ELSEVIER

NeuroImage

www.elsevier.com/locate/ynimg
NeuroImage 22 (2004) 676–687

Nonstationary cluster-size inference with random field and permutation methods

Satoru Hayasaka,[a] K. Luan Phan,[b] Israel Liberzon,[c,d] Keith J. Worsley,[e] and Thomas E. Nichols[a,*]

[a] Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109-2029, USA
[b] Department of Psychiatry and Behavioral Neurosciences, Wayne State University School of Medicine, Detroit, MI 48201, USA
[c] Department of Psychiatry, University of Michigan, Ann Arbor, MI 48109-0704, USA
[d] Psychiatry Service, Ann Arbor VAMC, Ann Arbor, MI 48105, USA
[e] Department of Mathematics and Statistics, McGill University, Montréal, QC, Canada

Received 11 September 2003; revised 23 January 2004; accepted 23 January 2004

Available online 30 April 2004

Because of their increased sensitivity to spatially extended signals, cluster-size tests are widely used to detect changes and activations in brain images. However, when images are nonstationary, the cluster-size distribution varies depending on local smoothness. Clusters tend to be large in smooth regions, resulting in increased false positives, while in rough regions, clusters tend to be small, resulting in decreased sensitivity. Worsley et al. proposed a random field theory (RFT) method that adjusts cluster sizes according to local roughness of images [Worsley, K.J., 2002. Nonstationary FWHM and its effect on statistical inference of fMRI data. Presented at the 8th International Conference on Functional Mapping of the Human Brain, June 2–6, 2002, Sendai, Japan. Available on CD-ROM in NeuroImage 16 (2) 779–780; Hum. Brain Mapp. 8 (1999) 98]. In this paper, we implement this method in a permutation test framework, which requires very few assumptions, is known to be exact [J. Cereb. Blood Flow Metab. 16 (1996) 7] and is robust [NeuroImage 20 (2003) 2343]. We compared our method to stationary permutation, stationary RFT, and nonstationary RFT methods. Using simulated data, we found that our permutation test performs well under any setting examined, whereas the nonstationary RFT test performs well only for smooth images under high *df*. We also found that the stationary RFT test becomes anticonservative under nonstationarity, while both nonstationary RFT and permutation tests remain valid under nonstationarity. On a real PET data set we found that, though the nonstationary tests have reduced sensitivity due to smoothness estimation variability, these tests have better sensitivity for clusters in rough regions compared to stationary cluster-size tests. We include a detailed and consolidated description of Worsley nonstationary RFT cluster-size test.

Keywords: Cluster size inference; Permutation; Stationary

Introduction

Whether functional or structural, detecting changes in brain images is a central problem in neuroimaging. Cluster-size tests, pioneered by Poline and Mazoyer (1993) and Friston et al. (1994), have been widely used in such investigations because of increased sensitivity to spatially extended signals, compared to voxel-intensity tests (Friston et al., 1996; Poline et al., 1997). Different implementations of cluster-size tests have been developed, including simulation-based tests (Forman et al., 1995; Ledberg et al., 1998; Poline and Mazoyer, 1993; Roland et al., 1993), random field theory-based (RFT) tests (Cao and Worsley, 2001; Worsley et al., 1996), and permutation tests (Holmes et al., 1996; Nichols and Holmes, 2002).

One of the assumptions usually required in a cluster-size test is stationarity or uniform smoothness. This assumption is crucial because when it is violated, the sensitivity and the specificity of the test can depend on local smoothness of images (Worsley et al., 1999). In smooth regions, clusters tend to be large even in the absence of true signals, thus resulting in increased false positives. On the other hand, in rough regions, clusters tend to be small, and even a true positive cluster may be too small to be detected, resulting in reduced power. Because of such bias, Ashburner and Friston (2000) discourage use of cluster-size tests in voxel-based morphometry (VBM) data that are known to exhibit nonstationary. Even in a typical BOLD fMRI data set, the stationarity assumption is questionable (see Fig. 1), yet this assumption is not routinely assessed, and in our experience, rarely true.

To address this problem associated with nonstationarity, Worsley et al. (1999) suggested adjusting cluster sizes according to the local smoothness at each voxel. With the RFT framework by Cao (1999), this approach has been implemented (Worsley, 2002). However, this test is subject to various random field assumptions: Images have to be a lattice approximation of a smooth random field, the cluster defining threshold needs to be sufficiently high, and the cluster-size distribution is considered to approximately follow a known parametric form. Though nonstationarity is accounted for, the test is still restricted by such stringent assumptions.

[*] Corresponding author. Department of Biostatistics, School of Public Health, University of Michigan, SPH II, M4065, 1420 Washington Height, Ann Arbor, MI 48109-2029. Fax: +1-734-763-2215.
E-mail address: nichols@umich.edu (T.E. Nichols).
Available online on ScienceDirect (www.sciencedirect.com.)

1053-8119/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2004.01.041

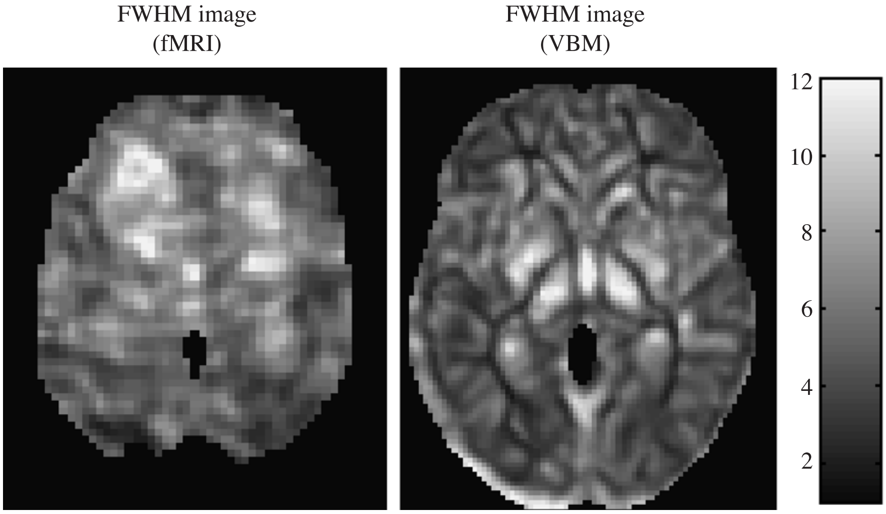

Fig. 1. Examples of local FWHM images, from a multisubject fMRI study comparing emotional responses between schizophrenics and controls (Taylor et al., 2003) (left), and from a VBM study comparing gray matter images from AIDS patients and controls (Varhola et al., 2000) (right). In both cases, smoothness is not uniform, with the VBM data showing dramatic structure in local smoothness.

Therefore, as an alternative to the nonstationary RFT test, we propose a nonstationary permutation test. Using Worsley et al.'s roughness adjusted cluster sizes, we adjust sensitivity of the test according to image smoothness. Since the permutation cluster-size test requires few assumptions and it is robust and exact (Hayasaka and Nichols, 2003), we avoid stringent restrictions of the RFT test. We validate the nonstationary permutation test by noise image simulations, both stationary and nonstationary, and compare its performance to the RFT counterpart. We further validate both nonstationary tests by applying them to a PET data set and examine their sensitivity relative to stationary cluster-size tests.

This paper is structured as following: In Methods and materials, we describe briefly both nonstationary RFT and permutation tests. Furthermore, we describe the validation of both tests in simulations and in data analysis. In Results, findings from the simulations and the data analysis are presented. In Discussion, we examine the findings from the simulation and draw conclusions.

Methods and materials

Implementation of nonstationary cluster-size tests

Model

We assume that voxel intensities of a brain image can be expressed as a linear model

$$Y(v) = X\beta(v) + \sigma(v)\varepsilon(v) \qquad (1)$$

where $v = (x,y,z) \in \mathbb{R}^3$ is an index for voxels, $Y(v) = \{Y_1(v), Y_2(v), \ldots, Y_n(v)\}'$ is a vector of observed voxel intensities at v from n scans, X is a known design matrix of size $n \times p$, $\beta(v)$ is a p-dimensional vector of unknown parameters, $\sigma(v)$ is an unknown standard deviation at v, and $\varepsilon(v) = \{\varepsilon_1(v), \varepsilon_2(v), \ldots, \varepsilon_n(v)\}'$ is a vector of unknown random errors with unit variance. Images are denoted by omitting the index v, so that, for example, ε_i denotes the error image from the ith scan. In this study, we primarily focus on data whose error images ε are uncorrelated across subjects or scans, such as PET, second-level fMRI, and VBM data.

Let $\hat{\beta}(v)$ be an unbiased estimate of $\beta(v)$, then the residuals are defined as

$$e(v) = Y(v) - X\hat{\beta}(v)$$

and the residual variance can be estimated by

$$\hat{\sigma}^2(v) = \frac{1}{\eta} e(v)' e(v)$$

where η is the degrees of freedom for errors. If $\varepsilon_i(v)$'s are independent and identically normally distributed, then the statistic image T can be calculated as

$$T(v) = \frac{\boldsymbol{c}\hat{\beta}(v)}{\sqrt{c(X'X)^{-1}c'\hat{\sigma}(v)}}$$

where $\boldsymbol{c}$ is a row vector expressing the contrast of interest. Based on the T image, clusters are formed as a set of contiguous voxels with their $T(v)$ exceeding a fixed cluster defining threshold u_c sharing at least one common edge. For a 3D image, this is known as the 18 connectivity scheme; in a 3 × 3 × 3 voxel cube, all the voxels except eight corner voxels are considered connected to the voxel at the center.

Roughness estimation

For the data described in Eq. (1), the underlying image roughness is calculated as a variance–covariance matrix of the spatial partial derivatives of ε,

$$\Lambda(v) = \mathrm{Var}(\dot{\varepsilon}(v)) \qquad (2)$$

where $\dot{\varepsilon}(v)$ is a 3 × 1 vector of spatial derivatives, typically estimated with first-order differences (Kiebel et al., 1999). This $\Lambda(v)$ matrix is related to a widely used measure of smoothness, full-width at half-maximum (FWHM) of a Gaussian kernel required to smooth a white noise image to have roughness Λ,

$$\mathrm{FWHM}(v) = (4\log 2)^{D/2}\,|\Lambda(v)|^{-1/(2D)} \qquad (3)$$

where D is the spatial dimensionality of the data. FWHM or $|\Lambda|$ can be considered as a global parameter for the entire image, by pooling $\varepsilon(v)$ from all the voxels in Eq. (2) assuming stationarity (Kiebel et al., 1999; Forman et al., 1995), or as a local parameter FWHM(v) or $|\Lambda(v)|$ varying at each voxel (Worsley, 2002; Worsley et al., 1999). Using FWHM, Worsley et al. (1992) developed a concept of resolution elements (RESELs) as a sampling element, which is defined by

$$\mathrm{RESELs} = \frac{V}{\mathrm{FWHM}^D} \tag{4}$$

where V is the search volume (in voxel units). The RESEL, by expressing the volume relative to its smoothness, is a useful metric in nonstationary methods. By reducing the search volume to a single voxel in Eq. (4) and estimating FWHM at each voxel, the local roughness can be expressed as RESEL per voxel (RPV) or RESEL density. RESELs depend on $\Lambda(v)$ through $|\Lambda(v)|^{1/2}$, and Worsley et al. (1999) developed an estimator of $|\Lambda(v)|^{1/2}$ as

$$|\hat{\Lambda}(v)|^{1/2} = |\Delta u(v)'\Delta u(v)|^{1/2} \tag{5}$$

where u is a normalized residual image defined by

$$u(v) = \frac{e(v)}{(e(v)'e(v))^{1/2}}$$

and $\Delta u(v)$ is its spatial gradient vector in the principal axes' directions at v computed with the first-order differences. Alternatively, using the standardized residuals u^*

$$u^*(\eta) = \frac{e(v)}{\left(\frac{1}{v}e(v)'e(v)\right)^{1/2}}$$

Eq. (5) can be written as

$$|\hat{\Lambda}(v)|^{1/2} = \left|\frac{1}{\eta}\Delta u^*(v)'\Delta u^*(v)\right|^{1/2}$$

showing the estimator to be a determinant of matrices averaged over scans. Using $|\hat{\Lambda}(v)|^{1/2}$, an estimate of RPV is obtained as,

$$\widehat{\mathrm{RPV}}(v) = (4\ln 2)^{-D/2}\,|\hat{\Lambda}(v)|^{1/2} \tag{6}$$

A more refined estimate of $|\Lambda(v)|^{\alpha}$ for any α can be obtained using bias corrections in Appendix A. These corrections include a small *df* correction. The *df* correction for $\alpha = 1/2$, as in our RPV estimate, is 1 and thus not necessary. However, when FWHM is to be estimated with $\alpha = 1/2D$ as in Eq. (3), the *df* correction should be employed.

We assume that the smoothness is the same across subjects or scans. However, it is possible that, in practice, there is some intersubject difference in smoothness. Even under such difference, we have previously found that the permutation test is more robust compared to RFT methods (Hayasaka and Nichols, 2003).

Cluster sizes under nonstationarity

Once RPV is calculated at each voxel, the roughness-adjusted cluster-size R (in RESEL units) can be calculated by simply summing RPV values over voxels within each cluster. In other words, for cluster C, the cluster size in terms of RESELs is

$$R = \sum_{v \in C} \mathrm{RPV}(v) \tag{7}$$

Worsley et al. (1999) explain that measuring cluster size in this manner is equivalent to distorting a nonstationary image to stationarity by adjusting the distance between voxels and measuring the cluster volume in the resulting image. Furthermore, a new result by Taylor and Adler (2003) shows that the existence of the warp is no longer necessary and that the random field results work under almost any form of nonstationarity. Such distortion stretches rough areas and shrinks smooth areas, so that on average, the cluster sizes are about the same and stationarity can be achieved (see Fig. 2 for a visual illustration of this). An advantage of measuring cluster sizes in RESELs is the ability to calculate cluster volumes in the distorted or transformed stationary image without carrying out the actual distortion.

The true distribution of R is unknown, thus needs to be approximated by various methods such as RFT or permutations. As in a stationary cluster-size test, the uncorrected P value, or the P value for a single cluster, can be obtained from the approximated null distribution of cluster-size R as the probability of observing a cluster of certain size or larger. In practice, multiple clusters could occur at a given threshold, and testing all the cluster sizes simultaneously using uncorrected P values creates a multiple comparison problem. For instance, if there are 20 clusters and each of which is tested at 0.05 significance level, then, on average, the null hypothesis is rejected at one of these clusters by chance alone even if there is no signal. To account for this problem, family-wise error (FWE) correction is often employed, which controls type I error (false rejection) rates for all the clusters collectively. The FWE correction is achieved by calculating P values based on the distribution of the largest cluster-size $R_{\max}$. The rationale for using the distribution of $R_{\max}$, as well as detailed explanation of its implementation, is found in Holmes et al. (1996) and Hayasaka and Nichols (2003).

Nonstationary RFT cluster-size test

In the nonstationary RFT cluster-size test for t images (Worsley, 2002; Worsley et al., 1999), the distribution of R is approximated by

$$R \sim cB^{1/2}\left(\frac{U_0^D}{\prod_{b=1}^{D} U_b}\right)^{1/2} \tag{8}$$

where B is a beta random variable with parameters $(1,(\eta - D)/2)$, U_0 is a χ^2 random variable with $df = \eta$, and U_b's ($b = 1,2,\ldots,D$) are independent χ^2 random variables with $df = \eta + 2 - b$ (Cao, 1999). The constant c is chosen so that

$$\boldsymbol{E}[R]\boldsymbol{E}[L] = \boldsymbol{E}[N] \tag{9}$$

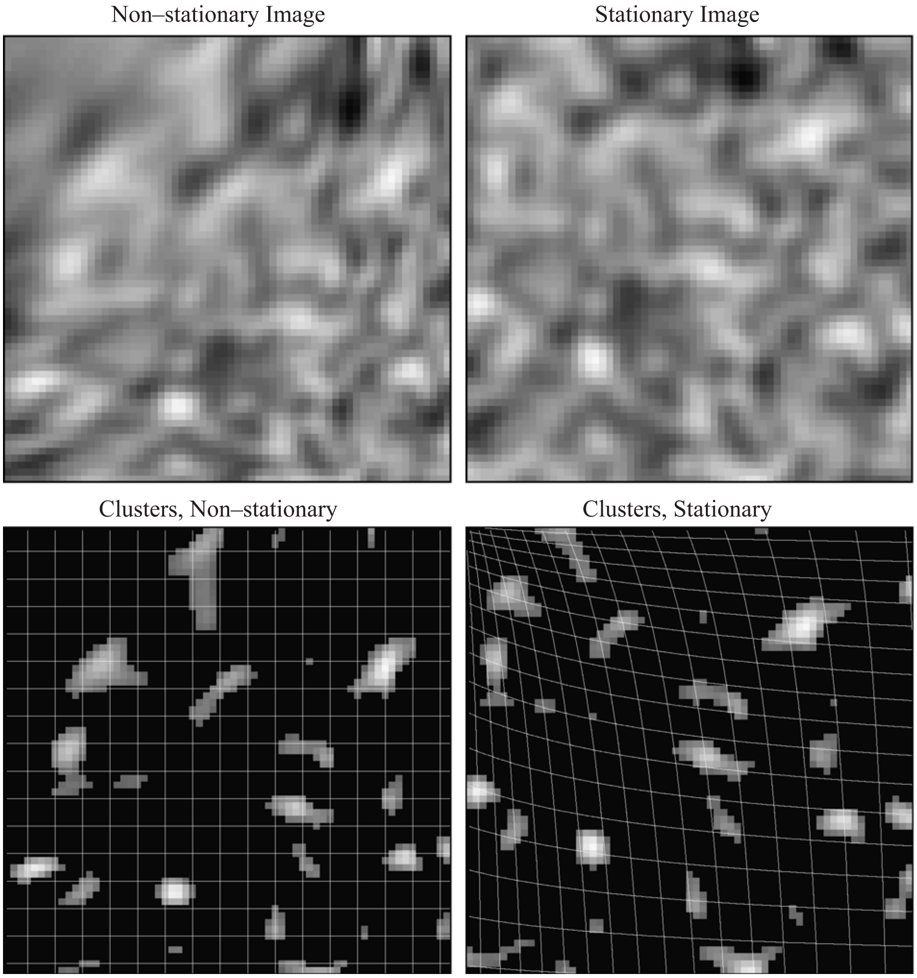

Fig. 2. In a nonstationary image (top left), there is a smooth region at the top left corner and a rough region at the bottom right corner. When thresholded (bottom left), clusters in a smooth (rough) region tend to be large (small). However, if the smooth (rough) region is shrank (stretched), then the resulting image (top right) appears to be stationary, and the size of clusters should follow the same distribution anywhere within the image (bottom right).

where L is the number of clusters and N is the volume (in RESEL units) above u_c. For the search volume Q (RESELs), $\mathbf{E}[N]$ can be found by $QPr(T > u_c)$, and for a sufficiently large search volume and high u_c, $\mathbf{E}[L] = Q\rho(u_c)$ where $\rho(u_c)$ is the Euler characteristic (EC) density of a random field thresholded at u_c (Worsley et al., 1996). In practice, the true cluster size in terms of RESELs R is never known, thus its estimate $\hat{R}$ is used, which is obtained by replacing RPV in Eq. (7) by its estimate $\widehat{\mathrm{RPV}}$ in Eq. (6). $\hat{R}$ is an estimate, assumed to have the distribution

$$\hat{R} \sim R\left(\frac{\prod_{a=1}^{D} V_a}{V_0^D}\right)^{1/2} \tag{10}$$

where V_0 is a χ^2 random variable with $df = \eta$ and V_a's (a = 1,2,. . .,D) are independent χ^2 random variables with $df = \eta - a$. An outline of the derivation of Eq. (10) is found in Appendix B. $\hat{R}$ is more variable than R since it measures cluster volume using the noisy $\widehat{\mathrm{RPV}}$.

Once the uncorrected cluster-size distribution—the distribution of a single cluster size—is obtained from Eqs. (8) and (10), the Poisson clumping heuristic (Aldous, 1989 cited in Cao and Worsley, 2001) is used to compute the distribution of the largest cluster size

$$Pr(R_{\max} > r) = 1 - \exp\{-\boldsymbol{E}[L](1 - F_R(r))\} \tag{11}$$

where $F_R(r)$ is the cumulative cluster-size distribution from which uncorrected P values are calculated. This yields FWE-corrected P values. More details on implementation of the nonstationary RFT test is found in Appendix B.

Nonstationary permutation cluster-size test

The permutation test is based on the idea of exchangeability. Under the null hypothesis, exchangeability asserts that data labels can be permuted without altering the distribution of a test statistic. In this case, the largest cluster size is used as a test statistic in order to control the FWE rate (FWER). Permutation allows an empirical distribution of the largest cluster size to be generated; for each labeling, a statistic image is created and thresholded, and the largest cluster size is recorded. The corrected P value is then calculated by comparing the largest cluster size from the original labeling to this empirical distribution. Under nonstationarity, it is necessary to account for the

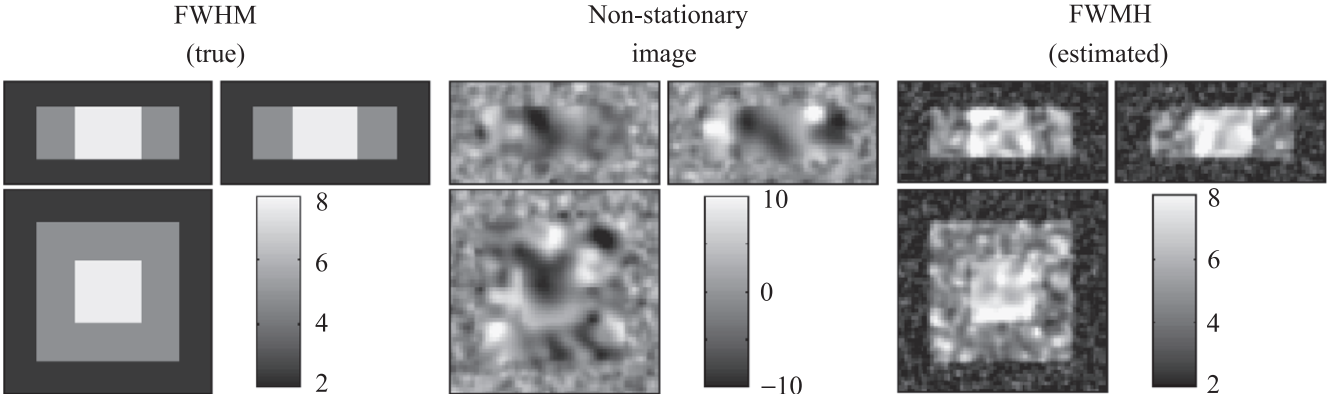

Fig. 3. An example of a nonstationary image. The left panel shows its true FWHM. The rough outer layer (FWHM 2.5 voxels) surrounds the middle layer (FWHM 4.9 voxels), by which encircles the smooth core (FWHM 7.8 voxels). The middle panel shows an actual realization of a nonstationary image with the above smoothness. The right panel shows the FWHM image estimated from a set of 20 nonstationary images.

estimation uncertainty in $\widehat{\mathrm{RPV}}$. In the RFT method, this is done by Eq. (10), but in our permutation test, this is done implicitly by calculating $\widehat{\mathrm{RPV}}$ for each permutation.

Besides calculating cluster sizes in terms of RESELs, our test is the same as the permutation test under stationarity. We used the MATLAB-based SnPM toolbox[1] to implement our test.

Note that this permutation framework is justified only when error images ε are uncorrelated across subjects or scans, as in a second-level fMRI data set, a PET data set, or a VBM data set. To apply this method to correlated data, such as a BOLD fMRI time series, the data must be decorrelated before permutation (Bullmore et al., 1996, 2001).

Simulation-based validation

To validate our permutation test and to compare its performance to the nonstationary RFT test, we carried out Monte-Carlo simulations generating t random noise images, both stationary and nonstationary.

For the stationary t noise simulation, for each realization, two-groups of 32 × 32 × 32 Gaussian images (5 and 5, 10 and 10, or 15 and 15) were generated and a two-sample t test statistic image (df = 8, 18, or 28, respectively) was calculated. Our use of a two-sample t statistic image was motivated by our collaborators' data of comparing controls and patients (Taylor et al., 2003), and the results should be similar to that of a one-sample test with the same degrees of freedom. Each Gaussian image in each realization was generated from a 104 × 104 × 104 white noise image convolved with a Gaussian smoothing kernel (FWHM 0 (no smoothing), 1.5, 3, 6, and 12 voxels), and its outer 36 voxels were truncated to avoid nonstationarity at the edge. The same white noise image was used for different kernel widths in order to reduce computation time.

For the nonstationary t noise image simulation, for each realization, a two-sample t test statistics image with df = 18 was generated from two sets of ten 64 × 64 × 32 nonstationary Gaussian noise images. Each of these nonstationary Gaussian images was created from a single 100 × 100 × 68 white noise image. The white noise image was smoothed with three different 3D Gaussian kernels, producing three images with low, medium, and high smoothness. These images were combined in a way that a rough outer layer encloses a medium-smoothness middle layer, which encircles a smooth core (see Fig. 3, left). The center core was 20 × 20 × 16 voxels, centered within a 44 × 44 × 16 voxel middle layer, which itself was centered in the volume. The combined image was smoothed again with a 3D Gaussian filter with FWHM 2 voxels. This secondary smoothing was applied in order to eliminate discontinuities at the borders of different smoothness, producing a smooth image. The outer 18 voxels of the smoothed image were truncated, yielding a 64 × 64 × 32 nonstationary image with 6400 core voxels, 24576 middle-layer voxels, and 106496 outer-layer voxels. The middle panel in Fig. 3 displays an example of a nonstationary image, and the right panel shows the local FWHM image estimated from a set of 20 nonstationary images. Table 1 shows different smoothness settings for the nonstationary simulation.

In both stationary and nonstationary simulations, each generated t image was thresholded and clusters were formed. For the stationary simulation, three cluster defining thresholds u_c were used (corresponding to t critical values of 0.01, 0.001, and 0.0001), while only one threshold (0.01) was used in the nonstationary simulation. Three thousand realizations were generated for each sample size in the stationary simulation, and 2000 realizations were generated for the nonstationary simulation. Both RFT and permutation test with 100 permutations were applied.

Table 1
Different settings for the nonstationary simulations

Primary Smoothing				
Outer layer	4.5	4.0	3.0	1.5
Middle layer	4.5	4.5	4.5	4.5
Core	4.5	5.0	6.0	7.5
Secondary Smoothing	2	2	2	2
True Smoothness				
Outer layer	4.9	4.5	3.6	2.5
Middle layer	4.9	4.9	4.9	4.9
Core	4.9	5.4	6.3	7.8

The outer layer was set to be rougher than the middle layer, whereas the core was set to be smoother than the middle layer. Smoothness is in terms of FWHM in voxels. The theoretical smoothness as a combination of the primary and secondary smoothing is the square root of squared sum of these smoothings (Holmes, 1994).

[1] Andrew Holmes and Thomas Nichols, http://www.fil.ion.ucl.ac.uk/spm/snpm/.

Each test's rejection rates were recorded and the corresponding 95% confidence intervals (CIs) were obtained by normal approximation of a binomial proportion $\hat{p}\pm1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n_{\mathrm{r}}}}$, where $\hat{p}$ is the observed rejection rate and n_{r} is the number of realizations. The significance level of tests was set to 0.05; thus, the CIs should cover 0.05. Since a large number of CIs were examined, some may not capture 0.05 purely by chance. However, it is impossible to calculate the expected number of CIs outside of 0.05 by chance alone, since simulation results are correlated due to the same white noise image being used in multiple smoothness settings. If the simulated rejection rate is smaller than 0.05, then the test is conservative but still considered as valid. On the other hand, if the rejection rate is greater than 0.05, then the test is anticonservative, or liberal, and no longer considered as valid.

Table 2
Results from the nonstationary simulation

	Smoothness [FWHM voxels] (outer, middle, core)			
	(4.9, 4.9, 4.9)	(4.5, 4.9, 5.4)	(3.6, 4.9, 6.3)	(2.5, 4.9, 7.8)
Estimated overall FWHM	4.86	4.54	3.84	2.85
Rejection rates				
Stationarity assumed				
RFT (SPM)	0.031	0.038	0.137	0.538
Perm (SnPM)	0.050	0.045	0.046	0.052
Nonstationarity assumed				
RFT	0.008	0.004	0.001	0.000
Perm	0.046	0.052	0.052	0.048

Rejection rates for the RFT and permutation tests, both when stationarity and nonstationarity are assumed, along with the smoothness estimates in terms of FWHM assuming stationarity. The 95% Monte-Carlo confidence interval for rejection rate α = 0.05 is 0.040–0.060.

Data analysis

The PET data set used in this work is a subset of a study comparing emotional responses between combat veterans with post-traumatic stress disorder (PTSD) and controls. Male participants were recruited by advertisement in veterans affairs hospitals and by local newspapers. The patients had been diagnosed with PTSD according to the DSM-IV criteria. Both patients and controls were screened for chronic illness, dementia, substance abuse, and structural abnormalities in the head. Written informed consent was obtained from all participants, approved by the University of Michigan and Ann Arbor VAMC IRBs. Sixteen PTSD patients (PP) and 14 normal controls without combat experience (NC) were included in the data set for this work. For each subject, two neutral scripts (ns) and two traumatic/stressful scripts (ss) were formulated based on his past experiences described during an interview. These scripts were presented in a counterbalanced design of two ns and two ss, and the subjects were instructed to listen, reexperience, and maintain the evoked state. PET scans were acquired by Siemens ECAT-EXACT or ECAT-HR+ scanner. Obtained PET images were realigned for head movement correction, anatomically standardized to the Montréal Neurological Institute (MNI) coordinates, and smoothed with a 12-mm FWHM filter using the SPM99 computer package (Wellcome Department of Imaging Neuroscience, Univer-

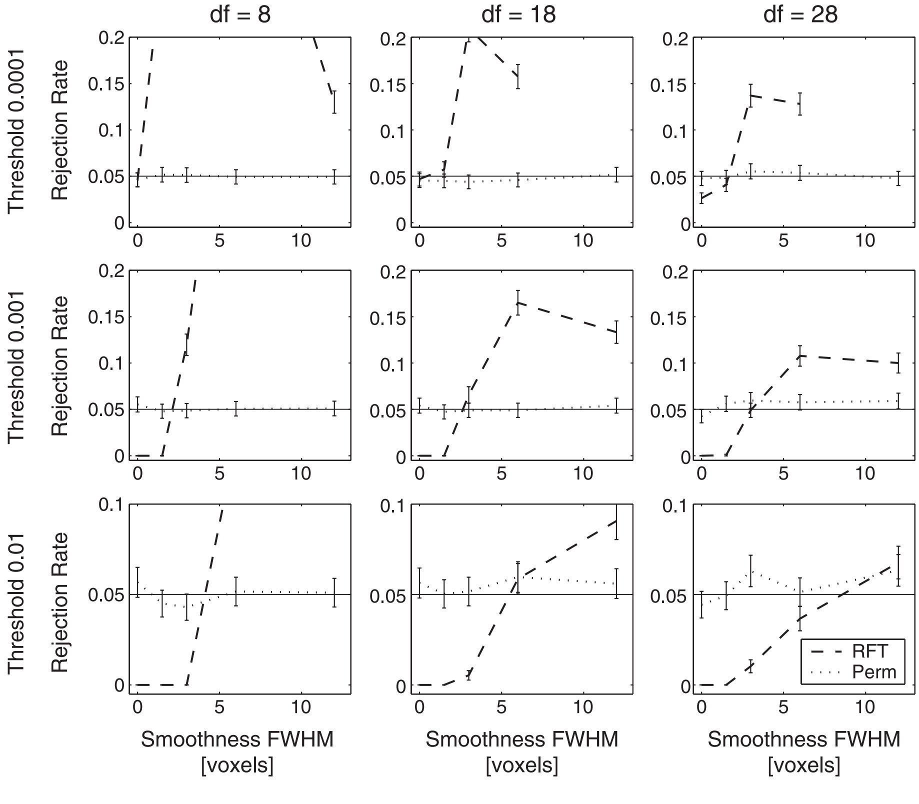

Fig. 4. Family-wise rejection rates of the nonstationary tests on stationary data. The RFT and permutation tests are compared for different sample sizes (5 and 5, 10 and 10, and 15 and 15, from left to right) when thresholded at *t* critical values 0.01, 0.001, and 0.0001 (from bottom to top), along with their 95% confidence intervals. Fine solid lines indicate the desired FWER (0.05) of the test.

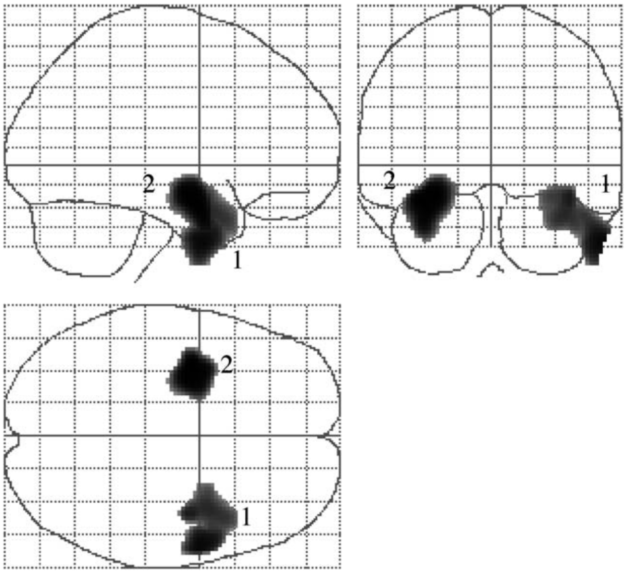

Fig. 5. The two largest clusters found at 0.01 threshold (t_{28} = 2.47).

sity College London, UK). For each subject, a contrast image of ss–ns, of size 79 × 95 × 68 with 2 × 2 × 2 mm voxels, was obtained for a random effect analysis (Holmes and Friston, 1999). To the resulting contrast images, a group comparison of NC-PP was made by the nonstationary RFT and permutation tests, as well as by the stationary RFT test, as implemented in SPM99, and by the stationary permutation test, as implemented in the SnPM toolbox. A Dell PC with dual 2.4 GHz Xeon processors and 2 GB of RAM with MATLAB version 6.5 (MathWorks Inc., Natick, MA) running on a Linux platform was used in the analysis, which took 9 h to perform the nonstationary permutation test with 1000 permutations with one of the two processors.

Results

Simulation-based validation

The results from the stationary simulation are shown in Fig. 4. The RFT test is found to be anticonservative for the highest threshold (0.0001) for any smoothness or *df*. For lower thresholds, the test is found to be conservative for low smoothness, and as smoothness increases, the test becomes less conservative. However, this increase in the rejection rates surpasses the designed significant level 0.05, and the test becomes anticonservative for high smoothness. This trend is particularly apparent in low *df*s, but for high *df*, the rejection rates are close to 0.05 for high smoothness.

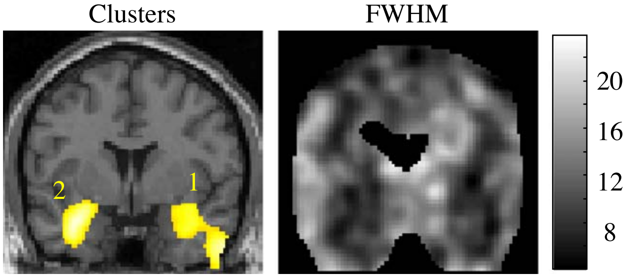

Fig. 6. Clusters 1 and 2 from the analysis results (left), and the FWHM image at the corresponding location (right). Cluster 1 includes both smooth and rough regions, so its smoothness is similar to the overall image smoothness. On the other hand, Cluster 2 lies in a rough region, and consequently, its smoothness is smaller than the overall image smoothness.

The rejection rates from the permutation test are close to 0.05 in any setting, indicating that the test performs well under any settings examined.

The results from the nonstationary simulation are shown in Table 2. As images become more nonstationary, the stationary RFT becomes more anticonservative; this is due to the underestimation of the smoothness in the middle and core layers. On the other hand, the nonstationary RFT becomes conservative as images become more nonstationary. This may be because a large portion of a simulated image is the rough outer layer where, as seen in the stationary simulation, the nonstationary RFT test is generally very conservative at a low threshold (0.01 in this case). Both stationary and nonstationary permutation tests produced rejection rates close to 0.05 in all the settings in this simulation. It is not surprising that the stationary permutation test remains exact, since stationarity assumption is not required in this test (see Discussion).

PET data analysis

At 0.01 threshold (or t_{28} = 2.47), nine clusters are found, of which the two largest are shown in Fig. 5. These clusters are located in bilateral amygdala (with some extension into middle and inferior temporal gyrus), and represent differential activation associated with processing of trauma/stress-related emotional content. The corrected *P* values for these clusters by the RFT and permutation tests, both under stationarity and nonstationarity, are shown in Table 3. These results are consistent with findings from other functional activation studies of emotion and emotion-based recall and imagery (Phan et al., 2002).

The average smoothness within Cluster 1 (average FWHM 9.5 voxels) is very similar to the overall smoothness of the entire image (FWHM 9.4 voxels). For this cluster, the corrected *P* value for the permutation test is larger in the nonstationary permutation test (*P* =

Table 3
The two largest clusters and their spatial extent under stationarity and nonstationarity, as well as their *P* values, corrected for the whole brain volume, for the RFT and permutation tests

Cluster	Extent (voxels)	Stationarity assumed			Nonstationarity assumed			Peak	
		Extent (RESELs)	$p_{\text{FWE-corr}}$		Extent (RESELs)	$p_{\text{FWE-corr}}$		*T*	P_{uncorr}
			RFT	Perm		RFT	Perm		
1	1274	1.552	0.224	0.116	1.463	0.228	0.159	3.69	<0.001
2	940	1.145	0.418	0.205	1.768	0.159	0.107	4.20	<0.001

0.159) compared to the one under stationarity (P = 0.116). This reduction in sensitivity is likely due to increased uncertainty in the permutation distribution caused by the variability in smoothness estimation in the nonstationary method.

Cluster 2 is located in a relatively rough region of the image (see Fig. 6), with its average smoothness FWHM 8.1 voxels, which is smaller than the overall average FWHM for the entire image. For this cluster, both nonstationary tests show increased sensitivity, with the corrected P values decreasing dramatically (from 0.418 to 0.159 in RFT and from 0.205 to 0.107 in permutation).

To verify the accuracy of these results using 1000 permutations, we reran the analysis with 10 000 permutations. The P values obtained from 10 000 permutations are very similar to that obtained from 1000 permutations. For Cluster 1, P values were 0.159 for 1000 permutations and 0.179 for 10 000 permutations, and for Cluster 2, P values were 0.107 and 0.122, respectively.

Discussion

Nonstationary permutation test

The nonstationary permutation test shows excellent control of FWER for all smoothness, *df*, and cluster defining thresholds considered. Note that it shows even more accurate control of FWER than the stationary permutation test. At very low smoothness, the stationary permutation test controls FWER conservatively because of discreteness in the cluster-size distribution: Most of clusters have the same size, 1, 2, or 3 voxels, and there is no critical cluster size that can exactly control FWER (though P values are still accurate) (Hayasaka and Nichols, 2003). On the other hand, when cluster sizes are measured in RESELs, the cluster-size distribution is continuous: Clusters rarely have the same size in RESELs, even though they may have the same number of voxels. As a result, the nonstationary permutation test controls FWER exactly.

With exchangeability under the null hypothesis and modest smoothness, both stationary and nonstationary tests should be exact, as seen in Table 2. However, under nonstationarity, the stationary test is not uniformly sensitive, as the maximum cluster-size distribution is most influenced by smooth areas. Fig. 7 shows images of critical cluster sizes for both tests, in terms of voxels and RESELs, from a realization in the nonstationary simulation. Since the stationary test has a uniform critical cluster size in voxel units (Fig. 7, top left), in units of RESELs, the critical cluster size is smaller in the core and the middle layer compared to the outer layer (Fig. 7, bottom left). Under nonstationarity, the null cluster-size distribution in RESELs should be homogeneous, this means that the stationary test has greater sensitivity in the smooth center of the image, at the expense of reduced sensitivity at the rough edge. In contrast, the nonstationary test has a uniform critical cluster size in RESEL units (Fig. 7, bottom right); thus, the test had uniform sensitivity under nonstationarity.

Nonstationary RFT test

For the nonstationary RFT test, even though the test is designed to overcome nonstationarity, it is still prone to violations in other assumptions in RFT, such as failure in lattice approximation of a smooth random field, biases in its approximated cluster-size

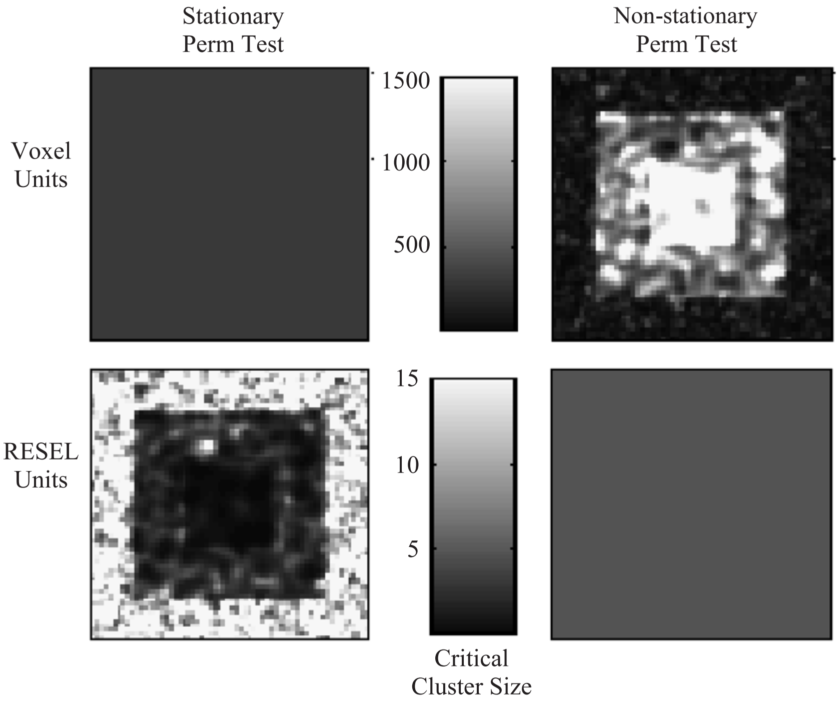

Fig. 7. Critical cluster sizes from the stationary (left) and nonstationary (right) permutation tests for a realization from the nonstationary noise simulation with FWHM = 2.5 (outer), 4.9 (middle), and 7.8 (core) voxels. When critical cluster sizes are expressed in voxel units (top), the stationary test is uniformly sensitive. However, when the critical cluster sizes are expressed in RESEL units (bottom), the nonstationary test is uniformly sensitive, whereas the stationary test is more sensitive in the core or the middle layer than the outer layer. Despite these differences, both permutation tests are exact, they just apportion sensitivity differently.

distribution, and clusters being truncated by the edge, as discussed in Hayasaka and Nichols (2003). However, the RFT test appears to perform well in low thresholds with sufficient smoothness, as *df* increases. We were unable to carry out simulations with $df > 28$ because of limitation in our computational resources, but for higher *df* and sufficient smoothness, the RFT test may be more suitable than the permutation test, considering that the computational burden is more severe in the permutation test for high *df*.

Sensitivity and smoothness estimation

The results from the data analysis indicate that the nonstationary permutation test has reduced sensitivity when the smoothness of a cluster is about equal to the overall smoothness. This reduced sensitivity is due to increased variability in smoothness estimation. Under an assumption of stationarity, smoothness is estimated pooling over all the voxels. On the other hand, in a nonstationary method, smoothness has to be estimated at each voxel separately without pooling, resulting smoothness estimates having much smaller effective *df* and a larger variation. When cluster sizes are measured in terms of RESELs, such uncertainty in RPV estimation is propagated into cluster sizes.

The impact of local smoothness estimation can be gauged by comparing stationary and nonstationary methods on stationary data. Fig. 8 shows results from the t_{18} stationary noise simulation with 0.01 cluster defining threshold. The left panel in Fig. 8 shows the 95th percentile of the maximum cluster size, computed assuming stationarity (dashed line) or not (dotted line). Since stationarity holds, the true cluster sizes are equivalent, but estimation variability in RPV increases the 95th percentile as much as one RESEL. In order to correct for such variability, both RFT and permutation tests have larger critical cluster sizes under nonstationarity, compared to that under stationarity (the middle and right panels, Fig. 8).

To overcome this problem, the variability in smoothness estimation needs to be reduced. One way to achieve this is by smoothing the RPV image to implicitly increase *df*, in a similar manner as the variance smoothing in the permutation test (Holmes et al., 1996). However, smoothing RPV image blurs rough regions, which results in adjusted RESEL cluster sizes in rough regions becoming smaller, reducing the sensitivity of the test in such regions. Also, the fine structure found in the local smoothness image from a VBM data set (Fig. 1, right) would be lost with smoothing.

Despite the overall reduced sensitivity mentioned above, in the data analysis, the cluster-size tests under nonstationarity demonstrated their increased sensitivity in a rough region. Another perspective in this conclusion is that when stationarity is falsely assumed in a nonstationary data set, a stationary cluster-size test may not have enough sensitivity to detect clusters in rough regions. Furthermore, underestimation in smoothness caused by rough regions could lead to anticonservativeness, especially in the RFT test, as seen in the nonstationary simulation. In order to avoid the biases in stationary cluster-size tests mentioned above, the image roughness should be examined.

Practical recommendations

For a practitioner wishing to apply cluster-size inference to his or her data set, the first step should be to examine whether or not images are stationary. This step is necessary because nonstationary methods' sensitivity can be lower than stationary methods' when applied to stationary images, due to extra uncertainty in smoothness estimation. One way to examine stationarity is to examine the RPV image generated by the SPM99 package after an analysis. The RPV image may not be readily interpretable, but it can be easily converted to FWHM by

$$\mathrm{FWHM}(v) \simeq \frac{1}{(\mathrm{RPV}(v))^{1/3}}$$

The FMRISTAT package (http://www.math.mcgill.ca/keith/fmristat) computes both RPV(v) and the bias corrected FWHM(v).

Once local FWHM is known, then stationarity can be more easily examined. If a gray matter probability map is readily available, as in a VBM data set, FHWM(v) or RPV(v) can be plotted against the gray matter probability to see if there is any pattern. Even if the smoothness does not systematically vary with gray matter density, nonstationarity may be present. If stationarity assumption is in doubt, then a nonstationary cluster-size test should be used.

If a nonstationarity test is to be used, a choice needs to be made on which test to be used, either the permutation or the RFT test. For low *df* (<30), the permutation test is more desirable. As seen in our simulations, it performs well for any threshold and smoothness, and though time consuming, it is computationally feasible for a small number of scans. Since the permutation test is exact regardless of threshold, widely used 0.01 threshold can be used. For high *df* (>30) and sufficient smoothness, with majority of voxels with FWHM > 3, the RFT test is preferred, since the permutation test becomes more computationally intensive as *df* increases. For the choice of threshold, though conservative for the RFT test, a 0.01 threshold is

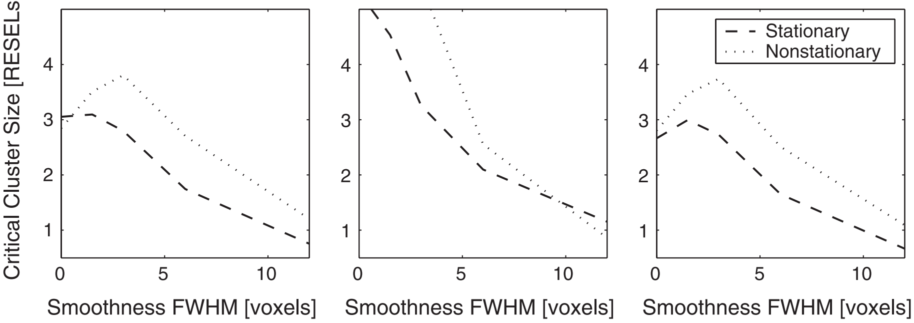

Fig. 8. Corrected critical cluster sizes from t_{18} simulation with 0.01 threshold. The true critical cluster sizes, or the 95th percentiles from the simulation (left), as well as the ones from the RFT test (middle) and the ones from the permutation test (right).

statistically valid, since it keeps false rejections under the desired level, as seen from the t_{28} simulation. For a sufficiently large *df* (30), widely used 0.001 threshold may also produce a statistically valid test, and may be less conservative than 0.01 threshold. However, we were unable to verify this since we were unable to simulate images with higher *df*. Even with high *df*, for images with large areas of low smoothness (FWHM < 3), the permutation test is preferred if computational resources permit, since the RFT test is considerably conservative for low smoothness.

Conclusion

We developed a cluster-size permutation test under nonstationarity using the framework developed by Worsley et al. (1999) to adjust cluster sizes according to local roughness. Through simulations and the data analysis, it was found that applying stationary cluster-size tests to nonstationary images could lead to reduced sensitivity in rough regions and increased false positives in smooth regions. Our nonstationary permutation test was found to be exact under any simulation settings examined, producing the desired significance level. The nonstationary RFT test was found to be conservative in our nonstationary noise simulation. However, the nonstationary RFT test remains valid while the stationary RFT test could become anticonservative under nonstationarity. For low *df*, where the RFT test is unstable, our permutation test is very stable, producing rejection rates close to the desired significance level. Thus, our test is desirable for low *df* settings. For high *df* and high smoothness, the RFT test seems to perform reasonably well, thus it is more suitable than the computationally intensive permutation test. Nonstationary cluster-size tests have increased sensitivity in rough regions compared to stationary cluster-size tests, despite the loss of overall sensitivity due to smoothness estimation variability. These nonstationary cluster-size tests can be more widely applied to data known to be nonstationary, such as VBM data.

Acknowledgments

This Human Brain Project/Neuroinformatics research is funded by the National Institute of Mental Health, National Institute on Aging, and the National Institute of Biomedical Imaging and Bioengineering. This research was also supported in part by a VA/DoD collaborative program grant to IL. The authors would like to thank Dr. Stephan F Taylor and Dr. James Becker for providing us the data sets for the illustration in Fig. 1.

Appendix A. Bias corrections in smoothness estimation

Since FWHM(v) and RPV(v) are functions of $|\Lambda(v)|$, in order to estimate FWHM(v) or RPV(v), $|\Lambda(v)|$ needs to be estimated accurately. From Kiebel et al. (1999), it can be shown that

$$|\dot{u}(v)'\dot{u}(v)| \sim |\Lambda(v)| \frac{\prod_{a=1}^{D} V_a}{V_0^D}. \tag{12}$$

Thus, $|\dot{u}(v)'\dot{u}(v)|$ can be used as an estimate of $|\Lambda(v)|$ or its power, with appropriate bias corrections below.

For large degrees of freedom η, it is clear that $|\dot{u}(v)'\dot{u}(v)| \to |\Lambda(v)|$ since the powers of V's sum to zero. For small degrees of freedom, we can obtain an unbiased estimate by dividing by a correction factor c:

$$|\hat{\Lambda}(v)|^{\alpha} = |\dot{u}(v)'\dot{u}(v)|^{\alpha}/c,$$

where

$$c = \boldsymbol{E}\left[\frac{|\dot{u}(v)'\dot{u}(v)|^{\alpha}}{|\Lambda(v)|^{\alpha}}\right] = \boldsymbol{E}\left[\left(\frac{\prod_{a=1}^{D} V_a}{V_0^D}\right)^{\alpha}\right]$$

$$= \frac{\Gamma\left(\frac{\eta}{2} - \alpha D\right)}{\Gamma\left(\frac{\nu}{2}\right)} \prod_{a=1}^{D} \frac{\Gamma\left(\frac{\eta - a}{2} + \alpha\right)}{\Gamma\left(\frac{\nu - a}{2}\right)},$$

provided that the arguments of all gamma functions are positive. In practice, this means $\eta > D$. Interestingly, $c = 1$ for RPV ($\alpha = 1/2$); thus, in our simulations and analysis, the *df* correction was not employed.

If $\eta \leq D$, then the bias in the above estimator cannot be corrected. In this case, we assume that $\Lambda(v)$ is diagonal, and base our estimator on the product of the diagonal elements $\prod_{k=1}^{D} diag_k(\dot{u}(v)'\dot{u}(v))$ which has the same distribution as (12) but the *df* of V_a is $\eta - 1$ instead of $\eta - a$, $a = 1, 2, \ldots, D$. The correction factor c is the same as above but with $a = 1$ inside the product.

Appendix B. RFT *P* value approximation

The *P* value for a cluster C in the nonisotropic case depends on the cluster RESELs R rather than the cluster volume measured in voxels. Since the cluster volume is small, the random error in the estimated cluster RESELs $\hat{R}$ can make a substantial contribution. From Eqs. (6) and (7), the distribution of $\hat{R}$ now depends on the local properties of the random field M, defined by

$$\hat{R} = \sum_{v \in C} \widehat{RPV}(v) = \sum_{v \in C} RPV(v) M(v)$$

where

$$M(v) = \frac{|\hat{\Lambda}(v)|^{1/2}}{|\Lambda(v)|^{1/2}}.$$

From (12), the distribution of $M(v)$ at each point is

$$M(v) \sim \left(\frac{\prod_{a=1}^{D} V_a}{V_0^D}\right)^{1/2}$$

where V_a's are independent χ^2 random variables with $\eta - a$ degrees of freedom. However, the distribution of weighted sums of $M(v)$, as above, depends on the spatial covariance function of $M(v)$, which in turn depends on the spatial covariance function of ε through more than just Λ, specifically, the variances of fourth derivatives of ε.

However, we can make a simple approximation as follows. First, note that $\boldsymbol{E}(M(v))$ and $\boldsymbol{Sd}(M(v))$ are the same at every voxel.

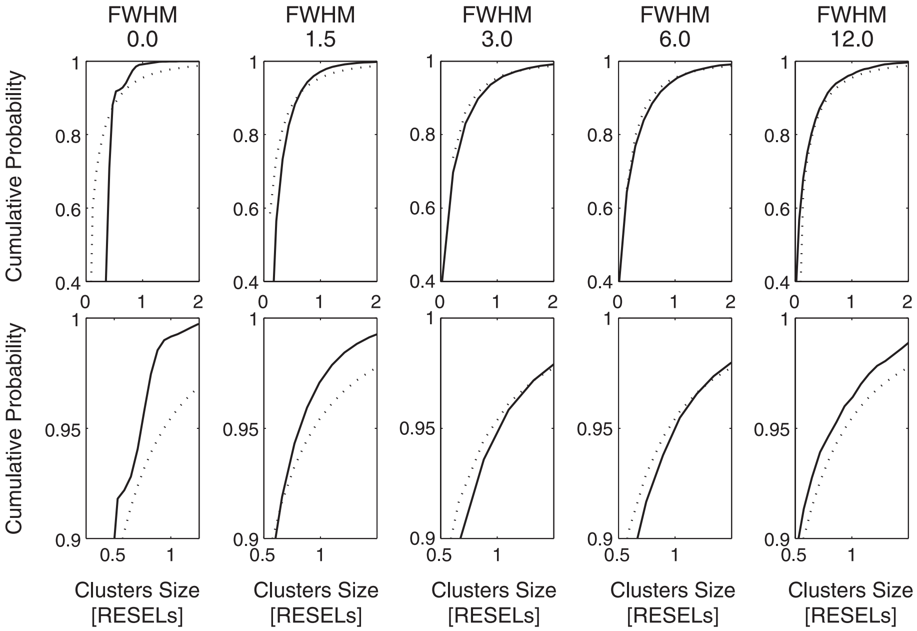

Fig. 9. The observed distribution of cluster RESELs R at 0.01 threshold for two-sample t_{18} images (solid lines) and its theoretical approximation $\hat{R}$ based on RFT (dashed lines). The top row shows the overall shape of the distributions from the 40th to 100th percentiles, while the bottom row shows the shape of the distributions around the 95th percentiles, the uncorrected critical cluster sizes.

Then, since $\boldsymbol{Cov}(M(u),M(v)) \leq \boldsymbol{Sd}(M(u))\boldsymbol{Sd}(M(v))$, it is straightforward to show that

$$\boldsymbol{Sd}(\hat{R}) \leq R\,\boldsymbol{Sd}(M(v))$$

with equality if $M(v)$ is flat over the cluster C. This suggests that we can bound $\hat{R}$ stochastically by R times the distribution of M at a single voxel. This bound would be close if M were very smooth, or if C were very small. This leads us to our approximation

$$\hat{R} \approx R\left(\frac{\prod_{a=1}^{D} V_a}{V_0^D}\right)^{1/2}.$$

Combining this with the approximate distribution for R from Cao (1999) requires no extra computational effort. The distribution $\hat{R}$ simply multiplies the distribution of R by a few more independent χ^2 random variables raised to various powers. In practice, the distribution function of $\hat{R}$ is best calculated by first taking logarithms, so that log $\hat{R}$ is then a sum of independent random variables. The density of a sum is the convolution of the densities, whose Fourier transform is the sum of the Fourier transforms. It is easier to find the upper tail probability of log $\hat{R}$ by replacing the density of one of the random variables by its upper tail probability before doing the convolution. The obvious choice is the Beta random variable, since its upper tail probability has a simple closed form expression. This method has been implemented in the stat_threshold.m function of FMRISTAT, available from http://www.math.mcgill.ca/keith/fmristat.

Fig. 9 shows the observed and theoretical distribution of cluster RESELs R. The RFT does not approximate the distribution well for low smoothness, but for sufficient smoothness, the RFT approximation is close to the observed distribution.

References

Aldous, D., 1989. Probability approximations via the Poisson clumping heuristic. Springer, New York.

Ashburner, J., Friston, K.J., 2000. Voxel-based morphometry—The methods. NeuroImage 11, 805–821.

Bullmore, E., Brammer, M., Williams, S.C.R., Rabe-Hesketh, S., Janot, N., David, A., Mellers, J., Howard, R., Sham, P., 1996. Statistical methods of estimation and inference for functional MR image analysis. Magn. Reson. Med. 35, 261–277.

Bullmore, E., Long, C., Suckling, J., 2001. Colored noise and computational inference in neurophysiological (fMRI) time series analysis: resampling methods in time and wavelet domains. Hum. Brain Mapp. 12, 61–78.

Cao, J., 1999. The size of the connected components of excursion sets of χ^2, t, and F fields. Adv. Appl. Probab. 31, 579–595.

Cao, J., Worsley, K.J., 2001. Applications of random fields in human brain mapping. In: Moore, M. (Ed.), Spatial Statistics: Methodological Aspects and Applications. Springer Lect. Notes Stat., vol. 159. Springer, New York, pp. 169–182.

Forman, S.D., Cohen, J.D., Fitzgerald, J.D., Eddy, W.F., Mintun, M.A., Noll, D.C., 1995. Improved assessment of significant activation in functional magnetic resonance imaging (fMRI): use of a cluster-size threshold. Magn. Reson. Med. 33, 636–647.

Friston, K.J., Worsley, K.J., Frackowiak, R.S.J., Mazziotta, J.C., Evans, A.C., 1994. Assessing the significance of focal activations using their spatial extent. Hum. Brain Mapp. 1, 210–220.

Friston, K.J., Holmes, A., Poline, J.-B., Price, C.J., Frith, C.D., 1996. Detecting activations in PET and fMRI: levels of inference and power. NeuroImage 4, 223–235.

Hayasaka, S., Nichols, T.E., 2003. Validating cluster size inference: random field and permutation methods. NeuroImage 20, 2343–2356.

Holmes, P., 1994. Statistical issues in functional brain mapping. PhD thesis, University of Glasgow.

Holmes, A.P., Friston, K.J., 1999. Generalizability, random effects, and population inference. Proceedings of the 4th International Conference on Functional Mapping of the Human Brain, June 7–12, 1998, Montréal, Canada. NeuroImage, vol. 7, p. S754.

Holmes, A.P., Blair, R.C., Watson, J.D.G., Ford, I., 1996. Nonparametric analysis of statistic images from functional mapping experiments. J. Cereb. Blood Flow Metab. 16 (1), 7–22.

Kiebel, S.J., Poline, J.-B., Friston, K.J., Holmes, A.P., Worsley, K.J., 1999. Robust smoothness estimation in statistical parametric maps using standardized residuals from the general linear model. NeuroImage 10, 756–766.

Ledberg, A., Åkerman, S., Roland, P.R., 1998. Estimation of the probability of 3D clusters in functional brain images. NeuroImage 8, 113–128.

Nichols, T.E., Holmes, A.P., 2002. Nonparametric permutation tests for functional neuroimaging: a primer with examples. Hum. Brain Mapp. 15, 1–25.

Phan, K.L., Wager, T., Taylor, S.F., Liberzon, I., 2002. Functional neuroanatomy of emotion: a meta-analysis of emotion activation studies in PET and fMRI. NeuroImage 16, 331–348.

Poline, J.-B., Mazoyer, B.M., 1993. Analysis of individual positron emission tomography activation maps by detection of high signal-to-noise-ratio pixel clusters. J. Cereb. Blood Flow Metab. 13, 425–437.

Poline, J.-B., Worsley, K.J., Evans, A.C., Friston, K.J., 1997. Combining spatial extent and peak intensity to test for activations in functional imaging. NeuroImage 5, 83–96.

Roland, P.E., Levin, B., Kawashima, R., Åkerman, S., 1993. Three-dimensional analysis of clustered voxels in 15-*O*-butanol brain activation images. Hum. Brain Mapp. 1, 3–19.

Taylor, J.E., Adler, R.J., 2003. Euler characteristics for Gaussian fields on manifolds. Ann. Probab. 31, 533–563.

Taylor, S.F., Welsh, R.C., Phan, K.L., Liberzon, I., 2003. Medial prefrontal cortex dysfunction in schizophrenia. Annual Meeting of the American College of Neuropsychopharmacology Abstracts. San Juan, Puerto Rico, December, 2003, vol. 57, p. 146.

Varhola, M., Caldararo, R., Miller, S., Becker, J., 2000. Voxel-based analysis of brain morphology in HIV/AIDS. Proceedings of the 6th International Conference on Functional Mapping of the Human Brain, June 12–16, 2000, San Antonio, TX, USA. NeuroImage, vol. 11, p. S141.

Worsley, K.J., 2002. Non-stationary FWHM and its effect on statistical inference of fMRI data. Presented at the 8th International Conference on Functional Mapping of the Human Brain, June 2–6, 2002, Sendai, Japan. NeuroImage, vol. 16 (2), pp. 779–780. Available on CD-ROM.

Worsley, K.J., Evans, A.C., Marrett, S., Neelin, P., 1992. Three-dimensional statistical analysis for CBF activation studies in human brain. J. Cereb. Blood Flow Metab. 12, 900–918.

Worsley, K.J., Marrett, S., Neelin, P., Vandal, A.C., Friston, K.J., Evans, A.C., 1996. A unified statistical approach for determining significant signals in images of cerebral activation. Hum. Brain Mapp. 4, 58–73.

Worsley, K.J., Andermann, M., Koulis, T., MacDonald, D., Evans, A.C., 1999. Detecting changes in nonisotropic images. Hum. Brain Mapp. 8, 98–101.