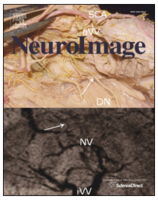
NeuroImage xxx (2013) xxx–xxx

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg



Randomized parcellation based inference

Benoit Da Mota [a,b,⁎,1], Virgile Fritsch [a,b,1], Gaël Varoquaux [a,b], Tobias Banaschewski [e,f], Gareth J. Barker [d], Arun L.W. Bokde [j,k], Uli Bromberg [g], Patricia Conrod [d,h], Jürgen Gallinat [i], Hugh Garavan [j,s,t], Jean-Luc Martinot [l,m], Frauke Nees [e,f], Tomas Paus [n,o,p], Zdenka Pausova [r], Marcella Rietschel [e,f], Michael N. Smolka [q], Andreas Ströhle [i], Vincent Frouin [b], Jean-Baptiste Poline [b,c], Bertrand Thirion [a,b,⁎], the IMAGEN consortium [2]

[a] Parietal Team, INRIA Saclay-Île-de-France, Saclay, France
[b] CEA, DSV, I²BM, Neurospin bât 145, 91191 Gif-Sur-Yvette, France
[c] Henry H. Wheeler Jr. Brain Imaging Center, University of California at Berkeley, USA
[d] Institute of Psychiatry, Kings College London, UK
[e] Central Institute of Mental Health, Mannheim, Germany
[f] Medical Faculty Mannheim, University of Heidelberg, Germany
[g] Universitaetsklinikum Hamburg Eppendorf, Hamburg, Germany
[h] Department of Psychiatry, Universite de Montreal, CHU Ste. Justine Hospital, Canada
[i] Department of Psychiatry and Psychotherapy, Campus Charité Mitte, Charité Universitätsmedizin Berlin, Germany
[j] Institute of Neuroscience, School of Medicine, Trinity College Dublin, Dublin, Ireland
[k] Discipline of Psychiatry, School of Medicine, Trinity College Dublin, Dublin, Ireland
[l] Institut National de la Santé et de la Recherche Médicale, INSERM CEA Unit 1000 "Imaging & Psychiatry", University Paris Sud, Orsay, Maison de Solenn, University Paris Descartes, Paris, France
[m] AP-HP Department of Adolescent Psychopathology and Medicine, Maison de Solenn, University Paris Descartes, Paris, France
[n] Rotman Research Institute, University of Toronto, Toronto, Canada
[o] School of Psychology, University of Nottingham, UK
[p] Montreal Neurological Institute, McGill University, Canada
[q] Neuroimaging Center, Department of Psychiatry and Psychotherapy, Technische Universität Dresden, Germany
[r] The Hospital for Sick Children, University of Toronto, Toronto, Canada
[s] Department of Psychiatry, University of Vermont, USA
[t] Department of Psychology, University of Vermont, USA

ARTICLE INFO

Article history:
Accepted 5 November 2013
Available online xxxx

Keywords:
Group analysis
Parcellation
Reproducibility
Multiple comparisons
Permutations

ABSTRACT

Neuroimaging group analyses are used to relate inter-subject signal differences observed in brain imaging with behavioral or genetic variables and to assess risks factors of brain diseases. The lack of stability and of sensitivity of current voxel-based analysis schemes may however lead to non-reproducible results. We introduce a new approach to overcome the limitations of standard methods, in which active voxels are detected according to a consensus on several random parcellations of the brain images, while a permutation test controls the false positive risk. Both on synthetic and real data, this approach shows higher sensitivity, better accuracy and higher reproducibility than state-of-the-art methods. In a neuroimaging–genetic application, we find that it succeeds in detecting a significant association between a genetic variant next to the *COMT* gene and the BOLD signal in the left thalamus for a functional Magnetic Resonance Imaging contrast associated with incorrect responses of the subjects from a *Stop Signal Task* protocol.

Introduction

Analysis of brain images acquired on a group of subjects makes it possible to draw inferences on regionally-specific anatomical properties of the brain, or its functional organization. The major difficulty with such studies lies in the inter-subject variability of brain shape and vasculature. In functional studies, a task-related variability of subject performance is also observed. The standard-analytic approach is to register and normalize the data in a common reference space. However a perfect voxel-to-voxel correspondence cannot be attained, and the impact of anatomical variability is tentatively reduced by smoothing (Frackowiak et al., 2003). This problem holds for any statistical test, including those associated with multivariate procedures. In the absence of ground truth, choosing the best procedure to analyze the data is a challenging problem. Practitioners as well as methodologists tend to prefer

⁎ Corresponding authors at: CEA, DSV, I²BM, Neurospin bât 145, 91191 Gif-Sur-Yvette, France.
E-mail addresses: benoit.da_mota@inria.fr (B. Da Mota), bertrand.thirion@inria.fr (B. Thirion).
[1] These authors contributed equally to this work.
[2] URL: http://www.imagen-europe.com

1053-8119/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.neuroimage.2013.11.012

models that maximize the sensitivity of a test under a given control for false detections. The level of sensitivity conditional to this control is indeed informative on the usefulness of a model.

Classic statistical tests for neuroimaging

The reference approach in neuroimaging is to fit and test a model at each voxel (univariate voxelwise method), but the large number of tests performed yields a multiple comparison problem. The statistical significance of the voxel intensity test can be corrected with various statistical procedures. First, Bonferroni correction consists in adjusting the significance threshold by dividing it by the number of tests performed. This approach is known to be conservative, especially when non-independent tests are involved, which is the case of neighboring voxels in neuroimaging. Another approach consists in a permutation test to perform a family-wise correction of the p-values (Nichols and Holmes, 2002). Although computationally costly, this method has been shown to yield more sensitive results than studies involving Bonferroni-corrected experiments (Petersson et al., 1999). A good compromise between computation cost and sensitivity can be found in analytic corrections based on Random Field Theory (RFT), in which the smoothness of the images is estimated (Worsley et al., 1992). However, this approach requires both high threshold and data smoothness to be really effective (Hayasaka et al., 2004).

Another widely used method is a test on cluster size, which aims to detect spatially extended effects (Friston et al., 1993; Poline and Mazoyer, 1993; Roland et al., 1993). The statistical significance of the size of an activation cluster can be obtained with theoretical corrections based on the RFT (Hayasaka et al., 2004; Worsley et al., 1996b) or with a permutation test (Holmes et al., 1996; Nichols and Holmes, 2002). Cluster-size tests tend to be more sensitive than voxel-intensity tests, especially when the signal is spatially extended (Friston et al., 1996; Moorhead et al., 2005; Poline et al., 1997), at the expense of a strong statistical control on all the voxels within such clusters. This approach however suffers from several drawbacks. First, such a procedure is intrinsically unstable and its result depends strongly on an arbitrary cluster-forming threshold (Friston et al., 1996). The threshold-free cluster enhancement (TFCE) addresses this issue, by avoiding the choice of an explicit, fixed threshold (Salimi-Khorshidi et al., 2011; Smith and Nichols, 2009) but leads to other arbitrary choices: the TFCE statistic mixes cluster-extent and cluster-intensity measures in proportions that can be defined by the user. More generally, tests that combine cluster size and voxel intensity have been proposed (Hayasaka and Nichols, 2004; Poline et al., 1997). Second, the correlation between neighboring voxels varies across brain images, which makes detection difficult where the local smoothness is low. Combining permutations and RFT to adjust for spatially-varying smoothness leads to more sensitive procedures (Hayasaka et al., 2004; Salimi-Khorshidi et al., 2011). A more complete discussion of the limitations and comparisons of these techniques can be found in (Moorhead et al., 2005; Petersson et al., 1999).

Spatial models for group analysis in neuroimaging

Spatial models try to overcome the lack of correspondence between individual images at the voxel level. The most straightforward and widely used technique consists of smoothing the data to increase the overlap between subject-specific activated regions (Worsley et al., 1996a). In the literature, several approaches propose more elaborate techniques to model the noise in neuroimaging, like Markov Random Fields (Ou et al., 2010), wavelet decomposition (Ville et al., 2004), spatial decomposition or topographic methods (Flandin and Penny, 2007; Friston and Penny, 2003) and anatomically informed models (Keller et al., 2009). These techniques are not widely used probably because they are computationally costly and not always well-suited for analysis of a group of subjects. A popular approach consists of working with subject-specific Regions of Interest (ROIs), that can be defined in a way that accommodates inter-subject variability (Nieto-Castanon et al., 2003). The main limitation of such an approach (Bohland et al., 2009) is that there is no widely accepted standard for partitioning the brain, especially for the neocortex. Data-driven parcellation was proposed by Thirion et al. (2006) to overcome this limitation: they improve the sensitivity of random effect analysis by considering parcels defined at the group level.

Neuroimaging–genetic studies

While most studies investigate the difference of activity between groups or the level of activity within a population, neuroimaging studies are often concerned by testing the effect of exogeneous variables on imaging target variables, and there is increasing interest in the joint study of neuroimaging and genetics to improve understanding of both normal and pathological variability of the brain organization. Single nucleotide polymorphisms (SNPs) are the most common genetic variants used in such studies: They are numerous and represent approximately 90% of the genetic between-subject variability (Collins et al., 1998). Voxel intensity and cluster size methods have been used for genome-wide association studies (GWAS) (Stein et al., 2010), but the multiple comparison problem does not permit finding significant results, despite efforts to estimate the effective number of tests (Gao et al., 2010) or by running computationally expensive, but accurate permutation tests (Da Mota et al., 2012). Recently, important efforts have been done to design more sophisticated multivariate methods (Floch et al., 2012; Kohannim et al., 2011; Vounou et al., 2010), the results of which are more difficult to interpret; another alternative is to work at the gene level instead of SNPs (Ge et al., 2012; Hibar et al., 2011).

The randomized parcellation approach

The parcellation model (Thirion et al., 2006) has several advantages: (*i*) it is a simple and easily interpretable method, (*ii*) by reducing the number of descriptors, it reduces the multiple comparisons problem, and (*iii*) the choice of the parcellation algorithm can lead to parcels adapted to the local smoothness. But parcellations, when considered as spatial functions, highly depend on the data used to construct them and the choice of the number of parcels. In general, a parcellation defined in a given context might not be a good descriptor in a slightly different context, or may generalize poorly to new subjects. This implies a lack of reproducibility of the results across subgroups, as illustrated later in Fig. 7. The weakness of this approach is the large impact of a parcellation scheme that cannot be optimized easily for the sake of statistical inference; it may thus fail to detect effects in poorly segmented regions. We propose to solve this issue by using several randomized parcellations (Bühlmann et al., 2012; Varoquaux et al., 2012) generated using resampling methods (bootstrap) and average the corresponding statistical decisions. Replacing an estimator such as parcel-level inference by means of bootstrap estimates is known to *stabilize* it; a fortunate consequence is that the *reproducibility* of the results (across subgroups of subjects) is improved. Formally, this can be understood as handling the parcellation as a hidden variable that needs to be integrated out in order to obtain the posterior distribution of statistical values. The final decision is taken with regard to the stability of the detection of a voxel (Alexander and Lange, 2011; Meinshausen and Bühlmann, 2010) across parcellations, compared to the null hypothesis distribution obtained by a permutation test.

A multivariate problem: the detection of outliers

The benefits of the randomized parcellation approach can also be observed in multivariate analysis procedures, such as predictive modeling (Varoquaux et al., 2012) or outlier detection. In this work, we focus on

Please cite this article as: Da Mota, B., et al., Randomized parcellation based inference, NeuroImage (2013), http://dx.doi.org/10.1016/j.neuroimage.2013.11.012

the latter: neuroimaging datasets often contain atypical observations; such *outliers* can result from acquisition-related issues (Hutton et al., 2002), bad image processing (Wu et al., 1997), or they can merely be extreme examples of the high variability observed in the population. Because of the high dimensionality of neuroimaging data, screening the data is very time consuming, and becomes prohibitive with large cohort studies. Covariance-based outlier detection methods have been proposed to perform statistically-controlled inclusion of subjects in neuroimaging studies (Fritsch et al., 2012) and yield a good detection accuracy. These methods rely on prior reduction of the data dimension which is obtained by taking signal averages within predefined brain parcels. As a consequence, the results depend on a fixed brain parcellation and are unstable. Randomization might thus improve the procedure.

Outline

In "Materials and methods", we introduce methodological prerequisites and we describe the randomized parcellation approach. In "Experiments", we provide the description of the experiments used to assess the performances of our procedure. We evaluate our approach on simulations and on real fMRI data for the random effect analysis problem. Then, we illustrate the interest of the approach for neuroimaging–genetic studies, on a gene candidate (*COMT*) which is widely investigated in the context of brain diseases. Finally, we show that this technique is suitable for detecting outliers in neuroimaging data, thus extending the application scope of randomized parcellations to multivariate analysis procedures. In "Results", we report the results of the experiments and finally we discuss different aspects and choices that can influence the method performance.

Materials and methods

Statistical modeling for group studies

Neuroimaging studies are often designed to test the effect of miscellaneous variables on imaging target variables. For a study involving n subjects, neuroscientists generally consider the following model:

$$Y = X\beta + \epsilon,$$

where Y is a $n \times p$ matrix representing the signal of n subjects described each by p descriptors (e.g. voxels or parcels of an fMRI contrast image) and X is the $n \times (q_1 + q_2)$ set of q_1 explanatory variables, a predefined linear combination of which is to be tested for a non-zero effect, and q_2 covariables that explain some portion of the signal but are not to be tested for an effect. β are the coefficients of the model to be estimated, and ϵ is some Gaussian noise. Variables in X can be of any type (genetic, artificial, behavioral, experimental…). A standard univariate analysis technique consists in fitting p Ordinary Least Square (OLS) regressions, one for each column of Y, as a target variable, and each time perform a non-zero significance test on the $c^T\beta$ quantity, where $c \in \mathbb{R}^{q_1+q_2}$ is the *contrast vector* that defines the linear combination of the variables to be tested. This test involves the estimated coefficients of the model $\hat{\beta}$ and the noise estimate $\hat{\sigma}$ to compute a standard t- or F-statistic.

Parcellation and Ward algorithm

In functional neuroimaging, brain atlases are often used to provide a low-dimensional representation of the data by considering signal averages within groups of voxels (regions of interest). If those groups of voxels do not overlap and every voxel belongs to one group, the term *parcel* is employed, and the atlas is called a *parcellation*. In this work, we restrict ourselves to working with parcellations, although our methodology could be applied to any kind of brain partition (set of ROIs). We construct parcellations from the images that we work on, because this data-driven approach better takes into account the unknown spatial data structure. Following (Michel et al., 2012; Varoquaux et al., 2012), we use spatially-constrained Ward hierarchical clustering (Ward, 1963) to cluster the voxels in K parcels, yielding what we will refer to as a K-parcellation. This approach creates a hierarchy of parcels represented as a tree. The root of the tree is the unique parcel that gathers all the voxels, the leaves being the parcels with only one voxel. When merging two clusters, the Ward criterion chooses the cluster that produces a supra-cluster with minimal variance. Any cut of the tree corresponds to a unique parcellation. This algorithm has several advantages: (*i*) It captures well local correlations into spatial clusters, (*ii*) efficient implementations exist (Pedregosa et al., 2011), and (*iii*) obtained parcellations are invariant by permutation of the subjects and sign of the input data. A gives a formal description of Ward's clustering algorithm. We also show some examples of parcellations and discuss the geometric properties of the parcels.

Randomized parcellation based inference

Randomized parcellation based inference (*RPBI*) performs several standard analyses based on different parcellations and aggregates the corresponding statistical decisions. Let $\mathcal{P}$ be a finite set of parcellations, and V be the set of voxels under consideration. Given a voxel v and a parcellation P, the parcel-based thresholding function θ_t is defined as:

$$\theta_t(v, P) = \begin{cases} 1 \text{ if } F(\Phi_P(v)) > t \\ 0 \text{ otherwise} \end{cases} \tag{1}$$

where $\Phi_P : V \to P$ is a mapping function that associates each voxel with a parcel from the parcellation P $(\forall v \in P^{(i)}, \Phi_P(v) = P^{(i)})$. For a predefined test, F returns the F-statistic associated with the average signal of a given parcel (a t or other statistic is also possible). Finally, the aggregating statistic at a voxel v is given by the counting function C_t:

$$C_t(v, \mathcal{P}) = \sum_{P \in \mathcal{P}} \theta_t(v, P). \tag{2}$$

$C_t(v, \mathcal{P})$ represents the number of times the voxel v was part of a parcel associated with a statistical value larger than t across the folds of the analysis conducted on the set of parcellations $\mathcal{P}$. We set the parameter t to ensure a Bonferroni-corrected control at $p < 0.1$[3] in each of the parcel-level analyses. In practice, the results are weakly sensitive to mild variations of t. In order to assess the significance of the counting statistic at each voxel, we perform a permutation test, i.e. we tabulate the distribution of $C_t(v, \mathcal{P})$ under the null hypothesis that there is no significant correlation between the voxels' mean signal and the target variable. Depending on the comparison to be performed, we switch labels (comparison between groups) or we swap signs (testing that the mean is non-zero). As a result, we get a voxel-wise p-value map similar to a standard group analysis map (see Fig. 1). We obtain family-wise error control by tabulating the maximal value across voxels in the permutation procedure. The θ_t function can be replaced by any function that is convex with respect to t. In particular, the natural choice $\theta_t(v,P) = F(\Phi_P(v))$ yields similar results (not shown in the paper) but its computation requires much more memory since the $v \to \theta_t(v,P)$ mapping and bootstrap averages are no longer sparse. An important prerequisite for our approach is to generate several parcellations that are different enough from each other to guarantee that the analysis

[3] We determine this value empirically to obtain a well-behaved null distribution of the counting statistic. With 1 target and 1000 parcels, it corresponds to a raw p-value $<10^{-4}$.

Please cite this article as: Da Mota, B., et al., Randomized parcellation based inference, NeuroImage (2013), http://dx.doi.org/10.1016/j.neuroimage.2013.11.012

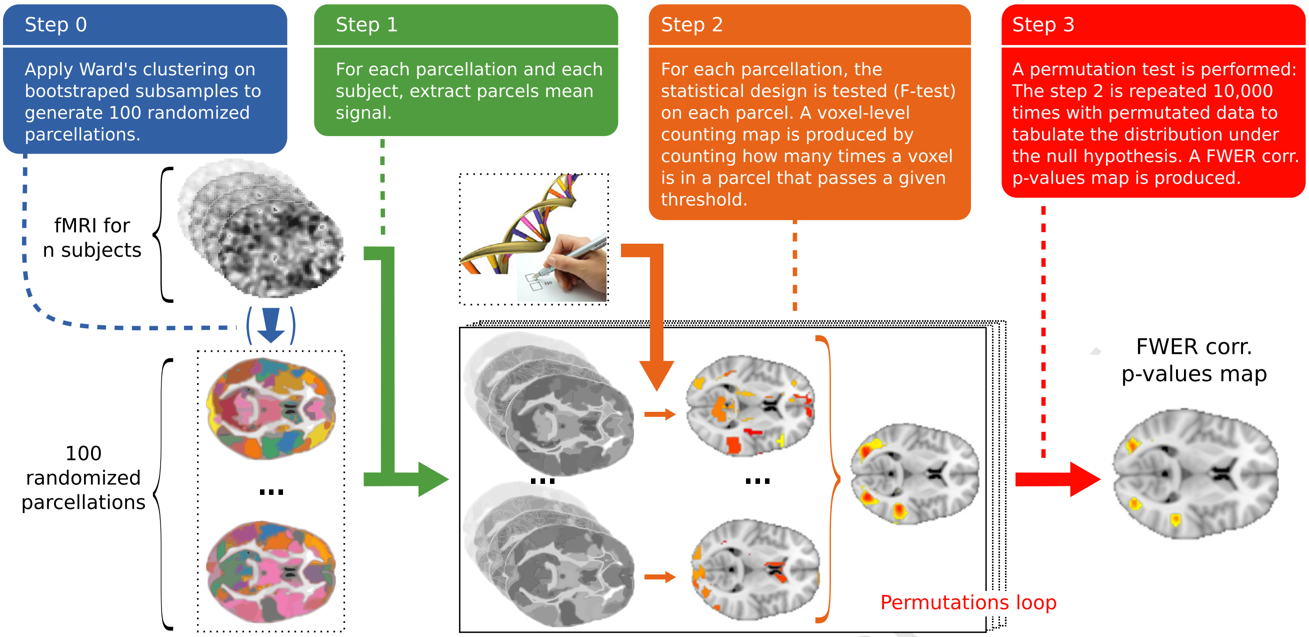

Step 0
Apply Ward's clustering on bootstraped subsamples to generate 100 randomized parcellations.
Step 1
For each parcellation and each subject, extract parcels mean signal.
Step 2
For each parcellation, the statistical design is tested (F-test) on each parcel. A voxel-level counting map is produced by counting how many times a voxel is in a parcel that passes a given threshold.
Step 3
A permutation test is performed: The step 2 is repeated 10,000 times with permutated data to tabulate the distribution under the null hypothesis. A FWER corr. p-values map is produced.
fMRI for n subjects
100 randomized parcellations
...
FWER corr. p-values map
Permutations loop

Fig. 1. Overview of the randomized parcellation based inference framework on an example with few parcels. The variability of the parcel definition is used to obtain voxel-level statistics.

conducted with each of those parcellations samples correctly the set of regions that display some activation for the effect considered. One way to achieve this is to take bootstrap samples of subjects and apply Ward's clustering algorithm to their contrast maps, to build brain parcellations that best summarize the data subsamples, i.e. so that the parcel-level mean signal summarizes the signal within each parcel, in each subject. If enough subjects are used, all the parcellations offer a good representation of the whole dataset. It is important that the bootstrap scheme generates parcellations with enough entropy (Varoquaux et al., 2012). Spatial models try to address the problem of imperfect voxel-to-voxel correspondence after coregistration of the subjects in the same reference space. Our approach is clearly related to anisotropic smoothing (Sol et al., 2001), in the sense that obtained parcels are not spherical and in the aggregation of the signals of voxels in a given parcel, certain directions are preferred. Unlike smoothing or spatial modeling applied as a preprocessing, our statistical inference embeds the spatial modeling in the analysis and decreases the number of tests and their dependencies. In addition to the expected increase of sensitivity, the randomization of the parcellations ensures a better reproducibility of the results, unlike inference on one fixed parcellation. Last, the $C_t(v, \mathcal{P})$ statistic is reliable in the sense that is does not depend on side effects such as the parcel size. This is formally checked in Appendix B.

Sensitivity and accuracy assessments

We want to assess the sensitivity of our approach at a fixed level of specificity and compare it to the other methods. Thus, we are interested in whether or not a significant effect was reported according to the different methods. Under the assumption that the method specificity is controlled with a given false positive rate, the method with the highest number of detections is the most sensitive.

Note that a direct comparison of the sensitivity of the different procedures (voxel-level, cluster-level, TFCE, parcel-based), i.e. their rate of detections, is not very meaningful. Indeed, only voxel-level statistics provide a strong control on false detections. The other procedures violate the subset pivotality condition, namely that the rejection of the null at a given location does not alter the distribution of the decision statistics under the null at other locations (see e.g. Westfall and Troendle, 2008). This means that the rejection of the null at a given location is not independent of the rejection at the null at nearby locations; specifically, the rejection of the null at a given voxel is bound to the voxel in voxel-based tests, while it is not for other kinds of inferences considered here. Strictly speaking, those only reject a global null. Note however, that such a weak control on false detections is still useful in problems with small effect sizes (see "Neuroimaging–genetic study"). The ideal method would be able to detect small effects, but would be also quite specific about their location. That is why an analysis of the sensitivity should always be considered with an analysis of the accuracy.

In our experiments, to estimate a method's accuracy, we construct Receiver Operating Characteristic (ROC) curves (Hanley and McNeil, 1982) by reporting the proportion of true positives in the detections for different levels of false positives. The true/false positives are determined according to a *ground truth* that is defined based on the simulation setup or empirically when dealing with real data. In practice, we are interested in low false positive rates, so we present the ROC curves in logarithmic scale.

Use of randomized parcellation in multivariate models

Various neuroimaging methods rely on a prior dimension reduction of the data, and can therefore benefit from a randomized parcellation approach that stabilizes the ensuing statistical procedure. Beyond the specific case of group analysis investigated in this manuscript, we apply the randomized parcellation technique to the outlier detection task. Unlike group analysis, outlier detection can be formulated as a multivariate problem, especially because we consider covariance-based outlier detection (Fritsch et al., 2012), where an estimate of the data covariance matrix is computed and then used to provide an outlier score for each observation, i.e. correlations between features are taken into account in the final decision about whether or not an image should be considered an outlier.

IMAGEN, a neuroimaging–genetic study

IMAGEN is a European multicentric study involving adolescents (Schumann et al., 2010). It contains a large functional neuroimaging

Please cite this article as: Da Mota, B., et al., Randomized parcellation based inference, NeuroImage (2013), http://dx.doi.org/10.1016/j.neuroimage.2013.11.012

database with fMRI associated with 99 different contrast images for 4 protocols in more than 2000 subjects, who gave informed signed consent. Regarding the functional neuroimaging data, the faces protocol (Grosbras and Paus, 2006) was used, with the [*angry faces–control*] contrast, i.e. the difference between watching angry faces and non-biological stimuli (concentric circles). We also use the Stop Signal Task protocol (Logan, 1994) (SST), with the activation during a [*go wrong*] event, i.e. when the subject pushes the wrong button. Images from the Modified Incentive Delay task (Knutson et al., 2000) (MID) were used to construct alternative randomized parcellations.

Eight different 3 T scanners from multiple manufacturers (GE, Siemens, Philips) were used to acquire the data. Standard preprocessing, including slice timing correction, spike and motion correction, temporal detrending (functional data), and spatial normalization (anatomical and functional data), were performed using the SPM8 software and its default parameters; functional images were resampled at 3 mm resolution. All images were warped in the MNI152 coordinate space using a study-specific template. Obvious outliers detected using simple rules such as large registration or segmentation errors or very large motion parameters were removed after this step. BOLD time series was recorded using Echo-Planar Imaging, with TR = 2200 ms, TE = 30 ms, flip angle = 75° and spatial resolution 3 mm × 3 mm × 3 mm. Gaussian smoothing at 5 mm-FWHM was finally added.[4] Contrasts were obtained using a standard linear model, based on the convolution of the time course of the experimental conditions with the canonical hemodynamic response function, together with standard high-pass filtering (period = 120 s) and temporally auto-regressive noise model. The estimation of the first-level was carried out using the SPM8 software. T1-weighted MPRAGE anatomical images were acquired with spatial resolution 1 mm × 1 mm × 1 mm, and gray matter probability maps were available for 1986 subjects as outputs of the SPM8 "New Segmentation" algorithm applied to the anatomical images. A mask of the gray matter was built by averaging and thresholding the individual gray matter probability maps. More details about data preprocessing can be found in (Thyreau et al., 2012). Genotyping was performed genome-wide using Illumina Quad 610 and 660 chips, yielding approximately 600,000 autosomic SNPs. 477,215 SNPs are common to the two chips and pass *plink* standard parameters (Minor Allele Frequency > 0.05, Hardy–Weinberg Equilibrium $P < 0.001$, missing rate per SNP > 0.05).

Experiments

Random effect analysis on simulated data

We simulate fMRI contrast images as volumes of shape 40 × 40 × 40 voxels. Each contrast image contains a simulated 4 × 4 × 4 activation patch at a given location, with a spatial jitter following a three-dimensional $N(0,I_3)$ distribution (coordinates of the jitter are rounded to the nearest integers). The strength of the activation is set so that the signal to noise ratio (SNR) peaks at 2 in the most associated voxel. The background noise is drawn from a (0,1) distribution, Gaussian-smoothed at σ_{noise} isotropic and normalized by its global empirical standard deviation. After superimposing noise and signal images, we optionally smooth at $\sigma_{\text{post}} = 2.12$ voxels isotropic, corresponding to a 5 voxel Full Width at Half Maximum (FWHM). Voxels with a probability above 0.1 to be active in a large sample test are considered as part of the ground truth. Ten subsamples (or groups) of 20 images are then generated to perform analyses. Each time, RPBI was conducted with one hundred 1000-parcellations built from a bootstrapped selection of the 20 images involved. For each of the 10 groups, we expect to obtain a p-value map that shows a significant effect at the mean location of generated artificial activations in the contrast images.

We investigate the ability of four methods to actually recover the region of activation:

(i) voxel-level group analysis, which is the standard method in neuroimaging;
(ii) cluster-size group analysis, which is known to be more sensitive than voxel-intensity group analysis;
(iii) threshold-free cluster enhancement (TFCE) (Smith and Nichols, 2009);
(iv) RPBI, which is our contribution.

We control the specificity of each procedure by permutation testing. In order to ensure an accurate type 1 error control, we generate 400 sets of 20 images with no activation (i.e. the images are only noise with $\sigma_{\text{noise}} = 1$, and SNR = 0). We evaluate the false positive rate at voxel level for RPBI. We perform the same simulated data experiment with a more complex activation shape (shown in Fig. 2) as we think it better corresponds to activations encountered in real data. The rest of the experimental design remains the same and we perform the same comparison between methods.

Random effect analysis on real fMRI data

In this experiment, we work with an [*angry faces–control*] fMRI contrast. We kept data from 1430 subjects after removal of the subjects with missing data and/or bad or missing covariables. After standard preprocessing of the images, including registration of the subjects onto the same template, we test each voxel for a zero mean across the 1430 subjects with an OLS regression, including handedness and sex as covariables, yielding a reference voxel-wise p-value map. We threshold this map in order to keep 5% of the most active voxels (corresponding to $-\log_{10}P > 77.5$), and we consider it the ground truth. Since we use a voxel based threshold, the ground truth may be biased to voxel-level statistics (thus disadvantaging our method).

Our objective is to retrieve the population's reference activity pattern on subsamples of 20 randomly drawn subjects and compare the performance of several methods in this problem. Because of the reduced number of subjects used, we cannot expect to retrieve the same activation map as in the full-sample analysis due to a loss in statistical

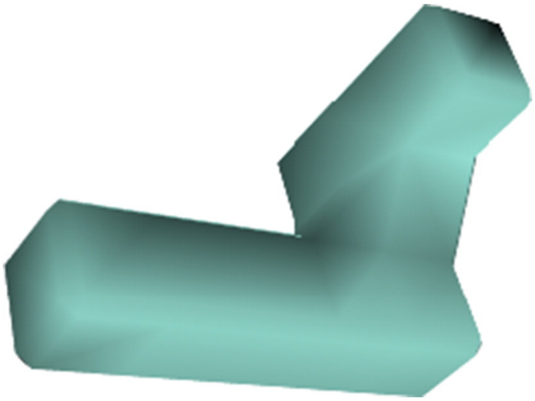


Fig. 2. Complex activation shape used for simulations. This activation shape is more scattered than a cube, and potentially better reflects the complex shape of real data activations. Note that, according to its original publication, TFCE performance is independent of the activation shape (Smith and Nichols, 2009).

[4] Smoothing is applied only in the first-level analysis in order to improve the sensitivity of the General Linear Model that yields the contrast maps.

power. We therefore measure the sensitivity and we build ROC curves to assess the performance of the methods. We perform our experiment on 10 different subsamples and we use the same analysis methods as the previous experiment. We propose to observe the behavior of our method with the use of parcellations of different kinds. We perform analysis of the 10 different subsamples with the following parcellation schemes:

(i) RPBI (sh. parcels) with parcellations built on bootstrapped subsamples of 150 images among the 1430 images corresponding to the fMRI contrast under study;
(ii) RPBI (alt. parcels) with shared parcellations built on images corresponding to another, independent fMRI contrast;
(iii) RPBI (rand. parcels) with shared parcellations built on smoothed Gaussian noise;

We also assess the stability of all these methods by counting how many times each voxel was associated to a significant effect across subsamples. We present the inverted cumulative normalized histogram of this count for each method, restricting our attention to the voxels that were reported at least once. A method is considered to be more stable than another if the same voxels appear more often, that is if its histogram shows many high values.

Neuroimaging–genetic study

The aim of this experiment is to show that RPBI has the potential to uncover new relationships between neuroimaging and genetics. We consider an fMRI contrast corresponding to events where subjects make motor response errors ([*go wrong*] fMRI contrast from a Stop Signal Task) and its associations with *Single-Nucleotide Polymorphisms* (*SNPs*) in the *COMT* gene. This gene codes for the Catechol-O-methyltransferase, an enzyme that catalyzes transfer of neurotransmitters like dopamine, epinephrine and norepinephrine, making it one of the most studied genes in relation to brain (Puls et al., 2009; Smolka et al., 2007). Subjects with too many missing voxels in the brain mask or with bad task performance were discarded. Regarding genetic variants, we kept 27 SNPs in the *COMT* gene (±20 kb) that pass *plink* standard parameters (Minor Allele Frequency > 0.05, Hardy–Weinberg Equilibrium $P > 0.001$, missing rate per SNP < 0.05). The ±20 kb window includes some SNPs in the ARVCF gene, that are in *linkage disequilibrium* with SNPs in *COMT*. Age, sex, handedness and acquisition center were included in the model as confounding variables. Remaining missing data were replaced by the median over the subjects for the corresponding variables. After applying all exclusion criteria 1372 subjects remained for analysis.

For each of the 27 SNPs, we perform a massively univariate voxelwise analysis with the algorithm presented in (Da Mota et al., 2012), including cluster-size analysis (Hayasaka and Nichols, 2003), and RPBI through 100 different Ward's 1000-parcellations. To assess significance with a good degree of confidence we performed 10,000 permutations.

Outlier detection

We finally apply the concept of randomized parcellations to outlier detection. We work with a cohort of 1886 fMRI contrast images. In a first step, we randomly select 300 subjects and summarize the dataset by computing a 500-parcellation (obtained by Ward's) and averaging signal over each parcel. We perform a reference outlier detection on this dataset with a regularized version of a robust covariance estimator *RMCD-RP* (Fritsch et al., 2012). This outlier detection algorithm consists of fitting robust covariance estimators to random data projections. For the outlier detection we use the average of the Mahalanobis distances of the observations to the population mean in every projection subspace. In a second step, we perform outlier detections with RMCD-RP on random subsamples: We randomly draw a subsample of n subjects and perform 100 outlier detections with RMCD-RP on 100 different p-dimensional representations of the data defined by 100 Ward's p-parcellations built on 300 bootstrapped subjects from the whole cohort. Following the model of RPBI, we report how many times each subject was reported as an outlier through these 100 outlier detections and we use that number as an outlier score. We hence construct two Receiver Operating Characteristic (ROC) curves (Hanley and McNeil, 1982): one for randomized parcellation-based (RPB) outlier detection and the other as the average ROC curve of the 100 inner outlier detections used to obtain the RPB outlier detection. Finally, we report the rate of correct detections when 5% of false detections are accepted, to control the sensitivity of this test when wrongly rejecting few non-outlier data. These statistics make it possible to easily measure the accuracy improvement of RPB outlier detection across several experiments performed with different subsamples of n subjects (keeping the same reference decision obtained at the first step). In our experiment, we choose to work with

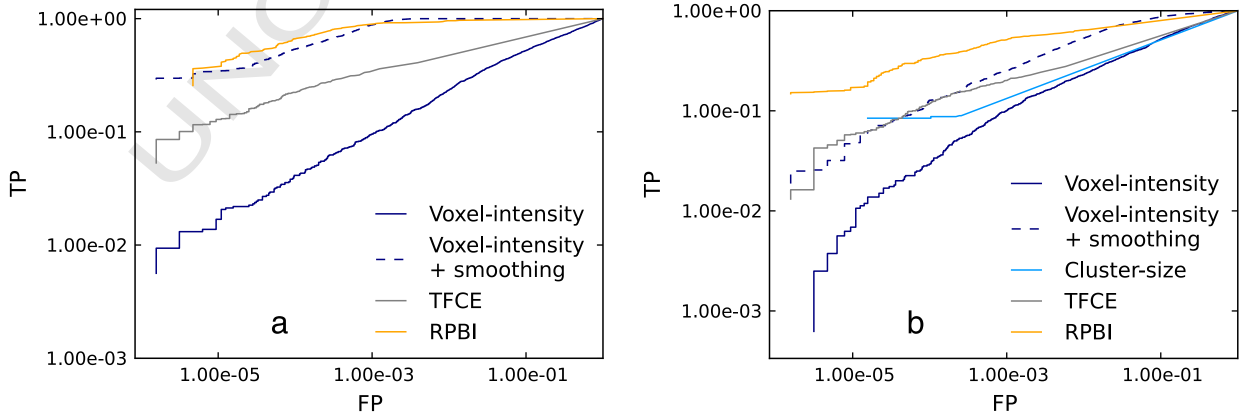


Fig. 3. Simulated data (cubic effect). ROC curves for various analysis methods across 10 random subsamples containing 20 subjects. SNR = 2 and noise spatial smoothness: (a) $\sigma_{\text{noise}} = 0$, (b) $\sigma_{\text{noise}} = 1$. The curves are obtained by thresholding the statistical brain maps at various levels, yielding as many points on the curves. The x-axis is the expected number of false positives per image. The curve for cluster-size inference could not be built for $\sigma_{\text{noise}} = 0$ because the detections correspond either to true positives only, or to false positives only. RPBI outperforms other methods.

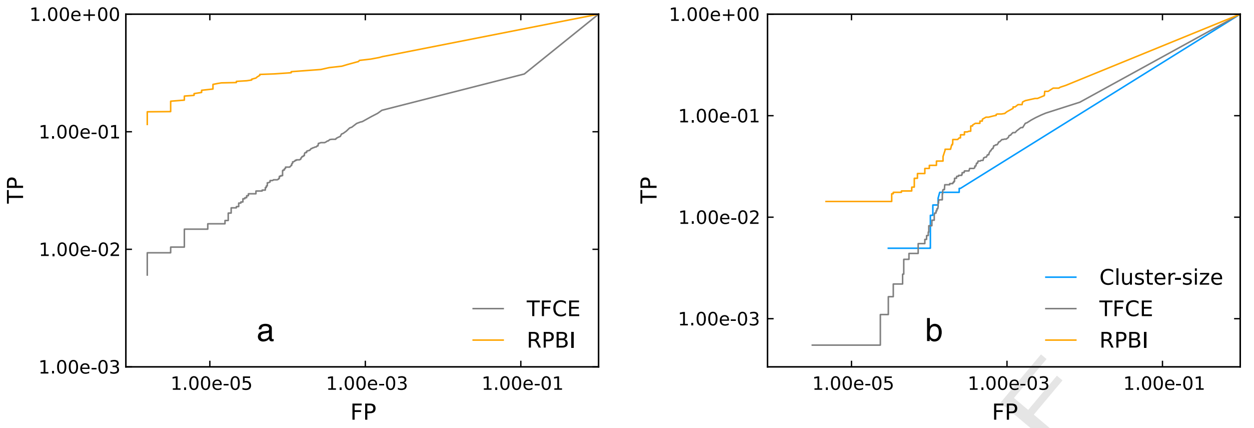


Fig. 4. Simulated data (complex activation shape). ROC curves for various analysis methods across 10 random subsamples containing 20 subjects. SNR = 2 and noise spatial smoothness: (a) $\sigma_{\text{noise}} = 0$, (b) $\sigma_{\text{noise}} = 1$. The curves are obtained by thresholding the statistical brain maps at various levels, yielding as many points on the curves. The *x*-axis is the expected number of false positives per image. The curve for cluster-size inference could not be built for $\sigma_{\text{noise}} = 0$ because the detections correspond either to true positives only, or to false positives only. For the same reason, voxel-intensity performance could not be presented in any of the plots. RPBI outperforms other methods.

$p = 100$ and $n = \{80,100,200,300,400\}$, yielding p/n configurations that correspond to various problem difficulties. For a fixed (n,p) couple, we run the experiment on 50 different subsamples and we present the rate of correct detections in a box-plot.

Results

Random effect analysis on simulated data

Voxel-intensity group analysis is the only method that benefits from a posteriori smoothing, while spatial methods lose sensitivity and accuracy when the images are smoothed. This is in agreement with the theory and the results of (Worsley et al., 1996a). Figs. 3 and 4 show that detections made by spatial methods (cluster-size group analysis, TFCE and RPBI) do not come with wrongly reported effects in voxels close to the actual effect location. This would be the case for a method that simply extends a recovered effect to the neighboring voxels and would wrongly be thought to be more sensitive because it points out more voxels. RPBI offers the best accuracy as its ROC curve dominates in Fig. 3. We could not always build ROC curves for the cluster-size method. This illustrates an issue of the cluster-forming threshold: most voxels do not pass the threshold and then were discarded by the method, leading to a true positive rate equal to zero. The cluster-forming threshold directly acts on the recovery capability of the method, but lowering the threshold does not increase the sensitivity of this approach in general. By integrating over multiple thresholds, the TFCE partially addresses this issue. We also encountered an issue in the construction of ROC curves for voxel-intensity based analysis in our simulations with a complex-shaped activation (see Fig. 4): either there were only true positives, or there were only false positives in our results, hence a lack of point for the construction of the ROC curves. When no signal is put in the data (SNR = 0), RPBI reports an activation 37 times over 400 at $P < 0.1$ FWER corrected, 20 times at $P < 0.05$ FWER corrected, and 4 times at $P < 0.01$ FWER corrected. In all cases, it corresponds to the nominal type I error rate.

Random effect analysis on real fMRI data

Fig. 5a shows the sensitivity improvement relative to cluster-size for various analysis methods under control for false detections at 5% FWER. Cluster-size was taken as the reference because it is the method that yields the most sensitivity among state-of-the-art methods to which we compare RPBI to. RPBI achieves the best

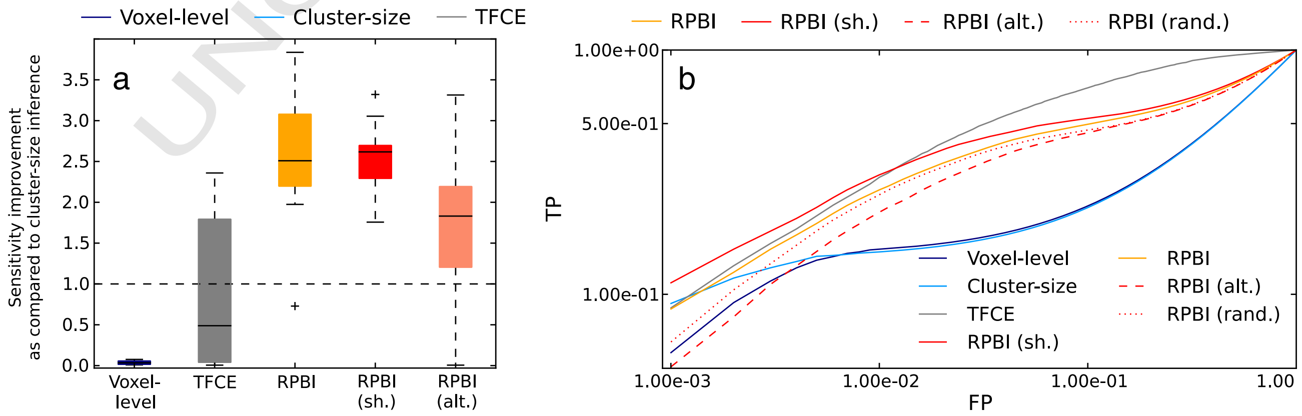


Fig. 5. Real fMRI data. Evaluation of the performances for various analysis methods across 10 random subsamples containing 20 subjects, on a [*angry faces–control*] fMRI contrast from the *faces* protocol. (a) Sensitivity improvement relative to cluster-size under control of the specificity at 5% FWER. (b) ROC curves built with a pseudo ground truth where 5% of the most active voxels across 1430 subjects are kept. RPBI and TFCE have similar performance for low false positive rates ($<10^{-2}$), although TFCE performs slightly better.

sensitivity improvement, and RPBI with shared, alternative or random parcels are always more sensitive than TFCE. Voxel-level group analysis yields poor performance while cluster-size analysis is comparable to TFCE. These gains in sensitivity should be linked with a measure of accuracy (see “Materials and methods”). Fig. 5b shows the ROC curves associated with the performance of the methods under comparison. For acceptable levels of false positives ($<10^{-2}$), RPBI almost equals TFCE when we use parcellations that have been built on the contrast under study. RPBI with alternative or random parcels yields poor recovery although these approaches are based on the randomized parcellation scheme. This demonstrates that the sensitivity is not a sufficient criterion and that the choice of parcellations plays an important role in the success of RPBI. Unlike simulations, real data may contain outliers, which reduce the effectiveness of all the presented methods. One benefit of RPBI with shared parcels is that the impact of bad samples in the test set is lowered, because the parcellations are informed by potentially abundant side data. This requires other data from a similar protocol, but Fig. 5b shows that this approach outperforms other methods by finding more true positives.

The lack of stability of group studies is a well-known issue, yet it depends on the analysis performed (Strother et al., 2002; Thirion et al., 2007). RPBI has better reproducibility than the other methods, as shown in Fig. 7. The histogram of the RPBI method dominates, which means that significant effects were reported more often at the same location (i.e. the same voxel) across subgroups when using RPBI than when using the other methods. For RPBI with shared parcels, it is even more pronounced and this is explained by the fact that parcellations are shared across subgroups, which is another advantage to this method.

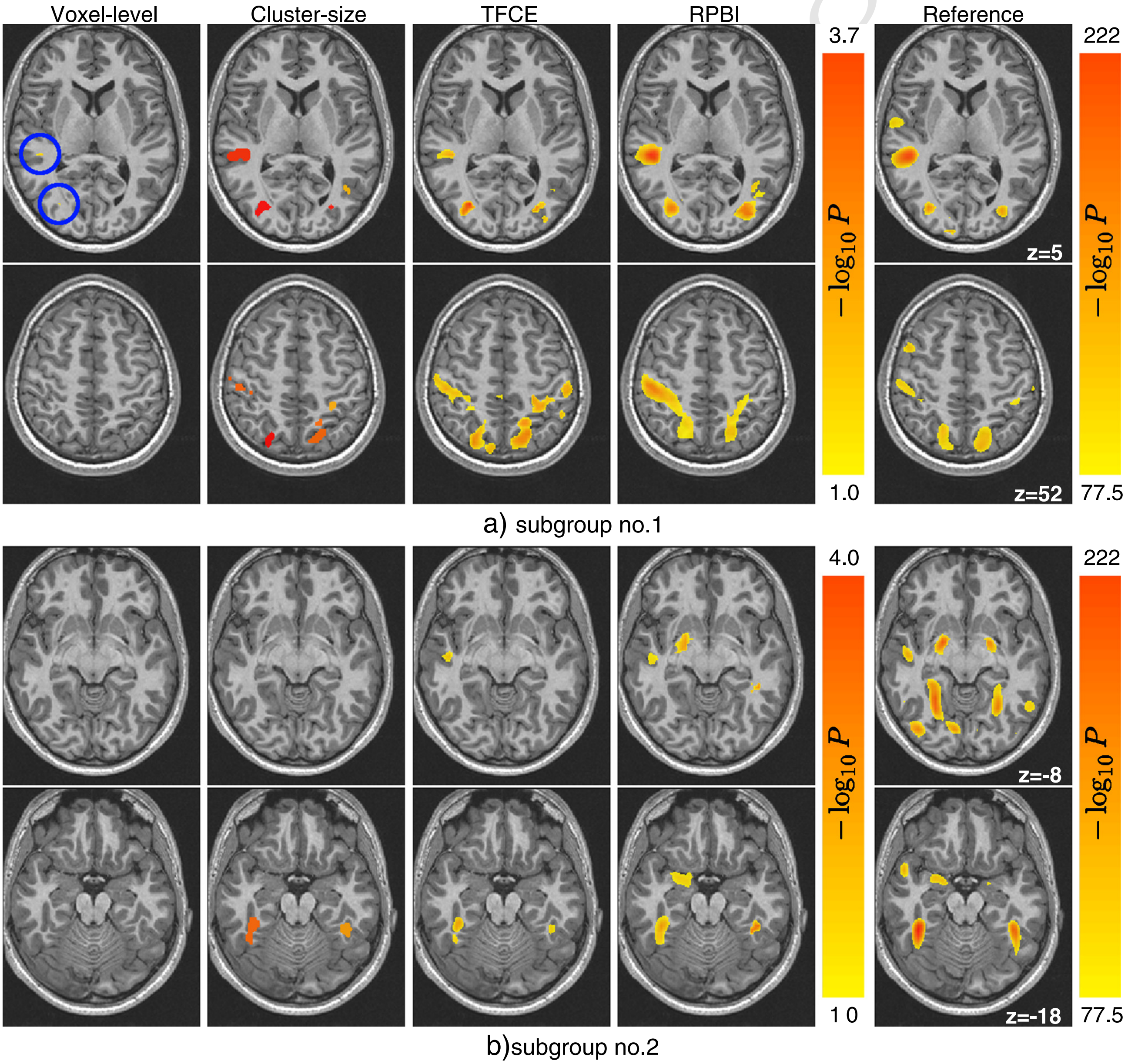


Fig. 6. Negative logp-value associated with a non-zero intercept test with confounds (handedness, site, sex), on a [*angry faces–control*] fMRI contrast from the *faces* protocol. The subgroups maps are thresholded at $-\log_{10}P > 1$ FWER corrected and the reference map at $-\log_{10}P > 77.5$ (i.e. 5% of the most active voxels). Small activation clusters are surrounded with a blue circle in order to make them visible.

Please cite this article as: Da Mota, B., et al., Randomized parcellation based inference, NeuroImage (2013), http://dx.doi.org/10.1016/j.neuroimage.2013.11.012

In general, the same activation peaks rise from the cluster-size, the TFCE and the RPBI maps (see Fig. 6). The TFCE slightly improves the results of cluster-size and provides voxel-level information. As can be seen in Fig. 6, the map returned by RPBI better matches the patterns of the reference map and is less scattered. Voxel-based group analysis clearly fails to detect some of the activation peaks.

Neuroimaging–genetic study

The SNP rs917478 yields the strongest correlation with the phenotypes and lies in an intronic region of ARVCF. The number of subjects in each genotype group is balanced: 523 homozygous with major allele, 663 heterozygous and 186 homozygous with minor allele. For RPBI, 31 voxels (resp. 81) are significantly associated with that SNP at $P < 0.05$ FWER corrected (resp. $P < 0.1$) in the left thalamus, a region involved in sensory-motor cognitive tasks. The association peak has a p-value of 0.016 FWER corrected. Cluster-size inference finds this effect but with a higher p-value ($P = 0.046$). Voxel-based inference does not find any significant effect. A significant association for rs917479 is reported only by RPBI; Fig. 8 shows that this SNP is in high linkage disequilibrium (LD) with rs917478 ($D' = 0.98$ and $R^2 = 0.96$). As shown in Fig. 8, those SNPs are also in LD with rs9306235 and rs9332377 in *COMT*, the targeted gene for this study. Fig. 8 shows the thresholded p-value maps obtained with RPBI with rs917478.

The ARVCF gene has already been found to be associated with intermediate brain phenotypes and neurocognitive error tests in a study about schizophrenia (Sim et al., 2012). We applied our method on this gene, for which we have 33 SNPs, and did not find any effect except from rs917478 and SNPs in LD with it.

Outlier detection

Fig. 9 illustrates the accuracy of RPB outlier detection as compared to standard outlier detection performed on data issued from a single parcellation. We present the rate of correct detections when 5% false detections are accepted. Since the experiment is conducted on 50 subsamples of n subjects, we present the results for various values of n ($n \in \{80,100,200,300,400\}$) with box-plots. For a large number of subjects (low-dimensional settings: $n < p$) RPB outlier detection performs slightly better than standard outlier detection, while in high-dimensional settings ($p > n$) it clearly outperforms the classic approach. Relative results are the same when allowing for any proportion of false detection comprising between 0% and 10%.

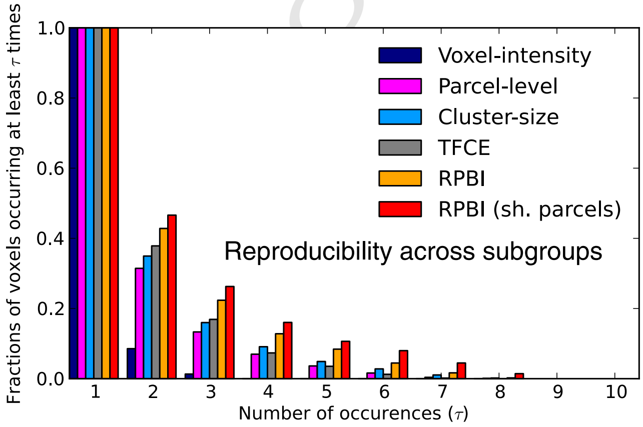


Fig. 7. Real fMRI data. Inverse cumulative histograms of the relative number of voxels that were reported as significant several times through the 10 subsamples ($P < 0.05$ FWER corrected), on a [*angry faces–control*] fMRI contrast from the *faces* protocol. Parcel-level inference yields results that are less reproducible than those of RPBI.

Discussion

In this work, we introduce a new method for statistical inference on brain images (RPBI) based on a randomized version of the parcellation model (Thirion et al., 2006) that is stabilized by a bootstrap procedure. In both simulation and real data experiments, RPBI shows better performance (sensitivity, recovery and reproducibility) than standard methods. The strength of this method is that the decision statistic takes into account the spatial structure of the data. Also, the randomization of the parcellations yields more reproducible results in view of between-subject variability and lowers the effect of inaccurate parcellation. Our experiments with simulated and real data show that the choice of the parcellations can greatly influence the success of RPBI. In this section, we discuss this choice. We also discuss some factors that can influence the method performance, such as images properties or tested features characteristics and computational aspects.

Brain parcellations

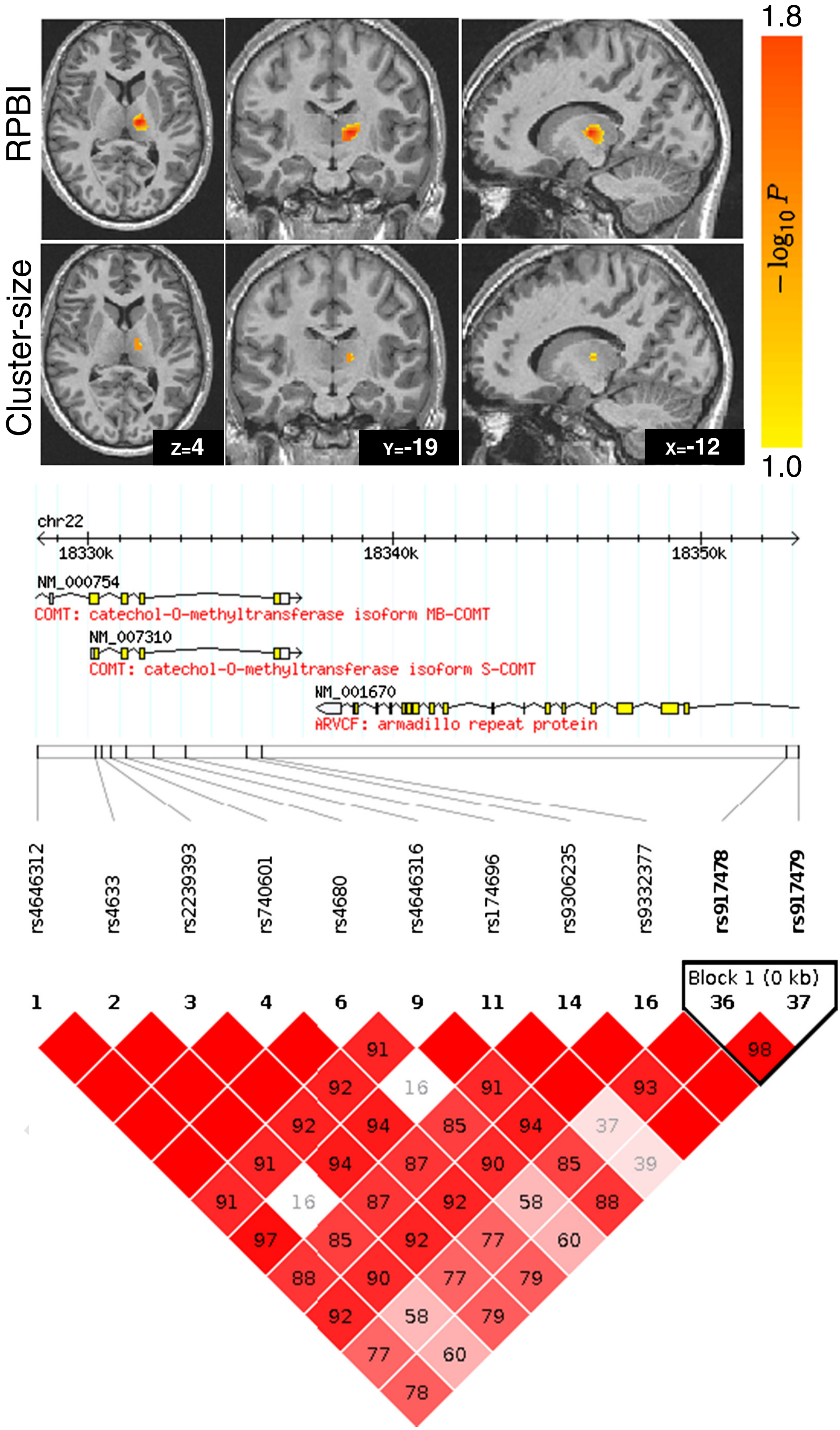
In our experiments, we used Ward clustering to build brain parcellations. The main advantage of this clustering algorithm is that it has the ability to take into account spatial pattern similarities between a set of input images, which acts as a spatial regularization. In addition, the Ward criteria is designed such that, taking the mean signal within each parcel as new features to describe one subject image gives the optimal data representation in terms of preserved information (for a fixed dimension corresponding to the number of parcels). Importantly, the variability of the parcellations is directly related to the variability and number of the images on which they are built. We determined empirically that using 1000 parcels is a good trade-off between accurate parcellations and dimension reduction. This choice leads to using an average of 50 voxels per parcel, which is a good order of magnitude to describe the activation clusters. Note that, this number of parcels is far from standard brain atlases with, at best, a few hundred ROIs, suggesting that atlases are not well-suited for such studies. Our first real data experiment demonstrates that it is beneficial that the parcellations reflect the group spatial activity pattern of the fMRI contrast under study: when the parcellations are built on another fMRI contrast or on random noise, the final performance of persistence analysis drops back to the level of state-of-the-art methods in terms of accuracy.

RPBI and images properties

Our first experiment shows that RPBI performance drops when images are smoothed a posteriori. Unlike voxel-intensity analysis, cluster-size analysis, TFCE and RPBI, which are spatial methods, suffer from data smoothing. In the presence of smooth noise, this experiment also shows that RPBI outperforms other methods. Our experiment on real data shows that RPBI can recover activations clusters of various size and shape, as visible on the effect maps reported in Fig. 6. Yet, the use of parcels clearly helps in focusing on activations with a spatial extent of the order of the average parcel size. Cluster-size group analysis also focuses more easily on some activations with a given size, according to internal parameters such as the cluster forming threshold or an optional data smoothing. TFCE is designed to address this issue and clearly enhances the results of the cluster-size inference.

Sensitivity and reproducibility

Usually, the sensitivity of a procedure is compared under a given control for false positives. Under this criterion, RPBI outperforms



Please cite this article as: Da Mota, B., et al., Randomized parcellation based inference, NeuroImage (2013), http://dx.doi.org/10.1016/j.neuroimage.2013.11.012

voxel-intensity, cluster-size analysis and TFCE (Fig. 5.a). By aggregating 100 × 1000 measurements, RPBI drastically reduces the multiple comparison problem and stabilizes parcel-based statistics. Neuroimaging studies are subject to a lack of reproducibility and using the most sensitive procedure does not guarantee the unveiling of reproducible results (Strother et al., 2002; Thirion et al., 2007). Experiments on real data show the gain in terms of reproducibility of RPBI compared to other methods when the subset of subjects changes (Fig. 7). RPBI with shared parcels has a better recovery and yields more reproducible results across various analysis settings.

Randomized parcellation can be applied to various neuroimaging tasks. However, sensitivity improvement is not straightforward and may depend on problem-specific settings. In particular, our experiment about outlier detection suggests that multivariate statistical algorithms require a more subtle use of randomized parcellation in order to get significant sensitivity improvement.

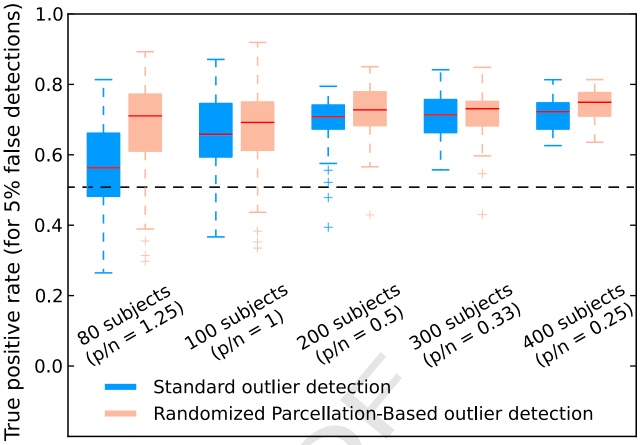


Fig. 9. Proportion of observations correctly tagged as outliers when 5% errors are accepted. Results are represented as boxes according to the number of subjects present in the subsamples in which we seek outliers. Chance level is given by the dashed black line. RPB outlier detection always outperforms standard outlier detection, although the difference between both is small and may not worth the implementation and computation costs. It is larger in the case where there are more features than subjects.

Computational aspects

Our goal here is not to provide an exhaustive study of the computational performance, but to report on our experience of the experiments performed. The procedure is separated into two distinct steps: (*i*) the generation of the 100 Ward *K*-parcellations and extraction of the signal means, then (*ii*) the statistical inference. The generation of parcellations is optional (parcellations can be replaced by precomputed ones), but Ward's hierarchical clustering algorithm is fast and this step takes only a few minutes on a desktop computer for 100 parcellations. The second step involves a permutation test. Our implementation fits a Massively Univariate Linear Model (Da Mota et al., 2012; Stein et al., 2010) in an optimized version adapted to permutation testing and our application. As a result, in our experiments with 20 subjects and 10,000 permutations, the statistical inference takes only 1 min × cores, i.e. 5 s on a 12-core computer. The total computation time thus amounts to a few minutes on a desktop computer and is limited by the construction of the parcellations. Asymptotically, the computation time increases only linearly with the number of subjects and the number of variables to test, which is a desirable property to scale to larger problems like neuroimaging–genetic studies.

Conclusion

RPBI is a general detection method based on a consensus across bootstrap estimates that can be applied to various neuroimaging problems such as group analyses or outlier detection. In our work, we use randomized parcellations to benefit from many ROI-based descriptions of our datasets that we construct with Ward's clustering. Simulations and real-data experiments show that RPBI is more sensitive and stable than state-of-the-art analysis methods. This is the case for various types of problems, including neuroimaging–genetic associations. We also demonstrate that the RPBI framework can be applied to outlier detection problem and improves detections accuracy.

Acknowledgments

This work was supported by Digiteo grants (HiDiNim project and ICoGeN project) and an ANR grant (ANR-10-BLAN-0128), as well the Microsoft INRIA joint center grant *A-brain*.

The data were acquired within the IMAGEN project, which received research funding from the E.U. Community's FP6, LSHM-CT-2007-037286. This manuscript reflects only the authors' views and the Community is not liable for any use that may be made of the information contained therein.

Appendix A. Formal description of Ward's clustering algorithm

Ward's clustering algorithm is a particular case of *hierarchical agglomerative clustering* (Johnson, 1967). Let $Y = \{y_1, \ldots, y_p\} \in \mathbb{R}^{n \times p}$ be a set of *n* fMRI volumes described by *p* voxels each. For two clusters of voxels *c* and *c′*, we define the distance:

$$\Delta(c,c') = \frac{|c||c'|}{|c|+|c'|}\left\|\langle Y\rangle_c - \langle Y\rangle_{c'}\right\|_2^2, \qquad \text{(A.1)}$$

where $\langle Y\rangle_c = \frac{1}{|c|}\sum_{j \in c} y^j$. For each partition $C = \{c_1, \ldots, c_k\}$ of the set of voxels *Y* (i.e. $\cup_{c \in C} = Y$ and $c_i \cap c_j = \emptyset \; \forall (c_i,c_j) \in C^2$), we note C^* the set of all pairs of clusters that share at least one neighboring voxel. Ward's clustering algorithm starts with an initial partition of *p* clusters $C = \{\{y_1\}, \ldots, \{y_p\}\}$ that correspond to one singleton cluster per voxel. At each iteration, we merge the two clusters c_i and c_j of C^* that minimize the distance Δ:

$$\left(c_i, c_j\right) = \underset{(c,c') \in C*}{\text{argmin}} \Delta(c,c'). \qquad \text{(A.2)}$$

The spatial constraint comes from the fact that we restrict the solution of the minimization criterion to C^*. When constructing a *K*-parcellation, the algorithm stops when card $(C) = K$.

Fig. A.10 shows some example parcellations, while Fig. A.11 shows the size and compactness of the parcels. In "Materials and methods", we use various Ward's clustering scheme that simply correspond to different choices for *Y*.

Fig. 8. Association study between 27 SNPs from the *COMT* gene (±20 kb) and fMRI contrast phenotypes. Family wise corrected p-values map (thresholded at $P < 0.1$) obtained with RPBI (top row) and cluster-size inference (bottom row) for rs917478, the SNP with the strongest reported effect. Linkage disequilibrium reported by HapMap for SNPs with $MAF > 0.05$ in a European population (CEU + TSI). For the sake of readability, other SNPs in ARVCF are hidden. Red boxes without values correspond to maximum linkage disequilibrium, i.e. $D' = 1$. The found SNPs (rs917478 and rs917479) are in high linkage disequilibrium with two SNPs at the end of *COMT*, namely rs9306235 and rs9332377.

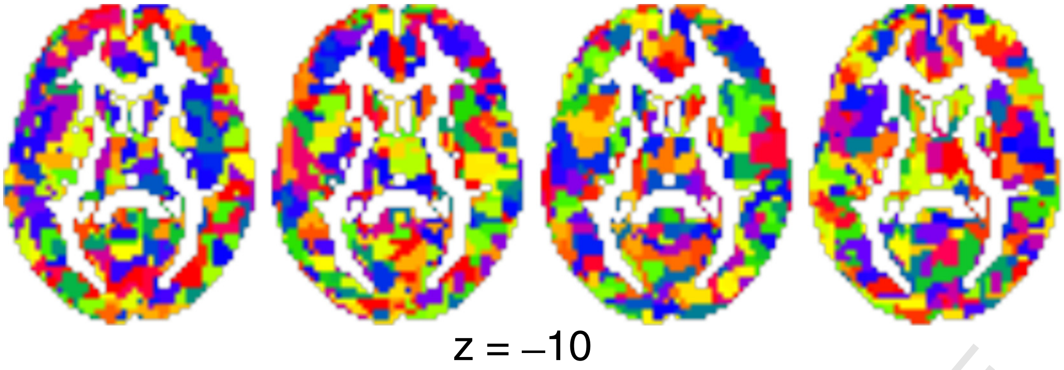


Fig. A.10. Example parcellations obtained with Ward's clustering algorithm. The [*angry faces–control*] fMRI contrast maps of 20 bootstrapped subjects were used.

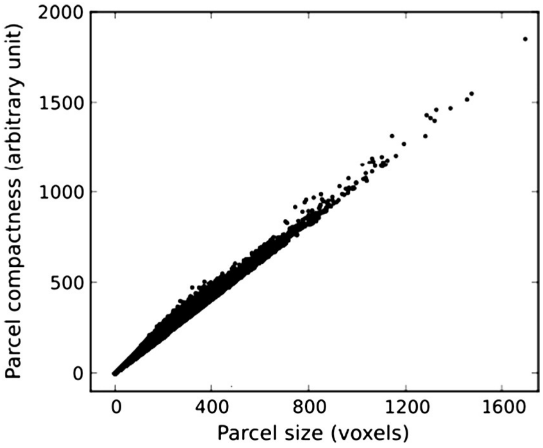


Fig. A.11. Size and compactness of the parcels obtained with Ward's clustering algorithm on fMRI contrast maps. For each parcel, the compactness is measured as the difference between a mask of the parcel and its 1-eroded image). One can observe a great variability in parcel size/compactness, which reflects the structure of the individual fMRI contrast maps.

Appendix B. Pivotality of the counting statistic

An important question is whether the counting statistic introduced in Eq. (1) is a valid statistic to detect activated voxels. One essential criterion for this is to check the pivotality, i.e. the convergence – under the null hypothesis – of the statistic distribution toward a law that is invariant under data distribution parameters. In the present case, the main deviation from pivotality could result from a distribution of (extreme) statistical values that depends on the parcel size: large parcels would represent fMRI signal averaged over larger domains, and thus would get typically lower values. This is indeed typically the case for the mean statistic (see Fig. B.12 (b)); however, we show for instance that the t statistic used in "Materials and methods" is very weakly influenced by the parcel size: we repeated the experiment described in "Materials and methods", i.e. computing the t statistic on parcels obtained by Ward's algorithm, based on 100 random batches of 20 subjects, after permutation by random sign swap. We tabulate the t distribution according to the parcel size by using 10 size bins. The result, shown in Fig. B.12 (a), is that the effect, if any, is not detectable by visual inspection.

To test more precisely the independence on the t distribution with respect to the parcel size, we tested the equality of the mean, median and variance of the size-specific distributions using the One-way (mean), Kruskal (median), Bartlett (variance), Levene (variance) and Fligner (variance) tests as implemented in the SciPy library.[5] All the tests are performed on the 10 bins jointly. We obtain the following p-values: One-way, $P = 0.36$; Kruskal, $P = 0.27$; Bartlett: $P = 0.95$; Levene: $P = 0.016$; Fligner: $P = 0.06$. This means that there is only a small effect on the variance, as reported by the Levene test, that is more sensitive than Fligner (which is non-parametric) and Bartlett, which assumes Gaussian distributions. However this effect is very small, and has no obvious consequence on the number of peak values of the statistic; in particular, we do not observe monotonic trends with size. Note that the small effect fades out when using larger number of subjects (here, only $n = 20$ subjects per groups were used). Finally, we did not find any significant correlation between the number of detections above a given threshold (using uncorrected p-values of 10^{-2}, 10^{-3}, 10^{-4}) and the parcel size.

In conclusion, the effect of parcel size is too small to jeopardize the usefulness of the counting statistic.

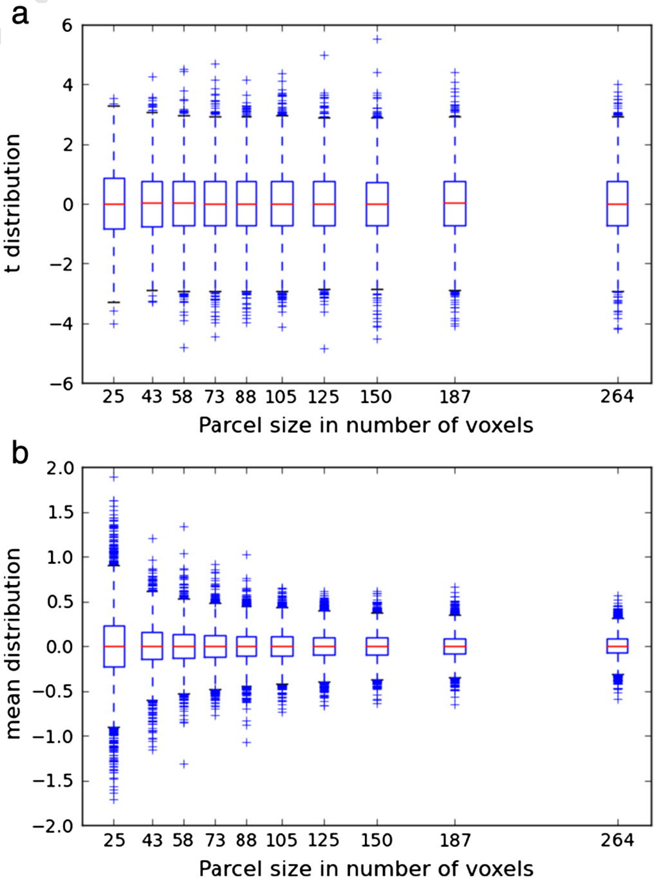


Fig. B.12. Impact of the parcel size on the distribution of the second-level one-sample t statistic (a) and of the mean value (b). While there is an obvious effect on the mean, there is no conspicuous effect on the t distribution.

[5] http://www.scipy.org/.

Appendix C. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.neuroimage.2013.11.012.

References

Alexander, D.H., Lange, K., 2011. Stability selection for genome-wide association. Genet. Epidemiol. 35, 722–728.

Bohland, J.W., Bokil, H., Allen, C.B., Mitra, P.P., 2009. The brain atlas concordance problem: quantitative comparison of anatomical parcellations. PLoS One 4, e7200.

Bühlmann, P., Rütimann, P., van de Geer, S., Zhang, C.H., 2012. Correlated Variables in Regression: Clustering and Sparse Estimation (ArXiv e-prints).

Collins, F.S., Brooks, L.D., Chakravarti, A., 1998. A DNA polymorphism discovery resource for research on human genetic variation. Genome Res. 8, 1229–1231.

Da Mota, B., Frouin, V., Duchesnay, E., Laguitton, S., Varoquaux, G., Poline, J.B., Thirion, B., 2012. A fast computational framework for genome-wide association studies with neuroimaging data. 20th International Conference on Computational Statistics.

Flandin, G., Penny, W.D., 2007. Bayesian fMRI data analysis with sparse spatial basis function priors. NeuroImage 34, 1108–1125.

Floch, E.L., Guillemot, V., Frouin, V., Pinel, P., Lalanne, C., Trinchera, L., Tenenhaus, A., Moreno, A., Zilbovicius, M., Bourgeron, T., Dehaene, S., Thirion, B., Poline, J.B., Duchesnay, E., 2012. Significant correlation between a set of genetic polymorphisms and a functional brain network revealed by feature selection and sparse partial least squares. NeuroImage 63, 11–24.

Frackowiak, R., Friston, K., Frith, C., Dolan, R., Price, C., Zeki, S., Ashburner, J., Penny, W., 2003. Human Brain Function, second ed. Academic Press.

Friston, K.J., Penny, W., 2003. Posterior probability maps and spms. NeuroImage 19, 1240–1249.

Friston, K.J., Worsley, K.J., Frackowiak, R.S.J., Mazziotta, J.C., Evans, A.C., 1993. 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.

Fritsch, V., Varoquaux, G., Thyreau, B., Poline, J.B., Thirion, B., 2012. Detecting outliers in high-dimensional neuroimaging datasets with robust covariance estimators. Med. Image Anal. 16, 1359–1370.

Gao, X., Becker, L.C., Becker, D.M., Starmer, J.D., Province, M.A., 2010. Avoiding the high Bonferroni penalty in genome-wide association studies. Genet. Epidemiol. 34, 100–105.

Ge, T., Feng, J., Hibar, D.P., Thompson, P.M., Nichols, T.E., 2012. Increasing power for voxel-wise genome-wide association studies: the random field theory, least square kernel machines and fast permutation procedures. NeuroImage 63, 858–873.

Grosbras, M.H., Paus, T., 2006. Brain networks involved in viewing angry hands or faces. Cereb. Cortex 16, 1087–1096.

Hanley, J., McNeil, B., 1982. The meaning and use of the area under a receiver operating (ROC) curve characteristic. Radiology 143, 29–36.

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

Hayasaka, S., Nichols, T.E., 2004. Combining voxel intensity and cluster extent with permutation test framework. NeuroImage 23, 54–63.

Hayasaka, S., Phan, K.L., Liberzon, I., Worsley, K.J., Nichols, T.E., 2004. Nonstationary cluster-size inference with random field and permutation methods. NeuroImage 22, 676–687.

Hibar, D.P., Stein, J.L., Kohannim, O., Jahanshad, N., Saykin, A.J., Shen, L., Kim, S., Pankratz, N., Foroud, T., Huentelman, M.J., Potkin, S.G., Jack, C.R., Weiner, M.W., Toga, A.W., Thompson, P.M., Initiative, A.D.N., 2011. Voxelwise gene-wide association study (vGeneWAS): multivariate gene-based association testing in 731 elderly subjects. NeuroImage 56, 1875–1891.

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

Hutton, C., Bork, A., Josephs, O., Deichmann, R., Ashburner, J., Turner, R., 2002. Image distortion correction in fMRI: a quantitative evaluation. NeuroImage 16, 217–240.

Johnson, S.C., 1967. Hierarchical clustering schemes. Psychometrika 32, 241–254.

Keller, M., Lavielle, M., Perrot, M., Roche, A., 2009. Anatomically informed Bayesian model selection for fMRI group data analysis. Med. Image Comput. Comput. Assist. Interv. 12, 450–457.

Knutson, B., Westdorp, A., Kaiser, E., Hommer, D., 2000. Fmri visualization of brain activity during a monetary incentive delay task. NeuroImage 12, 20–27.

Kohannim, O., Hibar, D.P., Stein, J.L., Jahanshad, N., C. R. J.J., Weiner, M.W., Toga, A.W., P.M.T., 2011. Boosting power to detect genetic associations in imaging using multi-locus, genome-wide scans and ridge regression. Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium on, pp. 1855–1859.

Logan, G.D., 1994. On the ability to inhibit thought and action: a users' guide to the stop signal paradigm. Psychol. Rev. 91, 295–327.

Meinshausen, N., Bühlmann, P., 2010. Stability selection. J. R. Stat. Soc. Ser. B Stat Methodol. 72, 417–473.

Michel, V., Gramfort, A., Varoquaux, G., Eger, E., Keribin, C., Thirion, B., 2012. A supervised clustering approach for fMRI-based inference of brain states. Pattern Recogn. 45, 2041–2049.

Moorhead, T.W.J., Job, D.E., Spencer, M.D., Whalley, H.C., Johnstone, E.C., Lawrie, S.M., 2005. Empirical comparison of maximal voxel and non-isotropic adjusted cluster extent results in a voxel-based morphometry study of comorbid learning disability with schizophrenia. NeuroImage 28, 544–552.

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

Nieto-Castanon, A., Ghosh, S.S., Tourville, J.A., Guenther, F.H., 2003. Region of interest based analysis of functional imaging data. NeuroImage 19, 1303–1316.

Ou, W., Wells, W.M., Golland, P., 2010. Combining spatial priors and anatomical information for fMRI detection. Med. Image Anal. 14, 318–331.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al., 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Petersson, K.M., Nichols, T.E., Poline, J.B., Holmes, A.P., 1999. Statistical limitations in functional neuroimaging. II. Signal detection and statistical inference. Philos. Trans. R. Soc. Lond. B Biol. Sci. 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. 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.

Puls, I., Mohr, J., Wrase, J., Vollstädt-Klein, S., Leménager, T., Vollmert, C., Rapp, M., Obermayer, K., Heinz, A., Smolka, M.N., 2009. A model comparison of comt effects on central processing of affective stimuli. NeuroImage 46, 683–691.

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.

Salimi-Khorshidi, G., Smith, S.M., Nichols, T.E., 2011. Adjusting the effect of nonstationarity in cluster-based and tfce inference. NeuroImage 54, 2006–2019.

Schumann, G., Loth, E., Banaschewski, T., Barbot, A., Barker, G., Bchel, C., Conrod, P.J., Dalley, J.W., Flor, H., Gallinat, J., Garavan, H., Heinz, A., Itterman, B., Lathrop, M., Mallik, C., Mann, K., Martinot, J.L., Paus, T., Poline, J.B., Robbins, T.W., Rietschel, M., Reed, L., Smolka, M., Spanagel, R., Speiser, C., Stephens, D.N., Strhle, A., Struve, M., IMAGEN consortium, 2010. The IMAGEN Study: reinforcement-related behaviour in normal brain function and psychopathology. Mol. Psychiatry 15, 1128–1139.

Sim, K., Chan, W.Y., Woon, P.S., Low, H.Q., Lim, L., Yang, G.L., Lee, J., Chong, S.A., Sitoh, Y.Y., Chan, Y.H., Liu, J., Tan, E.C., Williams, H., Nowinski, W.L., 2012. Arvcf genetic influences on neurocognitive and neuroanatomical intermediate phenotypes in Chinese patients with schizophrenia. J. Clin. Psychiatry 73, 320–326.

Smith, S.M., Nichols, T.E., 2009. Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. NeuroImage 44, 83–98.

Smolka, M., Bühler, M., Schumann, G., Klein, S., Hu, X., Moayer, M., Zimmer, A., Wrase, J., Flor, H., Mann, K., et al., 2007. Gene–gene effects on central processing of aversive stimuli. Mol. Psychiatry 12, 307–317.

Sol, A.F., Ngan, S.C., Sapiro, G., Hu, X., Lpez, A., 2001. Anisotropic 2-d and 3-d averaging of fMRI signals. IEEE Trans. Med. Imaging 20, 86–93.

Stein, J.L., Hua, X., Lee, S., Ho, A.J., Leow, A.D., Toga, A.W., Saykin, A.J., Shen, L., Foroud, T., Pankratz, N., Huentelman, M.J., Craig, D.W., Gerber, J.D., Allen, A.N., Corneveaux, J.J., Dechairo, B.M., Potkin, S.G., Weiner, M.W., Thompson, P., Alzheimer's Disease Neuroimaging Initiative, 2010. Voxelwise genome-wide association study (vGWAS). NeuroImage 53, 1160–1174.

Strother, S.C., Anderson, J., Hansen, L.K., Kjems, U., Kustra, R., Sidtis, J., Frutiger, S., Muley, S., LaConte, S., Rottenberg, D., 2002. The quantitative evaluation of functional neuroimaging experiments: the npairs data analysis framework. NeuroImage 15, 747–771.

Thirion, B., Flandin, G., Pinel, P., Roche, A., Ciuciu, P., Poline, J.B., 2006. Dealing with the shortcomings of spatial normalization: multi-subject parcellation of fMRI datasets. Hum. Brain Mapp. 27, 678–693.

Thirion, B., Pinel, P., Mriaux, S., Roche, A., Dehaene, S., Poline, J.B., 2007. Analysis of a large fMRI cohort: statistical and methodological issues for group analyses. NeuroImage 35, 105–120.

Thyreau, B., Schwartz, Y., Thirion, B., Frouin, V., Loth, E., Vollstdt-Klein, S., Paus, T., Artiges, E., Conrod, P.J., Schumann, G., Whelan, R., Poline, J.B., Consortium I.M.A.G.E.N., 2012. Very large fMRI study using the IMAGEN database: sensitivity-specificity and population effect modeling in relation to the underlying anatomy. NeuroImage 61, 295–303.

Varoquaux, G., Gramfort, A., Thirion, B., 2012. Small-sample brain mapping: sparse recovery on spatially correlated designs with randomization and clustering. In: John, L., Joelle, P. (Eds.), International Conference on Machine Learning. Andrew McCallum, Edimbourg, United Kingdom.

Ville, D.V.D., Blu, T., Unser, M., 2004. Integrated wavelet processing and spatial statistical testing of fMRI data. NeuroImage 23, 1472–1485.

Vounou, M., Nichols, T.E., Montana, G., Initiative, A.D.N., 2010. Discovering genetic associations with high-dimensional neuroimaging phenotypes: a sparse reduced-rank regression approach. NeuroImage 53, 1147–1159.

Ward, J., 1963. Hierarchical grouping to optimize an objective function. J. Am. Stat. Assoc. 58, 236–244.

Westfall, P.H., Troendle, J.F., 2008. Multiple testing with minimal assumptions. Biom. J. 50, 745–755.

Worsley, K.J., Evans, A.C., Marrett, S., Neelin, P., 1992. A 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., Evans, A.C., 1996a. Searching scale space for activation in PET images. Hum. Brain Mapp. 4, 74–90.

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

Wu, D.H., Lewin, J.S., Duerk, J.L., 1997. Inadequacy of motion correction algorithms in functional MRI: role of susceptibility-induced artifacts. J. Magn. Reson. Imaging 7, 365–370.

Please cite this article as: Da Mota, B., et al., Randomized parcellation based inference, NeuroImage (2013), http://dx.doi.org/10.1016/j.neuroimage.2013.11.012