# Enhancing the Reproducibility of Group Analysis with Randomized Brain Parcellations

Benoit Da Mota[1,2], Virgile Fritsch[1,2], Gaël Varoquaux[1,2], Vincent Frouin[2],
Jean-Baptiste Poline[2,3], and Bertrand Thirion[1,2]

[1] Parietal Team, INRIA Saclay-Île-de-France, Saclay, France
benoit.da_mota@inria.fr
[2] CEA, DSV, I$^2$BM, Neurospin bât 145, 91191 Gif-Sur-Yvette, France
[3] Henry H. Wheeler Jr. Brain Imaging Center, University of California at Berkeley

**Abstract.** Neuroimaging group analyses are used to compare the inter-subject variability observed in brain organization with behavioural 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. A new approach is introduced 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 syntetic and real data, this approach shows higher sensitivity, better recovery and higher reproducibility than standard methods and succeeds in detecting a significant association in an imaging-genetic study between a genetic variant next to the COMT gene and a region in the left thalamus on a functional Magnetic Resonance Imaging contrast.

**Keywords:** neuroimaging, group analysis, parcellation, reproducibility.

## 1 Introduction

Statistical analyses of functional brain images recorded on a group of subjects are used to infer some regional characteristics of brain cognitive organization and to assess the correlation of their variability with other information. For instance, massively univariate two-sample t-tests are used to compare groups of subjects. The major difficulty with such studies lies in the inter-subject variability of brain shape, regional functional organization and vasculature. The standard analytic approach is to register and normalize the data in a common reference space, yet a perfect voxel-to-voxel correspondence cannot be achieved. As many parameters settings or approaches exist to alleviate the issue, practicioners as well as methodologists tend to choose the one that maximizes the sensitivity under a given control for false detections. While the choice of a significance level is arbitrary, the sensitivity of the test for a given control of the specificity is indeed informative on the appropriateness of a model.

The most straightforward technique consists in smoothing the data to increase the overlap between subject-specific activated regions [14,9]. A popular approach

K. Mori et al. (Eds.): MICCAI 2013, Part II, LNCS 8150, pp. 591–598, 2013.

consists in working with pre-defined Regions of Interest (ROIs), but several difficulties arise for such an approach, the main one being that there is no widely accepted standard for partitioning the brain, and especially the neocortex [1]. Several attempts have been made to use spatial information to improve statistical inference, like with Markov Random Fields [6], wavelets decomposition [11], and the widely used cluster-size inference [3]. Amongst those, the parcellation model in [10] has several advantages: *(i)* it is a simple and easily interpretable method, *(ii)* it lowers the impact of the multiple comparisons problem by reducing the number of descriptors, and *(iii)* the parcellation algorithm yields parcels that are adapted to the local smoothness. But parcellations, when considered as spatial functions, are very unstable and highly depend on the data used to construct them. In general, a parcellation defined for a given dataset might not be a good model in a slightly different context, and it can generalize poorly to other subjects. Thus, the weakness of the parcellation-based approach is its dependence to a possibly imperfect parcellation that fails to detect effects in poorly segmented regions.

*The Randomized Parcellation Approach.* We propose to robustify parcel-based approaches by using several random parcellations [12] and aggregate the corresponding statistical decisions. Formally, this can be understood as handling the parcellation as a hidden variable, which needs to be integrated out in order to obtain the posterior distribution of statistics values. The final decision is taken with regard to the stability of the detection of a voxel [4] across parcellations, compared to a null distribution obtained by a permutation test. We evaluate this new approach on simulations, then on real data for the random effect analysis problem. Then, we illustrate the interest of the approach for neuroimaging-genetic studies on a candidate gene (COMT) which is widely investigated in the context of brain diseases.

## 2 Methods and Materials

*Parcellation and Ward algorithm.* In functional neuroimaging, brain atlases are mainly used to provide low dimensional representations of the data by considering signal averages within groups of neighboring voxels. Let $V$ be the set of all brain voxels, a $K-$parcellation $P$ is a partition of $V$, $P = \{P^{(i)}, i \in \{1 \dots K\}\}$, with $P^{(i)} \subset V, \forall i \in \{1 \dots K\}$; $P^{(i)} \cap P^{(j)} = \emptyset, \forall i, j, i \neq j$ and $\cup_{i=1}^{K} P^{(i)} = V$. Following [5,12], we apply spatially-constrained Ward hierarchical clustering [13] to different subgroups of subjects in order to build slightly variable $K$-parcellations. The clustering step may involve the same data as the subsequent analysis, as data-driven brain parcellations better take into account the unknown data structure. Ward's algorithm has several advantages : *(i)* it captures well local correlations to form small and connected clusters, *(ii)* efficient implementations exist, and *(iii)* obtained parcellations are invariant by permutation and sign swap of the data.

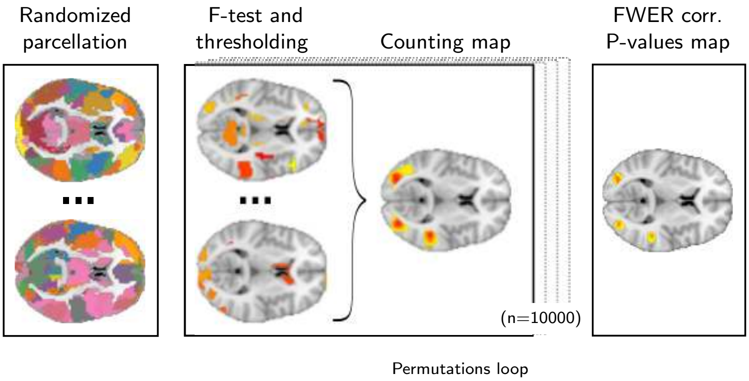


**Fig. 1.** Overview of the randomized parcellation based inference framework on an example with few parcels

*Randomized parcellation based inference (RPBI)* consists in performing several standard analyses based on different parcellations and aggregating the corresponding statistical decisions. In practice, we perform these decisions from F-tests at a fixed threshold $t$, that we set so that it ensures a Bonferroni-corrected control at $p < 0.1$. Let $\mathcal{P}$ be a finite set of parcellations. Given a voxel $v$ and a parcellation $P$, the parcel-based thresholding function $\theta_t$ 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 $\Phi_P : V \rightarrow P$ is a mapping function that associates each voxel with a parcel from the parcellation $P$ ($\forall v \in P^{(i)}$, $\Phi_P(v) = P^{(i)}$), and $F$ returns the $F$-statistic associated with a given parcels average signal for a pre-defined test. 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 $\mathcal{P}$. 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 test, we switch labels or we swap signs. We obtain family-wise error control by tabulating the maximal value across voxels in the permutation procedure. As a result, we get a voxelwise p-values map similar to a standard group analysis map (see Figure 1).

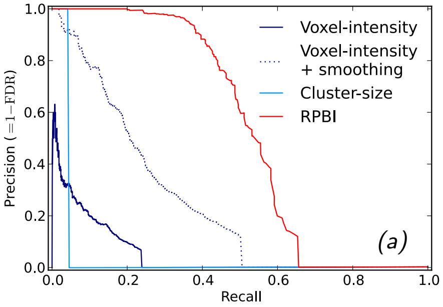
*Dataset.* We use a dataset from a European multicentric study involving adolescents [7]. It contains a large functional neuroimaging database with fMRI associated with 99 different contrast images for 4 protocols. Standard preprocessings

SPM8 software and its default parameters; functional images were resampled at 3mm resolution. All images were warped in the MNI152 coordinate space. 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 procedure and temporally auto-regressive noise model. An additional Gaussian smoothing at 5mm-FWHM was performed. The estimation of the model parameters was carried out using the SPM8 software.

## 3 Experiments

*Random Effect Analysis on Simulation.* We generate a set of 1000 simulated fMRI contrast images as volumes of shape $40 \times 40 \times 40$ voxels. Each contrast image contains a simulated $4 \times 4 \times 4$ activation at a precise location, with a spatial jitter following a three-dimensional $\mathcal{N}(\mathbf{0}, \boldsymbol{I}_3)$ distribution. 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 $\mathcal{N}(0, 1)$ distribution, Gaussian-smoothed at $\sigma_{\text{noise}}$ isotropic and normalized by its global empirical standard deviation. After superimposing noise and signal image, we optionally smooth the images at $\sigma_{\text{post}} = 2.12$ voxels isotropic, corresponding to a 5 voxels Full Width at Half Maximum (FWHM). We define the ground truth as the $p < 0.05$ (corrected) thresholded p-value map resulting from a voxel-level non-zero intercept test on the full set of 1000 images. Ten subsets of 20 randomly drawn images are then used to conduct the same analysis with three methods: *(i)* voxel-intensity group analysis, *(ii)* cluster-size group analysis, and *(iii)* RPBI. Each time, RPBI was conducted with one hundred 1000-parcellations built from a bootstrapped selection of the 20 images involved. The goal of this experiment is to compare the sensitivity and the recovery of the methods. First, the sensitivity is assessed by counting the number of detections at a fixed level of specificity. Then to estimate the recovery, precision-recall curves are constructed by reporting the proportion of true positives in the detections (precision) for different levels of recovery of the ground truth (recall).

*Random Effect Analysis on Real Data.* In this experiment, we work with an fMRI contrast with 1567 available images after removal of the subjects with too many missing data and/or bad/missing covariates. We test each voxel for a zero mean across the 1567 subjects with an OLS regression, including handedness and gender as covariables, yielding a reference voxelwise p-values map that we consider as the ground truth. Our objective is to retrieve that reference activity pattern considering only subsamples of 20 randomly drawn subjects and compare the performance of several methods in this problem. We perform our experiment on 10 different subsamples and we use the same analysis methods as the previous experiment, plus: *iv)* RPBI (other contrast) with parcellations built on images from an independent fMRI protocol, and *v)* RPBI (fully random) with parcellations built on smoothed Gaussian noise. These two methods aim to show how data-driven parcellations are important to RPBI.



| $\sigma_{\text{noise}}$ | 0 | | 1 | | 2 | |
|---|---|---|---|---|---|---|
| post-smoothing | ✗ | ✓ | ✗ | ✓ | ✗ | ✓ |
| Voxel-intensity | 3 | **10** | 3 | **4** | 3 | **2** |
| Cluster-size | 9 | **10** | 6 | 3 | 1 | 0 |
| RPBI | **10** | **10** | **7** | **4** | **4** | **2** |

*(b)*

**Fig. 2.** *(a)* Precision-recall curves for various analysis methods across 10 random subsamples containing 20 subjects for the simulation with $\sigma_{\text{noise}} = 1$. *(b)* Number of detections (over 10 simulations) of a significant effect for each analysis method.

*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 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. Subjects with too many missing voxels in the brain mask or with bad task performance were discarded. Regarding genetic variants, we keep only 27 SNPs in the COMT gene (±20kb) that pass standard filters (Minor Allele Frequency $< 0.05$, Hardy-Weinberg Equilibrium $p < 0.001$, missing rate per SNP $< 0.05$). 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. Our experiment involves 1,372 subjects. For each of the 27 SNPs, we perform a massively univariate voxelwise analysis with the algorithm presented in [2], including cluster-size analysis, and RPBI.

## 4 Results

*Random Effect Analysis on Simulations.* Figure 2b gives the number of times that a significant effect was reported according to the different methods. The specificity of the detections is controlled at 5% (corrected for multiple comparisons) for all the methods. Thus the results indicate that RPBI is more sensitive since it always achieves more detections. Voxel-intensity group analysis is the only method that benefits from data smoothing, while spatial methods lose sensitivity when the images are smoothed. Figure 2a shows that detections made by spatial methods (cluster-size group analysis and RPBI) does not come with wrongly reported effects in voxels close to the actual effect location. That 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 precision-recall compromise as its precision-recall curve dominates.

*Random Effects Analysis on Real Data.* We build precision-recall curves by thresholding the reference map at several arbitrary levels so that we can compute as many precision and recall scores for the thresholded ($p < 0.05$ corrected) maps corresponding to each subgroup and each analysis method (Figure 3a). RPBI outperforms other methods when we use parcellations that have been built on the contrast under study. Voxel-intensity group analysis yields poor performance while the other methods always have a much better recall at a given precision. RPBI (other contrast) or RPBI (fully random) yield poorer recovery although they are both based on the randomized parcellation scheme. This demonstrates that the choice of parcellations plays an important role in the success of RPBI. We also estimate the reproducibility of the methods findings by counting how many times each voxel is associated with a significant effect across subgroups (Figure 3b). RPBI results are the most stable. As Figure 4 illustrates, the map returned by RPBI better matches the patterns of the reference map.

*Neuroimaging-genetic Study.* The SNP rs917478 yields the strongest correlation with the phenotypes and lies in an intronic region of the ARVCF gene, that has already been found to be associated with intermediate brain phenotypes and neurocognitive error tests [8]. This SNP is also in high linkage disequilibrium (LD) with rs9332377 in COMT. The number of subjects in each genotype group is balanced. For RPBI, 31 voxels in the left thalamus, a region involved in sensory-motor cognitive tasks, are significantly associated with that SNP at $p < 0.05$ corrected. The cluster-size inference also finds this effect but with a higher p-value. A significant associations for rs917479, an SNP in high LD with rs917478, is only reported by RPBI. Figure 5 shows the thresholded p-values maps obtained with RPBI on SNP rs917478.

## 5 Conclusion

RPBI is a general decision method based on voting that can be applied to various image-based statistical inference problems such as group analyzes in neuroimaging. Our approach is clearly related to anisotropic smoothing [9], in the sense that obtained parcels are not spherical and by computing the mean of the voxels within a parcel, certain directions are preferred. Unlike smoothing or spatial modeling methods, our statistical inference embeds the spatial modeling 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, which is a benefit of the stabilization inherent to the aggregating procedure. Simulations and real-data experiments shows that RPBI has better sensitivity, recovery and stability than state-of-the-art analysis methods.

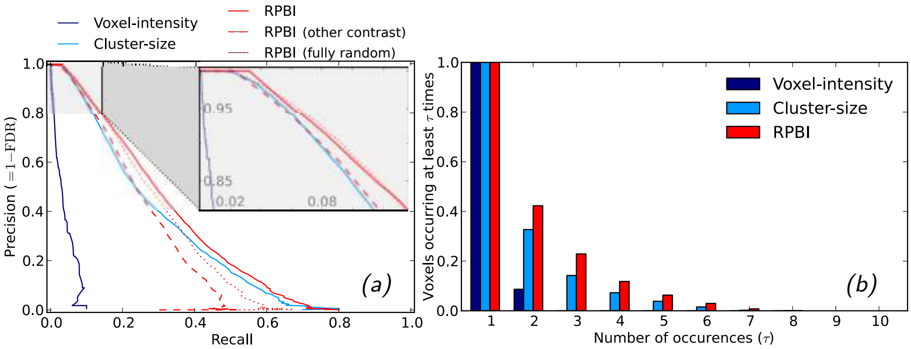


**Fig. 3.** *(a)* Precision-recall curves for various methods across 10 random subsamples containing 20 subjects. *(b)* Inverse cumulative histograms of the relative number of voxels that were reported as significant several times through the 10 subsamples.

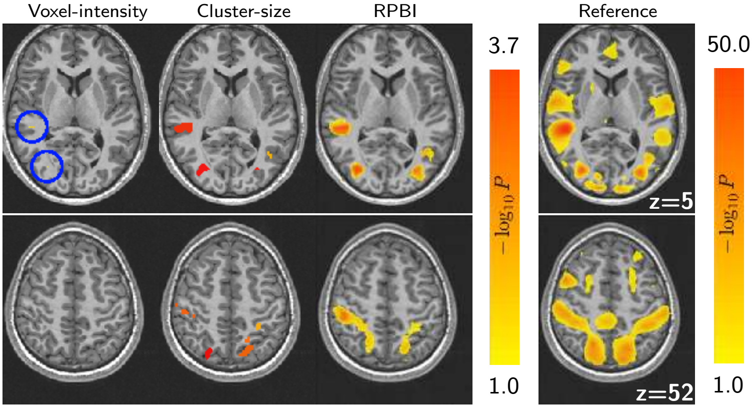


**Fig. 4.** Results of the one-sample test in one subgroup of subjects: Negative logarithm of the family-wise corrected p-values associated with a non-zero intercept test with confounds (handedness, site, gender)

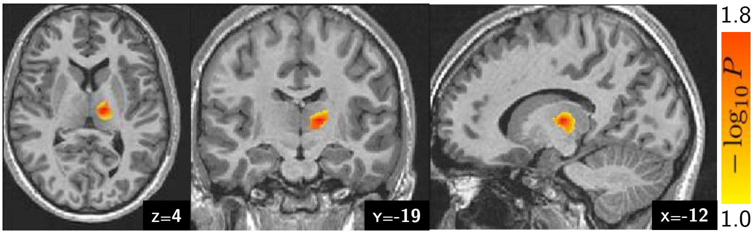


**Fig. 5.** Map obtained with RPBI in the neuroimaging genetic study: Negative logarithm of the family-wise corrected p-values for rs917478, the SNP with the strongest reported effect

**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. JBP was partly funded by the Imagen project, which receives research funding from the E.U. Community's FP6, LSHM-CT-2007-037286. This manuscript reflects only the author's views and the Community is not liable for any use that may be made of the information contained therein.

## References

1. Bohland, J.W., Bokil, H., Allen, C., Mitra, P.: The brain atlas concordance problem: quantitative comparison of anatomical parcellations. PLoS One 4(9) (2009)
2. Da Mota, B., Frouin, V., Duchesnay, E., Laguitton, S., Varoquaux, G., Poline, J.B., Thirion, B.: A fast computational framework for genome-wide association studies with neuroimaging data. In: 20th Int. Conf. on Comp. Stat. (2012)
3. Hayasaka, S., Phan, K.L., Liberzon, I., Worsley, K.J., Nichols, T.E.: Nonstationary cluster-size inference with random field and permutation methods. Neuroimage 22(2), 676–687 (2004)
4. Meinshausen, N., Bühlmann, P.: Stability selection. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72(4), 417–473 (2010)
5. Michel, V., Gramfort, A., Varoquaux, G., Eger, E., Keribin, C., Thirion, B.: A supervised clustering approach for fMRI-based inference of brain states. Pattern Recogn. 45(6), 2041–2049 (2012)
6. Ou, W., Wells, W.M., Golland, P.: Combining spatial priors and anatomical information for fMRI detection. Med. Image Anal. 14(3), 318–331 (2010)
7. Schumann, G., et al.: The imagen study: reinforcement-related behaviour in normal brain function and psychopathology. Mol. Psychiatry 15(12), 1128–1139 (2010)
8. Sim, K., et al.: Arvcf genetic influences on neurocognitive and neuroanatomical intermediate phenotypes in chinese patients with schizophrenia. J. Clin. Psychiatry 73(3), 320–326 (2012)
9. Solé, A.F., Ngan, S.C., Sapiro, G., Hu, X., López, A.: Anisotropic 2-d and 3-d averaging of fMRI signals. IEEE Trans. Med. Imaging 20(2), 86–93 (2001)
10. Thirion, B., Flandin, G., Pinel, P., Roche, A., Ciuciu, P., Poline, J.B.: Dealing with the shortcomings of spatial normalization: multi-subject parcellation of fMRI datasets. Hum. Brain Mapp. 27(8), 678–693 (2006)
11. Van De Ville, D., Blu, T., Unser, M.: Integrated wavelet processing and spatial statistical testing of fMRI data. Neuroimage 23(4), 1472–1485 (2004)
12. Varoquaux, G., Gramfort, A., Thirion, B.: Small-sample brain mapping: sparse recovery on spatially correlated designs with randomization and clustering. In: John, L., Joelle, P. (eds.) Int Conf on Machine Learning (2012)
13. Ward, J.: Hierarchical grouping to optimize an objective function. J. Am. Stat. Assoc. 58(301), 236–244 (1963)
14. Worsley, K.J., Marrett, S., Neelin, P., Evans, A.C.: Searching scale space for activation in PET images. Hum. Brain Mapp. 4(1), 74–90 (1996)