$$T_{\mathrm{OR}}(s) = P_\Psi\{\theta(s) = 0 | \mathbf{X}^n\} \tag{3.4}$$

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

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

$$\delta(s) = I\{T_{\mathrm{OR}}(s) < (1+\lambda)^{-1}\}. \tag{3.5}$$

(b) Let $\mathbf{T}_{\mathrm{OR}} = \{T_{\mathrm{OR}}(s) : s \in S\}$. Then $\mathbf{T}_{\mathrm{OR}}$ satisfies the MRC (3.3).
(c) There is an oracle threshold

$$t_{\mathrm{OR}}(\alpha) = \sup\{t : \mathrm{mFDR}(t) \leqslant \alpha\} \tag{3.6}$$

such that the oracle testing procedure

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

has the smallest MDR among all α-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 α) is the solution to an equivalent weighted classification problem with loss function (3.2) and penalty $\lambda(\alpha) = \{1 - t_{\mathrm{OR}}(\alpha)\}/t_{\mathrm{OR}}(\alpha)$. The procedure is called an 'oracle' procedure because it relies on knowledge of the true distributional information and the optimal threshold $t_{\mathrm{OR}}(\alpha)$, which are typically unknown in practice.

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

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

where $t_{\mathrm{OR,FDX}} = \arg\max_t\{\mathrm{FDX}_\tau(t) \leqslant \alpha\}$ is the oracle FDX threshold.

3.2. Oracle procedure for clusterwise analysis

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

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

where λ is the penalty for false positive results. As we would expect from remark 1, the FCR control problem can be solved by connecting it to a weighted classification problem with a suitably chosen λ. In practice λ is an unknown function of the FCR level α and needs to be estimated. In contrast, the weights w_k are prespecified. Let T_k be a clusterwise test statistic. Define $p_k = P(\vartheta_k = 1)$, $G_{jk}(t) = P(T_k < t | \vartheta_k = j)$ and $g_{jk}(t) = \mathrm{d}G_{jk}(t)/\mathrm{d}t$, $j = 0, 1$. Consider the generalized monotone ratio condition (GMRC)

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

The GMRC guarantees that the MCR is decreasing in FCR. Let $\mathcal{D}_c$ be the class of decision rules

of the form $\mathbf{\Delta} = \{I(T_k < t) : k = 1, \ldots, K\}$, where $\mathbf{T} = (T_1, \ldots, T_k)$ satisfies the GMRC (3.10). We have the following results.

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

$$T_{\mathrm{OR}}(C_k) = P_{\Psi}(\vartheta_k = 0 | \mathbf{X}^n) \tag{3.11}$$

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

(a) The classification risk with loss (3.9) is minimized by $\mathbf{\Delta} = \{\Delta_k : k = 1, \ldots, K\}$, where

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

(b) $\mathbf{T}_{\mathrm{OR}} = \{T_{\mathrm{OR}}(C_k) : k = 1, \ldots, K\}$ satisfies the GMRC (3.10).

(c) Define the oracle mFCR procedure

$$\mathbf{\Delta}_{\mathrm{OR}} = \{\Delta^k_{\mathrm{OR}} : k = 1, \ldots, K\} = \{I\{T_{\mathrm{OR}}(C_k) < t^{\mathrm{c}}_{\mathrm{OR}}(\alpha)\} : k = 1, \ldots, K\}, \tag{3.13}$$

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

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

4. False discovery controlling procedures and computational algorithms

The oracle procedures are difficult to implement because

(a) it is impossible to make an uncountable number of decisions when S is continuous and
(b) 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 we discuss how to calculate the test statistics.

4.1. FDR- and FDX-procedures for pointwise inference

To avoid making inference at every point, our strategy is to divide a continuous S into m 'pixels', to pick one point in each pixel and to 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 by a small constant. This condition is stated precisely as condition 2 when we study the asymptotic validity of the method proposed. Next take a point s_i from each S_i. In practice it is natural to use the centre point of S_i but we shall see that the choice of s_i is non-essential as long as condition 2 is fulfilled. Let $T^{(1)}_{\mathrm{OR}} \leqslant T^{(2)}_{\mathrm{OR}} \leqslant \ldots \leqslant T^{(m)}_{\mathrm{OR}}$ denote the ordered oracle statistics defined by equation (3.4) and $S_{(i)}$ the region corresponding to $T^{(i)}_{\mathrm{OR}}$. The following testing procedure is proposed for FDR control.

Procedure 1 (FDR-control): define $R_j = \cup_{i=1}^j S_{(i)}$ and

$$r = \max\left\{ j : \nu(R_j)^{-1} \sum_{i=1}^{j} T^{(i)}_{\mathrm{OR}} \nu(S_{(i)}) \leqslant \alpha \right\}. \tag{4.1}$$

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

Next we propose an FDX-procedure at level (γ, α) based on the same ranking and partition schemes. Let $R_j^m = \{s_1, \ldots, s_m\} \cap R_j$ be the set of rejected representation points. The main idea of the following procedure is first to obtain a discrete version of FDX_τ based on a finite approximation, then to estimate the actual FDX-level for various cut-offs and finally to choose the largest cut-off which controls FDX.

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

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

where $\theta(s_i)$ is a binary variable indicating the true state at location s_i. Let $r = \max\{j\colon \mathrm{FDX}_{\tau,j}^m \leqslant \alpha\}$; 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 $A = [A_\mathrm{l}, A_\mathrm{u}]$. To see why such a requirement is needed, define

$$\mu^m(s) = \sum_{i=1}^{m} \mu(s_i)\, I(s \in S_i),$$

$$\theta(s) = I\{\mu(s) \in A^\mathrm{c}\},$$

$$\theta^m(s) = I\{\mu^m(s) \in A^\mathrm{c}\}.$$

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 $\mu(s)$ pointwise as the partition becomes finer. 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 $A = [A_\mathrm{l}, A_\mathrm{u}]$ be the indifference region and ε a small positive constant. Then $\int_S P\{A_* - \varepsilon < \mu(s) < A_* + \varepsilon\}\, \mathrm{d}\nu(s) \to 0$ as $\varepsilon \to 0$, for $A_* = A_\mathrm{l}$ or $A_* = A_\mathrm{u}$.

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 occurs with only 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 is a sequence of partitions $\{S = \cup_{i=1}^{m} S_i : m = 1, 2, \ldots\}$ such that, for any given $\varepsilon > 0$, $\lim_{m\to\infty} \int_S P\{|\mu(s) - \mu^m(s)| \geqslant \varepsilon\}\, \mathrm{d}\nu(s) = 0$.

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 Appendix A.

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 $T_{\mathrm{OR}}(s)$ and $\mathrm{FDX}_{\tau,j}^m$ defined in equations (3.4) and (4.2) respectively. Let $\{\cup_{i=1}^{m} S_i : m = 1, 2, \ldots\}$ be a sequence of partitions of S satisfying conditions 1 and 2. Then

(a) the FDR-level of procedure 1 satisfies $\mathrm{FDR} \leqslant \alpha + o(1)$ when $m \to \infty$ and
(b) the FDX-level of procedure 2 satisfies $\mathrm{FDX}_\tau \leqslant \alpha + o(1)$ when $m \to \infty$.

When S is discrete, the FDR- or 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 $\mathrm{FDR} \leqslant \alpha$;
(b) the FDX-level of procedure 2 satisfies $\mathrm{FDX}_\tau \leqslant \alpha$.

4.2. FCR-procedure for clusterwise inference

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

Procedure 3 (FCR-control): let $T^{\mathrm{c}}_{(1)} \leqslant \ldots \leqslant T^{\mathrm{c}}_{(K)}$ be the ordered $T_{\mathrm{OR}}(C_k)$ values, and $\mathcal{H}_{(1)}, \ldots, \mathcal{H}_{(K)}$ and $w_{(1)}, \ldots, w_{(K)}$ the corresponding hypotheses and weights respectively. Let

$$r = \max\left\{ j : \frac{\sum_{k=1}^{j} w_{(k)} T^{\mathrm{c}}_{(k)}}{\sum_{k=1}^{j} w_{(k)}} \leqslant \alpha \right\}.$$

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

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

Theorem 4. Consider $T_{\mathrm{OR}}(C_k)$ defined in equation (3.11). Then the FCR of procedure 3 is controlled at the level α.

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

4.3. Data-driven procedures and computational algorithms

An important special case of model (1.1) is the Gaussian random field, where the signals and errors are generated as Gaussian processes with means $\bar{\mu}$ and 0, and covariance matrices Σ_1 and Σ_2 respectively. Let Ψ 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 Ψ. Denote by $\mathbf{X}^n = (X_1, \ldots, X_n)$ the collection of random variables that are associated with locations $s_1^*, \ldots, s_n^*$. Further let $f(\boldsymbol{\mu} | \mathbf{X}^n, \hat{\Psi}) \propto \pi(\boldsymbol{\mu})\, f(\mathbf{X}^n | \boldsymbol{\mu}, \hat{\Psi})$ be the posterior density function of $\boldsymbol{\mu}$ given $\mathbf{X}^n$ 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, which is used in both the simulation study and the 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 pointwise testing problem with $H_0(s) : \mu(s) \in A$ *versus* $H_1(s) : \mu(s) \notin A$, $s \in S$. Let $S^m = (s_1, \ldots, s_m)$ denote the collection of the representative points based on partition $S = \cup_{i=1}^{m} S_i$.

We discuss only the result for a continuous S (the result extends to a discrete S by simply taking $S^m = S$). Suppose that the MCMC samples are $\{\hat{\boldsymbol{\mu}}_b^m : b = 1, \ldots, B\}$, where $\hat{\boldsymbol{\mu}}_b^m = (\hat{\mu}_b^{m,1}, \ldots, \hat{\mu}_b^{m,m})$ is an m-dimensional posterior sample indicating the magnitudes of the signals at locations $s_1, \ldots, s_m$ 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

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

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

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

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

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

where j is the number of points in $\mathbf{s}^m$ 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 a subset of points in S^m which are rejected. Given the MCMC samples $\{\hat{\boldsymbol{\mu}}_b^m : b = 1, \ldots, B\}$, $\mathrm{FDX}_{\tau,j}^m$ can be estimated as

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

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

Next we turn to clusterwise 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 $\mathbf{s}^{m_k} = (s^{m_k,1}, \ldots, s^{m_k,m_k})$ be the collection of sampled points in cluster C_k, $m = \Sigma_{k=1}^K m_k$ be the count of points sampled in S and $\mathbf{s}^m = (\mathbf{s}^{m_1}, \ldots, \mathbf{s}^{m_K})$. If we are interested in testing partial conjunction of nulls $H_0(C_k) : \pi_k \leqslant \gamma$ *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 $\vartheta_k^m = I\{\Sigma_{i=1}^{m_k} \theta(s_i^k)\, \nu(S_i^k) > \gamma\, \nu(C_k)\}$ as an approximation to $\vartheta_k = I(\pi_k > \gamma)$. If the goal is to test average activation amplitude, i.e. $H_0(C_k) : \bar{\mu}(C_k) \leqslant \bar{\mu}_0$ *versus* $H_1(C_k) : \bar{\mu}(C_k) > \bar{\mu}_0$, then we can define $\vartheta_k^m = I\{\Sigma_{i=1}^{m_k} \mu(s_i^k)\, \nu(S_i^k) > \bar{\mu}_0\, \nu(C_k)\}$. Let $T_{\mathrm{OR}}^m(C_k) = P(\vartheta_k^m = 0 | \mathbf{X}^n)$.

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

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

Denote by $\hat{\boldsymbol{\mu}}_b^{m_k} = (\hat{\mu}_b^{m_k,1}, \ldots, \hat{\mu}_b^{m_k,m_k})$ the MCMC samples for cluster C_k at points $\mathbf{s}^{m_k}$ in replication b, $b = 1, \ldots, B$. Further let $\hat{\theta}_b^{m_k,i} = I(\hat{\mu}_b^{m_k,i} \notin A)$. Then $\int_{C_k} \theta(s)\, \mathrm{d}s$ in a particular replication b can be approximated by ${m_k}^{-1} \Sigma_{i=1}^{m_k} \hat{\theta}_b^{m_k,i}\, \nu(S_i^k)$ and the oracle statistic $T_{\mathrm{OR}}^m(C_k)$ can be estimated by

$$\hat{T}_{\mathrm{OR}}(C_k) = \frac{1}{B}\sum_{b=1}^{B} I\left\{\frac{1}{m_k}\sum_{i=1}^{m_k} \hat{\theta}_b^{m_k,i}\, \nu(S_i^k) < \gamma\, \nu(C_k)\right\}.$$

If the goal is to test average activation amplitude, $T_{\mathrm{OR}}^m(C_k)$ can be estimated as

$$\hat{T}_{\mathrm{OR}}(C_k) = \frac{1}{B}\sum_{b=1}^{B} I\left\{\nu(C_k)^{-1}\sum_{i=1}^{m_k}\hat{\mu}_b^{m_k,i}\,\nu(S_i^k) < \bar{\mu}_0\right\}.$$

5. Simulation

We conduct simulation studies to investigate the numerical properties of the methods proposed. A significant advantage of our method over conventional methods is that the procedure can carry 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 from `http://www-bcf.usc.edu/~wenguang/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 the errors being Gaussian processes. Let $\|\cdot\|$ denote the Euclidean distance. The signal process μ has mean $\bar{\mu}$ and powered exponential covariance $\mathrm{cov}\{\mu(s), \mu(s')\} = \sigma_\mu^2 \exp\{-(\|s-s'\|/\rho_\mu)^k\}$, whereas the error process ε has mean 0 and covariance $\mathrm{cov}\{\varepsilon(s), \varepsilon(s')\} = (1-r)\, I(s=s') + r\exp\{-(\|s-s'\|/\rho_\varepsilon)^k\}$ so $r \in [0,1]$ 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 $s_1, \ldots, s_n \sim^{\mathrm{IID}} \mathrm{uniform}([0,1]^2)$. 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 ρ_μ and ρ_ε. We present the results for only $k = 1$. The conclusions from simulations for $k = 2$ are similar in the sense that our methods control FDR more precisely and are more powerful than competitive methods. For each combination of spatial covariance parameters, we generate 200 data sets. For simulations studying the effects of varying ρ_μ we fix $\rho_\varepsilon = 0.05$, and for simulations studying the effects of varying ρ_ε we fix $\rho_\mu = 0.05$.

5.1.1. Pointwise analysis

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

We generate the MCMC samples by using a Bayes model, where we assume that k is known, and we select uninformative priors: $\bar{\mu} \sim N(0, 100^2)$, $\sigma_\mu^{-2} \sim \mathrm{gamma}(0.1, 0.1)$ and $r, \rho_\mu, \rho_\varepsilon \sim \mathrm{uniform}(0,1)$. The oracle FDR or oracle 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 $\alpha = \tau = 0.1$. Fig. 2 plots the averages of the FDPs and MDPs over the 200 data sets.

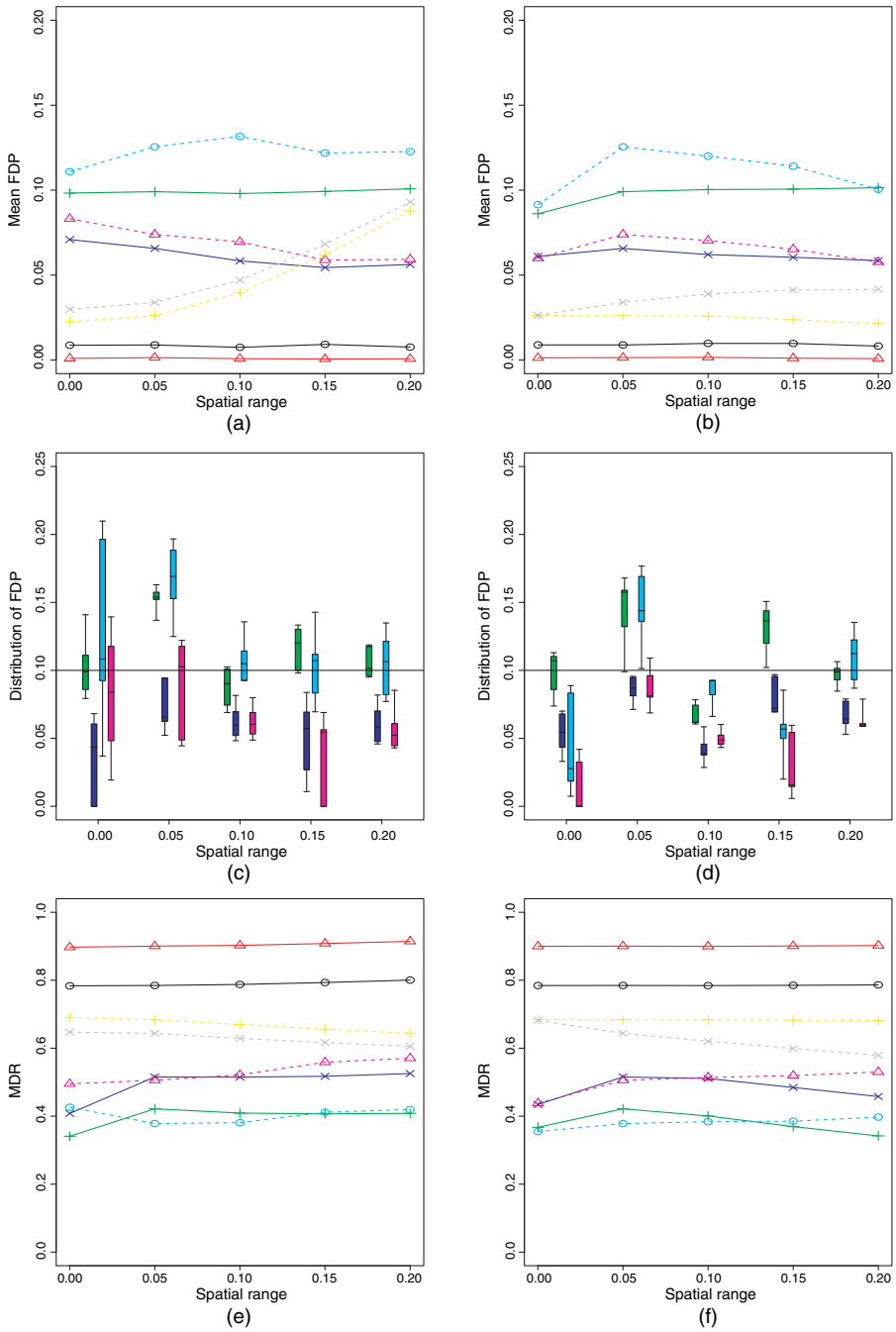

Fig. 2. Summary of the sitewise simulation study with exponential correlation: (a) FDR by spatial range of the signal (○, Benjamini–Hochberg; △, Genovese–Wasserman; +, oracle FDR; ×, oracle FDX; ○, MC FDR; △, MC FDX; +, PGVW FDR; ×, PGVW FDX); (b) FDR by spatial range of the error; (c) distribution of FDP by spatial range of the signal (▮, oracle FDR; ▮, oracle FDX; ▮, MC FDR; ▮, MC FDX; —, 0.10-, 0.25-, 0.50-, 0.75- and 0.90-quantiles of FDP); (d) distribution of FDP by spatial range of the error; (e) MDR by spatial range of the signal (○, Benjamini–Hochberg; △, Genovese–Wasserman; +, oracle FDR; ×, oracle FDX; ○, MC FDR; △, MC FDX; +, PGVW FDR; ×, PGVW FDX); (f) MDR by spatial range of the error

We can see that the oracle FDR procedure controls FDR nearly perfectly. The MC FDR procedure, with uninformative priors on the unknown spatial correlation parameters, also has good FDR control, between 10% and 12%. As expected, the oracle and MC FDX methods that are tuned to control FDX are more conservative than the FDR-methods, with observed FDR between 5% and 8%. The FDX-methods become increasingly conservative as the spatial correlation of the signal increases to adjust appropriately for higher correlation between tests. In contrast, the Benjamini–Hochberg, Genovese–Wasserman and PGVW procedures are very conservative, with much higher MDR-levels. The distribution of FDP is shown in Figs 2(c) and 2(d). 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 95th percentile of FDP is 0.15 for the smallest spatial range in Fig. 2(c), and less than 0.12 in all other cases.

5.1.2. Clusterwise analysis

We use the same data-generating schemes and MCMC sampling methods as in the sitewise 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 that the parameters are known, which is denoted by oracle FCR) and the corresponding MCMC method with non-informative priors (which is denoted by MC FCR). We compare our methods with the combined p-value approach that was 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 Fig. 3. We can see that the oracle FCR procedure controls 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 the oracle FCR and the MC FCR procedures have much lower missed cluster rates (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 κ_μ. The error process $\{\varepsilon(s) : s \in S\}$ has mean 0 and covariance $\mathrm{cov}\{\varepsilon(s), \varepsilon(t)\} = (1 - r)\, I(s = t) + r\, M(\|s - t\|; \rho_\varepsilon, \kappa_\varepsilon)$ so $r \in [0, 1]$ controls the proportion of the error variance with spatial correlation.

For each simulated data set, data are generated at n spatial locations $s_i \sim^{\mathrm{IID}} \mathrm{uniform}(\mathcal{D})$, where $\mathcal{D}$ is the unit square $\mathcal{D} = [0, 1]^2$. Predictions are made and tests of $\mathcal{H}_0 : \mu(s) \leqslant \mu_0$ *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 give a smoother spatial process than the exponential function but with roughly the same effective range (the distance at which correlation

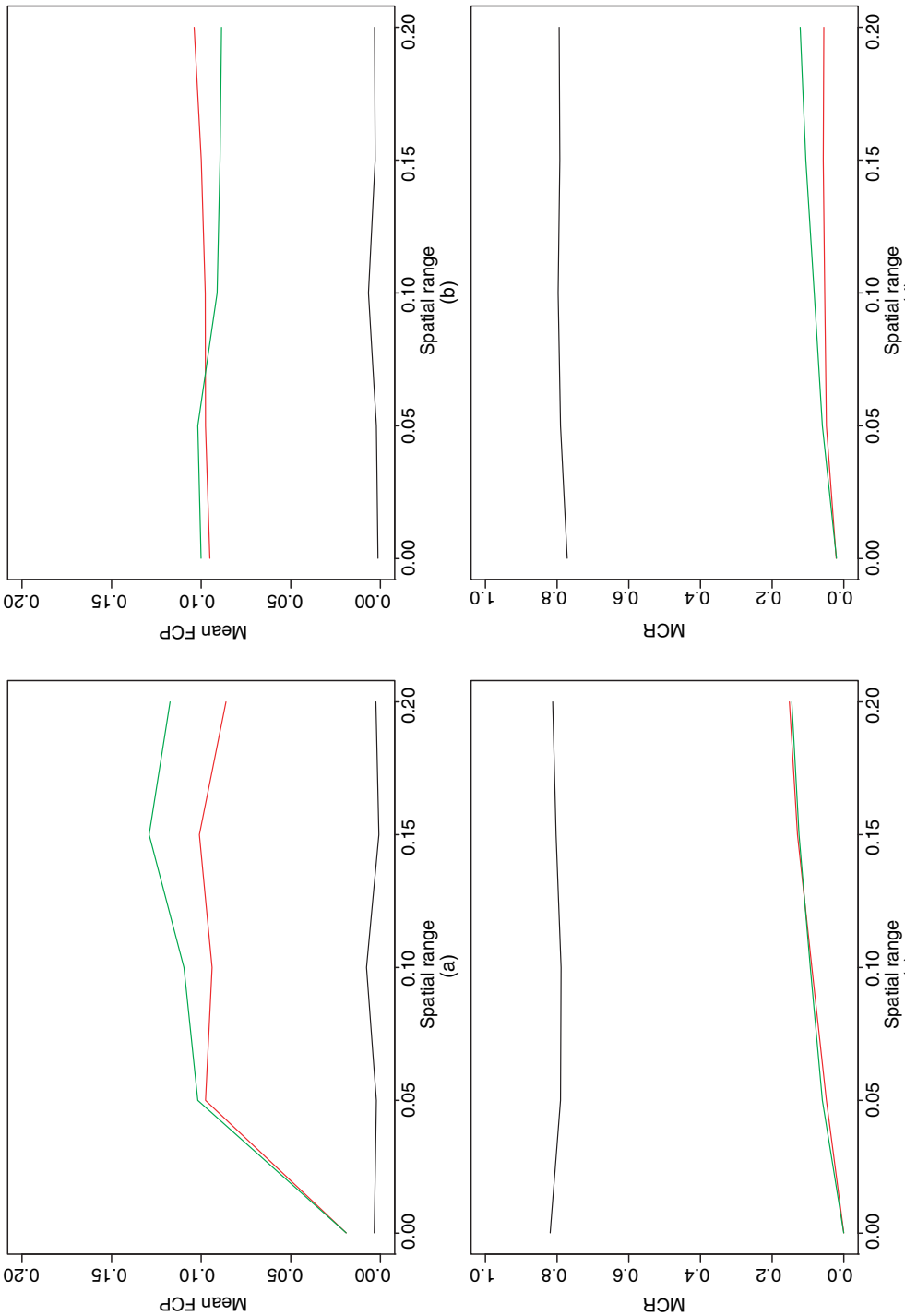

Fig. 3. Summary of the cluster simulation study (——, Benjamini–Hochberg; ——, oracle FDR; ——, MC FCR): (a) FDR by spatial range of the signal; (b) FDR by spatial range of the error; (c) MDR by spatial range of the signal; (d) MDR by spatial range of the error

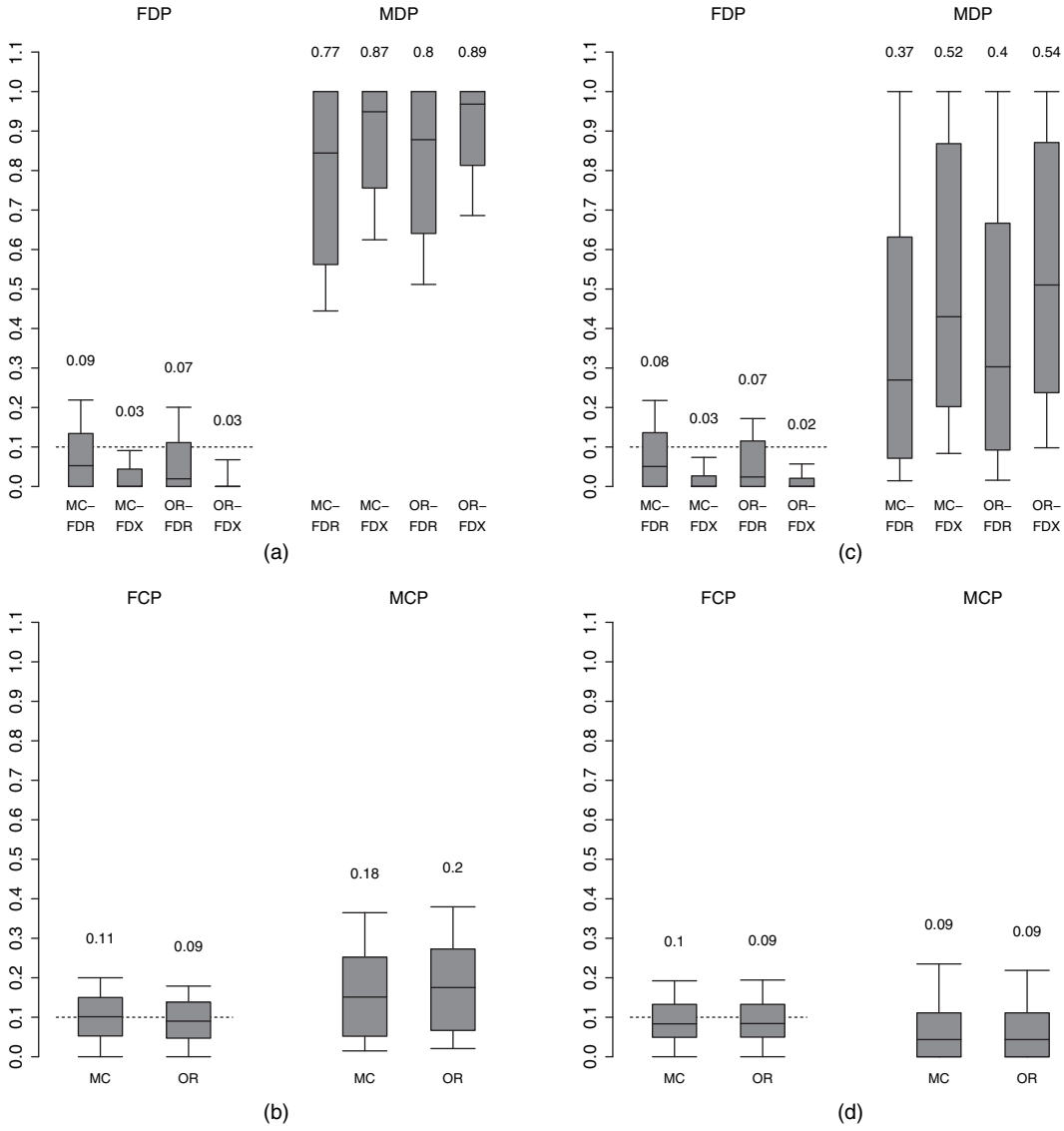

Fig. 4. Simulation results for FDP and MDP with $n = 200$ with data generated with (a), (b) exponential and (c), (d) Matérn spatial correlation (—, 0.10-, 0.25-, 0.50-, 0.75- and 0.90-quantiles of FDP and MDP; the numbers above the boxplots are the means of FDP or FDR and MDP or MDR): (a), (c) pointwise analysis; (b), (d) cluster analysis

is 0.05). For both correlation functions we generate 200 data sets and fit the model with Matérn correlation function and priors $\bar{\mu} \sim N(0, 1000^2)$, $\sigma_\mu^{-2} \sim \text{gamma}(0.01, 0.01)$, $r \sim \text{uniform}(0, 1)$ and $\kappa_\mu, \kappa_\varepsilon, \rho_\mu, \rho_\varepsilon \sim^{\text{IID}} N(-1, 1)$. For comparison we also fit the oracle model with hyperparameters $\bar{\mu}$, σ_μ, r, κ_μ, κ_ε, ρ_μ and ρ_ε fixed at their true values.

The results are summarized in Fig. 4. For data simulated with exponential correlation, both the data-driven procedure and the 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 by using this simulation design. Data were generated and the models were fitted as for the pointwise simulation. We define the spatial cluster regions by first creating a 10×10 regular partition of $\mathcal{D}$, and then combining the final two columns and final two rows to give unequal cluster sizes. This gives 81 clusters and between four 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. FDR and FDX are controlled in all cases, and the power is much higher for the smoother Matérn data. FDR and FDX for the data-driven procedures are comparable with the oracle procedure with these parameters fixed at their true values, suggesting that the proposed testing procedure is efficient even in this difficult setting.

6. Ozone data analysis

To illustrate the method proposed, we analyse daily surface level 8-h average ozone levels for the eastern USA. The data are obtained from the US Environmental Protection Agency's air explorer database (`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 to 2005, we compute the fourth highest daily value of 8-h average ozone level. 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 co-ordinates to the unit square by using a Mercator projection, the model for β and ε and the priors for all hyperparameters are the same as those in the simulation study in Section 5. The estimated slopes and corresponding z-values are plotted in Fig. 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 level is more than 1 ppb per decade, i.e. $H_0 : \beta(s) \geqslant -0.1$ *versus* $H_1 : \beta(s) < -0.1$.

We choose $k = 1$ (exponential correlation) and generate MCMC samples based on the posterior distribution of β on a rectangular 100×100 grid of points covering the spatial domain (including areas outside the USA), and we test the hypotheses at each grid cell in the USA. Comparing Figs 5(a) and 1(a), 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 $1 - \hat{T}_{\mathrm{OR}}(s_i)$ are plotted in Fig. 5(b). The estimated FDR- ($\alpha = 0.1$) and FDX- ($\alpha = \tau = 0.1$) thresholds for $\hat{T}_{\mathrm{OR}}$ are 0.30 and 0.16 respectively. Figs 5(c) and 5(d) show that the null hypothesis is rejected by 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 hypothesis is rejected for much of North Carolina and Virginia by using FDR, but not FDX.

We also conduct a clusterwise 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

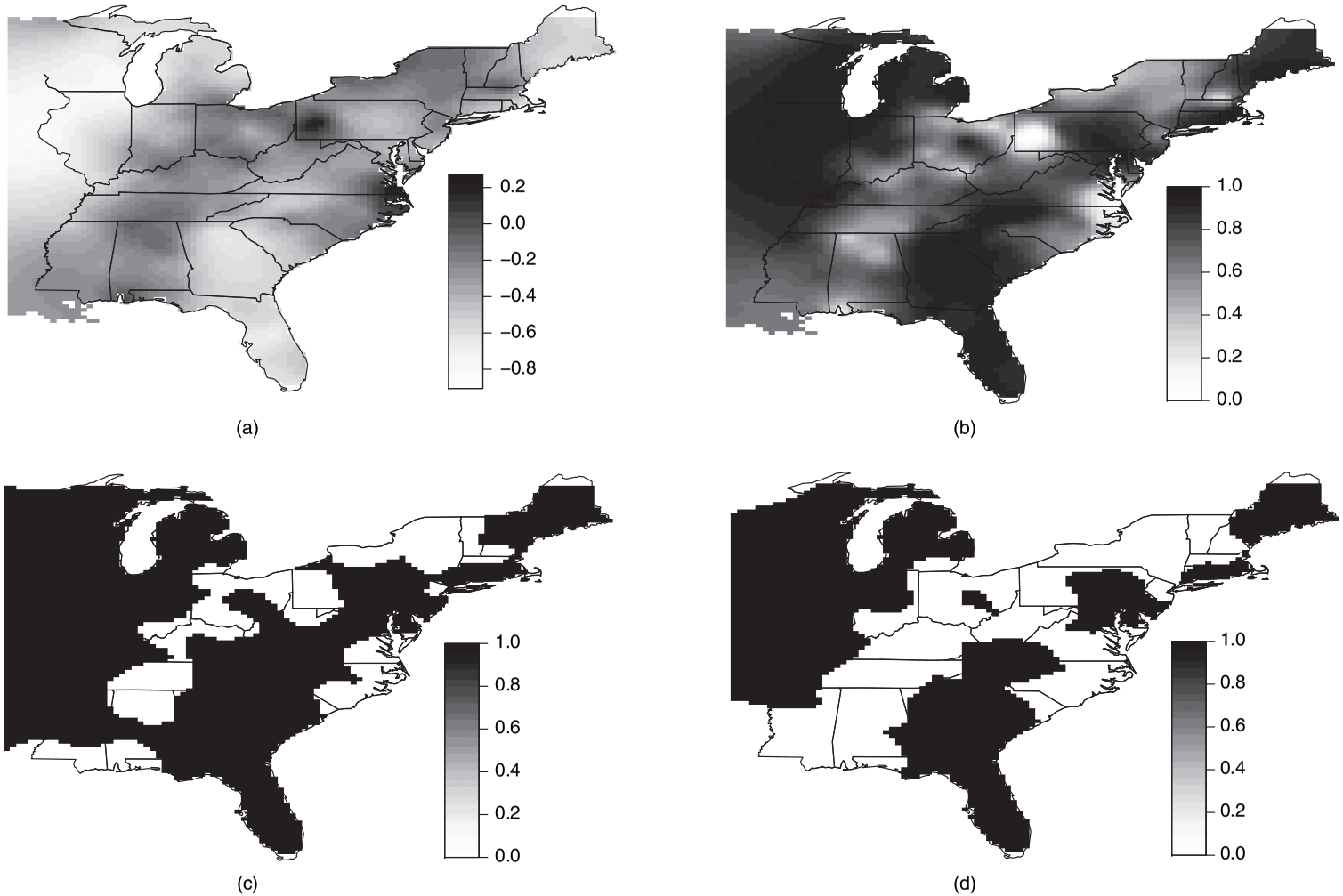

Fig. 5. Summary of the ozone data analysis: (a) posterior mean of $\beta(\mathbf{s})$; (b) posterior probability that $\beta(\mathbf{s}) < 0.1$; (c) rejection region by using FDR; (d) rejection region by using FDX (rejection plotted as a 1, and acceptance as 0)

Table 1. Cluster analysis for the ozone data†

State	*Number of monitors*	*Number of grid points*	*State average trend*	*Probability state average* < -0.1	*Proportion non-null*	*Posterior probability 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‡

†'State average trend' is the posterior mean of the average of the $\beta(\mathbf{s})$ at the grid cells in the state.
‡Significant at $\alpha = 0.1$.

$\beta(s)$ at the grid locations that were 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 1 summarizes the clusterwise analysis. We define the state to have a significant change in ozone level if at least 80% of the state has slope less than -0.1 ppb. Using this criterion gives $\hat{T}_{\mathrm{OR}}(C_k)$ a threshold of 0.27 for an FCR-analysis at level $\alpha = 0.1$, and 10 of the 26 states have a statistically significant trend in ozone level. An alternative way to perform clusterwise analysis is to define a cluster as active if its mean $\bar{\beta}_j < -0.1$. Table 1 gives the posterior probabilities that $\bar{\beta}_j < -0.1$ for each state. All 26 states have a statistically significant trend in ozone concentration by using an FCR-analysis at level 0.1.

Acknowledgements

Sun's research was supported in part by National Science Foundation grants DMS-CAREER 1255406 and DMS-1244556. Reich's research was supported by the US Environmental Protection Agency (grant R835228), National Science Foundation (grant 1107046) and National Institutes of Health (grants 5R01ES014843-02 and R21ES022795-01A1). Cai's research was supported in part by National Science Foundation 'Focused research groups' grant DMS-

0854973, National Science Foundation grant DMS-1208982 and National Institutes of Health grant R01 CA 127334. Guindani's research is supported in part by the National Institutes of Health–National Cancer Institute grant P30CA016672. Schwartzman's research is supported in part by National Institutes of Health grant R01CA157528. We thank the Associate Editor and two referees for detailed and constructive comments which led to a much improved paper.

Appendix A: Proofs

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

A.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 $\mathbf{T} = \{T(s) : s \in S\}$ 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

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

The posterior classification risk is

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

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

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

(b) We have assumed that $G_0(t) = \int_S P\{\theta(s) = 0, T_{\mathrm{OR}}(s) < t\}\mathrm{d}\nu(s)$ and $G_1(t) = \int_S P\{\theta(s) = 1, T_{\mathrm{OR}}(s) < t\}\, \mathrm{d}\nu(s)$ 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

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

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

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

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

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

and the result follows.

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

Let $\mathrm{ERA}\{\mathbf{T}, t(\alpha)\}$, $\mathrm{ETPA}\{\mathbf{T}, t(\alpha)\}$ and $\mathrm{EFPA}\{\mathbf{T}, t(\alpha)\}$ 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

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

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

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

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

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

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

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

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

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

Hence we always have $t_{\mathrm{OR}}(\alpha) - \alpha > 0$.

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

$$\mathrm{ERA}\{\mathbf{T}^*, t^*(\alpha)\} > \mathrm{ERA}\{\mathbf{T}_{\mathrm{OR}}, t_{\mathrm{OR}}(\alpha)\}. \tag{A.2}$$

Then, when $\boldsymbol{\delta}^*$ is used in the weighted classification problem with loss function (A.1), the classification risk of $\boldsymbol{\delta}^*$ is

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

The first equation holds because $\delta\{\mathbf{T}^*, t^*(\alpha)\}$ is also an α-level mFDR-procedure. This contradicts

the result in theorem 1, which claims that $\boldsymbol{\delta}_{\mathrm{OR}}$ minimizes the classification risk with loss function (A.1).

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

A.2. Proof of theorem 3

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

Lemma 2. Consider the discrete approximation based on a sequence of partitions of the spatial domain $\{S = \cup_{i=1}^m S_i : m = 1, 2, \ldots\}$. Then, under the conditions of theorem 3, we have $\int_S P\{\theta(s) \neq \theta^m(s)\} \mathrm{d}\nu(s) \to 0$ as $m \to \infty$.

The proof of theorem 3 is in two parts.

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

$$\begin{aligned}
\mathrm{FDR} &\leqslant E\left[\frac{\displaystyle\int_{S\{1-\theta(s)\}} \delta^m(s)\, \mathrm{d}\nu(s)}{\nu(R_r) \vee c_0}\right] \\
&= E\left(\frac{1}{\nu(R_r) \vee c_0}\left[\sum_{i=1}^m \delta(s_i) \int_{S_i} E\{1 - \theta(s) | \mathbf{X}^n\}\, \mathrm{d}\nu(s)\right]\right) \\
&= E\left(\frac{1}{\nu(R_r) \vee c_0}\left[\sum_{i=1}^m \delta(s_i) T_{\mathrm{OR}}(s_i)\, \nu(S_i) + \sum_{i=1}^m \delta(s_i) \int_{S_i} E\{\theta(s_i) - \theta(s) | \mathbf{X}^n\}\, \mathrm{d}\nu(s)\right]\right) \\
&\leqslant E\left\{\frac{1}{\nu(R_r) \vee c_0} \sum_{i=1}^r T_{\mathrm{OR}}^{(i)}\, \nu(S_{(i)})\right\} + Z_m,
\end{aligned}$$

where $Z_m = E[\{\nu(R_r) \vee c_0\}^{-1} \int_S E\{\theta(s) - \theta^m(s) | \mathbf{X}^n\}\, \delta^m(s)\, \mathrm{d}\nu(s)]$. The second equality follows from the 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,

$$\begin{aligned}
Z_m &\leqslant \frac{1}{c_0} \int_S E[\delta^m(s)\, E\{\theta(s) - \theta^m(s) | \mathbf{X}^n\}]\, \mathrm{d}\nu(s) \\
&\leqslant \frac{1}{c_0} \int_S P\{\theta(s) \neq \theta^m(s)\}\, \mathrm{d}\nu(s) \to 0.
\end{aligned}$$

Since the operation of procedure $\boldsymbol{\delta}^m$ guarantees that

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

for all realizations of $\mathbf{X}^n$, FDR is controlled at level α asymptotically.

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

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

where A and B are the corresponding terms on the left-hand 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

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

Let I denote an indicator function. Applying 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

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

It is easy to see that

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

The operation property of procedure 2 guarantees that $\mathrm{FDX}^m_{\tau,r} \leqslant \alpha$ for all realizations of $\mathbf{X}^n$. Therefore the first term in the expression of FDX_τ is less than α. The second term in the upper bound of FDX_τ satisfies

$$\begin{aligned} P\left[\{\nu(R_r) \vee c_0\}^{-1} \int_S \delta^m(s)\{\theta^m(s) - \theta(s)\} \, \mathrm{d}\nu(s) \geqslant \varepsilon_0 \right] & \leqslant (\varepsilon_0 c_0)^{-1} E\left[\int_S \delta^m(s) \, |\theta^m(s) - \theta(s)| \, \mathrm{d}\nu(s) \right] \\ & \leqslant (\varepsilon_0 c_0)^{-1} \int_S P\{\theta(s) \neq \theta^m(s)\} \, \mathrm{d}\nu(s) \to 0 \end{aligned}$$

and the desired result follows.

References

Benjamini, Y. and Heller, R. (2007) False discovery rates for spatial signals. *J. Am. Statist. Ass.*, **102**, 1272–1281.

Benjamini, Y. and Heller, R. (2008) Screening for partial conjunction hypotheses. *Biometrics*, **64**, 1215–1222.

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

Benjamini, Y. and Hochberg, Y. (1997) Multiple hypotheses testing with weights. *Scand. J. Statist.*, **24**, 407–418.

Benjamini, Y. and Hochberg, Y. (2000) On the adaptive control of the false discovery rate in multiple testing with independent statistics. *J. Educ. Behav. Statist.*, **25**, 60–83.

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

Bogdan, M., Gosh, J. and Tokdar, S. (2008) A comparison of the Benjamini-Hochberg procedure with some Bayesian rules for multiple testing. In *Beyond Parametrics in Interdisciplinary Research: Festschrift in Honor of Professor Pranab K. Sen* (eds N. Balakrishnan, E. Peña and M. Silvapulle), pp. 211–230. Beachwood: Institute of Mathematical Statistics.

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

Chen, M., Cho, J., and Zhao, H. (2011) Incorporating biological pathways via a markov random field model in genome-wide association studies. *PLOS Genet.*, **7**, article e1001353.

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

Efron, B. (2007) Correlation and large-scale simultaneous significance testing. *J. Am. Statist. Ass.*, **102**, 93–103.

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

Finner, H. and Roters, M. (2002) Multiple hypotheses testing and expected number of type i errors. *Ann. Statist.*, **30**, 220–238.

Gelfand, A. E., Diggle, P. J., Fuentes, M. and Guttorp, P. (2010) *Handbook of Spatial Statistics*. New York: Chapman and Hall–CRC.

Genovese, C. R., Lazar, N. A. and Nichols, T. (2002) Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *Neuroimage*, **15**, 870–878.
Genovese, C. and Wasserman, L. (2002) Operating characteristics and extensions of the false discovery rate procedure. *J. R. Statist. Soc.* B, **64**, 499–517.
Genovese, C. R. and Wasserman, L. (2006) Exceedance control of the false discovery proportion. *J. Am. Statist. Ass.*, **101**, 1408–1417.
Green, P. and Richardson, S. (2002) Hidden markov models and disease mapping. *J. Am. Statist. Ass.*, **97**, 1055–1070.
Guindani, M., Müller, P. and Zhang, S. (2009) A Bayesian discovery procedure. *J. R. Statist. Soc.* B, **71**, 905–925.
Heller, R. (2010) Comment: Correlated z-values and the accuracy of large-scale statistical estimates. *J. Am. Statist. Ass.*, **105**, 1057–1059.
Heller, R., Stanley, D., Yekutieli, D., Rubin, N. and Benjamini, Y. (2006) Cluster-based analysis of fmri data. *Neuroimage*, **33**, 599–608.
Lehmann, E. L. and Romano, J. P. (2005) *Testing Statistical Hypotheses*. New York: Springer.
Meinshausen, N., Bickel, P. and Rice, J. (2009) Efficient blind search: optimal power of detection under computational cost constraints. *Ann. Appl. Statist.*, **3**, 38–60.
Miller, C., Genovese, C., Nichol, R., Wasserman, L., Connolly, A., Reichart, D., Hopkins, A., Schneider, J. and Moore, A. (2007) Controlling the false-discovery rate in astrophysical data analysis. *Astron. J.*, **122**, 3492–3505.
Müller, P., Parmigiani, G. and Rice, K. (2007) Fdr and bayesian multiple comparisons rules. In *Bayesian Statistics 8* (eds J. M. Bernardo, M. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. F. M. Smith and M. West). Oxford: Oxford University Press.
Müller, P., Parmigiani, G., Robert, C. P. and Rousseau, J. (2004) Optimal sample size for multiple testing: the case of gene expression microarrays. *J. Am. Statist. Ass.*, **99**, 990–1001.
Newton, M. A., Noueiry, A., Sarkar, D. and Ahlquist, P. (2004) Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics*, **5**, 155–176.
Owen, A. B. (2005) Variance of the number of false discoveries. *J. R. Statist. Soc.* B, **67**, 411–426.
Pacifico, M. P., Genovese, C., Verdinelli, I. and Wasserman, L. (2004) False discovery control for random fields. *J. Am. Statist. Ass.*, **99**, 1002–1014.
Peng, G., Luo, L., Siu, H., Zhu, Y., Hu, P., Hong, S., Zhao, J., Zhou, X., Reveille, J. D., Jin, L., Amos, C. I. and Xiong, M. (2009). Gene and pathway-based second-wave analysis of genome-wide association studies. *Eur. J. Hum. Genet.*, **18**, 111–117.
Pyne, S., Futcher, B. and Skiena, S. (2006) Meta-analysis based on control of false discovery rate: combining yeast chip-chip datasets. *Bioinformatics*, **22**, 2516–2522.
Sarkar, S. K. (2002) Some results on false discovery rate in stepwise multiple testing procedures. *Ann. Statist.*, **30**, 239–257.
Schwartzman, A., Dougherty, R. F. and Taylor, J. E. (2008) False discovery rate analysis of brain diffusion direction maps. *Ann. Appl. Statist.*, **2**, 153–175.
Schwartzman, A. and Lin, X. (2011) The effect of correlation in false discovery rate estimation. *Biometrika*, **98**, 199–214.
Storey, J. D. (2002) A direct approach to false discovery rates. *J. R. Statist. Soc.* B, **64**, 479–498.
Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S. and Mesirov, J. P. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natn. Acad. Sci. USA*, **102**, 15545–15550.
Sun, W. and Cai, T. T. (2007) Oracle and adaptive compound decision rules for false discovery rate control. *J. Am. Statist. Ass.*, **102**, 901–912.
Sun, W. and Cai, T. T. (2009) Large-scale multiple testing under dependence. *J. R. Statist. Soc.* B, **71**, 393–424.
Wei, Z. and Li, H. (2007) A markov random field model for network-based analysis of genomic data. *Bioinformatics*, **23**, 1537–1544.
Wei, Z., Sun, W., Wang, K. and Hakonarson, H. (2009) Multiple testing in genome-wide association studies via hidden markov models. *Bioinformatics*, **25**, 2802–2808.
Wu, W. B. (2008) On false discovery control under dependence. *Ann. Statist.*, **36**, 364–380.
Zaykin, D. V., Zhivotovsky, L. A., Westfall, P. H. and Weir B. S. (2002) Truncated product method for combining p-values. *Genet. Epidem.*, **22**, 170–185.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Web appendix for "False discovery control in large-scale spatial multiple testing"'.