

# **Validating Cluster Size Inference: Random Field and Permutation Methods**

Satoru Hayasaka, Thomas E Nichols

*Department of Biostatistics, The University of Michigan, Ann Arbor, MI, USA*

May 9, 2003

*Running title: Validating Cluster Size Inference*

Address for correspondence:  
Thomas E. Nichols  
Department of Biostatistics  
University of Michigan  
1420 Washington Height,  
Ann Arbor, MI48109  
Phone: +1-734-936-1002  
Fax: +1-734-763-2215  
email: nichols@umich.edu

## Abstract

Cluster size tests used in analyses of brain images can have more sensitivity compared to intensity based tests. The random field (RF) theory has been widely used in implementation of such tests, however the behavior of such tests is not well understood, especially when the RF assumptions are in doubt. In this paper, we carried out a simulation study of cluster size tests under varying smoothness, thresholds, and degrees of freedom, comparing RF performance to that of the permutation test (Holmes *et al.*, 1996; Nichols & Holmes, 2002), which is known to be almost exact. For Gaussian images, it was found that the RF methods were generally conservative, especially for low smoothness and low threshold. For  $t$  images, the RF tests were found to be conservative at lower thresholds and do not perform well unless the threshold is high and images are sufficiently smooth. The permutation test performed well for any settings if the discreteness in cluster sizes is accounted for. We make specific recommendations on when permutation tests are to be preferred to RF tests.

# 1 Introduction

Central interest to neuroscientists is the detection of changes in brain images obtained from PET (Positron Emission Tomography) or MRI (Magnetic Resonance Imaging). Cluster size inference is one of the approaches used in such investigation. A typical cluster size test consists of two steps. First, clusters are defined as sets of contiguous voxels whose intensity exceeds a pre-selected cluster defining threshold  $u_c$ , then the null hypothesis is tested by examining whether or not the spatial extent of these clusters is unusually large by chance alone.

This test is known to have increased sensitivity compared to tests based on voxel intensity (Friston *et al.*, 1996) when the signal is spatially extended. However, it has not been validated under various conditions (smoothness, threshold, etc), in particular for  $t$  images. Furthermore, the sensitivity of cluster size tests outside of ideal conditions is not understood either. In this paper we seek to characterize under what condition cluster size tests perform well. In addition, when these tests do not perform well, we examine probable causes in detail.

The idea of cluster size inference was pioneered by Poline & Mazoyer (1993) and Roland *et al.* (1993); they generate the distribution of cluster sizes from simulated images having the same characteristics, such as spatial autocorrelation, as the observed data. This approach was further studied in fMRI by Forman *et al.* (1995) and in PET by Ledberg *et al.* (1998). The most widely used methods, however, are the ones based on the random field (RF) theory (Friston *et al.*, 1994; Cao & Worsley, 2001).

RF-based cluster size tests are derived from a distribution approximation of cluster sizes based upon various parametric distributions. Like any other parametric tests, several assumptions are required, such as smooth images, a sufficiently high threshold  $u_c$ , and the uniform smoothness of images (Worsley *et al.*, 1992; Worsley *et al.*, 1996; Petersson *et al.*, 1999). Despite such restrictions, there is only a vague guideline as to how smooth images should be (Petersson *et al.*, 1999). Furthermore, though the choice of threshold is made by investigators according to signals of interest (Poline *et al.*, 1997), there is virtually no consensus on how high the threshold  $u_c$  should be for

the RF theory to work.

There have been some simulation based validations on Gaussian RF results under reasonable smoothness and threshold. Friston *et al.* (1994) validated their RF test and found that, for sufficient smoothness and a high threshold, the test performs well. Holmes (1994) carried out simulations with different thresholds and found the RF test to be conservative for low thresholds. However this conservativeness was not observed in simulations by Poline *et al.* (1997). Rather, they found that the RF test is anti-conservative for low thresholds and becomes conservative for high thresholds. In the same simulations, they also found that the RF test becomes less conservative if images are smoother. One common feature in these validations is that the RF test was validated under ideal conditions, where images were sufficiently smooth and thresholds were reasonably high. In real data analyses, however, investigators prefer to use as little smoothing as necessary to avoid focal signals being blurred and various thresholds  $u_c$  to identify signals of their interest. Also many users focus on low degrees of freedom  $t$  images. Under such conditions, the behavior of the test has not been well-characterized, especially for  $t$  images.

An alternative to the RF test is the permutation test (Holmes *et al.*, 1996; Nichols & Holmes, 2002). Unlike the RF test, it requires almost no assumptions. The sole assumption is null hypothesis exchangeability, that is, exchanging or permuting the group labeling does not alter the distribution of the test statistic. The test exploits this assumption by shuffling or permuting data labels randomly, and generates the null distribution of the test statistic of interest from the data itself. No knowledge of the underlying distribution of image voxels is required. The test is exact for the family-wise error (FWE) rate, which means that the probability of one or more type I errors is the same as the significance level of the test. However, because of a large number of calculations required, the permutation test is more computationally intensive than the RF test. Furthermore, while the test is straightforward for simple designs, multi-condition designs or correlated data complicate the test (Bullmore *et al.*, 1996).

In this work, we compare these two approaches and determine which is to be preferred under various conditions. In particular, we simulated Gaussian random fields and  $t$  random fields with

different degrees of freedom, applied the aforementioned tests, and compared the performance of an RF test relative to the permutation test. We did not use a real data set for validation since the uniform smoothness assumption cannot be verified and is often questionable (Hayasaka & Nichols, 2002). Under non-uniform smoothness, or non-stationarity, there are relatively smooth and rough regions within the image, and that will alter the distribution of cluster sizes locally, resulting in biased inference (Worsley *et al.*, 1999). In our subsequent work, we will consider cluster size inference on non-stationary images (Hayasaka *et al.*, 2003).

One of the novelties in this study is the validation of these tests on  $t$  images, which is done with laborious  $t$  image simulations where a number of independent Gaussian images are simulated to form a  $t$  statistic image (there is no algorithm to directly generate smooth  $t$  random fields). While some authors (Poline *et al.*, 1997) use Fourier domain simulation to simulate periodic images (where the left edge is continuous with the right edge), we simulated images in the spatial domain to obtain the most realistic results. In addition, we estimated smoothness from the simulated data as done in real data analyses. This estimation process introduces an additional source of variation into the inference. Another notable aspect is that a permutation test was carried out for each realization and its performance was assessed as well.

This paper is structured as follows: Details regarding the tests, as well as simulations are explained in the Method section. Results from the simulations are presented in the Results section. Finally close examinations of findings from the simulations and conclusions are presented in the Discussion section. Appendices are included which summarize the RF theory in a consistent notation and address important details of the SPM and `fmrstat` implementations as well as smoothness estimation.

## 2 Methods

### 2.1 Model

In a brain image analysis a linear model can be written as

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

where  $v = (x, y, z) \in \mathbb{R}^3$  is an index for voxels,  $Y(v) = \{Y_1(v), Y_2(v), \dots, Y_n(v)\}'$  is a vector of observed image intensities at voxel  $v$  from  $n$  scans,  $X$  is a known  $n \times p$  design matrix,  $\beta(v)$  is a  $p$ -dimensional vector of unknown parameters,  $\sigma(v)$  is an unknown standard deviation at voxel  $v$ , and  $\varepsilon(v) = \{\varepsilon_1(v), \varepsilon_2(v), \dots, \varepsilon_n(v)\}'$  is a vector of unknown random errors. We denote images by omitting the voxel index  $v$  (e.g.,  $\varepsilon_i$  denotes the error image from the  $i$ th scan).

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

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

and an estimate of the residual variance is

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

where  $\nu$  is the error degrees of freedom. If  $\varepsilon_i(v)$ 's are independent among scans and identically normally distributed, then the statistic image  $T$  is defined as

$$T(v) = \frac{\mathbf{c}\hat{\beta}(v)}{\sqrt{\mathbf{c}(X'X)^{-1}\mathbf{c}'\hat{\sigma}(v)}} \quad (2)$$

where  $\mathbf{c}$  is a row vector contrast of interest.  $T$  is then used to define clusters. Each cluster is formed as a set of contiguous voxels with their  $T$  exceeding a fixed cluster defining threshold  $u_c$  and sharing at least one common edge (18 connectivity scheme).

### 2.2 Cluster Size Inference

Let the size of a cluster be  $S$ . The true distribution of  $S$  is unknown, thus approximated by various methods such as the RF theory and permutations. The uncorrected p-value, or the p-value of

a single cluster size, is defined as the probability of observing a certain cluster size or larger, and can be calculated from the approximated distribution of  $S$ . In a real data analysis, however, multiple clusters could occur at a given threshold, creating a multiple comparisons problem among cluster sizes. To correct for this problem, family-wise error (FWE) correction is widely used, which controls the rate of type I errors for all the clusters collectively. The FWE correction is implemented by calculating p-values based on the distribution of the largest cluster size  $S_{max}$ . The rationale behind using the distribution of  $S_{max}$  is that the probability of observing  $S_{max}$  larger than  $s$  is the same as the probability of at least one or more clusters are greater than  $s$ , thus correcting for multiple clusters. Detailed explanation of the FWE correction is found in Appendix A.

### 2.2.1 RF Test

The assumptions of the RF test include: That the images are lattice approximations of a smooth random field (lattice approximation), whose smoothness is uniform within the image (stationarity) and relatively large compared to the voxel size (smooth image), and the cluster defining threshold  $u_c$  is sufficiently high (high threshold).

Two versions of the RF method are considered in this study. The one based on an assumption that  $S$  raised to a power is exponentially distributed (Friston *et al.*, 1994), as implemented in the SPM package<sup>1</sup>, and the other based on an assumption that the distribution of  $S$  is approximated by the product of a beta and  $\chi^2$  random variables (Cao & Worsley, 2001) as implemented in the `fmristat` package<sup>2</sup>. To correct for the FWE rate, the distribution of  $S_{max}$  was used to obtain critical cluster sizes. Details on these methods have been reported in a number of publications. We collect them all in a consistent notation and present in Appendix A. The distribution approximation of the SPM RF test is for Gaussian images, whereas that of the `fmristat` RF test is a more refined for  $t$  images. The `fmristat` package uses the same cluster size distribution approximation as the SPM package for Gaussian images, but some calculations are done differently (see Appendix B).

---

<sup>1</sup>Wellcome Department of Cognitive Neurology, University College London. <http://www.fil.ion.ucl.ac.uk/spm>

<sup>2</sup>Keith J Worsley. <http://www.math.mcgill.ca/keith/fmristat>

### 2.2.2 Permutation Test

Since proposed by Holmes *et al.* (1996), the permutation test for brain image analyses has been further studied (Bullmore *et al.*, 1999; Nichols & Holmes, 2002) and implemented in the SPM package as the SnPM toolbox<sup>3</sup>. Unlike the RF test mentioned before, this test does not require any distributional assumption, and produces valid p-values even when the distribution of the image voxel is unknown. One of the few assumptions and the rationale for this test is the exchangeability assumption; that is, under the null hypothesis, scan labels can be permuted without altering the joint distribution of cluster sizes. In our test, we focus on the distribution of  $S'_{max}$  in order to correct the FWE rate. Stationarity is not assumed in the permutation test, though variable smoothness will result in non-uniform sensitivity.

The implementation of this test is explained in detail in Nichols & Holmes (2002). In this study, the permutation test is used as implemented in the SnPM toolbox.

### 2.3 Simulation

We carried out Gaussian and  $t$  image simulations to validate two RF tests, as implemented in the SPM package and in the `fmrstat` package, and the permutation test as implemented in the SnPM toolbox. For each simulation, rejection rates were recorded and their 95% confidence intervals (CIs) were calculated by normal approximation of a binomial proportion  $\hat{p} \pm 1.96 \sqrt{\frac{\hat{p}(1-\hat{p})}{3000}}$ , where  $\hat{p}$  is the observed rejection rate. The significance level of all the tests were set to 0.05, thus the CIs should cover 0.05. We essentially examine many CIs, so some, by chance alone, may not capture 0.05. However, it is impossible to compute the expected number of CIs not covering 0.05 by chance alone, since the results are correlated due to the same white noise image used in each realization of the simulations.

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 anti-conservative, or liberal, and no longer considered as valid.

---

<sup>3</sup>Andrew Holmes and Tom Nichols. <http://www.fil.ion.ucl.ac.uk/spm/snpm>

### 2.3.1 Gaussian Image Simulation

A smooth Gaussian image can be generated in three steps. First, a single  $104 \times 104 \times 104$  white noise image was generated for each realization, which was then smoothed with a 3D Gaussian kernel with different full-width at half-maximum (FWHM) (1.5, 3, 6, and 12 voxels). Finally the outer 36 voxels from the smoothed images were truncated in order to avoid non-uniform smoothness at the edge, yielding a  $32 \times 32 \times 32$  image. The resulting image was then thresholded with thresholds  $u_c$ 's with upper tail Gaussian probabilities of 0.01, 0.001 and 0.0001.

3,000 realizations of Gaussian images were generated, which yield the Monte-Carlo standard error of 0.4% at 0.05 rejection rate. Only the two RF tests were applied to the simulated data at 0.05 significance level, since there was only one Gaussian image generated in each realization, which cannot be permuted in the permutation test. In the RF test, the known smoothing kernel width is used instead of estimating smoothness from a single image in each realization, since there were no residuals from which to estimate smoothness.

### 2.3.2 $t$ Image Simulation

A  $t$  image can be simulated by calculating a  $t$ -statistic image (2) from a set of Gaussian images. In our simulation, for each realization, a set of 10, 20, or 30  $32 \times 32 \times 32$  Gaussian images were generated by the method described above, with smoothing kernel FWHM 0 (no smoothing), 1.5, 3, 6, and 12 voxels. Then a  $t$  image is calculated based on a model, either a one-sample  $t$ -test or a two-sample  $t$ -test with equal sample sizes. The degrees of freedom for the  $t$  image is 9, 19 or 29 for the one-sample test, or 8, 18, or 28 for the two-sample test corresponding to group sizes of 5 & 5, 10 & 10, and 15 & 15. Our use of a two-sample  $t$ -statistic image was motivated by our collaborators' data of comparing controls and schizophrenics, and the results should be similar to that of a one-sample test with the same degrees of freedom. The generated  $t$  image was thresholded at the quantiles of a  $t$ -random variable with appropriate degrees of freedom with the upper tail probabilities of 0.01, 0.001, and 0.0001, and clusters were defined.

The image smoothness was estimated from the data in this simulation. Details regarding

smoothness estimation is found in Appendix C.

For each sample size, 3,000 sets of Gaussian images were simulated to generate  $t$ -images, and both `SPM` and `fmrifat` RF tests and the permutation test with 100 permutations were applied at 0.05 significance level.

### 2.3.3 Quality of Gaussian Images Simulated

Gaussian images, both for the Gaussian simulation and the  $t$  simulation in this study, were generated by convolving a white noise image with a Gaussian smoothing kernel (Worsley *et al.*, 1992; Worsley, 1996). However, with decreasing smoothness, the Gaussian kernel is more coarsely sampled and it is unclear whether the nature of the dependence is affected by this. To investigate the quality of Gaussian images simulated, we carried out two additional simulations. In the first simulation, images of size  $96 \times 96 \times 96$  voxels, smoothed with a kernel of FWHM 9 voxels, were generated. Then they were down-sampled at every 3rd voxel so that the resulting image should be  $32 \times 32 \times 32$  with FWHM 3 voxels (down-sized simulation). The other simulation was done in the same manner as the down-sized simulation, except they were not down-sampled. Thus the simulated image size and its smoothness are three times that of the first simulation (over-sized simulation). For each simulation, 3,000 realizations of two-sample (5 & 5)  $t$  images were generated, and a comparison was made on the 95th percentiles of the peak intensity and the largest cluster size at 0.001 threshold, among the down-sized and over-sized simulations, as well as the conventional method.

Note that in this comparison, cluster sizes are measured in terms of RESELS (RESolution ELEMENTs), volume measured in units of smoothness:

$$\text{RESELS} = \frac{\# \text{ Voxels}}{\text{FWHM}^3}$$

We use RESELS, instead of voxels, because the search volumes in the three simulations in terms of RESELS are the same even though the search volumes in terms of voxels are different.

### 2.3.4 Robustness to Smoothness Outliers

The robustness of the permutation test against a violation in the exchangeability assumption was examined. In particular, in a two-sample  $t$  test setting, we investigated by simulations the effect of a single image with different smoothness (smoothness outlier), and also the effect of a systematic smoothness difference between two groups (smoothness difference).

For the smoothness outlier simulation, 19 images with the same smoothness (FWHM 0, 1.5, 6, or 12 voxels) and one image with a different smoothness (FWHM 12, 6, 1.5, or 0 voxels, respectively) were generated for each realization and a  $t$  image for a two-sample test was calculated. For the smoothness difference simulation, two groups of 10 images having different smoothness, FWHM 6 or 12 voxels for one group and FWHM 1.5 or 0 voxels for the other group, respectively, were generated for each realization and a  $t$  image for a two-sample test was calculated.

For both simulations, 3,000 realizations were generated and the SPM and `fmrstat` RF tests and the permutation test with 100 permutations were applied.

## 2.4 Computing Environment

Each simulation was divided into segments of 200 to 1,000 realizations to be run on several different computers separately, and the results were merged once all the segments were done. The fastest computer used in this study was a Dell PC with dual 2.4GHz Xeon processors and 2GB of RAM, on a Linux platform, with MATLAB version 6.5 (MathWorks Inc., Natick, MA). It took this computer 4 days to run a 1,000 realization segment of the  $t$  image simulation with  $df=28$ .

## 3 Results

### 3.1 Gaussian Image Simulation

Results from the Gaussian simulation is shown in Figure 1. The plots show that the RF tests are conservative in most settings. The tests are especially conservative when the threshold is low,  $u_c$  corresponding to  $\alpha = 0.01$  or 0.001. It was also found that the tests are conservative for low

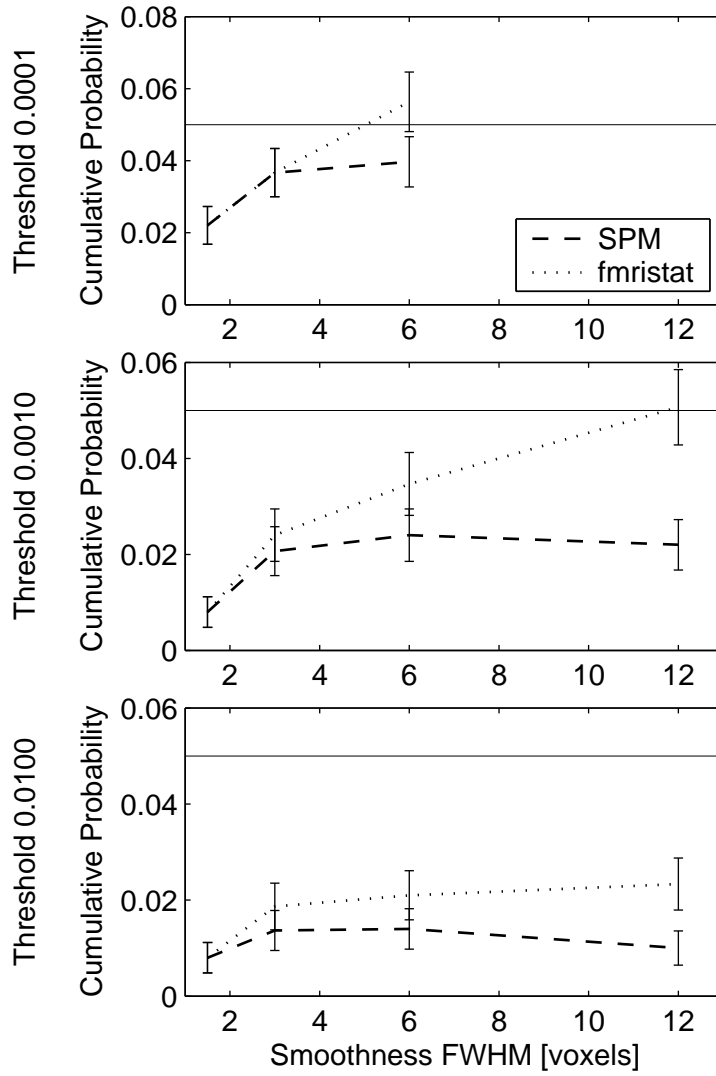


Figure 1: Results from the Gaussian simulation. Rejection rates of the RF tests when thresholded at upper-tail probabilities 0.01, 0.001, and 0.0001 (from bottom to top), along with their 95% confidence intervals. Fine solid lines indicate the desired type I error rate (0.05) of the test.

smoothness, and for high smoothness, the `fmristat` RF test becomes less conservative, while the significance level does not change dramatically for the SPM RF test. Both tests are unable to calculate a critical cluster size as a real number at 0.0001 threshold with smoothness 12 voxels FWHM. As explained in (10) in Appendix A, the critical cluster size cannot be calculated for a certain combination of  $u_c$  and smoothness such as this one.

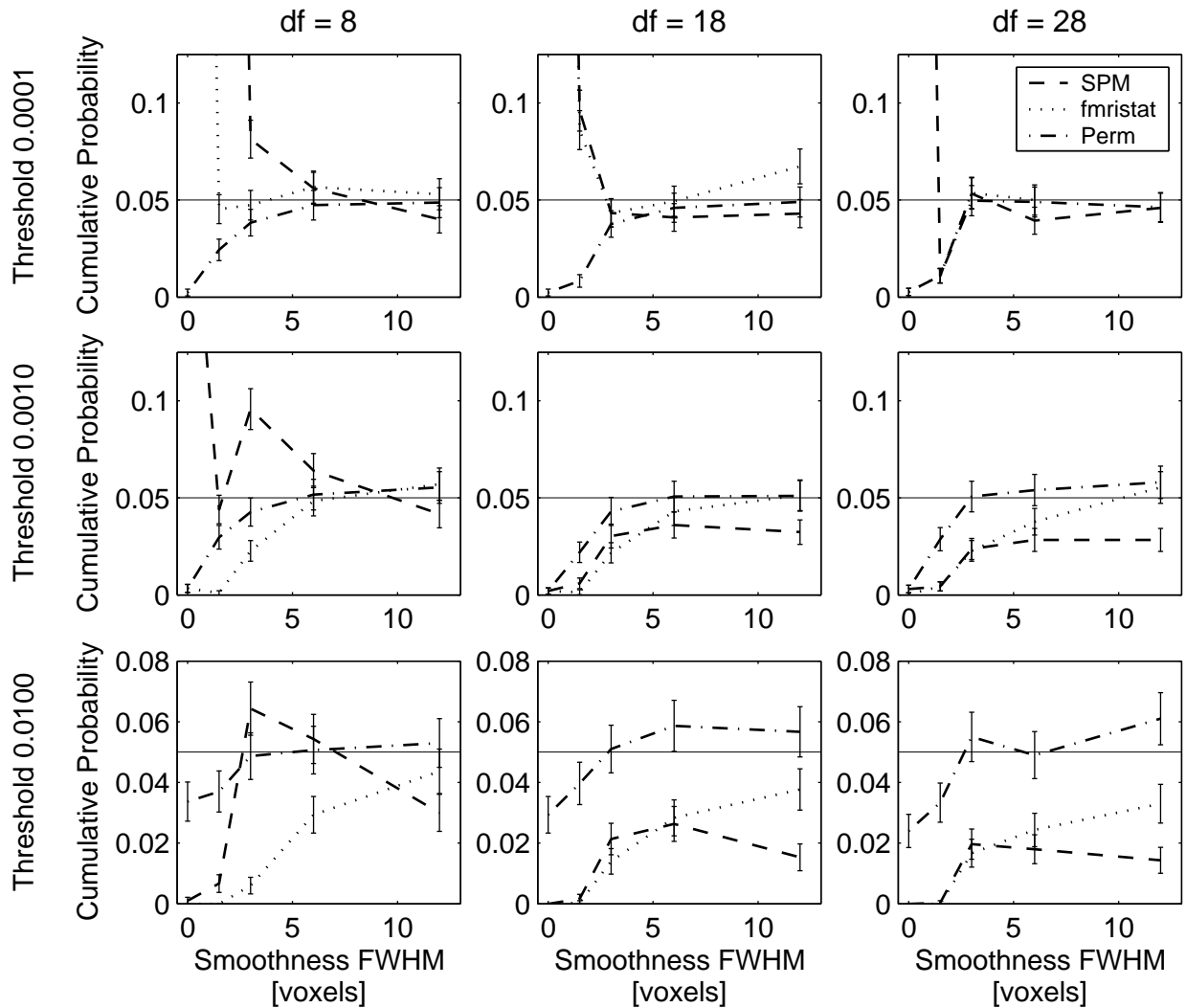


Figure 2: Results from the  $t$  image simulation. Rejection rates of the RF tests and the permutation test for different sample sizes (5 & 5, 10 & 10, and 15 & 15, from left to right) when thresholded at upper-tail probabilities 0.01, 0.001, and 0.0001 (from bottom to top), along with their 95% confidence intervals. Fine solid lines indicate the desired type I error rate (0.05) of the test.

## 3.2 $t$ Image Simulation

Since our one-sample and two-sample simulations produced similar results, we only present the results from the two-sample simulation.

### 3.2.1 RF Test

With a widely used threshold of 0.01, the RF tests seem generally conservative, especially for low smoothness. Figure 2 shows the rejection rates of the RF tests from  $t$  image simulation. The rejection rates do not approach to 0.05 unless the threshold is extremely high (0.0001) and images

Image Size	95th percentile peak intensity	95th percentile cluster size [RESELS]
$32 \times 32 \times 32$ (conventional method)	11.4727	0.6138
$32 \times 32 \times 32$ (down-sized)	11.7513	0.6601
$96 \times 96 \times 96$ (over-sized)	16.5730	0.7425

Table 1: Comparison of the 95th percentiles of the peak intensity and the largest cluster at 0.001 threshold from the conventional method, the down-sized simulation, and the over-sized simulation for a two-sample  $t_8$  image. All have the same RESEL volume but discrepancies between  $32^3$  and  $96^3$  volumes suggest that the lattice approximation does not hold for 3 voxel FWHM.

are smooth. In some cases, at a high threshold and low smoothness, the RF tests are extremely anti-conservative. For low thresholds, rejection rates decrease with increasing df. While it is unusual for performance to worsen with increasing df, the high df results do appear to converge to the Gaussian results (see Figure 1 and Figure 2 right column).

It was found that the `fmrstat` RF test is more conservative in low smoothness and less conservative in high smoothness, compared to the `SPM` approach.

### 3.2.2 Permutation Test

The permutation test in general works well for sufficiently smooth images at any threshold or any df (see Figure 2). However, for low smoothness ( $\text{FWHM} < 3$  voxels), the test is generally conservative, and it worsens as images become less smooth. This conservativeness is related to the discreteness in the  $S_{max}$  distribution, which is explained in detail in the Discussion section below.

### 3.2.3 Quality of Gaussian Images Simulated

Table 1 displays the results (95th percentiles) from the down-sized and over-sized simulations, along with the results from the conventional simulation method with appropriate parameters. The percentiles are the FWE-controlling intensity and cluster size thresholds. It can be seen that the conventional simulation and the down-sized simulation produce very similar results, indicating that discretization of the Gaussian kernel had little impact, at least for 3 voxel FWHM smoothness. Thus we believe that our simulation of  $t$  images was appropriately done.

Smoothness	0	1.5	3	6	12
Outlier smoothness	12	6	3	1.5	0
Smoothness estimate	1.21	1.69	3.04	4.66	4.77
<b>Rejection Rates</b>					
SPM	0.000	0.005	0.025	0.120	0.330
fmr <sub>i</sub> stat	0.000	0.001	0.021	0.112	0.323
Permutation	0.030	0.039	0.051	0.052	0.047

Table 2: Rejection rates of the two RF tests (SPM and fmr<sub>i</sub>stat) and the permutation test when a smoothness outlier is present for a two-sample  $t_{18}$  simulation thresholded at 0.01. Smoothness estimates are also shown, which are highly underestimated for smooth images, which explain the anti-conservativeness in the RF tests.

<b>Smoothness FWHM</b>			
Group 1	3	6	12
Group 2	3	1.5	0
Smoothness estimate	3.04	2.24	1.65
<b>Rejection rates</b>			
SPM	0.025	0.571	0.612
fmr <sub>i</sub> stat	0.021	0.486	0.571
Permutation	0.051	0.077	0.066

Table 3: Rejection rates and smoothness estimates from the smoothness difference simulation. Two groups of 10 Gaussian images with different smoothness were used to generate a  $t_{18}$  image and cluster size tests were applied at the 0.01 threshold.

In contrast, the over-sized simulation with equivalent RESEL volume had appreciably larger intensity threshold and 10% larger cluster size threshold. Such discrepancies indicate that  $96 \times 96 \times 96$  images are a better approximation of a smooth random field, compared to  $32 \times 32 \times 32$  images, and suggest that the lattice approximation is poor even for 3 voxel FWHM smoothness (for  $df=8$ ).

### 3.2.4 Robustness to Smoothness Outliers

Table 2 shows the rejection rates of the two RF tests and the permutation test when a smoothness outlier is present for a two-sample (10 & 10)  $t$  simulation thresholded at 0.01. While the results from the permutation test is somewhat close to that of without smoothness outliers, the RF test results become highly anti-conservative, especially for high smoothness images with a rough outlier, possibly due to underestimation in smoothness.

Table 3 shows the rejection rates when there is a systematic smoothness difference between two groups of 10 images in a two-sample test setting, thresholded at 0.01. Such smoothness difference influences the permutation test to be slightly anti-conservative. However, compared to the RF tests which are highly anti-conservative, the permutation is more robust when its null hypothesis exchangeability assumption is violated.

## 4 Discussion

We have simulated Gaussian images and  $t$  images and have applied three different cluster size tests to the simulated images, two RF tests and the permutation test. Their performances at different thresholds, smoothness, and dfs are recorded, which enable us to assess the specificity and robustness of these tests.

### 4.1 Comparison with Other Gaussian Simulation Results

There have been some simulation-based validations of the RF test on Gaussian images, comparable to our Gaussian image simulation. Friston *et al.* (1994) validated the RF test on 10,000 simulated images with size  $32 \times 32 \times 64$  with FWHM 5.7 voxels thresholded at 2.8 (upper-tail probability 0.0026). In this simulation, the cluster size distribution from the RF test was very close to the simulated cluster size distribution. This result was different from ours, where the RF test was found to be conservative. Some possible explanations for these discrepancies include: their use of estimated FWHM, their larger search volume, and their possible non-uniform smoothness at edges of simulated images.

Holmes (1994) simulated 10,000 images of  $65 \times 87 \times 26$  masked in the shape of a brain (72,410 voxels), smoothed with a non-isotropic Gaussian filter of FWHM  $5 \times 5 \times 2.5$  and thresholded at upper-tail probabilities  $p = 0.01, 0.001, \text{ and } 0.0001$ . His filter was also non-stationary, in that he truncated and renormalized the kernel when it contacted the mask. The RF test was done based on both known kernel FWHM and estimated FWHM. To be consistent with our results, we focus on the one with known FWHM. The results from Holmes' simulation are somewhat consistent

with our results, except at threshold  $p = 0.01$  where the result was less conservative compared to ours. Some possible explanations for this discrepancy include a brain-shaped search volume which reduces the chance of clusters touching the boundary (as a brain being more spherical than a box) and being truncated, and a larger search volume.

Poline *et al.* (1997) simulated 3,000 Gaussian images of size  $64 \times 64 \times 32$  with smoothing kernels FWHM  $4.7 \times 4.7 \times 3.9$ ,  $7.05 \times 7.05 \times 5.9$ , and  $9.4 \times 9.4 \times 7.85$ , thresholded at 2.0, 2.5, 3.0, and 3.5 (upper-tail probabilities 0.023, 0.006, 0.0013, and 0.00023, respectively). Their results indicate that higher the smoothness, less conservative the test becomes, which is consistent with our results in Gaussian simulation. However, contrary to our simulation, for lower thresholds (2.0 and 2.5), they found that the test was actually anti-conservative, and as the threshold is raised to 3.0, the test became conservative, approaching to the true significance level at threshold 3.5. A possible explanation for this discrepancy is the fact that they simulated images in a periodic manner, so that clusters were not truncated by the edge.

In general other authors have found that the RF test performs better at high thresholds in smooth images, which is consistent with our results.

## 4.2 t Simulation Results

### 4.2.1 RF Theory

In an RF cluster size test, the expected value (or the mean) of  $S$  is obtained from the expected value of the supra-threshold volume  $N$  and the number of clusters  $L$ , based on the identity

$$\mathbf{E}[S] = \frac{\mathbf{E}[N]}{\mathbf{E}[L]}$$

Details on the derivation of the expected values above are presented in Appendix A. Since each voxel in a statistic image is a  $t$  or  $Z$  statistic, the distribution of  $N$  can be easily approximated. We examine the estimation of the other quantities, namely  $\mathbf{E}[L]$  and  $\mathbf{E}[S]$ , to better understand the shortcomings of the RF tests.

The distribution of  $L$  is approximated by a Poisson distribution in the RF theory. As seen in Figure 3, the observed distribution (solid lines) of  $L$  can be well-approximated by a Poisson distri-

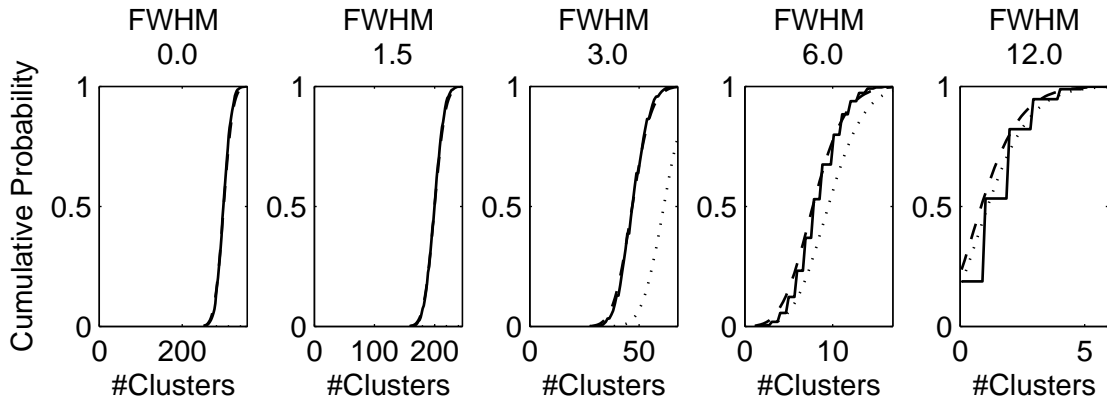


Figure 3: The distribution of the number of clusters at 0.01 threshold for two-sample  $t_{18}$  images (solid lines). The Poisson distribution having the same mean as that of the observed distribution (dashed lines) approximates the observed distribution quite well. However, the Poisson distribution with the mean based on the RF theory (dotted lines, off the plots for 0 & 1.5 FWHM) does not approach to the observed distribution unless images are smooth.

bution having the same mean (dashed lines). However, in practice, the mean of this distribution is unknown, thus estimated based on the RF theory using topological features of the supra-threshold volume (Worsley *et al.*, 1996). When the mean is estimated solely based on the RF theory, which is grossly overestimated, resulting approximated distribution (dotted lines) deviates from the observed. The left panel in Figure 4 shows the bias in the estimated  $\mathbf{E}[L]$ , which is substantial for low smoothness. A possible explanation for this overestimation is that the RF theory expects sub-voxel clusters (i.e., clusters whose volume is less than a voxel) to occur which cannot be observed in a real statistic image. Such sub-voxel clusters could occur more in low smoothness where lattice approximation is crude, resulting in a substantial overestimation seen in Figure 4.

The distribution of  $S$  in an RF test is approximated either by (7) for SPM or (8) for `fmrstat`. Figure 5 shows the observed cluster size distribution and approximated cluster size distributions used in SPM. Each plot shows the cumulative probability, which can be interpreted as a plot of percentiles. The point where the cumulative probability is 0.95 is the 95th percentile, or the uncorrected critical cluster size. The bottom row shows magnified cumulative probability plots around 0.95. Even when having the same mean as the observed distribution (solid lines), the theoretical distribution (dashed lines) does not approximate the observed distribution well unless images are

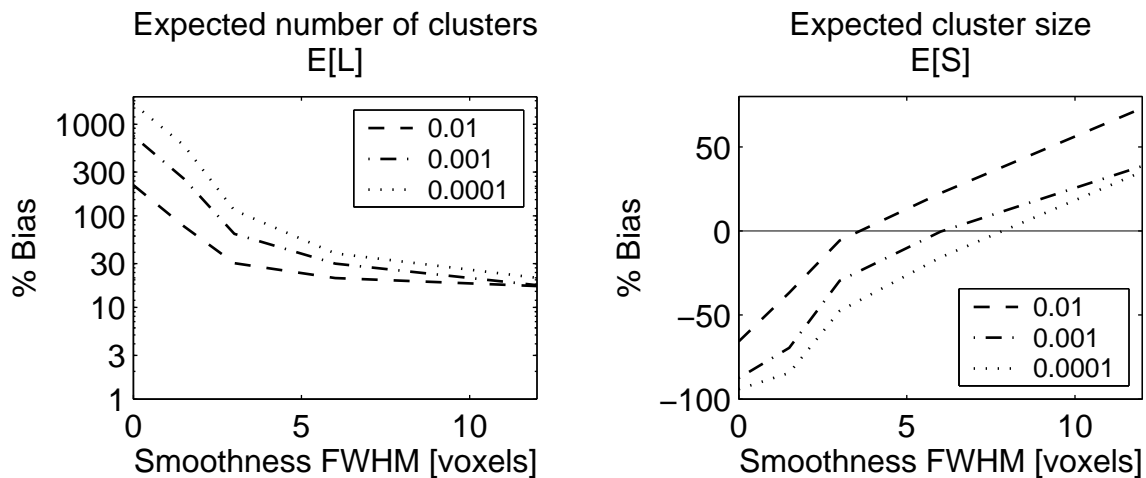


Figure 4: Bias in estimating the expected number of clusters (left) and the expected cluster size (right) by the RF theory compared to the observed values for two-sample  $t_{18}$  images thresholded at different thresholds.

smooth. When the mean of the theoretical distribution is derived solely using the RF theory (dotted lines), this deviation from the observed worsens. For low smoothness, the observed distribution is discrete, with the majority of cluster sizes being 1 or 2 voxels, whereas the theoretical distribution is continuous. Therefore the RF test can only be either extremely conservative or anti-conservative depending on where the majority of such small clusters lie relative to the theoretical critical cluster size.

Since  $\mathbf{E}[L]$  is overestimated, one might expect underestimation of  $\mathbf{E}[S]$ . However, such underestimation only occurs for low smoothness (see Figure 4 right panel). For high smoothness,  $\mathbf{E}[S]$  is actually overestimated possibly because the bias in  $\mathbf{E}[L]$  is small and at the same time some parts of clusters are truncated by the boundary of the search volume, yielding smaller clusters than expected by the RF theory. As it can be seen in the plots of cluster truncation rates in Figure 6, clusters are more likely to be truncated at low thresholds and in smooth images. Though the SPM method incorporates edge correction terms (Worsley *et al.*, 1996), the RF theory appears to not fully account for such cluster truncation, resulting in overestimation of  $S$  and ultimately conservativeness of RF tests.

In summary, the RF tests should not be used at low smoothness where the lattice approximation assumption fails and the cluster size distribution approximation is inaccurate. Even if images are

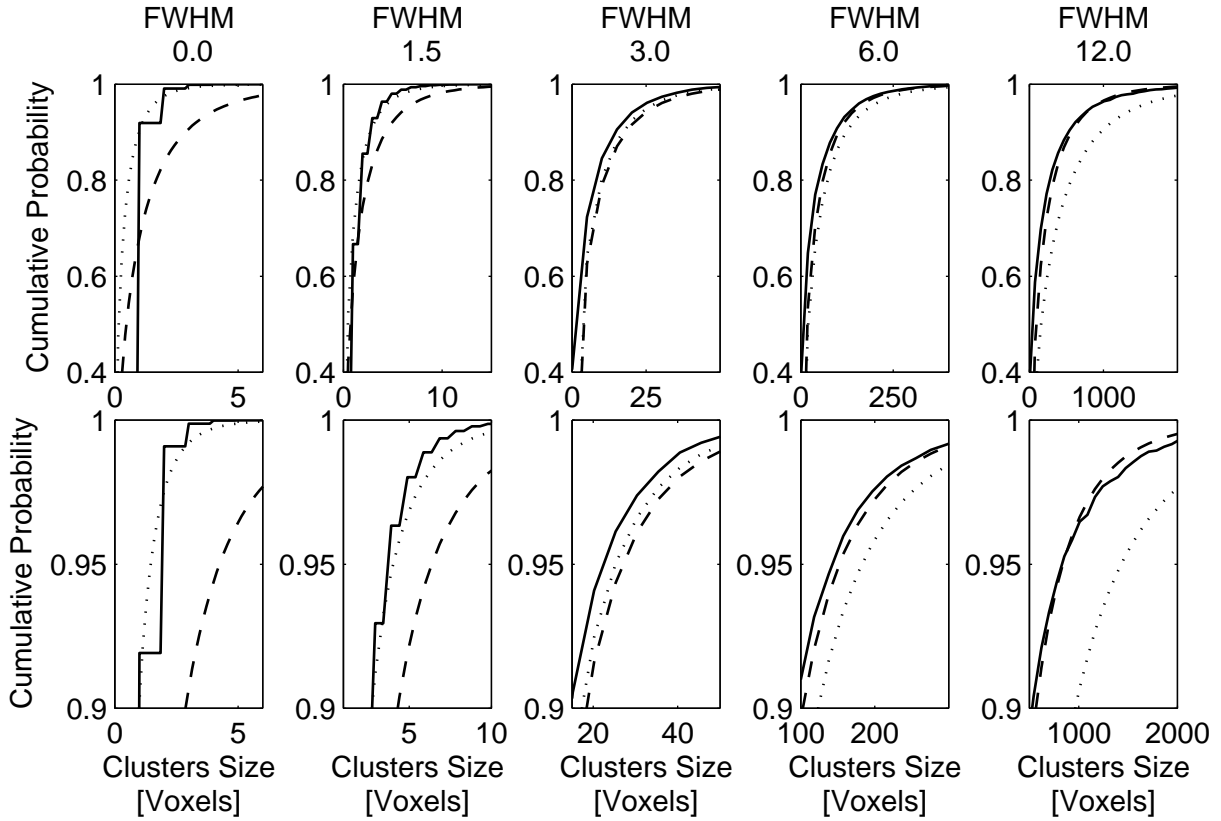


Figure 5: The distribution of cluster sizes at 0.01 threshold for two-sample  $t_{18}$  images (solid lines). The shape of the distribution based on the theory (SPM) does not approximate the observed distribution well unless images are smooth, even when the theoretical distribution is set to have the same mean as that of the observed distribution (dashes lines). The RF theory (dotted lines) is biased relative to the theoretical distribution with the observed mean, but happens to be close to the observed distribution for low smoothness. 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.

sufficiently smooth, say  $\text{FWHM} \geq 3$  voxels, then clusters being truncated at the edge could lead to conservativeness, which is particularly of concern at low thresholds.

#### 4.2.2 Permutation Test

For sufficient smoothness ( $\text{FWHM} \geq 3$  voxels), the permutation test seems to perform well for any thresholds and dfs. Figure 7 shows the observed  $S_{max}$  distribution (solid lines), along with approximated  $S_{max}$  distributions from the three tests examined, for two-sample (10 & 10)  $t$  images at 0.01 threshold. From the figure, it can be seen that the distribution from a single permutation test (dash-dot lines) is close to the observed distribution for any smoothness despite a small number

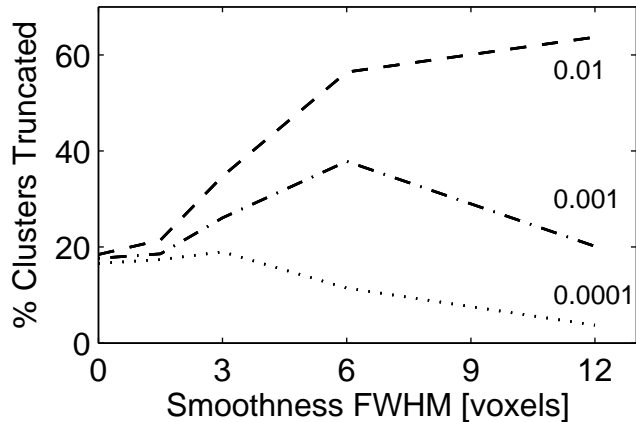


Figure 6: The proportion the clusters touching the boundary and being truncated by the boundary for different thresholds on two-sample  $t_{18}$  images. Cluster truncation is most frequent for low thresholds and high smoothness.

(100) of permutations, while the SPM RF test (dashed lines) and the `fmrstat` RF test (dotted lines) are conservative for some smoothness.

The conservativeness of the permutation test under low smoothness is due to the discreteness in the  $S_{max}$  distribution. Because of this discreteness, the 95th percentile in the observed distribution cannot be uniquely defined. Thus it is impossible to attain the rejection rate of 0.05, even when the approximated distribution is close to the observed distribution. Nevertheless, the permutation test produces accurate p-values even under such circumstances; for example, with no smoothing, for a cluster of size 5 voxels in a  $t$  image with  $df=28$  at 0.01 threshold, p-values were 0.045 for the truth, where as the average p-values from 3,000 realizations were 0.048 for the permutation, and 0.756 for SPM.

In summary, the permutation test performs well in all settings considered. Even when discreteness of cluster size distribution is an issue for low smoothness, the test yields accurate p-values.

### 4.3 Conclusions

In this study we carried out simulations to validate two cluster size inference methods, the RF test and the permutation test, in Gaussian images and  $t$  images. It was found that the RF tests do not perform well in some settings when theoretical approximations are not accurate. On the other hand, the permutation test works well for any threshold smoothness, and  $df$ , and showed great

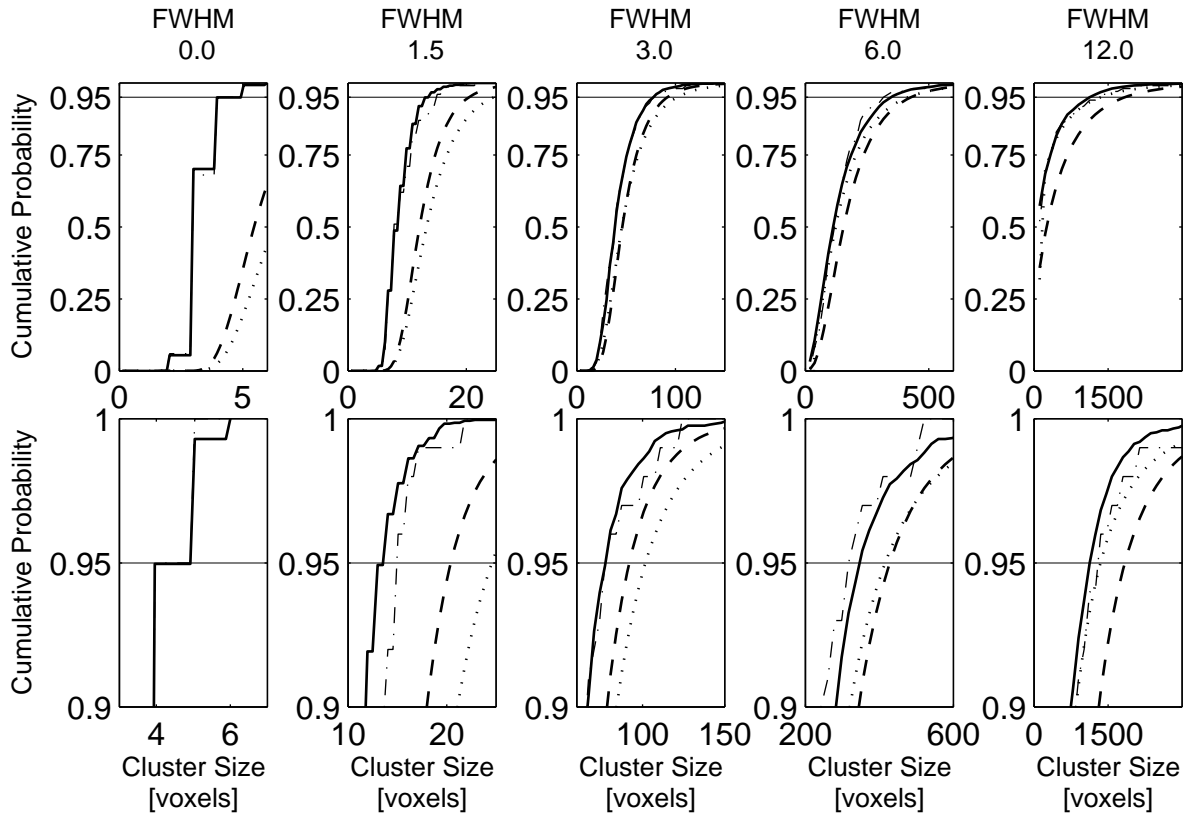


Figure 7: The observed distribution of the largest cluster sizes for the two-sample  $t_{18}$  images thresholded at 0.01 (solid lines), along with its approximations based on the SPM RF test (dashed lines), the *fmristat* RF test (dotted lines), and a single permutation test with 100 permutations (dash-dot lines). The top row shows the overall shape of the distributions from the 0th to 100th percentiles, while the bottom row shows the shape of the distributions around the 95th percentiles, the FWE corrected critical cluster sizes.

robustness when assumptions were violated. Thus, when possible, the permutation test should be used. If the permutation test cannot be used or the RF test is chosen, then the smoothness and the threshold should be chosen wisely. We suggest smoothing images so that the images have at least 3 voxels FWHM. As for the threshold, we caution that lower thresholds leads to conservativeness in highly smoothed images and anti-conservativeness in low-smoothness images.

In this study we did not simulate signals, so we are unable to make detailed comments on the power of these tests, though a conservative test will generally be less sensitive than an exact test.

## **5 Acknowledgments**

We would like to thank Dr. Keith Worsley for useful comments and advice. Also we would like to thank the fMRI Laboratory at the University of Michigan for use of their computing resources.

## References

- Adler, R J. 1980. *The Geometry of Random Fields*. New York: Wiley.
- 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. *Magnetic Resonance in Medicine*, **35**, 261–277.
- Bullmore, E T, Suckling, J, Overmeyer, S, Rabe-Hesketh, S, Taylor, E, & Brammer, M J. 1999. Global, Voxel, and Cluster Tests, by Theory and Permutation, for a Difference between Two Groups of Structural MR Images of the Brain. *IEEE Transactions on Medical Imaging*, **18**, 32–42.
- Cao, J. 1999. The Size of the Connected Components of Excursion Sets of  $\chi^2$ ,  $t$ , and  $F$  Fields. *Advances in Applied Probability*, **31**, 579–595.
- Cao, J, & Worsley, K J. 2001. Applications of Random Fields in Human Brain Mapping. *Pages 169–182 of: Moore, M (ed), Spatial Statistics: Methodological Aspects and Applications*. Springer Lecture Notes in Statistics, vol. 159. Springer.
- 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. *Magnetic Resonance in Medicine*, **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. *Human Brain Mapping*, **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. 2002. A Resel-Based Cluster Size Permutation Test for Non-Stationary Images. *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), 1062–1063.*
- Hayasaka, S, Nichols, T E, & Worsley, K J. 2003. Non-Stationary Cluster Size Inference with a Permutation Test. *in preparation.*
- Holmes, A P. 1994. *Statistical Issues in functional Brain Mapping.* Ph.D. thesis, University of Glasgow.
- Holmes, A P, Blair, R C, Watson, J D G, & Ford, I. 1996. Nonparametric Analysis of Statistic Images from Functional Mapping Experiments. *Journal of Cerebral Blood Flow and Metabolism, 16(1), 7–22.*
- Jenkinson, M. 2000. *Estimation of Smoothness from the Residual Field.* Tech. rept. Oxford Centre for Functional Magnetic Resonance Imaging of the Brain (FMRIB).
- 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. *Human Brain Mapping, 15, 1–25.*
- Nosko, V P. 1969. Local structure of Gaussian random fields in the vicinity of high level shines. *Soviet Mathematics: Doklady, 10, 1481–1484.*
- Petersson, K M, Nichols, T E, Poline, J-B, & Holmes, A P. 1999. Statistical Limitations in Functional Neuroimaging II. Signal Detection and Statistical Inference. *Philosophical Transaction of the Royal Society of London. Series B, 354, 1261–1281.*

- 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. *Journal of Cerebral Blood Flow and Metabolism*, **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. *Human Brain Mapping*, **1**, 3–19.
- Worsley, K J. 1996. The geometry of random images. *CHANCE*, **9**, 27–40.
- 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. Available on CD-ROM in NeuroImage*, **16**(2), 779–780.
- Worsley, K J, Evans, A C, Marrett, S, & Neelin, P. 1992. Three-Dimensional Statistical Analysis for CBF Activation Studies in Human Brain. *Journal of Cerebral Blood Flow and Metabolism*, **12**, 900–918.
- Worsley, K J, Marrett, S, Neeline, P, & Evans, A C. 1995. A Unified Statistical Approach for Determining Significant Signals in Location and Scale Space Images of Cerebral Activation. *Pages 327–333 of: Myers, R, Cunningham, V J, Bailey, D L, & T, Jones (eds), Quantification of Brain Function Using PET*. San Diego: Academic Press.
- 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. *Human Brain Mapping*, **4**, 58–73.
- Worsley, K J, Andermann, M, Koulis, T, MacDonald, D, & Evans, A C. 1999. Detecting Changes in Nonisotropic Images. *Human Brain Mapping*, **8**, 98–101.

# Appendix

## A RF Cluster Size Test

Under the RF theory, the distribution of the cluster size  $S$  is obtained based on the distribution of the supra-threshold volume  $N$ , the number of clusters in the search volume  $L$ , and the identity

$$\mathbf{E}[S] = \frac{\mathbf{E}[N]}{\mathbf{E}[L]} \quad (3)$$

The expected value of  $N$  is

$$\mathbf{E}[N] = V(1 - F(u_c)) \quad (4)$$

where  $V$  is the search volume and  $F(\cdot)$  is the cumulative distribution function (cdf) of an appropriate random variable. If the image is assumed to be a Gaussian random field, then  $F(\cdot)$  is the cdf of a Gaussian random variable, and if the image is considered as a  $t$ -random field with degrees of freedom  $\nu$ , then  $F(\cdot)$  is the cdf of a  $t$ -random variable with  $\nu$  degrees of freedom. If the image is  $D$ -dimensional (typically  $D = 3$ ), then the expected value of  $L$  is

$$\mathbf{E}[L] = \sum_{d=0}^D R_d \rho_d(u_c) \quad (5)$$

where  $R_d$ 's are  $d$ -dimensional RESEL counts and  $\rho_d$ 's are  $d$ -dimensional resel densities.  $R_d$ 's and  $\rho_d$  depends on the underlying random field (see Worsley *et al.* (1995) and Worsley *et al.* (1996)).

It is known that, for a Gaussian RF with a large  $u_c$ ,  $S^{2/D}$  is approximately distributed as an exponential random variable (Nosko, 1969; Friston *et al.*, 1994) with mean  $1/\psi$ , where

$$\psi = \left[ \frac{\Gamma(\frac{D}{2} + 1) \mathbf{E}[L]}{\mathbf{E}[N]} \right]^{2/D} \quad (6)$$

The distribution of  $S$  can then be approximated by

$$Pr(S > s) \simeq \exp(-\psi s^{2/D}) \quad (7)$$

Strictly, (7) is valid only for Gaussian images. A more refined approximation of the cluster size distribution in  $t$  RF was found by Cao (1999) and Cao & Worsley (2001); they found  $S$  was

distributed as

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

where  $B$  is a Beta random variable with parameters  $(1, (\nu - D)/2)$ ,  $U_0$  is a  $\chi^2$  random variable with degrees of freedom  $\nu + 1 - D$ , and  $U_b$ 's ( $b = 1, 2, \dots, D$ ) are independent  $\chi^2$  random variables with degrees of freedom  $\nu + 2 - b$ .  $c$  is a constant chosen so that (3) is satisfied.

Once the distribution of each cluster is found, either from (7) or (8), then the critical cluster size is to be found, adjusted for a desired family-wise error rate (FWER), or the probability of false rejections controlling for multiple comparisons among clusters. In this case, clusters are assumed to be independent (Adler, 1980; Friston *et al.*, 1994), and the number of clusters whose size exceeds  $s$ , say  $L_s$ , can be approximated by a Poisson distribution with the mean  $\mathbf{E}[L] \cdot Pr(S > s)$ . Using this result, the FWER can be found as the probability of at least one cluster exceeding  $s$ , or 1 minus the probability that no cluster exceeding  $s$ , which is

$$Pr(L_s \geq 1) \simeq 1 - \exp(-\mathbf{E}[L] \cdot Pr(S > s))$$

Note that the probability of at least one cluster exceeding  $s$  is equivalent to the probability that the largest cluster exceeding  $s$ . Thus, in the test, the largest cluster size  $S_{max}$  is used as the test statistic, instead of all the cluster sizes, and its distribution is expressed as

$$Pr(S_{max} > s) \simeq 1 - \exp(-\mathbf{E}[L] \cdot Pr(S > s)) \quad (9)$$

which is the FWE corrected p-value of a cluster of size  $s$ . The critical cluster size is obtained as the cluster size at which (9) yields the desired significance level.

If the distribution of  $S$  is assumed to be (7), then the FWER adjusted critical cluster size  $k_\alpha$  at a desired significance level  $\alpha$  can be easily calculated by

$$k_\alpha \simeq \left[ \frac{\ln\left(\frac{-\mathbf{E}[L]}{\ln(1-\alpha)}\right)}{\psi} \right]^{D/2} \quad (10)$$

Note that if the ratio  $\frac{-\mathbf{E}[L]}{\ln(1-\alpha)}$  is less than 1, then the natural log of the ratio becomes negative and  $k_\alpha$  cannot be calculated. Such instances could occur when the expected number of clusters  $\mathbf{E}[L]$

is too small due to an unrealistic combination of threshold  $u_c$  and the smoothness, or when  $\alpha$  is extremely large.

## B $t$ RF Cluster Size Test Implementations

In different software programs, theories described in Appendix A are implemented differently. In the SPM package, the cluster size distribution for  $t$  images is assumed to be in the form (7), whereas in the `fmrstat` package, the distribution is assumed to be (8). Another difference between these packages is in the calculation of  $\mathbf{E}[L]$ . Though  $\mathbf{E}[L]$  should be calculated with  $D$  different dimensional terms, in practice, if the search volume is large, the lower dimensional terms are often negligible. In the SPM package,  $\mathbf{E}[L]$  in (9) is calculated using all the dimensional terms, while the  $\mathbf{E}[L]$  in the  $\psi$  parameter in (6) is calculated only with the highest dimensional term  $R_d \rho_d(u_c)$ , thus the  $\psi$  parameter becomes

$$\psi = \left[ \frac{\Gamma(\frac{D}{2} + 1) Q_D \rho_D(u_c)}{\mathbf{E}[N]} \right]$$

In the `fmrstat` package,  $\mathbf{E}[L]$  is always calculated with the highest dimensional term  $R_d \rho_d(u_c)$  only.

For Gaussian images, both SPM and `fmrstat` approximate the cluster size distribution with (7), though as mentioned above, they calculate  $\mathbf{E}[L]$  differently.

## C Smoothness Estimation

In neuroimage analyses, there exist different ways to estimate image smoothness. Widely used methods are Kiebel *et al.* (1999) and Forman *et al.* (1995). Jenkinson (2000) explains both methods in detail. In our simulation, we used the approach by Kiebel *et al.*, the one used in the SPM package.

The smoothness of images are estimated in terms of FWHM from the variance-covariance matrix of partial derivatives of residual images. The variance-covariance matrix of spatial partial

derivatives of a random field  $G$  is defined as

$$\begin{aligned}
\Lambda &= \mathbf{Var} \left( \frac{\partial G}{\partial(x, y, z)} \right) \\
&= \begin{pmatrix} \mathbf{Var} \left( \frac{\partial G}{\partial x} \right) & \mathbf{Cov} \left( \frac{\partial G}{\partial x}, \frac{\partial G}{\partial y} \right) & \mathbf{Cov} \left( \frac{\partial G}{\partial x}, \frac{\partial G}{\partial z} \right) \\ \mathbf{Cov} \left( \frac{\partial G}{\partial y}, \frac{\partial G}{\partial x} \right) & \mathbf{Var} \left( \frac{\partial G}{\partial y} \right) & \mathbf{Cov} \left( \frac{\partial G}{\partial y}, \frac{\partial G}{\partial z} \right) \\ \mathbf{Cov} \left( \frac{\partial G}{\partial z}, \frac{\partial G}{\partial x} \right) & \mathbf{Cov} \left( \frac{\partial G}{\partial z}, \frac{\partial G}{\partial y} \right) & \mathbf{Var} \left( \frac{\partial G}{\partial z} \right) \end{pmatrix} \\
&= \begin{pmatrix} \lambda_{xx} & \lambda_{xy} & \lambda_{xz} \\ \lambda_{yx} & \lambda_{yy} & \lambda_{yz} \\ \lambda_{zx} & \lambda_{zy} & \lambda_{zz} \end{pmatrix}
\end{aligned} \tag{11}$$

In real data, the  $\Lambda$  matrix is estimated based on standardized residual images  $u$ , which is defined at each voxel  $v$  as

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

where  $\nu$  is the error degrees of freedom. Partial derivatives of  $u$  is calculated by taking the difference between  $u(v)$  and adjacent voxels in  $x$ ,  $y$ , and  $z$  directions and dividing it by the voxel dimension. Denote this by  $\Delta u(v) = \left( \frac{\Delta u(v)}{\Delta x}, \frac{\Delta u(v)}{\Delta y}, \frac{\Delta u(v)}{\Delta z} \right)$ . Then an estimate of  $|\Lambda|$ ,  $|\widehat{\Lambda}|$ , is given by

$$|\widehat{\Lambda}| = \frac{1}{V} \sum_v \left| \frac{1}{\nu} \Delta u(v)' \Delta u(v) \right| \tag{12}$$

where  $V$  is the number of voxels. This expression differs slightly from (Kiebel *et al.*, 1999) since we write it in terms of standardized residuals and not normalized residuals ( $u(v)/\sqrt{\nu}$ ).

FWHM is expressed in term of  $|\Lambda|$  by

$$\text{FWHM} = (4 \ln 2)^{1/2} |\Lambda|^{\frac{-1}{2D}} \tag{13}$$

(Worsley, 2002). Unfortunately the obvious estimate of FWHM, replacing  $|\Lambda|^{\frac{-1}{2D}}$  with  $|\widehat{\Lambda}|^{\frac{-1}{2D}}$  results in a biased estimator. Worsley (2002) showed that  $|\widehat{\Lambda}|^{\frac{-1}{2D}}$  needs to be divided by a bias correction which is a function of the degrees of freedom. In our case, the estimate of FWHM is to be used in the RF test in the form  $\frac{1}{\text{FWHM}^D}$ , instead of FWHM as it is (Worsley *et al.*, 1996). It turns out the correction factor for  $|\widehat{\Lambda}|$  in  $\frac{1}{\text{FWHM}^D}$  is 1 (Worsley, 2002), so  $\widehat{\text{FWHM}}$  can be obtained from (13) with  $\widehat{\Lambda}$  substituted for  $\Lambda$

$$\widehat{\text{FWHM}} = (4 \ln 2)^{1/2} |\widehat{\Lambda}|^{\frac{-1}{2D}}$$

For the calculation of (12), the *SPM* package assumes the off-diagonal elements of  $\widehat{\Lambda}$  to be zero and calculates  $\widehat{\text{FWHM}}$  accordingly. In this simulation, we calculated the  $\widehat{\text{FWHM}}$  in that manner. However, if the off-diagonals are assumed to be zero, then the bias correction factor is no longer 1. The `multistat.m` function in the `fmrstat` package calculates an appropriate bias correction factor according to the `df` and the dimensionality of the search space.