

## Which fMRI clustering gives good brain parcellations?

Bertrand Thirion, Gaël Varoquaux, Elvis Dohmatob, Jean-Baptiste Poline

► **To cite this version:**

Bertrand Thirion, Gaël Varoquaux, Elvis Dohmatob, Jean-Baptiste Poline. Which fMRI clustering gives good brain parcellations?. *Frontiers in Neuroscience*, Frontiers, 2014, 8 (167), pp.13. <<http://journal.frontiersin.org/Journal/10.3389/fnins.2014.00167/pdf>>. <10.3389/fnins.2014.00167>. <hal-01015172>

**HAL Id: hal-01015172**

**<https://hal.inria.fr/hal-01015172>**

Submitted on 25 Jun 2014

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



1

# Which fMRI clustering gives good brain parcellations?

Bertrand Thirion<sup>1,2,\*</sup>, Gaël Varoquaux<sup>1,2</sup>, Elvis Dohmatob<sup>1,2</sup> and Jean-Baptiste Poline<sup>2,3</sup>

<sup>1</sup> *Parietal project-team, INRIA, 1 rue Honoré d'Estienne d'Orves, 91120 Palaiseau, France*

<sup>2</sup> *CEA, DSV, I2BM, Neurospin, Bâtiment 145, 91191 Gif-sur-Yvette, France*

<sup>3</sup> *Henry H. Wheeler Jr. Brain Imaging Center, University of California at Berkeley, CA, US*

Correspondence\*:

Bertrand Thirion

Parietal project-team, INRIA, 1 rue Honoré d'Estienne d'Orves, 91120 Palaiseau  
91893 Orsay, France, bertrand.thirion@inria.fr

## 2 ABSTRACT

3 Analysis and interpretation of neuroimaging data often require one to divide the brain into a  
4 number of regions, or parcels, with homogeneous characteristics, be these regions defined in  
5 the brain volume or on on the cortical surface. While predefined brain atlases do not adapt  
6 to the signal in the individual subjects images, parcellation approaches use brain activity (e.g.  
7 found in some functional contrasts of interest) and clustering techniques to define regions with  
8 some degree of signal homogeneity. In this work, we address the question of which clustering  
9 technique is appropriate and how to optimize the corresponding model. We use two principled  
10 criteria: goodness of fit (accuracy), and reproducibility of the parcellation across bootstrap  
11 samples. We study these criteria on both simulated and two task-based functional Magnetic  
12 Resonance Imaging datasets for the Ward, spectral and K-means clustering algorithms. We  
13 show that in general Ward's clustering performs better than alternative methods with regard to  
14 reproducibility and accuracy and that the two criteria diverge regarding the preferred models  
15 (reproducibility leading to more conservative solutions), thus deferring the practical decision to  
16 a higher level alternative, namely the choice of a trade-off between accuracy and stability.

17 **Keywords:** Functional neuroimaging, Brain atlas, clustering, Model selection, cross-validation, group studies

## 1 INTRODUCTION

18 Brain parcellations divide the brain's spatial domain into a set of non-overlapping regions or modules that  
19 show some homogeneity with respect to information provided by one or several image modalities, such  
20 as cyto-architecture, anatomical connectivity, functional connectivity, or task-related activation. Brain  
21 parcellations are therefore often derived from specific clustering algorithms applied to brain images. Such  
22 approaches are generally useful because the voxel sampling grid of the reference space, e.g. the MNI  
23 template, is most often at a higher resolution than the brain structures of interest, or at a scale that is too  
24 fine for the problem under investigation, yielding an excessive number of brain locations and correlated  
25 data. In other words, the structures of interest are rarely at the level of a specific voxel, but at the level  
26 of many voxels constituting a (possibly small) brain region. Three strategies are commonly used to study

27 function beyond the voxel description: *i*) the use of anatomical or functional regions of interest (ROIs), *ii*)  
28 the use of a brain atlas, or *iii*) the use of data-driven parcellations.

29 *ROI-based analysis* has been advocated as a way to focus data analysis on some structures of interest  
30 and consists in building a summary of the signal in a predefined region (Nieto-Castanon et al., 2003). The  
31 choice of the region(s) can be based on prior experiments (e.g. Saxe et al. (2006)). Note that in extreme  
32 cases, the region can reduce to a single voxel, one reported in previous literature as the peak coordinate  
33 of a contrast image<sup>1</sup>. The obvious limitation of ROI-based analysis is that the signal present outside the  
34 region under consideration is ignored a priori; as a consequence, the results depend heavily on the choice  
35 of this ROI, which may not fit well the new data. In the hypothesis testing framework, the smaller number  
36 of tests performed may however increase the power of the analysis.

37 *Brain atlases* come into play to provide a set of ROIs that cover the brain volume (among many others  
38 see e.g. Mazziotta et al. (2001); Tzourio-Mazoyer et al. (2002); Shattuck et al. (2008)). An atlas  
39 generally accounts for a certain state of the knowledge of the brain structures (anatomically, functionally  
40 or based on connectivity), from which well-defined entities can be distinguished. In other words, an atlas  
41 represents a certain *labeling* of brain structures. Often this labeling is linked to an ontology representing  
42 the current knowledge (Cieslik et al., 2012; Eickhoff et al., 2011). In spite of their obvious usefulness,  
43 existing atlases are limited in two regards: *i*) There exist currently many different atlases, but they are  
44 mutually inconsistent (Bohland et al., 2009); *ii*) A given atlas may not fit the data well. Atlas misfits  
45 can be due to image characteristics and processing strategies that have evolved since the atlas creation, or  
46 because a given study deals with a population that is not well represented by the subjects used to construct  
47 the atlas, or because the information of interest is simply not mapped properly in the given atlas. Atlas  
48 misfit is often pronounced with regards to mapping brain function; for instance most anatomical atlases  
49 have large frontal brain regions that many researchers would rather divide into smaller ones with more  
50 precise functional roles.

51 Unlike brain atlases, also used to define regions of interest, brain parcellations are data-driven. They do  
52 not reflect a pre-defined ontology of brain structures –known anatomical names and concepts–, but they  
53 may much better represent the measurements or features of interest, i.e. they provide a better model of the  
54 signal (Flandin et al., 2002; Simon et al., 2004; Thirion et al., 2006; Lashkari et al., 2010, 2012). The  
55 (anatomical) labeling of these parcels can then be performed with the most appropriate atlas.

56 While functional parcellations can be used in different contexts, we focus here on finding a well-suited  
57 model to obtain local averages of the signal for group studies. These parcel averages can be thought of as  
58 a data reduction adapted to various tasks, such as the estimation of brain-level connectivity models (see  
59 e.g. Craddock et al. (2012); Yeo et al. (2011)), of physiological parameters (Chari et al., 2012), for  
60 group analysis (Thirion et al., 2006), the comparison of multiple modalities (Eickhoff et al., 2011) or  
61 in multivariate models (Michel et al., 2012). This is especially useful for the analysis of large cohorts  
62 of subjects, because this step can reduce the data dimensionality by several orders of magnitude while  
63 retaining most of the information of interest. We will show in this paper that common brain atlases, merely  
64 reflecting sulco-gyral anatomy, are not detailed enough to yield *adequate* models of the (functional) data.

65 Data-driven parcellations can be derived from various image modalities reflecting different  
66 neurobiological information, for instance T1 images with anatomical information, such as gyro-  
67 sulcal anatomy (Desikan et al., 2006; Klein and Tourville, 2012), post-mortem in vitro receptor  
68 autoradiography for cyto-architecture (Eickhoff et al., 2008; Fischl et al., 2008), anatomical connectivity  
69 (Roca et al., 2010) with diffusion imaging, or functional features with BOLD data. In this work, we focus  
70 on the latter, that we call *functional parcellations*. These parcellations are currently derived either from  
71 resting-state functional Magnetic Resonance Images (rs-fMRIs) (Yeo et al., 2011; Blumensath et al.,  
72 2012; Craddock et al., 2012; Kahnt et al., 2012; Wig et al., 2013), from activation data (Flandin  
73 et al., 2002; Lashkari et al., 2010, 2012; Michel et al., 2012), or from meta-analyses (Eickhoff et al.,  
74 2011). To investigate which parcellations are most appropriate, we restrict our work to activation data that

---

<sup>1</sup> Often, a small sphere will be drawn around this position to average signals locally.

75 have more tractable probabilistic models than resting-state data. We also omit edge-based parcellation  
76 methods, such as those described in **Wig et al. (2013)** or **Cohen et al. (2008)**: while these are certainly  
77 useful to segment the cortical surface by revealing abrupt changes in the functional connectivity patterns  
78 when crossing region boundaries, they do not lend themselves to model selection due to the absence  
79 of a probabilistic framework. This family of approaches is certainly an interesting competitor for future  
80 analyzes of functional parcellations performed on the cortical surface.

81 The most popular parcellation techniques are mixture models (**Golland et al., 2007; Lashkari et al.,**  
82 **2010, 2012; Tucholka et al., 2008**), variants of the k-means algorithm (**Flandin et al., 2002; Yeo**  
83 **et al., 2011; Kahnt et al., 2012**), hierarchical clustering (**Eickhoff et al., 2011; Michel et al., 2012;**  
84 **Orban et al., 2014**) and variants thereof (**Blumensath et al., 2012**), spectral clustering (**Thirion et al.,**  
85 **2006; Chen et al., 2012; Craddock et al., 2012**) and dense clustering **Hanson et al. (2007)**. Some of  
86 these approaches, but not all, impose spatial constraints on the model, and therefore provide spatially  
87 connected spatial components. In the multi-subject setting, some models adapt the spatial configuration  
88 to each subject (e.g. **Thirion et al. (2006); Lashkari et al. (2010, 2012)**), but most approaches do not.  
89 Parcellations can also be obtained from dictionary learning techniques such as independent components  
90 analysis (ICA) and variants of principal components analysis (PCA) (**Kiviniemi et al., 2009; Varoquaux**  
91 **et al., 2010, 2011, 2013; Abraham et al., 2013**). These rely on a linear mixing approach that changes the  
92 nature of the problem and implies other probabilistic models.

93 While parcellation techniques have great potential and can serve as the basis of many further analyses, it  
94 is important to assess their relative performance. To the best of our knowledge, no systematic comparison  
95 of parcellation methods has been carried out in previous work.

96 The comparison between clustering techniques is only relevant if for each technique the best possible  
97 model is selected. It turns out that model selection for clustering is a notoriously difficult problem, as  
98 is any unsupervised problem in which one wishes to identify some structure in noisy data. While in  
99 practice the choice of the model may depend on the context of the study (for instance, fitting a given  
100 target of interest using region-based signal averages **Ghosh et al. (2013)**), here we derive general rules to  
101 compare parcellation models from empirical observations. In the context of brain mapping, two criteria  
102 are particularly relevant for model selection: *i*) the *goodness of fit* or *accuracy* of a model, i.e. the  
103 ability of the parcellation extracted to model properly the signals of interest on observed and unobserved  
104 data, and *ii*) *stability*, i.e. the consistency of the parcellations obtained from different sub-groups of a  
105 homogeneous population. Importantly, there is a priori no reason why these two criteria should give  
106 consistent answers. There have been few attempts to tackle this, such as **Tucholka et al. (2008); Kahnt**  
107 **et al. (2012); Ghosh et al. (2013)**, but these approaches did not model the multi-subject nature of the  
108 signal; moreover **Tucholka et al. (2008)** were subdividing prior gyrus definition (hence not brain-wide)  
109 and they did not benchmark different clustering techniques. In the present work, we present experiments  
110 on simulated and real data using different clustering techniques and proper accuracy and reproducibility  
111 criteria. To make this tractable computationally and to obtain clear interpretation, we limit ourselves to  
112 nonlinear mixing models, i.e. clustering approaches. Note that methods comparison for clustering versus  
113 linear mixing models (ICA, variants of sparse PCA) has been addressed e.g. in **Abraham et al. (2013)**,  
114 while model order selection for linear model-based region extraction is still an open problem. For similar  
115 reasons, we consider the case in which parcels are identical for all subjects.

116 The remainder of the paper is organized as follows: in Section 2, we introduce the methods tested in  
117 this work and the criteria for model evaluation; in Section 3 we describe our experiments on simulated  
118 and real data, the results of which are given in Section 4. Conclusions on the choice of optimal processing  
119 algorithms and the selection of parcellation schemes are drawn in Section 5.

## 2 MATERIAL & METHODS

### 2.1 NOTATION

120 We start with a given set of  $n$  functional images that represent e.g. different contrasts in a given group of  
 121 subjects. We denote  $N$  to be the number of subjects and  $F$  the number of functional images (here contrasts)  
 122 per subject, such that  $n = NF$ . These images are typically the results of first-level analysis (standardized  
 123 effects) and are sampled on a grid of  $Q$  voxels. Starting from  $n$  fMRI volumes  $\mathbf{Y} = [\mathbf{y}^1, \dots, \mathbf{y}^Q] \in \mathbb{R}^{n \times Q}$   
 124 that consist of  $Q$  voxels, we seek to cluster these voxels so as to produce a reduced representation of  $\mathbf{Y}$ .

### 2.2 CLUSTERING METHODS FOR BRAIN FUNCTIONAL PARCELLATION

125 *K-means Algorithm* K-means is arguably the most used clustering technique for vector data. It consists of  
 126 an alternate optimization of *i*) the assignment  $u_{\text{k-means}}$  of samples to cluster and *ii*) the estimation of the  
 127 cluster centroids.

$$\forall j \in [1, Q], u_{\text{k-means}}(j) = \operatorname{argmin}_{c \in [1, \dots, K]} \|\langle \mathbf{Y} \rangle_c - \mathbf{y}^j\| \quad (1)$$

$$\langle \mathbf{Y} \rangle_c \triangleq \frac{1}{|c|} \sum_{u_{\text{k-means}}(j)=c} \mathbf{y}^j \quad (2)$$

128 It explicitly minimizes the inertia, i.e. the sum of squared differences between the samples and their  
 129 representative cluster centroid. We introduce an approximation for the sake of efficiency: the whole set of  
 130 feature data used in clustering (several contrasts from all the subjects) of dimension  $n = N(\text{subjects}) \times$   
 131  $F(\text{contrasts})$  is reduced by PCA to  $m = 100$  components prior to clustering, capturing about 50% of  
 132 the variance. It is important to note that k-means clustering of fMRI data are used without explicitly  
 133 considering their spatial structure, although spatial smoothing prior to clustering can indirectly provide  
 134 spatial regularization.

135 *Ward's Algorithm* As an alternative, we consider a *hierarchical agglomerative clustering* (**Johnson,**  
 136 1967). These procedures start with every voxels  $\mathbf{x}^j$  representing singleton clusters  $\{j\}$  and, at each  
 137 iteration, a pair of clusters, selected according to a criterion discussed below, is merged into a single  
 138 cluster. This procedure yields a hierarchy of clusters represented as a binary tree  $\mathcal{T}$ , also often called a  
 139 dendrogram (**Johnson,** 1967), where each non-terminal node is associated with the cluster obtained by  
 140 merging its two children clusters.

141 Among different hierarchical agglomerative clustering procedures, we use the variance-minimizing  
 142 approach of Ward's algorithm (**Ward,** 1963). In short, two clusters are merged if the resulting cluster  
 143 minimizes the sum of squared differences of the fMRI signal within all clusters. More formally, at each  
 144 step of the procedure, we merge the clusters  $c_1$  and  $c_2$  that minimize

$$\begin{aligned} \Delta(c_1, c_2) &= \sum_{j \in c_1 \cup c_2} \|\mathbf{y}^j - \langle \mathbf{Y} \rangle_{c_1 \cup c_2}\|_2^2 - \left( \sum_{j \in c_1} \|\mathbf{y}^j - \langle \mathbf{Y} \rangle_{c_1}\|_2^2 + \sum_{k \in c_2} \|\mathbf{y}^k - \langle \mathbf{Y} \rangle_{c_2}\|_2^2 \right) \\ &= \frac{|c_1||c_2|}{|c_1| + |c_2|} \|\langle \mathbf{Y} \rangle_{c_1} - \langle \mathbf{Y} \rangle_{c_2}\|_2^2, \end{aligned} \quad (3)$$

145 where  $\langle \mathbf{Y} \rangle_c$  is the average vector defined in (2). In order to take into account the spatial information, we  
 146 also add connectivity constraints in the hierarchical clustering algorithm, so that only neighboring clusters  
 147 can be merged together. In other words, we try to minimize the criterion  $\Delta(c_1, c_2)$  only for pairs of clusters  
 148 that share neighboring voxels. Given a number of parcels  $K$ , we stop the construction of the tree at the  
 149  $(Q - K)^{\text{th}}$  iteration and retain the corresponding assignment  $u_{\text{ward}}$ . Note that the data are subject to the  
 150 same PCA procedure as for k-means clustering.

151 *Spectral clustering* Spectral clustering (Shi and Malik, 2000; Ng et al., 2001) consists in performing  
 152 k-means clustering on a representation of the data that preserves the spatial structure yet represents the  
 153 functional features' similarity<sup>2</sup>. This representation is typically obtained by using the first eigenvectors of  
 154 the Laplacian matrix of the graph that encodes the spatial relationships weighted by the functional features  
 155 similarity between adjacent locations. For all voxel pairs  $(i, j) \in [1 \cdots Q]^2$ , Let

$$W_{ij} = \begin{cases} \exp(-\frac{\|\mathbf{y}^i - \mathbf{y}^j\|^2}{2\sigma_f^2}) & \text{if } i \text{ and } j \text{ are neighbors} \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

156 where we used  $\sigma_f^2 = \text{mean}_{i \sim j} \|\mathbf{y}^i - \mathbf{y}^j\|^2$ , where the averaging is performed over all pairs of adjacent  
 157 voxels.  $\sigma_f^2$  is thus the average squared distance between the data across neighboring voxels.  $W$  is therefore  
 158 an adjacency matrix weighted by the functional distance between voxels. We denote  $\Delta_W$  the diagonal  
 159 matrix that contains the sum of the rows of  $W$ .

160 Then, let  $(\xi_1, \dots, \xi_m)$  the first  $m$  solutions of  $W\xi = \lambda\Delta_W\xi$ . The spectral clustering of the dataset is  
 161 defined as:

$$u_{\text{spectral}} = \text{k-means}([\xi_1, \dots, \xi_m]), \quad (5)$$

162  $m = 100$  in our experiments. We also tried different (larger or smaller) values, but those did not yield  
 163 significantly better solutions.

164 *Geometric clustering* To provide a reference for comparison, we also use a clustering algorithm that does  
 165 not take into account the functional data, but only the spatial coordinates of the voxels. In practice, it is  
 166 obtained through a k-means clustering of the spatial coordinates.

### 2.3 A MIXED-EFFECTS MODEL OF THE SIGNAL WITHIN PARCELS

167 We introduce a probabilistic model of the signal of the voxels in a given (fixed) parcel  $\mathcal{P}_k$ ,  $k \in [1, \dots, K]$ ,  
 168 that includes a random subject effect. Let us first assume that we work with one functional image ( $F = 1$ ).  
 169 Let  $p$  be the number of voxels in the parcel, pooled across subjects: it is the size of the parcel multiplied  
 170 by  $N$ ; let  $\mathbf{y}$  be a  $p$ -dimensional vector that denotes the scalar signal in the voxels contained in  $\mathcal{P}_k$ ,  
 171 concatenated across subjects; we model it though the following mixed-effects model:

$$\mathbf{y} = \mu\mathbf{1} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad (6)$$

172 where  $\mu$  is the average signal within the parcel,  $\mathbf{1}$  is a vector of ones of length  $p$ ,  $\boldsymbol{\beta}$  is a vector of subject-  
 173 specific random effects parameters,  $\mathbf{X}$  the (known) matrix that maps subjects to voxels: for each row, a one  
 174 in the  $i^{\text{th}}$  column indicates that the value is from subject  $i$ .  $\boldsymbol{\varepsilon}$  represents the intra-subject variability of the  
 175 signal within a parcel. It is further assumed that  $\boldsymbol{\varepsilon}$  and  $\boldsymbol{\beta}$  are independent, normal and centered at 0, with  
 176 variance  $\sigma_1^2$  and  $\sigma_2^2$  that express respectively the within and between subject variance. The probabilistic  
 177 model of  $\mathbf{y}$  is thus:

$$\mathbf{y} \sim \mathcal{N}(\mu\mathbf{1}, \sigma_1^2\mathbb{I} + \sigma_2^2\mathbf{X}\mathbf{X}^T), \quad (7)$$

178 where  $\mathbb{I}$  is the  $p \times p$  identity matrix.

179 The generalization to non-scalar images (for instance,  $F > 1$  images per subject) is obtained by  
 180 assuming the independence of the observations conditional to the parcellation, hence it decouples into  
 181 multiple ( $F$ ) scalar models. The estimation of the parameters  $(\mu, \sigma_1, \sigma_2)$  is carried out in each parcel  
 182  $\mathcal{P}_k$ ,  $k \in [1, \dots, K]$  using the maximum likelihood principle; we use an Expectation-Maximization  
 183 algorithm to estimate the model parameters (Meng and van Dyk, 1998).

<sup>2</sup> A variant of spectral clustering replaces this k-means step by learning a rotation to discretize the representation Yu and Shi (2003). We used this approach, that outperforms k-means, in the case where the number of desired clusters  $K$  is smaller than the subspace dimension  $m$  (see next).



## 2.4 MODEL SELECTION FOR FUNCTIONAL PARCELLATIONS

184 A problem that comes naturally with clustering algorithms is the choice of the number  $K$  of clusters to  
 185 be used in the model. To guide this choice we consider four standard measures: BIC, cross-validated  
 186 likelihood, adjusted rand index, and normalized mutual information.

187 *Bayesian Information Criterion, BIC* The goodness of fit of a probabilistic model is given by the log-  
 188 likelihood of the data and the quality of the model is easily measured using the BIC criterion (**Schwarz,**  
 189 1978), that penalizes the negative log-likelihood by the number of parameters used. Within a given parcel  
 190  $\mathcal{P}_k$ , this yields the following:

$$bic(k) = -2 \log \mathcal{N}(\mathbf{y}; \mu \mathbf{1}, \sigma_1^2 \mathbb{I} + \sigma_2^2 \mathbf{X} \mathbf{X}^T) + 3 \log(p), \quad (8)$$

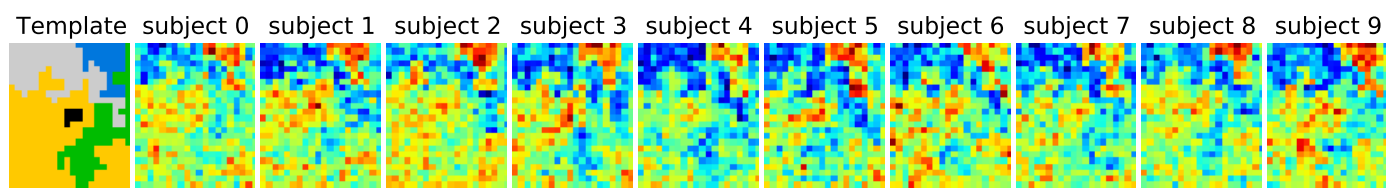
191 Where 3 is the number of parameters of the model  $(\mu, \sigma_1, \sigma_2)$ . Note that all the quantities in this formula  
 192  $(\mathbf{y}, \mu, \sigma_1^2, \sigma_2^2, \mathbf{X}, p)$  depend on  $k$ , the index of the parcel.  $bic(k)$  is summed across parcels in order to yield  
 193 a unique quantity that is comparable for different values of  $K$ , that we denote *BIC* henceforth.

194 The BIC is theoretically asymptotically optimal for model selection purpose (**Schwarz, 1978**), however,  
 195 it may fail in practice for several reasons. In particular, it relies on some hypotheses for the data, such  
 196 as the i.i.d structure of the residuals, which is violated in fMRI. This means that the goodness of fit of  
 197 over-parametrized models increases faster than it should in theory, and thus that more complex models,  
 198 i.e. with a large number of parcels, are systematically and spuriously preferred. In the case of brain  
 199 volume parcellation, the violation of the i.i.d. hypothesis might be related to different factors, such as data  
 200 smoothness or spatial jitter across individuals.

201 *Cross-validated likelihood* A nice feature of the model (7) is that it can be evaluated on test data, thus  
 202 making it possible to run a cross-validation procedure on different subjects; such a procedure does not  
 203 *overfit*, where overfit means *models non-reproducible noise, creating the optimistic bias inherent when*  
 204 *learning and evaluating a model on the same data*. We use the log-likelihood in a *shuffle-split* cross-  
 205 validation scheme: for each fold, the model is learned on the training set (i.e. a random subsample of  
 206 80 % of the data): this includes the estimation of the clustering and fitting the mixed-effects model; the  
 207 log-likelihood computed on the test data is then summed across parcels in order to yield a unique quantity,  
 208 denoted  $CV - LL$  in the following.

209 *Reproducibility by bootstrap* The two previous metrics only address the fit of the data by the model.  
 210 Another important criterion in neuroimaging is reproducibility (**LaConte et al., 2003**), which we define  
 211 in this context as the consistency of two clustering solutions across repeats on bootstrap samples taken  
 212 from the data, measured by assignment statistics of voxels to clusters. To estimate reproducibility, we  
 213 repeated the clustering by bootstrapping over subjects and assessed the stability of the clustering between  
 214 pairs of bootstrap samples using two standard metrics: adjusted mutual information or adjusted rand index.  
 215 The **adjusted Rand index** (ARI) is comprised between -1 and 1, and measures the consistency of the two  
 216 labellings while being invariant to a permutation of the labels (**Vinh et al., 2009**). A value of 1 means  
 217 perfect correspondence of the labeling, while a value of 0 implies that the correspondences are at chance.  
 218 An important feature of the ARI metric is that it scales well when the number of clusters  $K$  is large. See  
 219 [http://en.wikipedia.org/wiki/Rand\\_index](http://en.wikipedia.org/wiki/Rand_index) for more details.

220 **Adjusted mutual information** (AMI) upper bounded by 1, and possibly negative, is an estimate  
 221 of the mutual information of two discrete assignment of voxels to parcels, which is corrected for  
 222 chance: two statistically independent assignments should have an AMI value of 0, while two identical  
 223 assignments should have an AMI value of 1 [http://en.wikipedia.org/wiki/Adjusted\\_](http://en.wikipedia.org/wiki/Adjusted_mutual_information)  
 224 [mutual\\_information](http://en.wikipedia.org/wiki/Adjusted_mutual_information).



**Figure 1.** Example of simulated data used in the 2D simulation experiment. The *template* or ground truth labeling is shown on the left side, and 10 individual datasets are sampled according to the model, jittered spatially by 2 pixels and then smoothed with a kernel of fwhm 1.17 pixels.

## 2.5 IMPLEMENTATION

225 We use the algorithms and metrics from the scikit-learn toolbox (Pedregosa et al., 2011). In particular,  
 226 Ward's algorithm is very efficient on data size of typical brain images. The following version of the  
 227 software were used: Matlab R2013A, version 8.1.0.64, SPM8 v. 5242, scikit learn v. 0.14. The code used  
 228 in this work is available at [https://github.com/bthirion/frontiers\\_2014](https://github.com/bthirion/frontiers_2014).

## 3 EXPERIMENTS

### 3.1 SIMULATED DATA

229 Data are simulated according to model (6): on a 2D grid of shape  $20 \times 25$  pixels, 5 random clusters  
 230 are generated with a hierarchical clustering approach, by using Ward's parcellation on a set of random  
 231 signals; 10 individual datasets are sampled using the generative model: for each parcel the parameters  $\mu$   
 232 are sampled from  $\mathcal{N}(0, 1)$ ,  $\sigma_1 = 1$  and the random subject effect  $\beta$  are drawn from  $\mathcal{N}(0, 1)$ ,  $\sigma_2 = 1$ . Note  
 233 that the  $\beta$ s are kept constant across parcels. Data corresponding to a sample of 10 subjects are generated.  
 234 To make the data more realistic, we add a deformation to each individual dataset that has a magnitude  
 235 of 0, 1 or 2 pixels in each direction and smooth it -or not- with a kernel of full width at half maximum  
 236 (fwhm) of 1.17 pixel. Note however that this breaks (on purpose) the hypotheses of the generative model  
 237 and makes the simulations more realistic. An example is shown in Fig. 1.

238 The question that we address is whether we can hope to recover the true number of clusters from the  
 239 simulation; to do so, we can use one of the three selection criteria: BIC, cross-validation and bootstrap  
 240 reproducibility (we use B-AMI by default, but B-ARI yields similar results on this dataset). The recovery  
 241 is quantified through the adjusted rand index between the true labeling of voxels and the obtained one. The  
 242 results are based on 200 replications of the experiment, and the optimal number  $K$  of parcels is searched  
 243 in the  $\{2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 30\}$  set.

### 3.2 FUNCTIONAL LOCALIZER DATA

244 Data were acquired from 128 subjects who performed a *functional localizer* protocol as described in P  
 245 **inel et al. (2007)** and referred to as *Localizer* henceforth. This protocol is intended to activate multiple  
 246 brain regions in a relatively short time (128 brain volumes acquired in 5 minutes) with ten experimental  
 247 conditions, allowing the computation of many different functional contrasts: left and right button presses  
 248 after auditory or visual instruction, mental computation after auditory or visual instruction, sentence  
 249 listening or reading, passive viewing of horizontal and vertical checkerboards. The subjects gave informed  
 250 consent and the protocol was approved by the local ethics committee.

251 In 59 of the subjects, functional images were acquired on an 3T Siemens Trio scanner using an EPI  
 252 sequence (TR = 2400ms, TE = 60ms, matrix size =  $64 \times 64$ , FOV =  $19.2cm \times 19.2cm$ ). Each volume  
 253 consisted of 40 3mm-thick axial slices without gap. A session comprised 132 EPI scans, of which the first



254 four were discarded to allow the MR signal to reach steady state. The slices were acquired in interleaved  
255 ascending order. Anatomical fSPGR T1-weighted images were acquired on the same scanner, with a slice  
256 thickness of 1.0 mm, a field of view of 24 cm and an acquisition matrix of  $256 \times 256 \times 128$  voxels,  
257 resulting in 124 contiguous double-echo slices with voxel dimensions of  $(1.0 \times 1.0 \times 1.0) \text{ mm}^3$ .

258 In 69 of the subjects, functional and anatomical were acquired on a 3T Bruker scanner. Functional  
259 images were acquired using an EPI sequence ( $TR = 2400 \text{ ms}$ ,  $TE = 60 \text{ ms}$ , matrix size =  $64 \times 64$ ,  $FOV =$   
260  $19.2 \text{ cm} \times 19.2 \text{ cm}$ ). Each volume consisted of  $n_a$  3-mm- or 4-mm-thick axial slices without gap, where  
261  $n_a$  varied from 26 to 40 according to the session. A session comprised 130 scans. The first four functional  
262 scans were discarded to allow the MR signal to reach steady state. Anatomical T1 images were acquired  
263 on the same scanner, with a spatial resolution of  $(1. \times 1. \times 1.2) \text{ mm}^3$ .

264 The data was subject to a pre-processing procedure that includes the correction of the difference in slice  
265 timing, motion estimation and correction, co-registration of the EPI volumes to the T1 image, non-linear  
266 spatial normalization of T1 images, then of the fMRI scans to the SPM T1 template. All of these steps  
267 were performed using the SPM8 software. Optionally, we considered a 5mm isotropic smoothing the  
268 normalized images. In parallel, an average mask of the gray matter was obtained from the individual  
269 normalized anatomies, subsampled at the fMRI resolution, and used to mask the volume of interest in the  
270 functional dataset. This procedure yields approximately  $Q = 57,000$  voxels at 3mm resolution.

271 A General Linear Model (GLM) analysis was applied for the volume using the Nipy package <http://nipy.sourceforge.net/>. The model included the ten conditions of the experiments convolved  
272 with a standard hemodynamic filter and its time derivative, a high-pass filter (cutoff:128s); the procedure  
273 included an estimation of the noise auto-correlation using an AR(1) model.

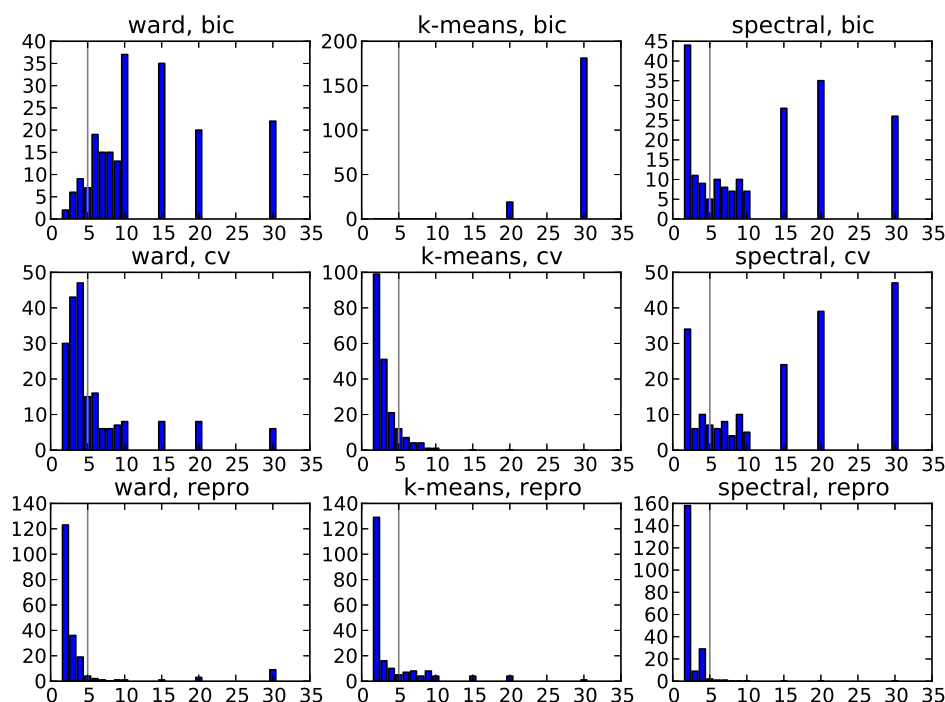
275 Activation maps were derived for six functional contrasts, that display the activations related to *left*  
276 *versus right button presses*, *motor versus non-motor tasks*, *sentence listening versus sentence reading*,  
277 *computation versus sentence reading*, *reading versus passive checkerboard viewing*, *vertical versus*  
278 *horizontal checkerboard viewing*. We consider that these six contrasts give the most usable summary of  
279 the topographic information conveyed by the initial ten conditions, without obvious redundancies, while  
280 avoiding non-specific effects.

281 The standardized effects related to these  $F = 6$  contrasts are used for parcellation fit and evaluation. We  
282 consider the possible range of values for  $K$ : 10, 20, 30, 40, 50, 70, 100, 150, 200, 300, 400, 500, 700,  
283 1000, 1500, 2000, 3000, 5000, 7000, 10000. We consider the value of the different criteria for different  
284  $K$  values.

### 3.3 HCP DATA

285 A set of  $N = 67$  subjects of the Human Connectome Project (HCP) dataset was also used in our  
286 experiments. These subjects are part of the Q2 release; we used the task-fMRI dataset, that comprises  
287 7 different sessions (see **Barch et al.** (2013) for details), all of which are used here. Starting from the  
288 preprocessed volume data provided by the HCP consortium, these dataset were analyzed similarly to the  
289 *Localizer* dataset, using the Nipy software for the GLM analysis, that was carried out using the paradigm  
290 information provided with the data. The same gray matter mask was used as for the *Localizer* dataset was  
291 used to facilitate comparisons between the two datasets.

292 In order to reduce computation time, a subset of  $F = 9$  functional contrasts were used: the *faces-shape*  
293 contrast of the *emotional* protocol, the *punish-reward* contrast of the *gambling* protocol, the *math-story*  
294 contrast of the *language* protocol, the *left foot-average* and *left hand-average* contrasts of the *motor*  
295 protocol, the *match-relation* contrast of the *relational* protocol, the *theory of mind-random* contrast of  
296 the *social* protocol and the *two back-zero back* contrast of the *working memory* protocol. this choice was  
297 meant to sample a significant set of cognitive dimensions tested in the protocol, without being exhaustive.



**Figure 2.** Results of the simulations: choice of the number parcels for different clustering methods and cluster selection techniques. Note that the range of possible values is [2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 30] and that the true value is 5. The results are based on data smoothed with a kernel of size 0.5 voxel, and under spatial jitter of 1 voxel isotropic, and are presented across 200 replications. bic, cv (cross validation) and repro (Adjusted Mutual Information) represent three model selection approaches, while ward, k-means and spectral represent three different clustering approaches.

## 4 RESULTS

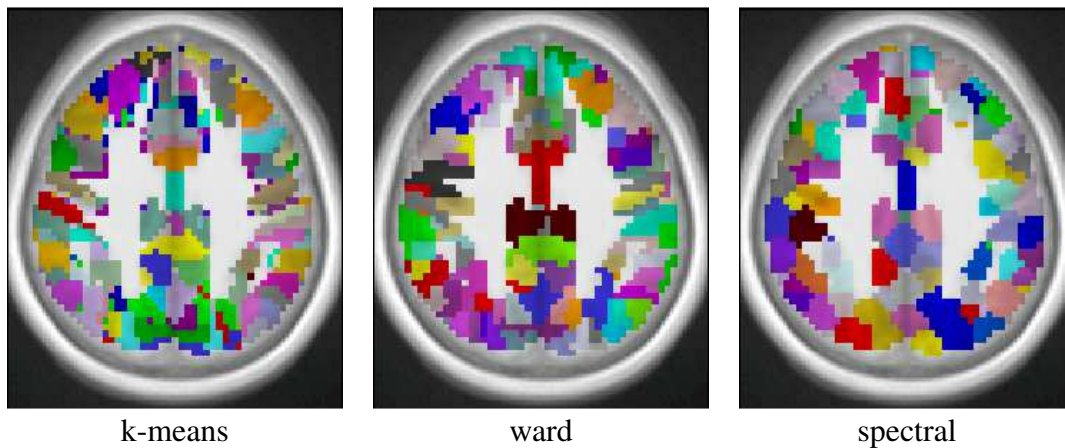
### 4.1 SIMULATIONS

298 Fig. 2 displays the selected  $K^*$  value, based on data smoothed with a kernel of size 0.5 voxel, and under  
 299 spatial jitter of 1 voxel isotropic; given that  $K^{true} = 5$  it shows that BIC tends to select too large number  
 300 of clusters, while, on the opposite, reproducibility, measured via bootstrapped AMI, is conservative;  
 301 cross-validated log-likelihood shows an intermediate behavior, as it is conservative for spectral clustering  
 302 and anti-conservative for k-means. However, the right model is not recovered in general, because the true  
 303 clustering is not in the solution path of the different methods (this is especially true for spectral clustering),  
 304 or because model selection fails to recover the right number of parcels.

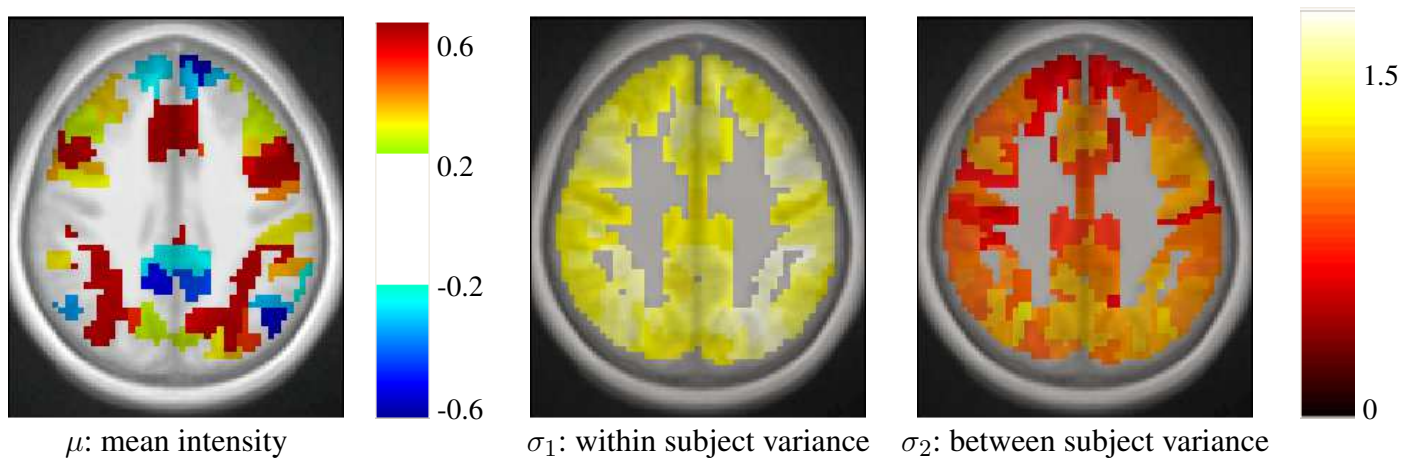
305 Our main observation is thus that reproducibility-based model selection criteria seem over-conservative,  
 306 while accuracy-based selection criteria are too liberal.

### 4.2 REAL DATA

307 *4.2.1 Qualitative assessment of the solutions* The spatial layout of the clusters can be observed in the  
 308 brain volume (see Fig. 3 for on axial slice), and it represents the characteristics of the competing clustering  
 309 algorithms: Spectral clustering yields a very geometrical parcellation of the volume, hinting at a lower  
 310 sensitivity to the functional input data, while k-means presents results with less spatial consistency (e.g.  
 311 disconnected clusters), yet a realistic representation of plausible functional patches, and Ward's algorithm  
 312 presents a compromise between the two solutions.

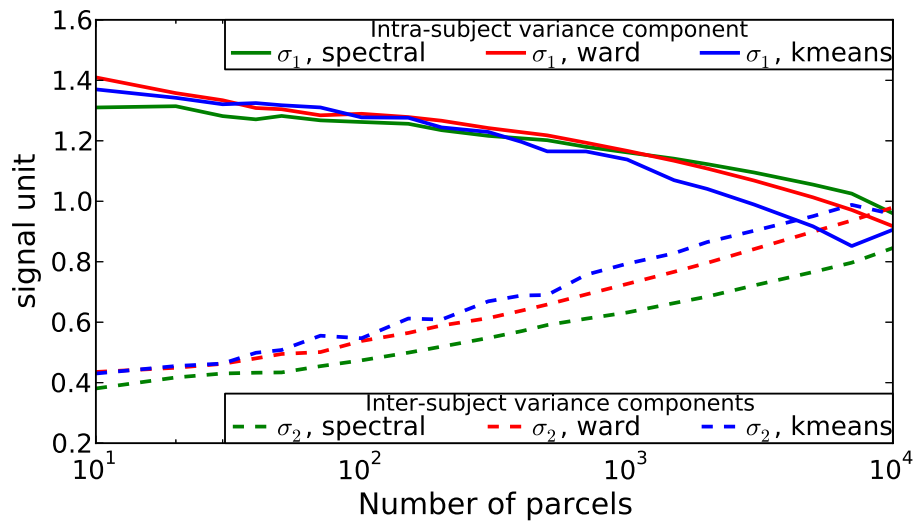


**Figure 3.** Example of parcellation with 500 parcels on the Localizer dataset.



**Figure 4.** Example of parameters estimated in a parcellation obtained with Ward's clustering and  $K = 500$  parcels. They are given in arbitrary units (percent of the baseline fMRI signal, squared for variance estimates). These parameters are those for the *computation-sentence reading* functional contrast.

313 After parcellation, the parameters of the model 6 are estimated in each parcel, for each functional  
 314 contrast and can be plotted in the brain volume; see Fig. 4. In particular, it can be seen that  $\sigma_1 > \sigma_2$   
 315 uniformly i.e. within-parcel variability dominates across-subject variability when  $K = 500$ . Moreover,  
 316 in the case of Ward's parcellation presented here, the within- and between-subject variance estimated  
 317 are quite homogeneous across the brain volume. Note however that there is a tendency for both to be  
 318 correlated with the absolute value of the mean signal. Next, we consider how the variance components,  
 319 averaged across parcels, change with  $K$  in Fig. 5. These values evolve monotonously with  $K$ : the intra-  
 320 subject parameter  $\sigma_1$  (that measures the cross-voxel variance within a given subject, averaged across  
 321 parcels) decreases monotonously with  $K$ , as expected; the inter-subject parameter  $\sigma_2$ , that characterizes  
 322 the cross-subject variability of the mean signal within a parcel, increases monotonously. Both parameters  
 323 come close to equality for large values of  $K$  (about 5000). These trends are similar across clustering  
 324 techniques. This actually means that changing the resolution yields a re-allocation of the variance from  
 325 the intra-subject to the inter-subject component of the mixed-effects model. More specifically, for low  
 326 values of  $K$ , the high within-subject variance shadows the between-subject variance, and a very large  
 327 value of  $K$  has to be used if one wants to estimate correctly the between-subject variability of the BOLD  
 328 signal within the parcellation framework.



**Figure 5.** Dependence on  $K$  of the variance components from model 6, averaged across parcels and contrasts: both  $\sigma_1$  and  $\sigma_2$  parameters show a monotonic behavior: the within subject variance decreases  $\sigma_1$  with  $K$ , while the between-subject variance  $\sigma_2$  increases with  $K$ .

Atlas	Summed log-likelihood	stdv
Harvard-Oxford atlas	$-6.642 \cdot 10^7$	$1.9 \cdot 10^5$
Geometric parcellation	$-6.589 \cdot 10^7$	$1.9 \cdot 10^5$
k-means parcellation	$-6.463 \cdot 10^7$	$1.8 \cdot 10^5$
Ward parcellation	$-6.513 \cdot 10^7$	$1.9 \cdot 10^5$
Spectral clustering parcellation	$-6.591 \cdot 10^7$	$1.8 \cdot 10^5$
(HCP) k-means parcellation	$-6.710 \cdot 10^7$	$1.9 \cdot 10^5$
(HCP) Ward parcellation	$-6.522 \cdot 10^7$	$1.9 \cdot 10^5$
(HCP) Spectral clustering parcellation	$-6.613 \cdot 10^7$	$1.9 \cdot 10^5$

**Table 1.** Summed log-likelihood of the Localizer data under different spatial models (the higher, the better): brain atlas (top), parcellation on the Localizer dataset (middle), parcellations from the HCP data (right). The number of parcels used is  $K = 158$  for all methods. The standard deviation is obtained by drawing 30 bootstrap samples.

329 *Comparison with an anatomical atlas* As a basis for comparison with anatomical atlases, we evaluated the  
 330 log-likelihood of the data with the most detailed atlas that we could find. We used the Harvard-Oxford atlas  
 331 both cortical and subcortical <http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/Atlases> together  
 332 with the cerebellar atlas (Diedrichsen et al., 2009). The version used was that of FSL 4.1. The regions  
 333 were systematically divided into left and right hemispheres by taking the sign of the  $x$  MNI coordinate  
 334 of the voxels. Using this procedure, we obtained 158 regions. This atlas was resampled at the resolution  
 335 of the test fMRI data, and the likelihood of the data summed over parcels was evaluated and compared  
 336 with that of data-driven parcellations with 158 parcels, obtained either from the Localizer dataset itself or  
 337 from the HCP dataset. Standard deviation of the log-likelihood are obtained by drawing  $B = 30$  bootstrap  
 338 samples. The results are shown in Table 1.

339 We show the corresponding results on the HCP dataset in Table 2.

340 It can be seen that the anatomical atlas achieves the poorest fit: summarizing the fMRI data on the  
 341 corresponding set of parcels loses a lot of information. Even a purely geometric parcellation performs  
 342 better, which can be understood given that it tends to create parcels with equal size, hence achieves a more

Atlas	Summed log-likelihood	stdv
Harvard-Oxford atlas	$-4.557 \cdot 10^7$	$3.3 \cdot 10^5$
Geometric parcellation	$-4.537 \cdot 10^7$	$3.4 \cdot 10^5$
k-means parcellation	$-4.459 \cdot 10^7$	$3.1 \cdot 10^5$
Ward parcellation	$-4.491 \cdot 10^7$	$3.4 \cdot 10^5$
Spectral clustering parcellation	$-4.543 \cdot 10^7$	$3.3 \cdot 10^5$
(Localizer) k-means parcellation	$-4.529 \cdot 10^7$	$3.3 \cdot 10^5$
(Localizer) Ward parcellation	$-4.530 \cdot 10^7$	$3.3 \cdot 10^5$
(Localizer) Spectral clustering parcellation	$-4.539 \cdot 10^7$	$3.3 \cdot 10^5$

**Table 2.** Summed log-likelihood of the HCP data under different spatial models (the higher, the better): brain atlas (top), parcellation on the HCP dataset (middle), parcellations from the Localizer data (right). The number of parcels used is  $K = 158$  for all methods. The standard deviation is obtained by drawing 30 bootstrap samples.

343 regular sampling of the volume of interest. For  $K = 158$  the best performing parcellation on the training  
344 set is obtained from k-means, but these parcellations do not generalize well from a dataset to another.  
345 Ward’s parcellation on the other hand, performs better than geometric clustering in all configurations.  
346 Finally, the bootstrap variability of these results is typically small with respect to between-method  
347 difference for the Localizer dataset, ensuring that the differences are significant. This is less so with the  
348 HCP dataset, for two reasons: the number of subjects is smaller, and the per-subject SNR seems relatively  
349 lower in that dataset (see **Barch et al.** (2013)).

350 *4.2.2 Analysis of the goodness of fit the models (Localizer dataset)* The goodness of fit of the model is  
351 given by the log-likelihood, which can be compared across methods for a fixed value of  $K$  in Fig. 6 (a).  
352 The main observations are:

- 353 • For all methods, the curve achieves an optimum value for a very large number of parcels ( $3000 \leq$   
354  $K \leq 7000$ ), which is much more than the number typically expected and used in neuroimaging  
355 experiments.
- 356 • *k-means* and *Ward’s* clustering achieve the lowest distortion –i.e. loss of information from the original  
357 signal–, with *k-means* performing better for low number of parcels and *Ward’s* clustering performing  
358 better for large number of clusters. Spectral clustering is inferior in terms of goodness of fit. It is even  
359 lower than a purely geometric parcellation of the brain volume for some values of  $K$ .
- 360 • The achieved log-likelihood is larger on smoothed data than on unsmoothed data, but the behavior is  
361 qualitatively similar. In this report, we present only results on unsmoothed data.

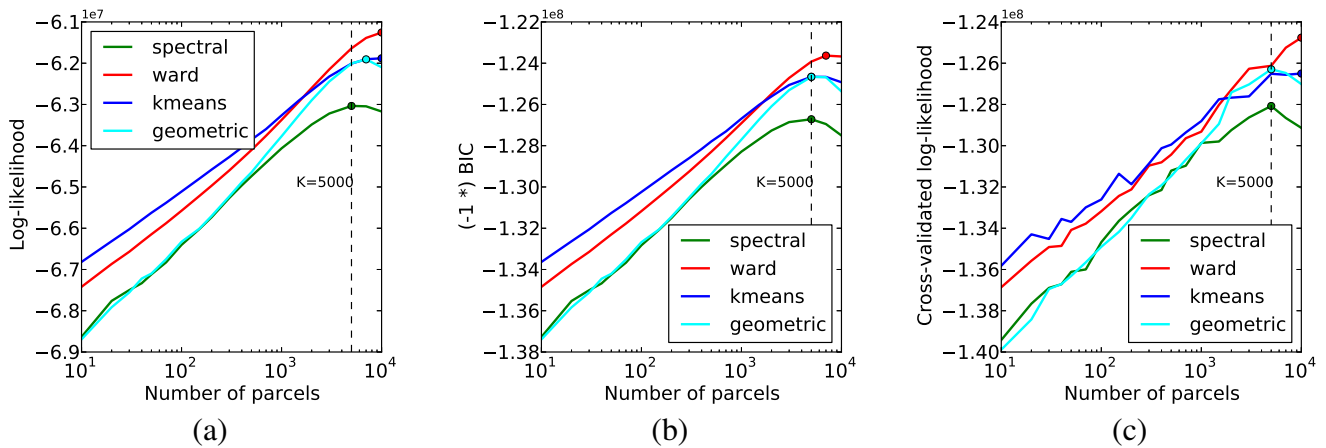
362

363 Second, we can observe that, unlike in our simulations, BIC and cross-validated log-likelihood (Fig. 6  
364 (b-c)) achieve their optimum at the same value of  $K$  as the data log-likelihood function, thus at very high  
365 values ( $3000 \leq K \leq 7000$ ).

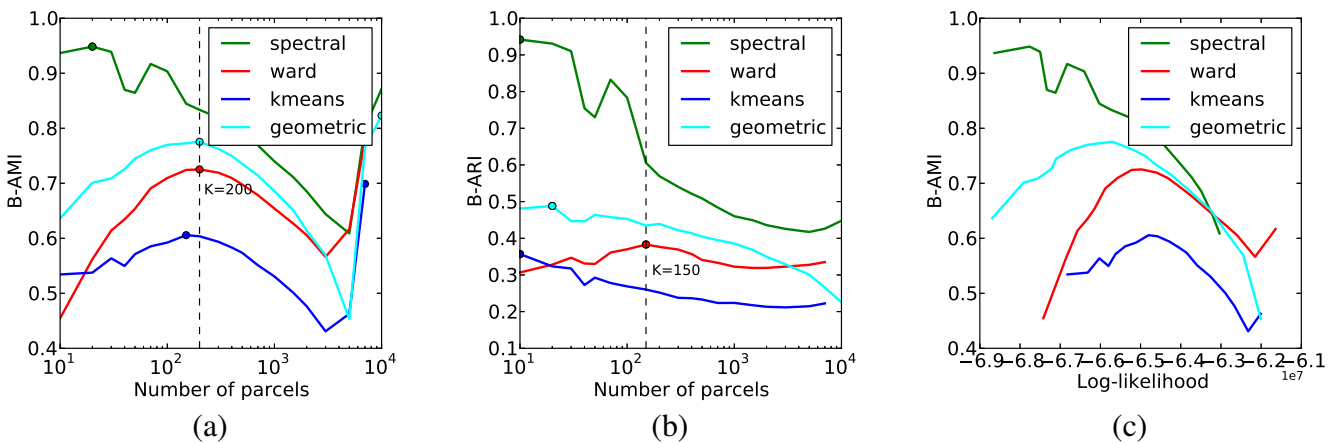
366 *4.2.3 Accuracy-reproducibility compromise (Localizer dataset)* The reproducibility of the clustering  
367 estimated by bootstrapping the data can be studied as a function of the number of clusters, or as a function  
368 of the likelihood. Both representations are presented in Fig. 7.

369 The reproducibility index displays a clear optimum value at  $K \simeq 200$  parcels. For larger values, the  
370 reproducibility index decreases slowly, but increases again for very large number of parcels  $K > 4000$ .  
371 This late increase can readily be interpreted as an artifact due to the fact that we are now observing a very





**Figure 6. Distortion metrics** (a) Accuracy of the model 6 measured through the summed Log-likelihood across parcels, as a function of the number  $K$  of clusters. The accuracy is maximized for very high values of  $K$ . The Bayesian Information Criterion (b) -with the sign flipped for the sake of visualization- and the cross-validated log-likelihood (c), that can be used to identify the right model show the same behavior as the log-likelihood function.

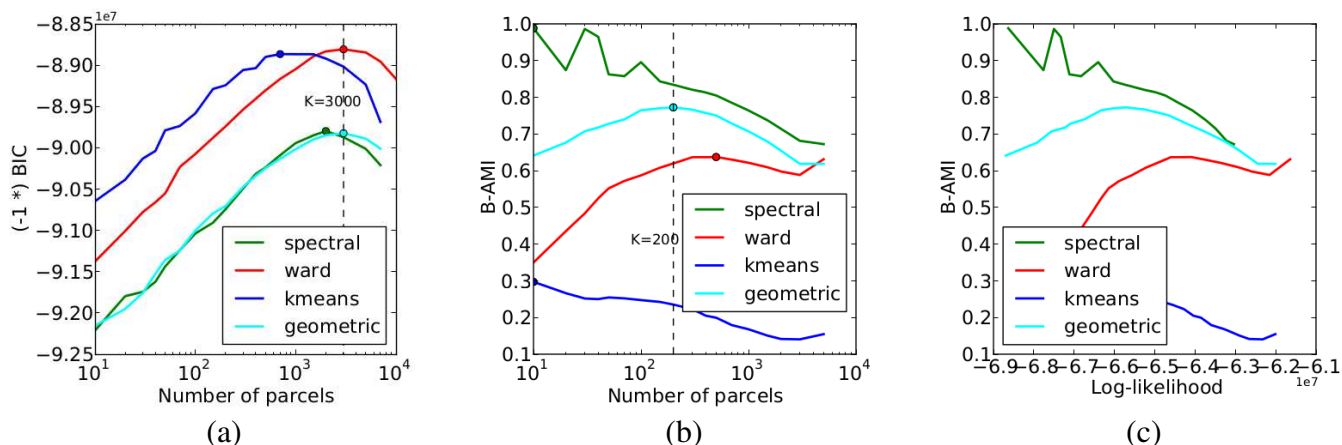


**Figure 7.** Analysis of the reproducibility index with respect to the number of parcels (a, b) and with respect to the negative log-likelihood (c). For all methods but one, the B-AMI (Bootstrapped Adjusted Mutual information) index (a) shows a (local) maximum for about 200 parcels and decreases against for larger numbers, until it increases again for very large number clusters ( $K \geq 5000$ ). By contrast, B-ARI (Bootstrapped Adjusted Rand Index) (b) only displays the local maximum on Ward's parcellation. If we consider B-AMI against accuracy there is thus a trade-off region, for a number of parcels comprised between 200 and 5000 (decreasing portion of the curves in the reproducibility-accuracy curve), in which each setting represents a different compromise. The two dominant techniques are spectral clustering, that maximizes the reproducibility index, and Ward's clustering, that yields higher accuracy overall.

372 large number of very small clusters, and that the reproducibility indexes are not well suited in this case.  
 373 It is also true that very small clusters tend to represent the spatial neighboring system, and thus this high  
 374 reproducibility is not very informative on the functional features carried by the data.

375 The spectral clustering outperforms the other alternatives regarding reproducibility, which means that it  
 376 is able to capture some stable features in the input data, although the overall representation is suboptimal  
 377 in terms of accuracy. Regarding the sensitivity/reproducibility compromise (see Fig. 7, right), the spectral  
 378 method is dominant in the low accuracy/high reproducibility region, while Ward's method dominates in  
 379 the high accuracy/low reproducibility region.





**Figure 8.** Results of the model selection experiments on the HCP dataset: (left) accuracy-based selection through the BIC score, (middle) reproducibility-based selection through Bootstrapped Adjusted Mutual Information, (right) ensuing sensitivity/reproducibility curve.

### 4.3 MODEL SELECTION RESULTS ON THE HCP DATASET

380 A summary of the results obtained by doing the same experiments on the HCP dataset is provided in  
 381 Fig. 8. In spite of weak changes of the optimal values  $K^* \leq 3000$  for accuracy,  $K^* \in [200, 500]$  for  
 382 reproducibility), this dataset reproduces exactly the trends observed with the Localizer dataset: Ward's  
 383 method outperforms the others in terms of accuracy and for high  $K$  values, spectral clustering yields a  
 384 poor fit and a high reproducibility, k-means a good fit, especially for small  $K$ , yet very low reproducibility.

## 5 DISCUSSION

385 Our experiments benchmark three methods to derive brain parcellations from functional data, using three  
 386 model selection criteria and two reproducibility measures. Though not exhaustive, these experiments are  
 387 very informative on the general behavior, the domain of optimality of the methods, and the issues that  
 388 limit the power of such approaches in neuroimaging data analysis.

### 5.1 GUIDELINES FOR FUNCTIONAL PARCELLATION EXTRACTION

389 *Which criterion to use for methods comparison?* To frame the problem, it is necessary to choose  
 390 the criterion used to guide model selection. Note that this is an important yet difficult aspect of  
 391 any unsupervised statistical learning procedure. We studied two different characteristics of functional  
 392 parcellation that are critical to their usage in brain mapping: how well they capture the functional signal  
 393 and how reproducible they are under perturbations of the data. To measure the goodness of fit of the  
 394 functional signal, it is important to distinguish within-subject variance from across-subject variance, as  
 395 only the first kind of variance is minimized when the number of parcels increases. Our probabilistic  
 396 model offers a natural goodness of fit criterion, the log-likelihood; by penalizing it (BIC criterion) or  
 397 by using cross-validation, it is possible to obtain a sound model selection. Our simulations show that  
 398 cross-validation almost systematically outperforms BIC, but we did not notice systematic differences in  
 399 the real dataset. The other important aspect of a brain description is its stability, and we also investigated  
 400 other criteria that the selected number of clusters based on the consistency of parcellations. This approach  
 401 behaved similarly as the others on synthetic data, but provided a much more conservative selection on  
 402 real data ( $K \sim 200$  to  $500$  parcels according to the dataset and method). The fact that reproducibility  
 403 and accuracy yield different decisions for model selection is well known, and has been illustrated in

404 neuroimaging by **LaConte et al.** (2003); this effect is tightly related to the classical bias /variance in  
405 statistics.

406 *Which algorithm to prefer?* Regarding the clustering algorithms themselves, our general finding is that  
407 Ward's algorithm should be preferred, unless a small number of parcels is required. Indeed, spatially-  
408 constrained Ward's clustering outperforms the other techniques in the large  $K$  regime (say,  $K \geq 500$ )  
409 in terms of goodness of fit, while having fair results in terms of reproducibility. With respect to k-  
410 means, it offers the additional advantage of providing spatially connected parcels. In theory, k-means  
411 algorithm should do better in terms of accuracy, but the optimization problem solved by k-means is hard  
412 (non-convex) and thus bound to sub-optimal solutions; as a consequence the greedy approach in Ward's  
413 algorithm outperforms it. Moreover, k-means based parcellations tend to fit data idiosyncrasies and thus  
414 do not generalize well across datasets, as shown in tables 1 and 2. We observed that mixture models  
415 would behave similarly to k-means, since k-means is in fact a constrained Gaussian mixture model with  
416 hard assignments. In a side experiment, we observed that Gaussian mixture models perform consistently  
417 better than k-means, but the difference is tiny and comes with high computational cost.

418 Spectral clustering is not a powerful approach to outline structures in the data. A simple geometric  
419 clustering procedure is as good, and sometimes better in terms of accuracy. The reason is that spectral  
420 clustering is efficient with high SNR data when clusters are easily discriminated, which is not the case  
421 with functional neuroimaging data, where it mostly outlines geometrical structures. Similar observations  
422 were made in **Craddock et al.** (2012). Note however that spectral clustering is even more stable than  
423 geometric clustering, meaning that it captures some structure of the input data.

424 *How many parcels?* It should first be emphasized that choosing the number of parcels in our model is not  
425 exactly the question of deciding how many functional regions can be found in the brain, but how many  
426 piecewise constant models can actually be fit to some fMRI data reliably. The distinction is important,  
427 because some regions, for instance V1, will contain internal functional gradients, such as those related to  
428 retinotopy, orientation sensitivity and ocular dominance. In theory, the function specificity could therefore  
429 be resolved at the level of columns in these regions, but this does not mean that larger structures do not  
430 exist. The conclusions that we draw here are bound to the data that we have used and generalization to  
431 different modalities or contrasts (resting-state fMRI, anatomical connectivity) is not guaranteed.

432 The goodness of fit-related criteria yields high numbers (up to  $K = 5000$  for Ward's clustering, slightly  
433 less for the others, but this may simply reflect a lack of sensitivity of these approaches in the large  $K$   
434 regime, in which Ward's clustering fits the data better), simply indicating that functional activations cannot  
435 easily be represented as piecewise constant models: whether this is an intrinsic feature of brain function  
436 or an impact of cross-subject spatial mismatch or pre-processing artifacts remains an open question. In  
437 the future, the use of brain registration algorithms based on functional data (**Sabuncu et al.**, 2010) may  
438 significantly affect model selection.

439 The reproducibility criterion, on the other hand, peaked at  $K \sim 200$ , meaning that there is probably a  
440 relevant level of description with such a resolution. Thus, when parcellations are used to obtain a model  
441 of brain organization that seeks to characterize individually each parcel, a conservative choice  $K \sim 200$   
442 to 500 should be preferred for the sake of reproducibility. Note that  $K = 200$  is a lower bound on the right  
443 dimensionality, i.e. models with a resolution lower than 200 regions are not flexible enough to represent  
444 functional signals without introducing severe distortions. In particular, anatomical atlases that propose  
445 a decomposition into about 100 regions, are not sufficient to summarize functional signals, some of the  
446 resulting ROIs being very large.

447 Yet, the problem of optimizing the number of parcels remains open and should be addressed in a data-  
448 driven fashion.

## 5.2 CHALLENGES AND FURTHER WORK

449 *The difficulty of model selection on noisy data* It is important to remember that discovering functionally  
 450 homogeneous structures is a hard problem, given that the SNR of the data is low, and that even  
 451 visual inspection would most often be insufficient to define a relevant parcellation. Besides this issue,  
 452 neuroimaging data come with additional difficulties: the data are smooth, which could be accounted for but  
 453 is not in model 6. The other difficulty is that the spatial jitter brought by imperfect spatial normalization,  
 454 and poor matching between functional organization and sulco-gyral anatomy across subjects makes  
 455 this an ill-posed problem, since regions with homogeneous functional characteristics may be slightly  
 456 displaced across individuals, which invalidates the model hypotheses. Both smoothing and jitter break  
 457 the hypotheses of BIC, which yields poor model selection. Cross-validation and reproducibility are more  
 458 resilient to this effect.

459 *Limitations of this experiment* Our experiments are based on two datasets, with a pre-defined set of  
 460 contrasts. We have been able to check that using any of the contrasts or all of them yields qualitatively  
 461 similar results (data not shown). The power of the experiment is that it is based on a relatively large  
 462 number of subjects (67 to 128), so that one can at least conjecture that the between-subject variability  
 463 observed in functional neuroimaging is correctly sampled. Note that the Localizer data come from two  
 464 different scanners, resulting in an un-modeled latent factor. We observed, however, that our conclusions  
 465 were unaltered when performed on a subset of subjects coming from the same scanner (data not shown).

466 The model that we use has several limitations:

- 467 • The parcellation itself is fixed across subjects. While a relaxation to individual dataset has been  
 468 proposed in **Thirion et al.** (2006), such a procedure loses some of the properties of clustering, and  
 469 make model selection much harder.
- 470 • Our model (eq. 6) does not account for spatial effects in the within-parcel covariance, which would  
 471 probably make it more robust to data smoothness and possibly to cross-subject spatial jitter, but the  
 472 computational price to pay for these models is high.
- 473 • It assumes that the true activation signal is piecewise constant. A smooth interpolation scheme  
 474 between parcels might make it more powerful, hence reducing the requirement of large  $K$  values.  
 475 Again, this would increase the complexity of the model fitting.

476 *Suggestions for population-level fMRI modeling* One of the observations made in this study is that the  
 477 problem of the spatial jitter across subjects remains the main limitation that needs to be overcome in order  
 478 to learn appropriate population-level atlases. This should be addressed using procedures such as those  
 479 presented in **Sabuncu et al.** (2010); **Robinson et al.** (2013). Other improvements of the model concern  
 480 the possibility of using not a single parcellation, but several different parcellations and to aggregate the  
 481 results (i.e. the significant effects across subjects) by marginalizing the parcellation as a hidden variable  
 482 of parametric models (**Da Mota et al.**, 2013; **Varoquaux et al.**, 2012). Besides, different parcellation  
 483 schemes could use different values for  $K$ . In particular, Ward's algorithm is a hierarchical algorithm,  
 484 that can actually be used to estimate multi-scale representations of brain activity (see e.g. **Michel et al.**  
 485 (2012); **Orban et al.** (2014)). Specifically **Orban et al.** (2014) suggest that the hierarchical organization  
 486 of nested clusterings obtained from hemodynamic response function would be stable in the population,  
 487 hinting at an intrinsic feature of brain organization. This is an additional asset of this procedure that has  
 488 not been considered in this work but could be used in future applications of brain parcellations.

## DISCLOSURE/CONFLICT-OF-INTEREST STATEMENT

489 The authors declare that the research was conducted in the absence of any commercial or financial  
 490 relationships that could be construed as a potential conflict of interest.

## ACKNOWLEDGMENT

491 We would like to thank Philippe Pinel for providing the Localizer data. The so-called HCP data were  
492 provided by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van  
493 Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the  
494 NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at  
495 Washington University

496 *Funding:* The authors acknowledge support from the ANR grant BrainPedia ANR- 2010-JCJC-1408-01  
497 and the Human Brain Project.

## REFERENCES

- 498 Abraham, A., Dohmatob, E., Thirion, B., Samaras, D., and Varoquaux, G. (2013), Extracting brain regions  
499 from rest fMRI with Total-Variation constrained dictionary learning, in MICCAI - 16th International  
500 Conference on Medical Image Computing and Computer Assisted Intervention - 2013 (Springer,  
501 Nagoya, Japon), nIDA R21 DA034954, SUBSample project from the DIGITEO Institute, France  
502 NiConnect project
- 503 Barch, D. M., Burgess, G. C., Harms, M. P., Petersen, S. E., Schlaggar, B. L., Corbetta, M., et al. (2013),  
504 Function in the human connectome: task-fMRI and individual differences in behavior., *Neuroimage*, 80,  
505 169–189
- 506 Blumensath, T., Behrens, T. E. J., and Smith, S. M. (2012), Resting-state FMRI single subject cortical  
507 parcellation based on region growing., *MICCAI*, 15, Pt 2, 188–195
- 508 Bohland, J. W., Bokil, H., Allen, C. B., and Mitra, P. P. (2009), The brain atlas concordance problem:  
509 quantitative comparison of anatomical parcellations., *PLoS One*, 4, 9, e7200, doi:10.1371/journal.pone.  
510 0007200
- 511 Chaari, L., Forbes, F., Vincent, T., and Ciuciu, P. (2012), Hemodynamic-informed parcellation of fMRI  
512 data in a joint detection estimation framework., *Med Image Comput Comput Assist Interv*, 15, Pt 3,  
513 180–188
- 514 Chen, H., Li, K., Zhu, D., Zhang, T., Jin, C., Guo, L., et al. (2012), Inferring group-wise consistent  
515 multimodal brain networks via multi-view spectral clustering., *Med Image Comput Comput Assist  
516 Interv*, 15, Pt 3, 297–304
- 517 Cieslik, E. C., Zilles, K., Caspers, S., Roski, C., Kellermann, T. S., Jakobs, O., et al. (2012), Is there "one"  
518 dlPFC in cognitive action control? evidence for heterogeneity from co-activation-based parcellation.,  
519 *Cereb Cortex*, doi:10.1093/cercor/bhs256
- 520 Cohen, A. L., Fair, D. A., Dosenbach, N. U. F., Miezin, F. M., Dierker, D., Essen, D. C. V., et al.  
521 (2008), Defining functional areas in individual human brains using resting functional connectivity MRI.,  
522 *Neuroimage*, 41, 1, 45–57, doi:10.1016/j.neuroimage.2008.01.066
- 523 Craddock, R. C., James, G. A., Holtzheimer, P. E., Hu, X. P., and Mayberg, H. S. (2012), A whole  
524 brain fMRI atlas generated via spatially constrained spectral clustering., *Hum Brain Mapp*, 33, 1914,  
525 doi:10.1002/hbm.21333
- 526 Da Mota, B., Fritsch, V., Varoquaux, G., Frouin, V., Poline, J.-B., and Thirion, B. (2013), Enhancing  
527 the Reproducibility of Group Analysis with Randomized Brain Parcellations, in MICCAI - 16th  
528 International Conference on Medical Image Computing and Computer Assisted Intervention - 2013  
529 (Nagoya, Japan)
- 530 Desikan, R. S., Sgonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., et al. (2006), An  
531 automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based  
532 regions of interest., *Neuroimage*, 31, 3, 968–980, doi:10.1016/j.neuroimage.2006.01.021
- 533 Diedrichsen, J., Balsters, J. H., Flavell, J., Cussans, E., and Ramnani, N. (2009), A probabilistic MR atlas  
534 of the human cerebellum., *Neuroimage*, 46, 1, 39–46, doi:10.1016/j.neuroimage.2009.01.045

- 535 Eickhoff, S. B., Bzdok, D., Laird, A. R., Roski, C., Caspers, S., Zilles, K., et al. (2011), Co-activation  
536 patterns distinguish cortical modules, their connectivity and functional differentiation., *Neuroimage*,  
537 57, 938, doi:10.1016/j.neuroimage.2011.05.021
- 538 Eickhoff, S. B., Rottschy, C., Kujovic, M., Palomero-Gallagher, N., and Zilles, K. (2008), Organizational  
539 principles of human visual cortex revealed by receptor mapping., *Cereb Cortex*, 18, 11, 2637–2645,  
540 doi:10.1093/cercor/bhn024
- 541 Fischl, B., Rajendran, N., Busa, E., Augustinack, J., Hinds, O., Yeo, B. T. T., et al. (2008), Cortical folding  
542 patterns and predicting cytoarchitecture., *Cereb Cortex*, 18, 8, 1973–1980, doi:10.1093/cercor/bhm225
- 543 Flandin, G., Kherif, F., Pennec, X., Malandain, G., Ayache, N., Poline, J.-B., et al. (2002), Improved  
544 detection sensitivity in functional MRI data using a brain parcelling technique., in *MICCAI* (1), 467–  
545 474
- 546 Ghosh, S., Keshavan, A., and Langs, G. (2013), Predicting treatment response from resting state fmri  
547 data: Comparison of parcellation approaches, in *Pattern Recognition in Neuroimaging (PRNI)*, 2013  
548 International Workshop on, 225–228, doi:10.1109/PRNI.2013.64
- 549 Golland, P., Golland, Y., and Malach, R. (2007), Detection of spatial activation patterns as unsupervised  
550 segmentation of fmri data., *Med Image Comput Comput Assist Interv*, 10, Pt 1, 110–118
- 551 Hanson, S. J., Rebecchi, R., Hanson, C., and Halchenko, Y. O. (2007), Dense mode clustering in brain  
552 maps., *Magn Reson Imaging*, 25, 9, 1249–1262, doi:10.1016/j.mri.2007.03.013
- 553 Johnson, S. C. (1967), Hierarchical clustering schemes, *Psychometrika*, 32, 3, 241–254
- 554 Kahnt, T., Chang, L. J., Park, S. Q., Heinzle, J., and Haynes, J.-D. (2012), Connectivity-based parcellation  
555 of the human orbitofrontal cortex., *J Neurosci*, 32, 18, 6240–6250, doi:10.1523/JNEUROSCI.0257-12.  
556 2012
- 557 Kiviniemi, V., Starck, T., Remes, J., Long, X., Nikkinen, J., Haapea, M., et al. (2009), Functional  
558 segmentation of the brain cortex using high model order group pica, *Human brain mapping*, 30, 12,  
559 3865–3886
- 560 Klein, A. and Tourville, J. (2012), 101 labeled brain images and a consistent human cortical labeling  
561 protocol., *Frontiers in neuroscience*, 6, December, 171, doi:10.3389/fnins.2012.00171
- 562 LaConte, S., Anderson, J., Muley, S., Ashe, J., Frutiger, S., Rehm, K., et al. (2003), The evaluation of  
563 preprocessing choices in single-subject BOLD fMRI using NPAIRS performance metrics., *Neuroimage*,  
564 18, 10
- 565 Lashkari, D., Sridharan, R., Vul, E., Hsieh, P.-J., Kanwisher, N., and Golland, P. (2012), Search for  
566 patterns of functional specificity in the brain: a nonparametric hierarchical bayesian model for group  
567 fMRI data., *Neuroimage*, 59, 1348, doi:10.1016/j.neuroimage.2011.08.031
- 568 Lashkari, D., Vul, E., Kanwisher, N., and Golland, P. (2010), Discovering structure in the space of fMRI  
569 selectivity profiles., *Neuroimage*, 50, 1085, doi:10.1016/j.neuroimage.2009.12.106
- 570 Mazziotta, J., Toga, A., Evans, A., Fox, P., Lancaster, J., Zilles, K., et al. (2001), A probabilistic atlas  
571 and reference system for the human brain: International consortium for brain mapping (icbm)., *Philos*  
572 *Trans R Soc Lond B Biol Sci*, 356, 1412, 1293–1322, doi:10.1098/rstb.2001.0915
- 573 Meng, X.-L. and van Dyk, D. (1998), Fast em-type implementations for mixed effects models, *Journal of*  
574 *the Royal Statistical Society. Series B (Statistical Methodology)*, 60, 3, 559–578
- 575 Michel, V., Gramfort, A., Varoquaux, G., Eger, E., Keribin, C., and Thirion, B. (2012), A supervised  
576 clustering approach for fMRI-based inference of brain states, *Pattern Recognition*, 45, 6, 2041–2049
- 577 Ng, A. Y., Jordan, M. I., and Weiss, Y. (2001), On spectral clustering: Analysis and an algorithm, in *NIPS*  
578 (MIT Press), 849–856
- 579 Nieto-Castanon, A., Ghosh, S. S., Tourville, J. A., and Guenther, F. H. (2003), Region of interest based  
580 analysis of functional imaging data., *Neuroimage*, 19, 4, 1303–1316
- 581 Orban, P., Doyon, J., Petrides, M., Mennes, M., Hoge, R., and Bellec, P. (2014), The richness of task-  
582 evoked hemodynamic responses defines a pseudohierarchy of functionally meaningful brain networks.,  
583 *Cereb Cortex*, doi:10.1093/cercor/bhu064
- 584 Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011), Scikit-learn:  
585 Machine learning in Python, *Journal of Machine Learning Research*, 12, 2825–2830



- 586 Pinel, P., Thirion, B., Meriaux, S., Jobert, A., Serres, J., Bihan, D. L., et al. (2007), Fast reproducible  
587 identification and large-scale databasing of individual functional cognitive networks., *BMC Neurosci*,  
588 8, 91, doi:10.1186/1471-2202-8-91
- 589 Robinson, E. C., Jbabdi, S., Andersson, J., Smith, S., Glasser, M. F., Van Essen, D. C., et al. (2013),  
590 Multimodal surface matching: Fast and generalisable cortical registration using discrete optimisation,  
591 in *Information Processing in Medical Imaging*, 475–486
- 592 Roca, P., Tucholka, A., Rivire, D., Guevara, P., Poupon, C., and Mangin, J.-F. (2010), Inter-subject  
593 connectivity-based parcellation of a patch of cerebral cortex., *Med Image Comput Comput Assist Interv*,  
594 13, Pt 2, 347–354
- 595 Sabuncu, M. R., Singer, B. D., Conroy, B., Bryan, R. E., Ramadge, P. J., and Haxby, J. V. (2010),  
596 Function-based intersubject alignment of human cortical anatomy., *Cereb Cortex*, 20, 1, 130–140,  
597 doi:10.1093/cercor/bhp085
- 598 Saxe, R., Brett, M., and Kanwisher, N. (2006), Divide and conquer: a defense of functional localizers.,  
599 *Neuroimage*, 30, 4, 1088–96; discussion 1097–9, doi:10.1016/j.neuroimage.2005.12.062
- 600 Schwarz, G. (1978), Estimating the Dimension of a Model, *The Annals of Statistics*, 6, 2, 461–464,  
601 doi:10.2307/2958889
- 602 Shattuck, D. W., Mirza, M., Adisetiyo, V., Hojatkashani, C., Salamon, G., Narr, K. L., et al. (2008),  
603 Construction of a 3d probabilistic atlas of human cortical structures., *Neuroimage*, 39, 3, 1064–1080,  
604 doi:10.1016/j.neuroimage.2007.09.031
- 605 Shi, J. and Malik, J. (2000), Normalized cuts and image segmentation, Technical report
- 606 Simon, O., Kherif, F., Flandin, G., Poline, J.-B., Rivire, D., Mangin, J.-F., et al. (2004), Automated  
607 clustering and functional geometry of human parietofrontal networks for language, space, and number.,  
608 *Neuroimage*, 23, 3, 1192–1202, doi:10.1016/j.neuroimage.2004.09.023
- 609 Thirion, B., Flandin, G., Pinel, P., Roche, A., Ciuciu, P., and Poline, J.-B. (2006), Dealing with the  
610 shortcomings of spatial normalization: multi-subject parcellation of fmri datasets., *Hum Brain Mapp*,  
611 27, 8, 678–693, doi:10.1002/hbm.20210
- 612 Tucholka, A., Thirion, B., Perrot, M., Pinel, P., Mangin, J.-F., and Poline, J.-B. (2008), Probabilistic  
613 anatomo-functional parcellation of the cortex: how many regions?, *MICCAI*, 11, Pt 2, 399–406
- 614 Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., et al. (2002),  
615 Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of  
616 the mni mri single-subject brain., *Neuroimage*, 15, 1, 273–289, doi:10.1006/nimg.2001.0978
- 617 Varoquaux, G., Gramfort, A., Pedregosa, F., Michel, V., and Thirion, B. (2011), Multi-subject dictionary  
618 learning to segment an atlas of brain spontaneous activity., *Inf Process Med Imaging*, 22, 562–573
- 619 Varoquaux, G., Gramfort, A., and Thirion, B. (2012), Small-sample brain mapping: sparse recovery on  
620 spatially correlated designs with randomization and clustering, in *ICML*
- 621 Varoquaux, G., Sadaghiani, S., Pinel, P., Kleinschmidt, A., Poline, J. B., and Thirion, B. (2010), A  
622 group model for stable multi-subject ica on fmri datasets., *Neuroimage*, 51, 1, 288–299, doi:10.1016/j.  
623 neuroimage.2010.02.010
- 624 Varoquaux, G., Schwartz, Y., Pinel, P., and Thirion, B. (2013), Cohort-level brain mapping: learning  
625 cognitive atoms to single out specialized regions, in *Information Processing in Medical Imaging*, 438–  
626 449
- 627 Vinh, N. X., Epps, J., and Bailey, J. (2009), Information theoretic measures for clusterings comparison: is  
628 a correction for chance necessary?, in *ICML*, 1073
- 629 Ward, J. H. (1963), Hierarchical grouping to optimize an objective function, *Journal of the American*  
630 *Statistical Association*, 58, 301, 236–244
- 631 Wig, G. S., Laumann, T. O., Cohen, A. L., Power, J. D., Nelson, S. M., Glasser, M. F., et al. (2013),  
632 Parcellating an individual subject’s cortical and subcortical brain structures using snowball sampling of  
633 resting-state correlations., *Cereb Cortex*, doi:10.1093/cercor/bht056
- 634 Yeo, B. T. T., Krienen, F. M., Sepulcre, J., Sabuncu, M. R., Lashkari, D., Hollinshead, M., et al.  
635 (2011), The organization of the human cerebral cortex estimated by intrinsic functional connectivity., *J*  
636 *Neurophysiol*, 106, 3, 1125–1165, doi:10.1152/jn.00338.2011
- 637 Yu, S. X. and Shi, J. (2003), Multiclass spectral clustering, in *Computer Vision, 2003. Proceedings. Ninth*  
638 *IEEE International Conference on*, 313–319