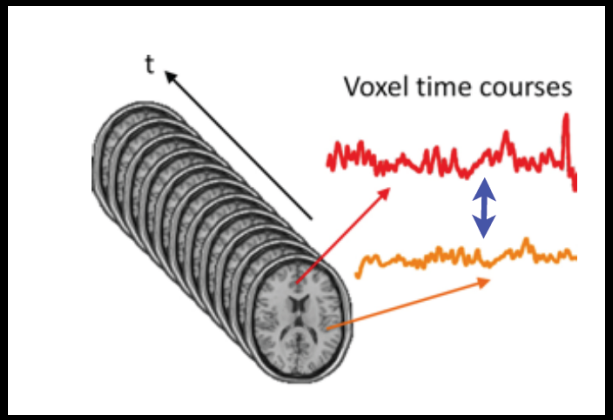
What a Cluster F… ailure!

Problems with cluster corrections for multiple comparisons in fMRI data

Computational Methods Lunch
Western University
Jody Culham
Nov. 7, 2016

Mega Multiple Comparisons Problem



Typical 3T Data Set

30 slices x 64 x 64
= 122,880 voxels of $(3\ mm)^3$

If we choose $p < 0.05$…

122,880 voxels x 0.05 = approx. 6144 voxels should be significant due to chance alone

We can reduce this number by only examining voxels inside the brain

~64,000 voxels (of $(3\ mm)^3$) x 0.05 = 3200 voxels significant by chance

Possible Solutions to Multiple Comparisons Problem

- Bonferroni Correction (Family-wise Error, FWE Correction)
  - small volume correction
- Cluster Correction
  - arbitrary threshold ($p<.001$) and cluster size (10×8 mm^3)
  - Monte Carlo simulations (e.g., AFNI)
  - Gaussian Random Field Theory (SPM)
- False Discovery Rate
- Test-Retest Reliability

Bonferroni (FWE) Correction

- divide desired p value by number of comparisons

  Example:

  desired p value: $p < .05$

  number of voxels in brain: 64,000

  required p value: $p < .05 / 64{,}000 \rightarrow p < .00000078$

- Variant: small-volume correction
  - only search within a limited space
    - brain
    - cortical surface
    - region of interest
  - reduces the number of voxels and thus the severity of Bonferroni

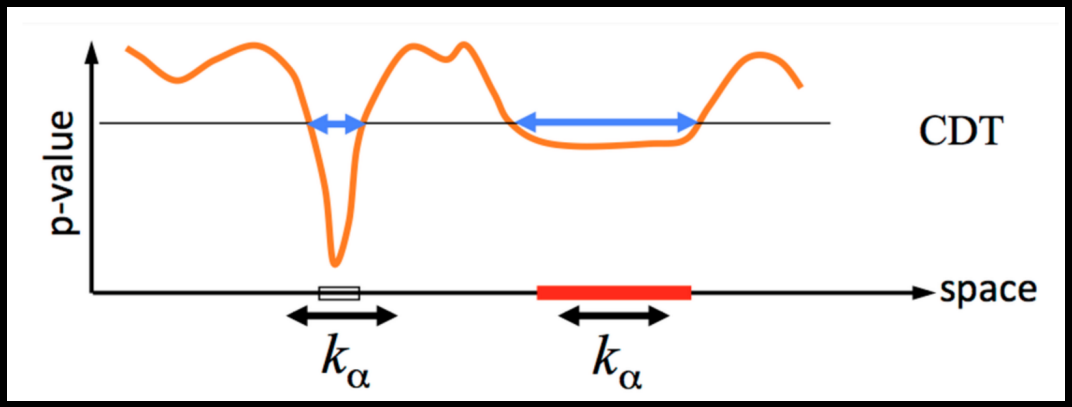
- Drawback: overly conservative
  - assumes that each voxel is independent of others
    - not true – adjacent voxels are more likely to be sig in fMRI data than non-adjacent voxels

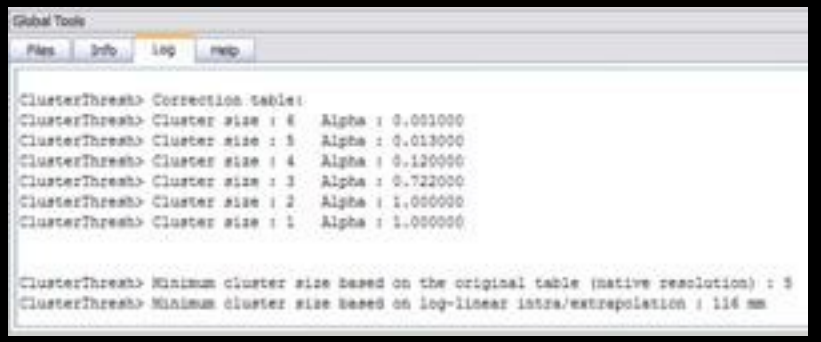
Cluster Correction

- falsely activated voxels should be randomly dispersed
- set minimum cluster size (k) to be large enough to make it unlikely that a cluster of that size would occur by chance
- some algorithms assume that data from adjacent voxels are uncorrelated (not true)
- some algorithms (e.g., Brain Voyager) estimate and factor in spatial smoothness of maps
  - cluster threshold may differ for different contrasts

- Drawbacks:
  - handicaps small regions (e.g., subcortical foci) more than large regions
  - researcher can test many combinations of p values and k values and publish the one that looks the best

How cluster correction works





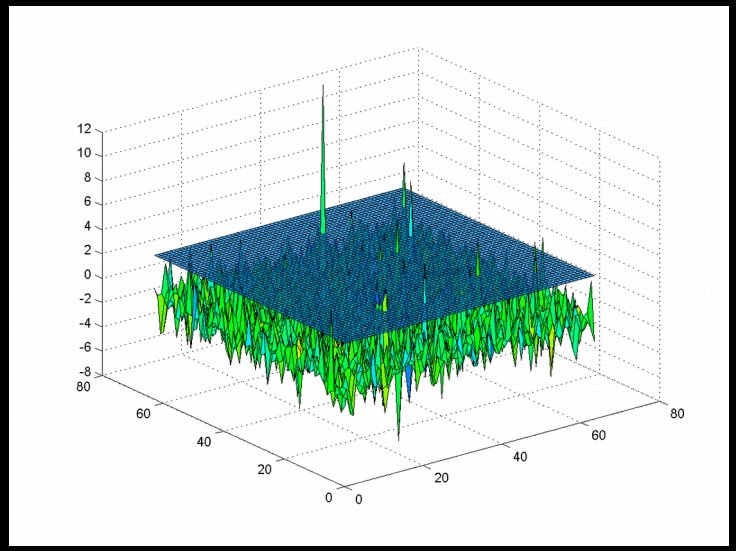
http://www.ohbmbrainmappingblog.com/blog/keep-calm-and-scan-on

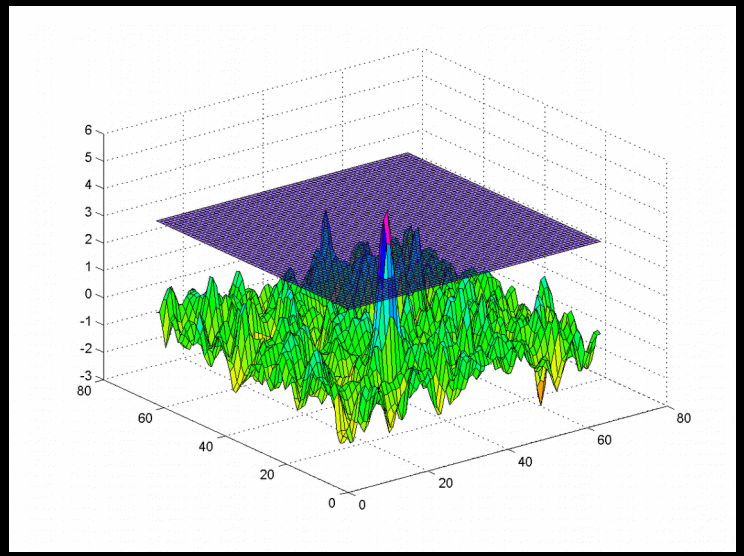
- Step 1: Choose a cluster-defining threshold (CDT)
- Step 2: Estimate smoothness of maps
- Step 3: Run Monte Carlo simulations on randomly generated maps with the smoothness determined in Step 2 to determine the likelihood of finding clusters of different sizes
- Step 4: Set a minimum cluster size (k) and exclude any clusters of voxels that are smaller

Gaussian Random Field Theory

- Fundamental to SPM
- If data are very smooth, then the chance of noise points passing threshold is reduced
- Can correct for the number of “resolvable elements” (“resels”) rather than number of voxels

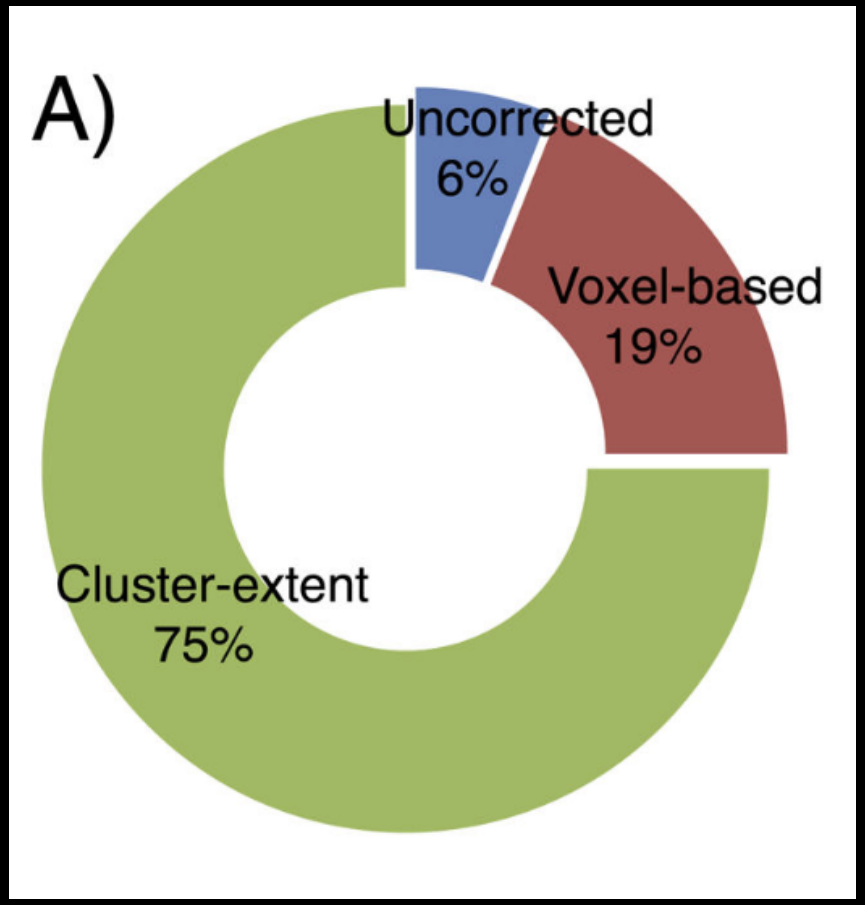
- Drawback: Requires smoothing



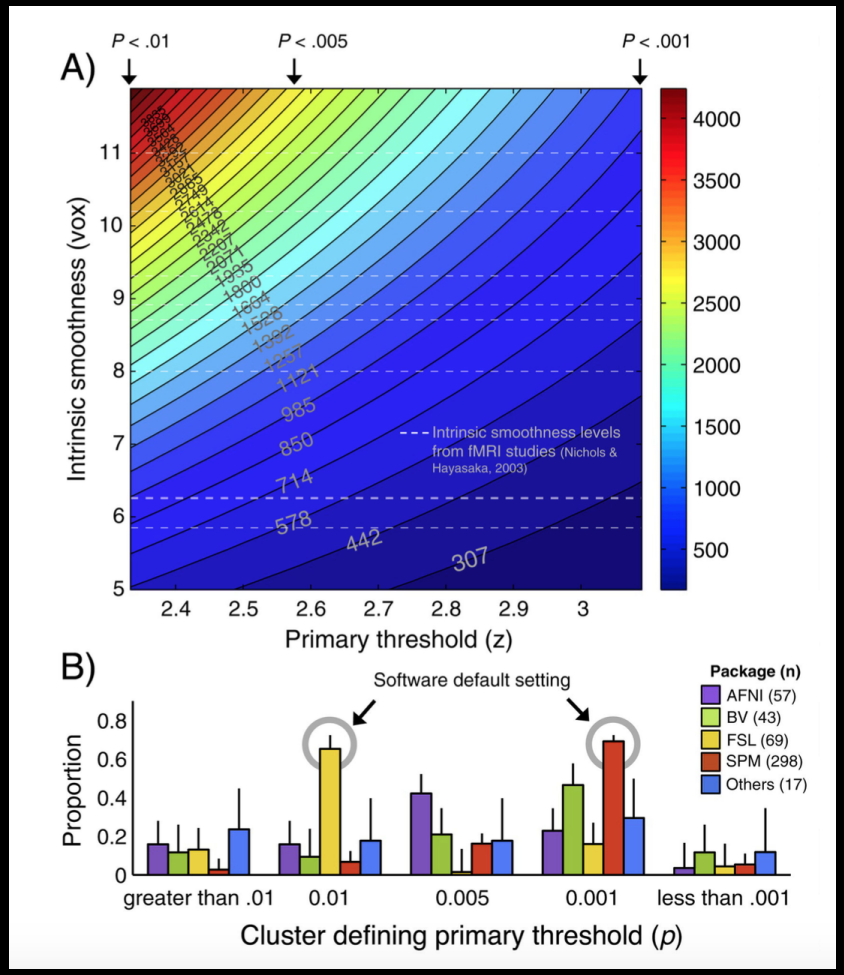


Slide modified from Duke course

Cluster Correction is Common



Sample of publications in Cerebral Cortex, Nature, Nature Neuroscience, NeuroImage, Neuron, PNAS, and Science (*N* = 814)



People use software defaults

Woo, Krishnan & Wager, 2014, NeuroImage

What a Clusterf… ailure!

PNAS

Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates

Anders Eklund[a,b,c,1], Thomas E. Nichols[d,e], and Hans Knutsson[a,c]

2016

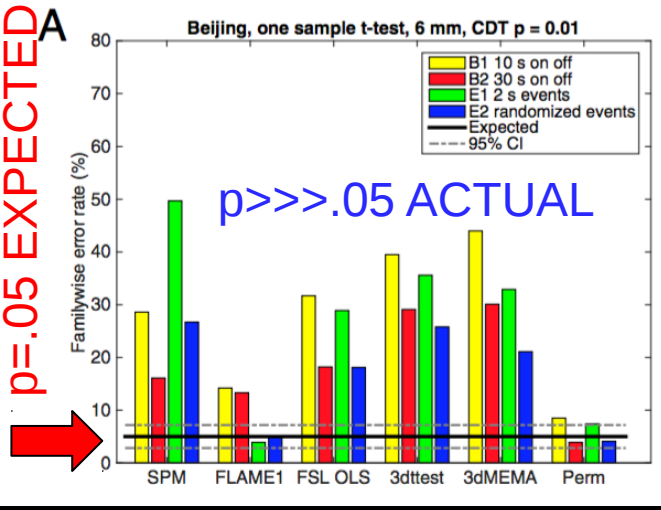
- resting-state data from Functional Connectomes Project
- applied stats to test for task-based “activation” using SPM, FSL and AFNI software packages
- since there was no real task-based activation, false positives for cluster correction should be 5%
- unlike previous tests of cluster correction algorithms, the resting-state data has real, not simulated, data properties

Table 1. Parameters tested for the different fMRI software packages, giving a total of 192 (3 × 2 × 2 × 4 × 2 × 2) parameter combinations and three thresholding approaches

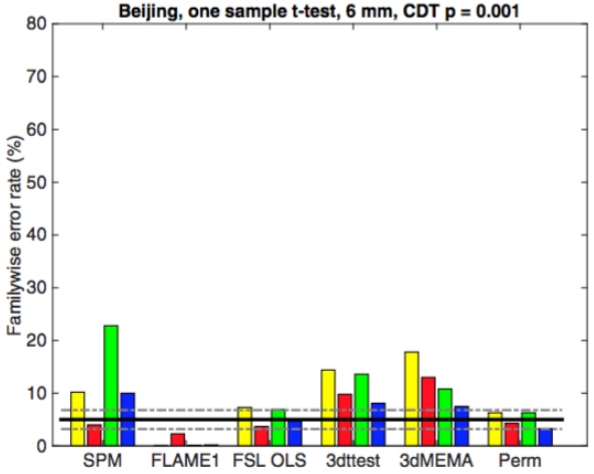
Parameter	Values used
fMRI data	Beijing (198 subjects), Cambridge (198 subjects), Oulu (103 subjects)
Block activity paradigms	B1 (10-s on off), B2 (30-s on off)
Event activity paradigms	E1 (2-s activation, 6-s rest), E2 (1- to 4-s activation, 3- to 6-s rest, randomized)
Smoothing	4-, 6-, 8-, 10-mm FWHM
Analysis type	One-sample *t* test (group activation), two-sample *t* test (group difference)
No. of subjects	20, 40
Inference level	Voxel, cluster
CDT	$P = 0.01$ ($z = 2.3$), $P = 0.001$ ($z = 3.1$)

One thousand group analyses were performed for each parameter combination.

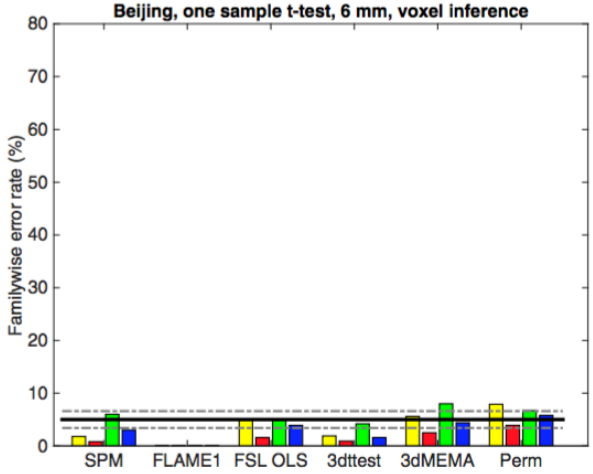
p = .50 not .05?!



n=20, FWHM = 6 mm
Cluster threshold p<.01



n=20 , FWHM = 6 mm
Cluster threshold p<.001



n=20 , FWHM = 6 mm
Voxel

Beware scientific click-bait

Significance

Functional MRI (fMRI) is 25 years old, yet surprisingly its most common statistical methods have not been validated using real data. Here, we used resting-state fMRI data from 499 healthy controls to conduct 3 million task group analyses. Using this null data with different experimental designs, we estimate the incidence of significant results. In theory, we should find 5% false positives (for a significance threshold of 5%), but instead we found that the most common software packages for fMRI analysis (SPM, FSL, AFNI) can result in false-positive rates of up to 70%. These results question the validity of some 40,000 fMRI studies and may have a large impact on the interpretation of neuroimaging results.

The Sky is Falling!

A bug in fMRI software could invalidate 15 years of brain research

This is huge.

BEC CREW 6 JUL 2016

Science alert

Cluster Failure: Biggest 'I Told You So' Yet. fMRI Stinks

JULY 6, 2016 / BRIGGS / 29 COMMENTS

Bug in fMRI software calls 15 years of research into question (Wired)

A bug in fMRI software could invalidate 15 years of brain research (Science alert)

Tens of thousands of FMRI brain studies may be flawed (Forbes)

Software faults raise questions about the validity of brain studies (Ars Technica)

15 years of brain research has been invalidated by a software bug, says Swedish scientists (International Business Times)

When big data is bad data (ZDNet)

Thousands of fMRI brain studies in doubt due to software flaws (New Scientist)

NEUROSCIENCE, STATISTICS

Correction for "Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates," by Anders Eklund, Thomas E. Nichols, and Hans Knutsson, which appeared in issue 28, July 12, 2016, of *Proc Natl Acad Sci USA* (113:7900–7905; first published June 28, 2016; 10.1073/pnas.1602413113).

The authors note that on page 7900, in the Significance Statement, lines 9–11, "These results question the validity of some 40,000 fMRI studies and may have a large impact on the interpretation of neuroimaging results" should instead appear as "These results question the validity of a number of fMRI studies and may have a large impact on the interpretation of weakly significant neuroimaging results."

Additionally, the authors note that on page 7904, left column, fifth full paragraph, lines 1–3, "It is not feasible to redo 40,000 fMRI studies, and lamentable archiving and data-sharing practices mean most could not be reanalyzed either" should instead appear as "Due to lamentable archiving and data-sharing practices, it is unlikely that problematic analyses can be redone."

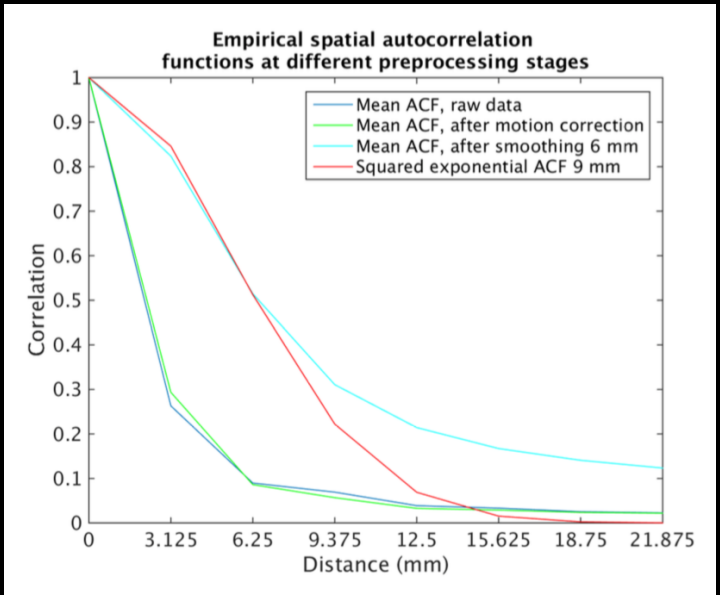
These errors do not affect the conclusions of the article. The online version has been corrected.

www.pnas.org/cgi/doi/10.1073/pnas.1612033113

According to Tom Nichols' blog, a more realistic estimate of # affected studies is 3,500

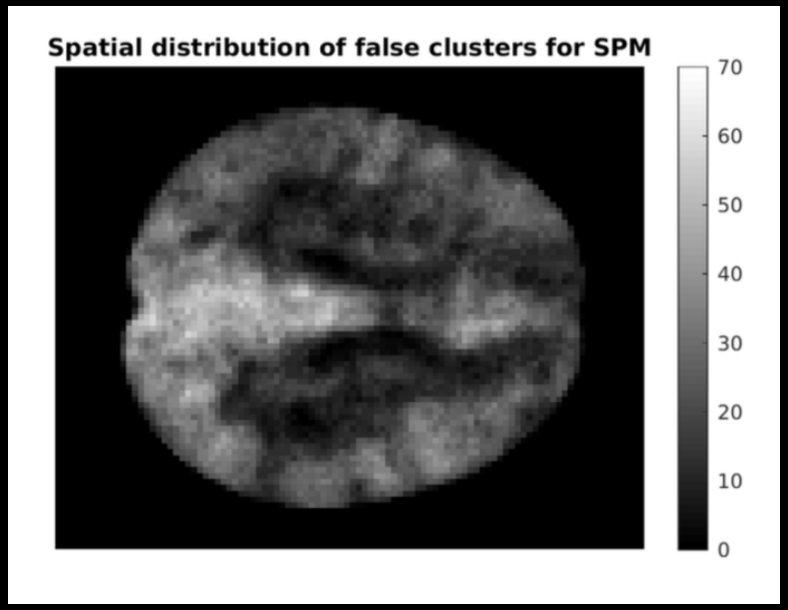
Why So Wrong?

- AFNI had a bug for 15 years
- Tests assume that spatial correlations have a particular shape (squared exponential distribution) – wrong!
- Tests assume constant spatial smoothness across the brain – wrong!



Actual shape of spatial correlation function has a longer tail than modelled function

- Also occurs for raw data and phantoms – inherent to MRI data



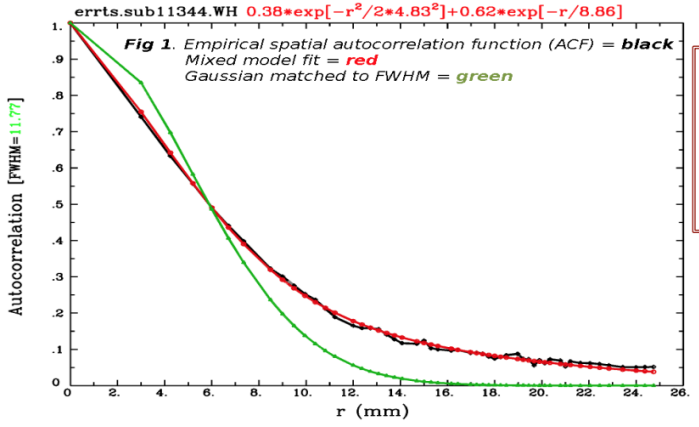
Some areas (esp. posterior cingulate) have higher-than-average smoothness and thus higher false positives

Algorithms improving

- AFNI is correcting cluster-thresholding algorithm to
  - take into account actual spatial correlation function
  - use better, local estimates of smoothness and a better algorithm

(2) Use new "global ACF" smoothness estimator

- Some improvement towards nominal 5% band
- **The first culprit**: long tails in spatial AutoCorrelation Function (ACF)



3 parameter model for ACF:

$$a \cdot e^{-r^2/(2b^2)} + (1-a) \cdot e^{-r/c}$$

Estimates (a,b,c) globally

Cox & Reynolds, HBM 2016

https://ww5.aievolution.com/hbm1601/files/content/abstracts/40760/2082_Cox.pdf

Algorithms improving

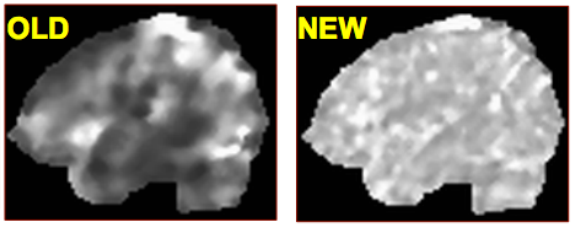
- AFNI is correcting cluster-thresholding algorithm to
  - take into account actual spatial correlation function
  - use better, local estimates of smoothness and a better algorithm

(3) + Use new "local ACF" variable blurring

➢ More improvement towards nominal 5% band

➢ **The second culprit**: strong spatial variability in smoothness

Local FWHM estimates for OLD and NEW blurring methods: Estimates (a,b,c) locally; Adjusts blurring locally



(4) Use median of "local ACF" smoothness estimator on *t*-test residuals

- Rather than use smoothness estimates from each subject (and then combine with median)

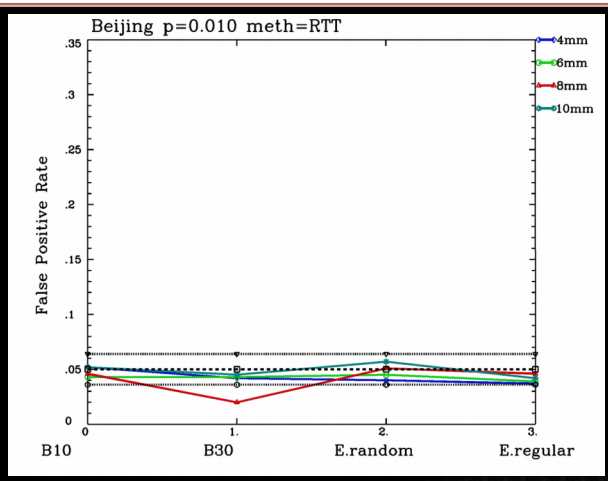
➢ More improvement towards nominal 5% band

- *local ACF estimator illustrated in images above*

(5) Direct randomization and clusterization from *t*-test residuals

- No estimation of smoothness needed
- Brute force computation
- Quick for *t*-tests
- Not easy for more complex statistics (*eg*, LME)

➢ However, the results look pretty good here

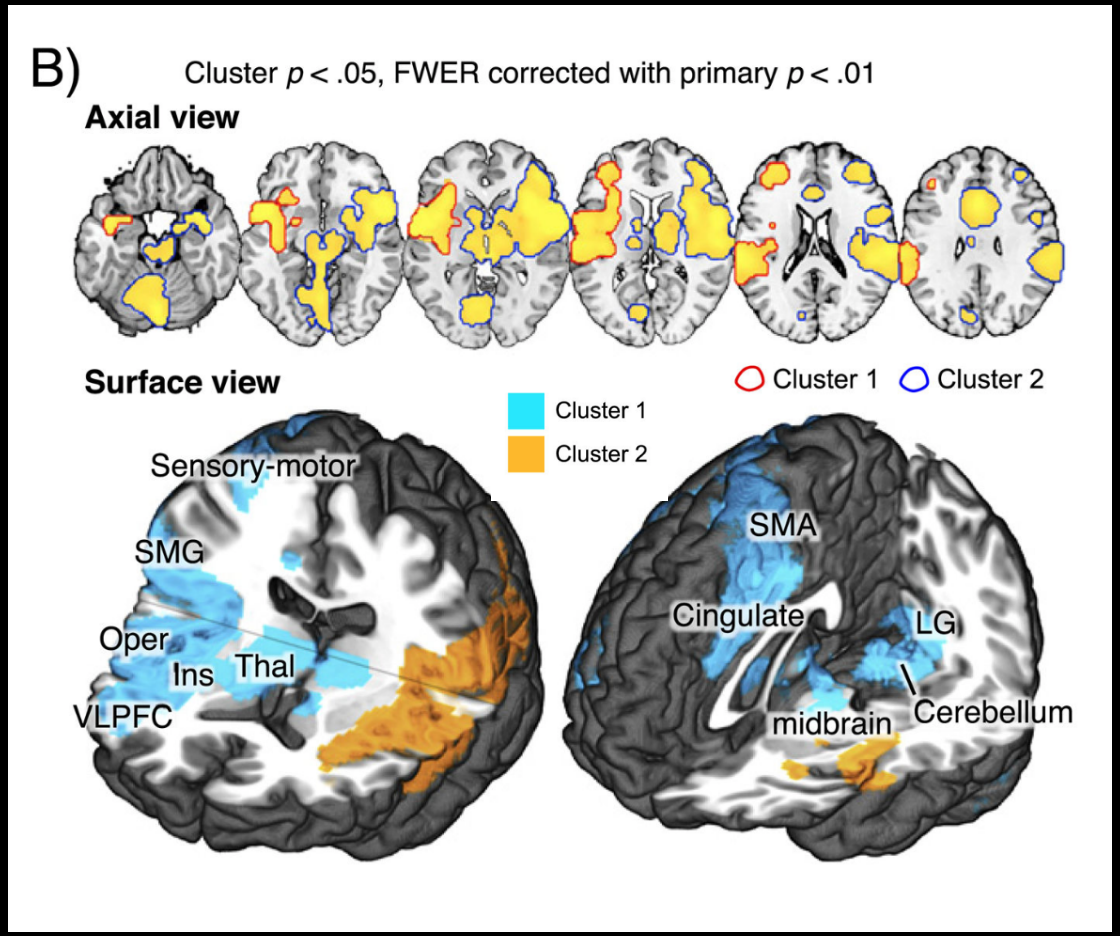


Cox & Reynolds, HBM 2016

https://ww5.aievolution.com/hbm1601/files/content/abstracts/40760/2082_Cox.pdf

Other problems with cluster correction

- problematic in “brain localizer” data
- can’t conclude that every sub-blob is signficant



- Not all blobs in all studies are vulnerable to this problem (many blobs sig above thresholds, not every study went fishing for blobs)
- False positives weren't too bad at p<.001
- Future software can incorporate better approaches such as non-parametric permutation testing (computationally intensive)

> Finally, it's important to remember that "70% chance of finding at least one false positive" does *not* imply that "70% of positives are false". If there are lots of true positives, only a minority of positives will be false. It's impossible to directly know the true positive rate, however.

http://www.ohbmbrainmappingblog.com/blog/keep-calm-and-scan-on

Five Arguments for Cluster Correction

- One pragmatic aspect can be computation resources/time available because permutation testing requires far more time.
- The effect size one is interested in and the sample size available play a role because for small effects the sensitivity is often better in parametric tests (using CDT p=0.001 in combination with an estimated cluster-extent), but potentially at the cost of slightly higher false positives. Importantly, the probability of finding false positives (i.e. the type I error rate) is not the same as the percentage of false positives for a given study! For instance, in experiments with large effects and many true positives, multiple testing correction associated with a 70% chance of finding false positives would overall still result in a relatively small percentage of false positives.
- The slightly higher false positive rates reported by Eklund and colleagues for CDT p=0.001 seem tolerable, one might want to risk/afford to not correct at the nominal 5% (i.e. slightly biased result, but greater sensitivity)
- This bias is probably less than the results in the paper reported for CDT p=0.001 suggest if they were now replicated with the corrected software packages that do not assume a Gaussian shape of the null-distribution and thus provide more stringent control.
- Lastly, a compelling recent analysis suggests that parametric cluster-inference with a CDT p=0.001 seem to correct at the nominal 5% when one takes the proportion (i.e. false discovery rate, FDR) rather than any false positive (family-wise error, FWER, as employed by Eklund and colleagues) as a benchmark.

http://www.cogneurosociety.org/journal-club-on-cluster-failure-why-fmri-inferences-for-spatial-extent-have-inflated-false-positive-rates/

Links of Interest

http://www.ohbmbrainmappingblog.com/blog/keep-calm-and-scan-on

http://blogs.discovermagazine.com/neuroskeptic/2015/12/07/false-positive-fmri-revisited/#.WA2DAZMrJsM
http://blogs.discovermagazine.com/neuroskeptic/2016/07/07/false-positive-fmri-mainstream

http://brainvoyager.com/bvresources/RainersBVBlog/files/a8a22212f9f1f01e4da11fef4ba91da8-34.html

http://www.cogneurosociety.org/journal-club-on-cluster-failure-why-fmri-inferences-for-spatial-extent-have-inflated-false-positive-rates
/

https://ww5.aievolution.com/hbm1601/files/content/abstracts/40760/2082_Cox.pdf