

Multi-level bootstrap analysis of stable clusters in resting-state fMRI

Pierre Bellec^{a,*}, Pedro Rosa-Neto^a, Oliver C. Lyttelton^a, Habib Benali^{b,c}, Alan C. Evans^a

^a McConnell Brain Imaging Center, Montreal Neurological Institute, McGill University, Montréal, Québec, Canada

^b Inserm, UMR_S 678, Laboratoire d'Imagerie Fonctionnelle, Paris, France

^c UPMC Univ Paris 06, UMR_S 678, Laboratoire d'Imagerie Fonctionnelle, Paris, France

ARTICLE INFO

Article history:

Received 12 November 2009

Revised 13 February 2010

Accepted 28 February 2010

Available online 10 March 2010

Keywords:

Bootstrap

Clustering

Functional MRI

Hierarchical clustering

k-Means

Multi-level analysis

Resting-state networks

Stability analysis

ABSTRACT

A variety of methods have been developed to identify brain networks with spontaneous, coherent activity in resting-state functional magnetic resonance imaging (fMRI). We propose here a generic statistical framework to quantify the stability of such resting-state networks (RSNs), which was implemented with *k*-means clustering. The core of the method consists in bootstrapping the available datasets to replicate the clustering process a large number of times and quantify the stable features across all replications. This bootstrap analysis of stable clusters (BASC) has several benefits: (1) it can be implemented in a multi-level fashion to investigate stable RSNs at the level of individual subjects and at the level of a group; (2) it provides a principled measure of RSN stability; and (3) the maximization of the stability measure can be used as a natural criterion to select the number of RSNs. A simulation study validated the good performance of the multi-level BASC on purely synthetic data. Stable networks were also derived from a real resting-state study for 43 subjects. At the group level, seven RSNs were identified which exhibited a good agreement with the previous findings from the literature. The comparison between the individual and group-level stability maps demonstrated the capacity of BASC to establish successful correspondences between these two levels of analysis and at the same time retain some interesting subject-specific characteristics, e.g. the specific involvement of subcortical regions in the visual and fronto-parietal networks for some subjects.

© 2010 Elsevier Inc. All rights reserved.

Introduction

Functional magnetic resonance imaging (fMRI) provides a measure of the vascular consequences of neuronal activity in the whole brain. In a seminal work on resting-state fMRI, Biswal et al. (1995) demonstrated that the spontaneous fMRI fluctuations exhibited a significant level of spatial coherence within the sensorimotor network. Since this initial experiment, functional connectivity analysis has been used to investigate other brain resting-state networks (RSNs), such as the visual, auditory and default-mode networks (Perlberg and Marrelec, 2008). The RSNs identified in fMRI were found to be largely consistent with other measurements of the brain organization such as task-evoked activations (Toro et al., 2008; Smith et al., 2009), diffusion imaging (Damoiseaux and Greicius, 2009), maps of anatomical connectivity derived using retrograde tracers in macaques (Vincent et al., 2007; Margulies et al., 2009) and electrophysiology either on the scalp (Laufs et al., 2003) or on the cortex (Shmuel and Leopold, 2008). The possibility opened by resting-state fMRI to fully characterize the functional organization of the brain with a single and simple experiment already proved very useful in

both clinical and basic neuroscience and is quickly growing popular (Fox and Raichle, 2007; Greicius, 2008; Broyd et al., 2009).

A variety of algorithms have been proposed to automatically identify RSNs in fMRI, including principal component analysis (Zhong et al., 2009), independent component analysis (ICA) (McKeown et al., 1998) and various clustering algorithms such as *k*-means (Baumgartner et al., 1998), hierarchical clustering (Cordes et al., 2002), normalized cut-graph (van den Heuvel et al., 2008), self-organizing maps and neural gas (Meyer-Baese et al., 2004). The cost function or the heuristic that actually drove these techniques varied across methods and implementations, yet they all resulted in a set of spatial maps which were loosely termed RSNs. An important word of caution is that ICA and clustering algorithms would find RSNs in randomly generated datasets. As a consequence, it is critical to verify that the RSNs derived in a particular experiment are real in an objective manner (Smith and Dubes, 1980). Unfortunately, RSNs derived in fMRI cannot be easily validated against some “ground truth” results, simply because there is no imaging technique other than fMRI and positron emission tomography that can access functional networks distributed in the whole human brain. As an alternative, it was proposed early in the clustering literature (Raghavan, 1982) to address the following question: *how stable would the RSNs be if the experiment was to be replicated?* That question actually entails two challenges. First, it may not be possible to replicate the experiment at all, or at least not a large number of times. Recently, some test-retest database has been used to investigate the

* Corresponding author.

E-mail address: pbellec@bic.mni.mcgill.ca (P. Bellec).

reproducibility of RSNs (Chen et al., 2008; Shehzad et al., 2009; Zuo et al., 2010) and, even though five sessions at most were acquired, the collection of such database represented a considerable effort. As most study will scan the subjects only once, some statistical techniques need to circumvent the actual replications of the datasets. Second, some measures have to be derived in order to quantify the stable features of the replicated RSNs. This task is difficult because the order of the RSNs is generally arbitrary within each particular experiment, preventing the direct comparison of RSNs across replications.

A number of stability methods were developed in the context of group analysis of ICA results in fMRI, see (Calhoun et al., 2009) for a comprehensive review on this issue. These methods were all based on a generic technique called component matching. Component matching is performed by applying a clustering algorithm to an ensemble of RSNs generated through multiple experiments. Each one of the resulting clusters is composed of RSNs from all experiments which shared similar spatial distributions. A voxelwise t -test could then be used to identify the stable regions within each cluster of RSNs, yielding one stable RSN per cluster. Esposito et al. (2005) and De Luca et al. (2006) applied component matching to study the stability of RSNs across different subjects. The stable RSNs thus formed a group-level summary of the individual results. To assess the stability of these group-level RSNs, Perlberg et al. (2008) generated multiple surrogate groups using bootstrap resampling of the subjects in the population. Component matching was then applied to the replicated group-level RSNs in order to study their stability across the surrogate groups. A similar approach (Damoiseaux et al., 2006; Smith et al., 2009) was employed for group-level RSNs generated through tensorial ICA (Beckmann and Smith, 2005). The stability methods based on component matching unfortunately suffers from two limitations. First, in the works we just reviewed, it was not possible to derive individual maps corresponding to each of the group-level RSNs. Some group-level ICA techniques have the potential to establish correspondences between individual and group-level analysis, i.e. back-reconstruction in temporally-concatenated group ICA (Calhoun et al., 2009) or dual-regression (Filippini et al., 2009). However, and despite some initial attempts (Himberg et al., 2004), the stability of the individual maps themselves was not assessed unless some test–retest database was available (Zuo et al., 2010). Deriving stable individual counterparts to the group-level RSNs would be of particular interest in clinical applications, where the RSNs could serve of biomarkers for various diseases (Greicius, 2008). A second limitation of component matching is that the stability is measured within a cluster of RSNs that were matched precisely because of their large spatial similarity. Such circularity in the definition of the stability yields an upward bias that has recently been evidenced and needs to be corrected using an additional statistical procedure (Langers, 2010).

In this paper, we propose a new framework called bootstrap analysis of stable clusters (BASC) to study the stability of RSNs in fMRI. Assessing stability is a central issue for exploratory methods in general, whether it be a component analysis or any type of clustering technique. We focussed here on the k -means clustering algorithm to demonstrate the feasibility of BASC because of the low computational cost of this technique. Our statistical approach is still versatile and could be applied to any algorithm working on individual time series. BASC belongs to the family of cluster ensemble methods, see (Fred and Lourenço, 2008) for a review, and is more specifically a generalization of the evidence accumulation algorithm (Fred and Jain, 2005). The originality of our approach is to provide a probabilistic formulation of the stability measure of the clustering process. In line with approaches developed in the field of phylogenetic analysis (Kerr and Churchill, 2001; Suzuki and Shimodaira, 2006), the stability measure can be estimated both for individual and group fMRI datasets using well-adapted bootstrap methods. The group-level BASC is actually based on the results of the individual-level BASC, making this approach analogous to the hierarchical multi-level framework used

for general linear model analysis where parameter and uncertainty estimates are passed from one level of the analysis to the other (Woolrich et al., 2004). It is moreover possible to combine both levels of analysis by generating individual stability maps associated with each group-level RSN. BASC also provides a principled way to select the parameters of the clustering algorithm, by maximizing a global stability contrast measure. The ability of the multi-level BASC method to accurately control for the stability of the RSNs was investigated using fully synthetic groups of fMRI time series. The method was then applied to a real 43 subject database of resting-state fMRI in order to compare the identified stable group RSNs to those previously reported in the literature. The individual stability maps associated with each group RSN were examined to demonstrate the ability of BASC to establish a successful correspondence between individual-level and group-level analysis.

Methods

We start by introducing the principle of BASC in a very general setting, before considering the particular context of individual-level and group-level fMRI time series. A procedure to generate maps of cluster stability for both levels of analysis is then provided. Strategies to choose the parameters of the multi-level BASC method are presented and the conditions of validity of the bootstrap approximation are discussed.

Bootstrap analysis of stable clusters (BASC)

In the most generic terms, a clustering algorithm is an operation which takes a dataset \mathbf{y} as input and produces a partition \mathcal{P} of the space into K non-overlapping subsets called clusters. The most famous example of a clustering process in neuroscience is arguably the experiment performed by Brodmann (Brodmann, 1909), in which case the dataset \mathbf{y} was a map of the cortical layers derived with Nissl stain, and the partition \mathcal{P} was the subdivision of the cortex into Brodmann's areas. The clustering was done in a subjective manner based on the observed features of the cortical layers. In modern days, this operation can be performed by one of the many automated clustering algorithms that have been proposed in the literature, see (Jain, 2010) for an excellent review. Such a clustering operation ϕ attempts to optimize the similarity of the data associated within the regions of each cluster in some sense which depends on the employed algorithm.

To assess statistically the random variations that might occur in the clustering results, the partition \mathcal{P} is modeled as one sample of a stochastic process consisting of two distinct steps:

$$\mathbf{Y} \xrightarrow{f} \mathbf{y} \xrightarrow{\phi} \Phi(\mathbf{y}). \quad (1)$$

The first step of the process is the generation of the dataset \mathbf{y} , which is typically done through a complex procedure such as an fMRI acquisition. This is formalized by assuming that \mathbf{y} is a sample of a random variable \mathbf{Y} with probability distribution function (pdf) f . The second step is the application of the clustering algorithm ϕ to \mathbf{y} , resulting in a partition which is represented through an adjacency matrix $\Phi(\mathbf{y})$ where $\Phi_{ij}(\mathbf{y})$ equals 1 if the regions i and j belong to the same cluster in the partition and equals 0 otherwise. The clustering process ϕ may itself have some stochastic behaviour, e.g. some algorithms are based on a random initialization.

The (pairwise) stability of a stochastic clustering process can be captured through the probability that a given pair of regions i and j belong to the same cluster:

$$S_{ij} = \Pr\left(\Phi_{ij}(\mathbf{y}) = 1 \mid \mathbf{Y} \xrightarrow{f} \mathbf{y}\right). \quad (2)$$

The matrix \mathbf{S} , called the stability matrix, quantifies the stable features of the stochastic clustering process. It is however not itself a clustering, in the sense that it does not provide a partition of the space into stable clusters. The key idea of the evidence accumulation algorithm (Fred and Jain, 2005) was to formulate the search for stable clusters as a clustering problem on the stability matrix. Stable clusters could indeed be defined as a partition of the space composed of regions that had a high probability of being clustering together, i.e. high values in the matrix \mathbf{S} . Many clustering algorithms, e.g. hierarchical agglomerative clustering (HAC), could be applied directly on an arbitrary similarity matrix and were thus suitable to solve this problem.

It may not be possible in general to derive a closed expression for the stability matrix because the pdf f and the clustering process ϕ do not follow simple parametric forms. Instead, if drawing independent samples $(\mathbf{y}^{(b)})_{b=1}^B$ from \mathbf{Y} was possible, the stability matrix \mathbf{S} could be approximated by Monte-Carlo estimation, which would simply consist of the average of all the sampled adjacency matrices:

$$\hat{S}_{ij}^{MC} = B^{-1} \sum_{b=1}^B \Phi_{ij}(\mathbf{y}^{(b)}) \doteq S_{ij}, \quad (3)$$

where \doteq means that the two terms are asymptotically equal as B tends toward infinity. This approximation of the stability matrix was already proposed by several authors under different names (Ben-Hur et al., 2002; Fred and Jain, 2005; Steinley, 2008; Fred and Lourenço, 2008), even though the relationship with the probabilistic formulation in Eq. 2 was not made explicitly.

In many applications, including individual fMRI resting-state data, it is not possible to derive independent data samples $(\mathbf{y}^{(b)})_{b=1}^B$ for B greater than 10, and it may even be that $B=1$. The Monte-Carlo estimate of stability from Eq. (3) would not achieve a reasonably accurate approximation in such situation. It may however be possible to resort to a non-parametric estimation of the distribution f in order to derive an estimate of the stability. The bootstrap is such a non-parametric technique which comes in different variations, adapted to different types of data structure (Efron and Tibshirani, 1994). The bootstrap uses a single data sample \mathbf{y} in order to build an approximation $\hat{f}_{\mathbf{y}}$ of the pdf f . It is then possible to derive a bootstrap estimation of the stability matrix:

$$\hat{S}_{ij}^{\text{boot}} = Pr\left(\Phi_{ij}(\mathbf{y}^{(*)}) = 1 \mid \mathbf{y} \xrightarrow{\hat{f}_{\mathbf{y}}} \mathbf{y}^{(*)}\right). \quad (4)$$

As for the real pdf f , the bootstrap pdf $\hat{f}_{\mathbf{y}}$ does not generally have a closed, analytical form. The pdf $\hat{f}_{\mathbf{y}}$ is rather defined implicitly through a procedure to draw independent samples from $\hat{f}_{\mathbf{y}}$. The classical bootstrap estimator is therefore a Monte-Carlo approximation of $\hat{S}_{ij}^{\text{boot}}$.

$$\hat{S}_{ij} = B^{-1} \sum_{b=1}^B \Phi_{ij}(\mathbf{y}^{(b)}) \doteq \hat{S}_{ij}^{\text{boot}}, \quad (5)$$

where $(\mathbf{y}^{(b)})_{b=1}^B$ are B independent samples drawn from $\hat{f}_{\mathbf{y}}$ using bootstrap.

Individual-level BASC

In the case of individual fMRI time series, the dataset \mathbf{y} is a time \times space array of size $T \times R$, where T is the number of time points and R is the number of regions. The data-generating process of the time series involves scanning a real subject in resting-state and a number of preprocessing steps performed on the raw fMRI volumes. Amongst the large collection of existing clustering algorithms available to search for individual clusters, we decided to investigate

the behaviour of the k -means algorithm¹ because it is very standard and fast to derive. The k -means algorithm directly applies to the time series and has a single parameter: the number of clusters K .

In order to build a non-parametric approximation of the data-generating process, it is important to note that the fMRI time series exhibit dependencies in both time and space (Bullmore, 2000). This makes the most standard version of the bootstrap non-applicable, yet a variant called circular block bootstrap (CBB) (Efron and Tibshirani, 1994) is adapted to this case. The CBB draws independently temporal blocks of length h from the time series \mathbf{y} to respect the temporal dependencies in the data. A block falling at the end of the time series is completed by points from the beginning, hence the ‘‘circular’’ in CBB. The resampled time blocks are pasted together in order to form a surrogate time series $\mathbf{y}^{(*)}$ with the same temporal dimension as the original. The same time blocks are used for all the regions to preserve spatial correlation and CBB formally leads to consistent confidence intervals of spatial correlations (Lahiri, 2003). This scheme has been demonstrated to be efficient in replicating the distribution of spatial correlations in real fMRI time series (Bellec et al., 2008). As the k -means algorithm is driven by the spatial correlation present in the data, that scheme should provide a well-behaved approximation of the individual stability. By applying the bootstrap approximation of Eq. 5, an individual stability matrix $\hat{\mathbf{I}}$ is estimated using B bootstrap samples generated through CBB.² Note that, for each replication of the individual clustering process, some new bootstrap surrogate time series are generated and a k -means algorithm is applied based on random initial points. As such, the bootstrap stability integrates all sources of random variations, mixing the generation of the dataset with the potential algorithmic variability. The BASC process for individual time series as well as the list of parameters involved is recapitulated in Fig. 1.

Group-level BASC

At the group level, an fMRI database is a collection of datasets $(\mathbf{y}^{(n)})_{n=1}^N$ acquired on a group of N different subjects. The subjects are recruited in the study following a data-generating process which controls for some factors of non-interest, e.g. the group is balanced regarding the number of men and women or the number of left- and right-handed subjects. A subpopulation of subjects that fits a particular profile regarding those factors is called a stratum, e.g. right-handed women. Within each of the strata, the subjects are simply independent samples of a population that meet some general criterion, e.g. healthy subjects with no history of psychological disorder within a certain age range.

We now describe a procedure to build a group-level clustering that will give an accurate picture of the most stable features of the individual-level clusterings. The key idea of the EAC algorithm, which is to run a clustering algorithm on the stability matrix in order to define stable cluster maps, can be applied to this end. Specifically, a group-level cluster should be composed of pairs of regions that maximize the average probability of belonging to the same cluster at the individual level. This last quantity is captured by the average individual stability matrix $\hat{\mathbf{J}}$:

$$\hat{J}_{ij} = N^{-1} \sum_{n=1}^N \hat{I}_{ij}^{(n)}, \quad (6)$$

¹ The clustering algorithm was an in-house implementation in Matlab of a k -means classification, as described in (Duda et al., 2000). The centroids of the clusters were initialized using a random subset of the time series. Whenever a cluster became empty in the classification process, it was replaced by the singleton time series located the further away from its centroid. To avoid falling too often in local minima, the k -means was actually iterated 5 times with different initializations and the result with the lowest within-cluster inertia was selected.

² In Eq. 5 the stability matrix is denoted by a generic notation \mathbf{S} , yet some specific notations \mathbf{I} and \mathbf{G} are introduced for the individual- and group-level stability matrices.

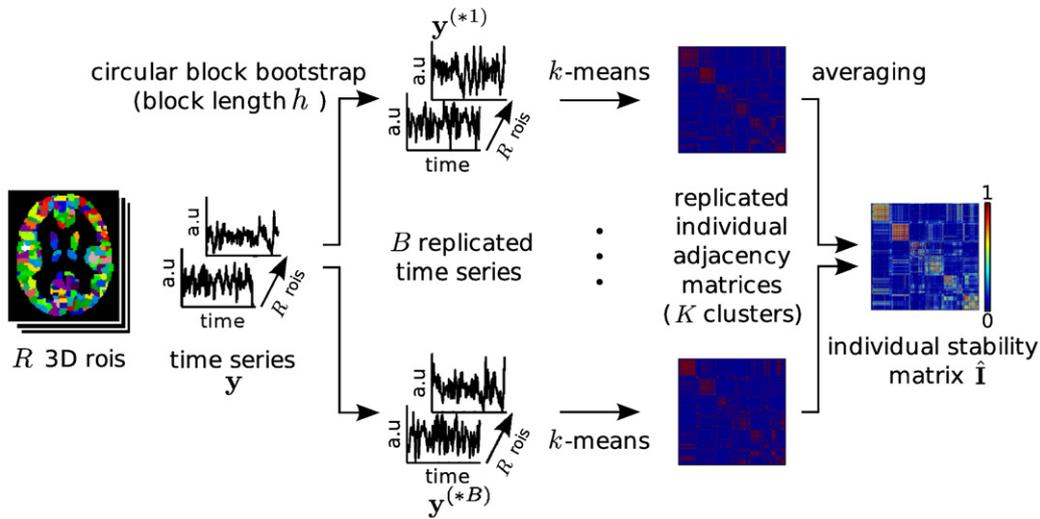


Fig. 1. Individual-level stability. Bootstrap estimation of the stability of the individual clustering of fMRI time series.

where $\hat{\mathbf{I}}^{(n)}$ is the bootstrap estimate of the individual stability matrix associated with the original dataset $\mathbf{y}^{(n)}$ of subject n . To build the group clusters, a hierarchical agglomerative clustering³ (HAC) is applied on the matrix $\hat{\mathbf{J}}$ as it would be on any arbitrary similarity matrix. The only parameter of the HAC algorithm is the number of clusters L .

To apply a BASC to the group-level clustering process, the bootstrap procedure has to approximate the distribution of the finite sample of subjects that is used to derive $\hat{\mathbf{J}}$ and the subsequent HAC. This means that an appropriate bootstrap scheme has to mimic the random variations of the subjects recruited within the group. In each stratum of the population, the subjects are independent and identically distributed. The standard bootstrap scheme (Efron and Tibshirani, 1994) can therefore be applied: it consists of drawing subjects with replacement from the real sample, in order to generate a surrogate stratum featuring the same number of subjects as the original. By repeating this step on every strata, the so-called stratified bootstrap (SB) scheme generates a surrogate database of fMRI time series $(\mathbf{y}^{(n,*)})_{n=1}^N$. The surrogate time series are then plugged in the individual-level clustering procedure to derive a replication $\hat{\mathbf{J}}^{(*)}$ of the average individual stability matrix, which in turn is used to replicate the group-level clustering. Note that, by contrast with the CBB scheme, the replicated time series in SB are identical to the original time series. The difference between the original and bootstrap datasets is that some subjects may be absent of a particular bootstrap sample, while others may be present multiple times. For this reason, the individual-level BASC does not have to be actually replicated on each bootstrap surrogate population. The individual stability matrices are rather generated once and the average individual stability matrix only is recomputed for each surrogate population based on the list of subjects included in that sample.

The SB will provide a consistent estimate of the distribution of the average individual stability $\hat{\mathbf{J}}$ (Sitter, 1992). More elaborate versions of the bootstrap have been developed for complicated stratified data, e.g. (Nigam and Rao, 1996), but the SB employed here was adapted to the

case of a small number of strata with a large number of individuals in each stratum. The SB is repeated with C bootstrap samples to generate an estimation $\hat{\mathbf{G}}$ of the group-level stability matrix. The BASC process for a group of fMRI time series as well as the list of parameters involved is recapitulated in Fig. 2.

Stable clusters and stability maps

Once a stability matrix has been estimated, it can be fed in a HAC to derive stable clusters, i.e. clusters that are optimally adapted to the stable features of the stochastic clustering process. This can be done both at the individual and group levels. The stable clusters do not convey by themselves any quantitative information regarding stability. A stability map can thus be derived by combining the clusters with the stability matrix. For each region in the brain, the stability score is defined as the average stability of that region with all the regions within the cluster. Iterating this process over every region yields a 3D stability map covering the whole brain. A low score in the map can be interpreted in two ways: (1) the values in the stability matrix are low or, (2) the cluster is poorly adapted to the stability matrix. The stable clusters are a good choice to generate stability maps, in the sense that they are built to minimize the effect of (2). Formally, at the group level, a HAC is applied on the stability matrix $\hat{\mathbf{G}}$ to derive M stable group clusters, or RSNs. A group stability map is generated for each cluster C :

$$w(C)_i = c_i^{-1} \sum_{j \in C, j \neq i} \hat{G}_{ij}, \quad (7)$$

for every region i in the brain.

At the individual level, the choice of the target clusters is not straightforward. Relying on the individual stable clusters is feasible, yet there would be no correspondence between clusters coming from different subjects. The alternative would be to select a single target clustering common to all subjects. The stable group clusters are natural candidates for this purpose, as they were built to reflect the average features of the individual clusters, which will minimize the effect of (2) on average. In this approach, the individual maps may still depart from the target group RSN. It is indeed possible that a subpart of a group RSN may exhibit zero stability for a particular subject, while a part of the brain outside of the group RSN may exhibit high stability with the regions of the group RSN. The ambiguity between (1) and

³ The HAC was an in-house implementation in Matlab of a sequential, agglomerative, non-overlapping, hierarchical algorithm based on a similarity matrix. The similarity between clusters was defined as the unweighted average stability between the regions of the clusters, see the so-called UPGMA criterion in (Day and Edelsbrunner, 1984). The algorithm started with single regions as clusters. At each step, the two most similar clusters were merged together, until a specified number of clusters was reached.

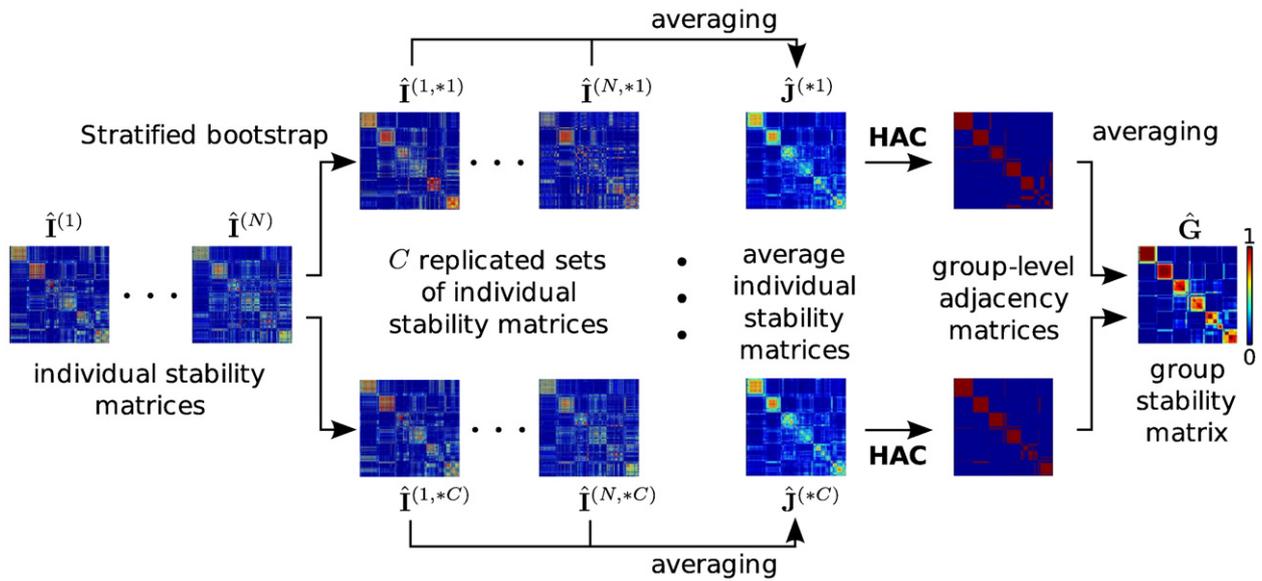


Fig. 2. Group-level stability. Bootstrap estimation of the stability of the group-level clustering.

(2) will still remain, in the sense that subjects that strongly depart from the group average will appear with low stability maps even though their stability matrix may exhibit large values. Individual stability maps based on the group clustering therefore achieve an implicit trade-off between the establishment of a clear correspondence between subjects and the possibility to capture subject-specific features.

Choice of the BASC parameters

The choice and impact of the parameters involved in the multi-level BASC procedure are reviewed below. The block length h in the CBB of individual time series needs to be adapted to the range of temporal dependencies present in the data as well as to the number of volumes T . It has been shown that this parameter has only minor impact on the bootstrap distribution of the spatial correlations as long as h is greater than 1, roughly of the order of \sqrt{T} (Bellec et al., 2008). Rather than setting a fixed value for h , the variability of the BASC results associated with h was included in the measure of stability. This was implemented by drawing randomly a different value of h for each bootstrap sample of time series with a uniform distribution within an interval of reasonable values.

The Monte-Carlo estimation of the bootstrap stability has a parameter which is the number of bootstrap samples B . The accuracy of the Monte-Carlo approximation (Eq. (5)) is different from the accuracy of the bootstrap approximation itself (Eq. (4)). The former level of approximation is a well studied problem (Li et al., 2009), while the latter is more difficult to assess as discussed in the next section. Specifically, B can be selected in order to achieve any desired level of accuracy. For example with $B = 100$ (resp. $B = 1000$) the variations on the stability estimate due to Monte-Carlo sampling are smaller than 0.1 (resp. 0.05) in 95% of the cases under a quite conservative approach (i.e. the variability is smaller in practice). See [Supplementary Material A](#) for more details on that issue.

The most important parameters of the multi-level BASC algorithm are the individual, group and final number of clusters, denoted by K , L and M respectively. It has long been proposed to select the number of clusters in order to maximize the stability of the clustering, e.g. (Jain, 1987; Ben-Hur et al., 2002), yet the way the stability was exactly measured varied across study, see (Jain, 2010) for a review. Stability

was demonstrated to be an effective principle to select the number of clusters in a number of practical situations and has established itself as a tool for model selection in cluster analysis. The theoretical foundations of this approach are however still subject to debate,⁴ e.g. (Ben-David et al., 2007). A global stability criterion was straightforward to implement in the BASC framework, which was designed to provide measures of stability. Specifically, a modified version of the silhouette criterion (Rousseeuw, 1987) was implemented on the group stability matrix using the stable group RSNs as a reference. Let i be a brain region belonging to the RSN C_i . The rationale of the criterion is to compare the average group-level stability of i with every other regions in C_i to the maximal average stability of i with regions from other RSNs. The stability contrast⁵ is formally defined as:

$$\sigma(K, L, M) = R^{-1} \sum_{i=1}^R \left(w(C_i)_i - \arg \max_{C \neq C_i} (w(C)_i) \right), \quad (8)$$

where $w(C)_i$ was defined in Eq. (7). The possible σ values range between -1 and 1 . If σ is close to 1 , it indicates that the within-cluster stability is much larger than the between-cluster stability and the clusters are well defined. On the contrary, a σ close to -1 indicates that the between-cluster stability is larger than the within-cluster stability, which means that the clusters are poorly adapted to $\hat{\mathbf{G}}$.

The stability contrast σ is a function of three integer parameters (K , L , M). In order to retain comparable numbers of clusters at all levels, the search for an optimal contrast with M final clusters is restricted to K and L belonging to a neighbourhood $\mathcal{N}(M)$ of M , for example the interval $[70\%M, 130\%M]$. The following locally maximal contrast function is thus defined:

$$\sigma_{\max}(M) = \max_{K, L \in \mathcal{N}(M)} \sigma(K, L, M). \quad (9)$$

⁴ A workshop was recently organized on that question, with lectures accessible online at http://videolectures.net/srmc07_tuebingen/.

⁵ The stability contrast is almost a particular case of the silhouette criterion (Rousseeuw, 1987), the only difference being a normalization factor which is absent in our criterion. This factor was originally designed to scale some generic unbounded similarity measures but is not necessary for the stability measure which is already bounded between 0 and 1.

The final number of clusters M_{opt} is selected as the one which maximizes σ_{max} , and the individual and group parameters K_{opt} and L_{opt} are selected as the one maximizing $\sigma(K, L, M_{\text{opt}})$ in the neighbourhood $\mathcal{N}(M_{\text{opt}})$.

Validity of the BASC approximation of stability

An important question regarding the BASC approach is the validity of the bootstrap estimate of the stability matrix. More specifically, would the estimator defined in Eq. (5) be unbiased, i.e. would the statistical expectation of the estimator be equal to the real stability measure defined in Eq. (2). There is a large literature regarding the theoretical asymptotical properties of bootstrap estimators (Shao and Tu, 1995). Unfortunately, those results are applicable under relatively strict conditions: the statistic would have to be a smooth function of the mean. In our case, the statistic of interest is the adjacency matrix resulting of a clustering algorithm, which departs from the “smooth function of the mean” category. The available theoretical results are thus quite limited (Field and Welsh, 2007). The standard bootstrap for independent data seems efficient in practice to estimate the stability in hierarchical clustering (Suzuki and Shimodaira, 2006), even though bias-correction techniques (Shimodaira, 2004) improve the behaviour. Acknowledging this current theoretical limitation of BASC, we resorted to bootstrap schemes that had an established satisfactory behaviour regarding the spatial dependencies of the surrogate bootstrap data. We also investigated the accuracy of BASC empirically on Monte-Carlo simulations of synthetic fMRI time series. This validation step would be necessary in any case, as the theoretical results hold asymptotically, i.e. with the number of time samples T and the number of subjects N tending towards infinity. Monte-Carlo simulations with synthetic datasets allow the behaviour of the method to be assessed for finite T and N .

Experiments on simulated time series

Simulation model

The multi-level BASC method has been evaluated on fully synthetic datasets. At the group level, brain regions were grouped into non-overlapping clusters formed of a fixed number of regions. Individual clusters were generated by applying a random perturbation of the

group-level clusters. More specifically, regions of the brains were treated as points on a circle. The group-level clusters were thus viewed as regular portions on the circle. Individual clusters were generated by randomly moving the edges between clusters. At the individual level, regions within a cluster were generated by adding a signal common to all regions, called cluster time series, and a noise following a Gaussian distribution independent in space and time with zero mean and unit variance. Each cluster time series was a sample of an auto-regressive process of order 1, with parameter $a = 0.5$, $T = 100$ samples and a variance σ_c^2 common to all clusters. The free parameters of the simulations were the following:

- The number of clusters Λ .
- The number of regions per group cluster Γ .
- The variance of cluster time series σ_c^2 (or equivalently the signal-to-noise ratio, SNR).
- The number of subjects N .
- The parameter p which defined the stability of the cluster edges position across subjects (a small p corresponded to a large variability).

Some preliminary experiments were done to determine eight different scenarios corresponding to various difficulties in terms of clustering stability, see [Supplementary Materials B](#). The parameters values of the different scenarios were $\Lambda = 5$, $\Gamma = 20$, SNR in $\{-10, -5\}$, N in $\{5, 20\}$ and p in $\{0.4, 0.1\}$.

Selection of the number of clusters

The multi-level BASC was applied on simulated group time series to select the number of clusters (K, L, M) with $B = 20$ bootstrap samples at the individual level and $C = 50$ at the group level. A number of 30 simulation experiments were performed. For each experiment, the number of clusters was investigated on a grid from 2 to 30. Fig. 3 represented the σ_{max} criterion as a function of M . For almost every scenario, the relationship was smooth with a global maximum located at $M = 5$ which corresponded to the ground truth. The only scenario departing from that behaviour was $(p = 0.1, \text{SNR} = -10, N = 5)$, i.e. the worst SNR with the maximal between-subjects variability and a very low number of subjects. Note that for $(p = 0.4, N = 20)$, the peak stability contrast was above 0.9, indicating a close-to-perfect clustering, regardless of the SNR at the individual level. Regarding

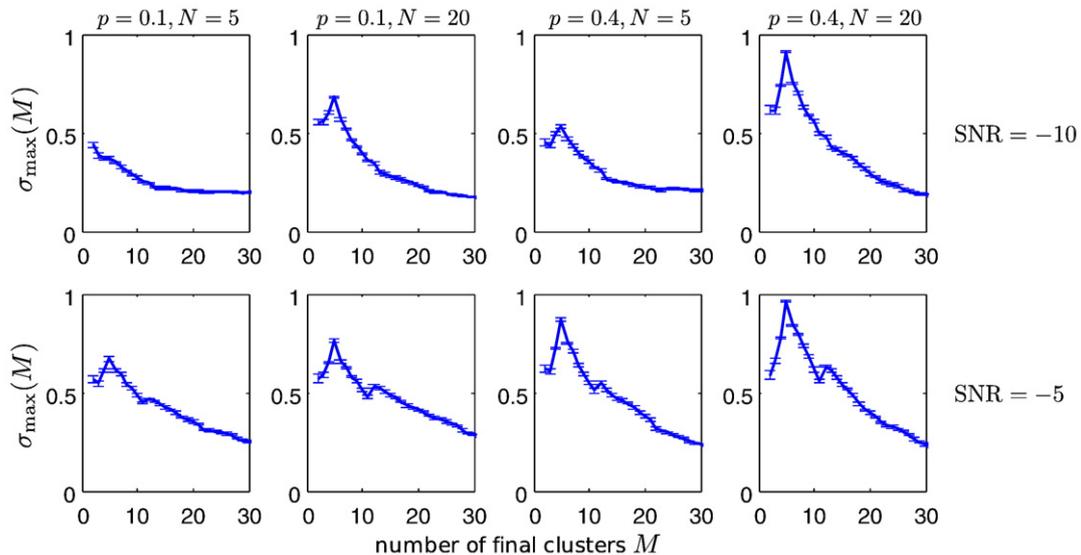


Fig. 3. Simulation study of the selection of the number of clusters. The stability contrast criterion σ_{max} was represented as a function of the number of final clusters M for different parameters of the group simulation: number of subjects N , the parameter p driving the variability of clusters across subjects and the signal-to-noise ratio (SNR). The curves were averaged over 30 experiments, and the error bars indicated the standard deviation of the mean.

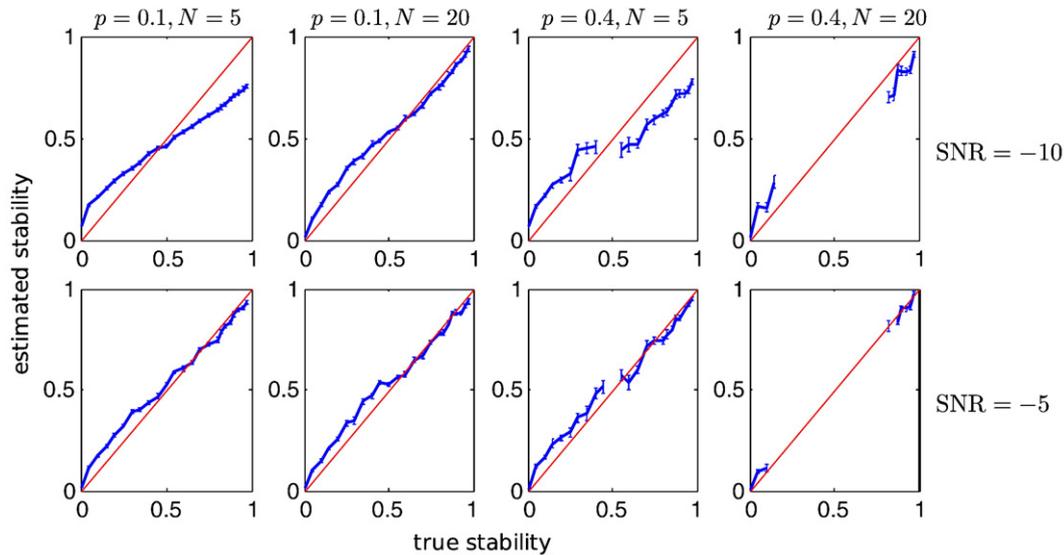


Fig. 4. Simulation study of the bias in the group-level stability estimation. The estimated group-level stability was represented as a function of the true stability for different parameters of the group simulation: number of subjects N , the parameter p driving the variability of clusters across subjects and the signal-to-noise ratio (SNR). The curves were averaged over 200 experiments, and the error bars indicated the standard deviation of the mean.

the final number of clusters M , the correct value $M=5$ was the most selected value in every scenario but for ($p=0.1$, $\text{SNR}=-10$, $N=5$), where it was $M=2$. With $N=20$ subjects, the correct number of clusters M was selected in more than 80% of the simulations, regardless of p and the SNR. The same behaviour was observed on the group number of clusters L . The individual number of clusters K had a much more widespread selection profile, even for ($p=0.4$, $N=20$). The most frequently selected values were {5,6,7}, but the range of selected number of clusters actually covered 2 to 8. This result is not actually an issue as the optimal number of individual clusters may not exactly match the real number of clusters extracted at the group level. This simulation study demonstrated that the stability contrast criterion was able to correctly identify the number of clusters at the group level in a reliable fashion as soon as the number of subjects was larger than 20.

Accuracy of the group-level stability estimation

A first simulation experiment was performed to examine a possible bias in the estimation of the stability using the BASC method. This was implemented by comparing the true stability measures, as defined through Eq. 3, and the bootstrap approximation at the group level, as defined through Eq. 5. The parameters of the BASC

approximation were selected to $K=L=M=4$, $B=50$ and $C=100$ bootstrap samples. A number of 200 experiments were used to derive the average relationship between the true and the estimated stability for each of the 8 scenarios of group fMRI time series, see Fig. 4. This analysis evidenced a small bias of the BASC approximation: stability above 0.5 is slightly underestimated, and stability below 0.5 is slightly overestimated. This bias was mostly present with $N=5$ and a low SNR of -10 dB. It was smaller for $N=20$ subjects, and actually negligible when the SNR was -5 dB. This experiment demonstrated that, in the context of employed simulation models, the multi-level BASC approach provided a satisfactory approximation of the group-level clustering stability.

A second simulation experiment quantified the performance of BASC in terms of detection of the real clustering structure by means of receiver-operating characteristic (ROC) curves (Sorenson and Wang, 1996). True findings in the stability were defined as pairs of regions that belonged to the regular clusters at the group level. For each possible threshold on the estimated group-level stability matrix, the associated sensitivity and specificity were derived and the relationship between those two parameters resulted into ROC curves presented in Fig. 5. The SNR at the individual level had only a minor impact on the results. The ROC analysis carried out on a small number of subjects ($N=5$) exhibited mitigated performance, yet BASC almost

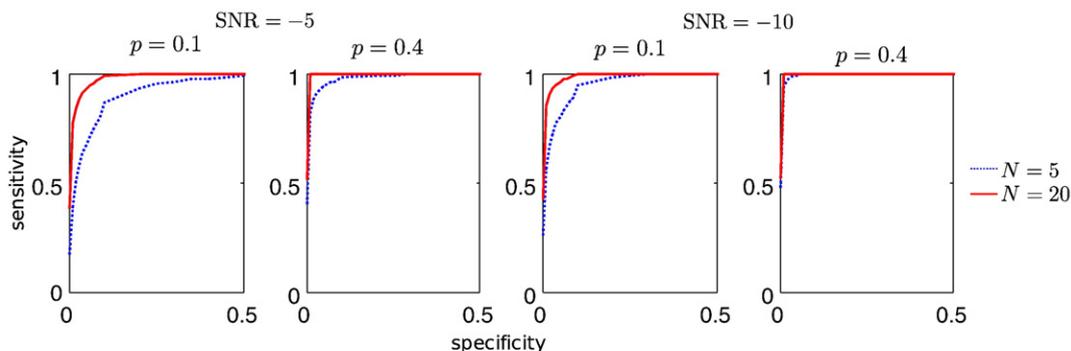


Fig. 5. Simulation study of the sensitivity and specificity of the group-level stability. The sensitivity of the group-level BASC was represented as a function of the specificity for different parameters of the group simulation: number of subjects N , the parameter p driving the variability of clusters across subjects and the signal-to-noise ratio (SNR).

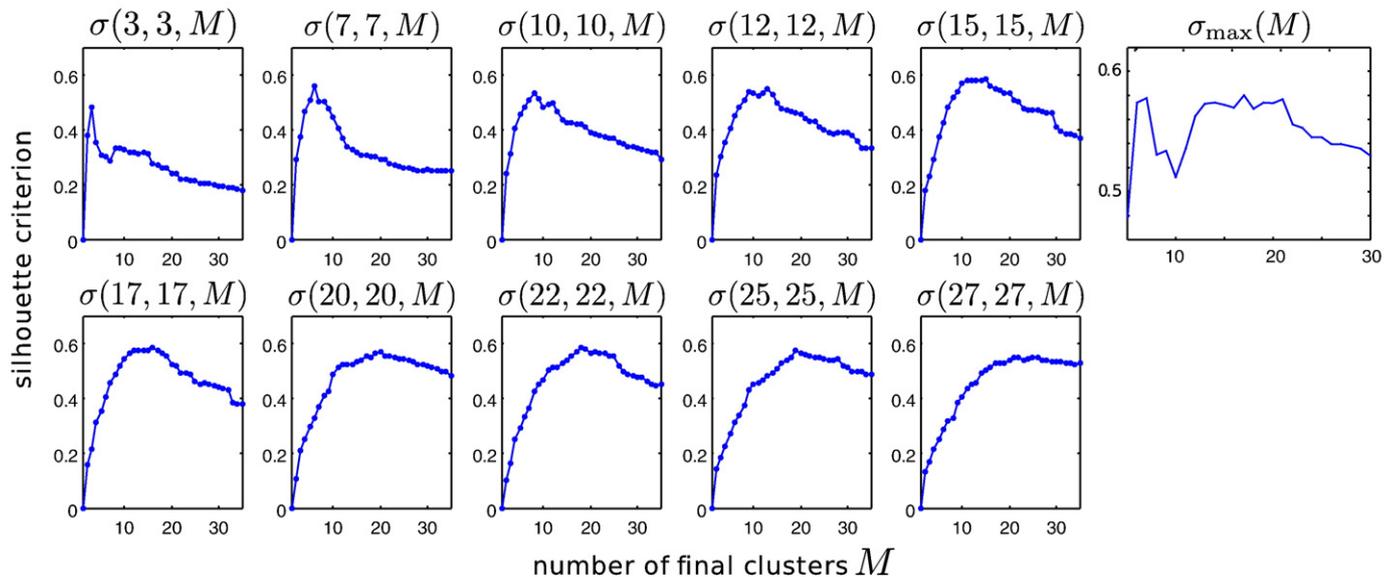


Fig. 6. Stability contrast criterion. The stability contrast criterion σ as a function of the clustering parameters (K,L,M) . The plots are for varying M and fixed K,L , except for the upper right plot which represents the maximal contrast σ_{\max} as a function of M for any (K,L) in a neighbourhood of M . See text for details.

behaved as a perfect classifier with $N=20$ subjects. This suggested that the statistical power of BASC as a detector of group-level clustering structures should be satisfactory in the conditions of application of most fMRI studies.

Experiment on a real resting-state fMRI database

Data acquisition

The multi-level BASC was applied on the younger subjects (age ranging from 19 to 44 years) of a neuroimaging database collected by the International Consortium for Brain Mapping (ICBM) and made publicly available as part of the 1000-connectome project.⁶ A cohort of $N=43$ healthy volunteers (21 men, 22 women) participated in this study which was approved by the local ethics committee. Subjects had no history of neurological or psychiatric disorders. Three functional runs were acquired for each subject under resting-state conditions i.e. the subjects were asked to remain still, eyes closed, and to refrain from any overt activity. For each functional run, 138 brain volumes of BOLD signals were recorded on a 1.5 T Siemens scanner using a 2D echoplanar BOLD MOSAIC sequence and the following parameters: TR/TE = 2 s/50 ms, 64×64 matrix with a 4×4 mm² resolution, 23 contiguous axial slices covering the cortex but not the cerebellum, slice thickness = 4 mm, flip angle = 90° and an 8-channel coil. A high-resolution anatomical T₁-weighted scan was also acquired: TR/TE = 0.022 s/0.0092 s, 256×256 matrix with a 1×1 mm² resolution, 176 contiguous sagittal slices covering the whole brain, slice thickness = 1 mm, flip angle = 30°.

Data preprocessing

The fMRI database was preprocessed using the pipeline implemented in the package called neuroimaging analysis kit⁷ (NIAK). The three first volumes of each run were suppressed to allow the magnetisation to reach equilibrium. Each dataset was corrected of inter-slice difference in acquisition time, rigid body motion, slow time drifts (high-pass filter with a 0.01 Hz cut-off) and physiological noise (Perlberg et al., 2007). Slow time drifts and physiological noise

correction were implemented in an attempt to reduce the spatially correlated noise present in the fMRI time series, which may introduce stable clusters unrelated to neural activity. Data-driven noise correction strategies have been reported to improve the detection of RSNs and are increasingly popular in functional connectivity analysis (Giove et al., 2009; Weissenbacher et al., 2009).

The following strategy was implemented to transform the individual brains into a common space of analysis. For each subject, the mean motion-corrected volume of all the datasets was coregistered with a T₁ individual scan using Minctracc⁸ (Collins et al., 1994), which was itself non-linearly transformed to the Montreal Neurological Institute (MNI) non-linear template using the CIVET⁹ pipeline (Zijdenbos et al., 2002). The functional volumes were resampled in the MNI space at a 2 mm isotropic resolution and spatially smoothed with a 6 mm isotropic Gaussian kernel. The spatial smoothing was implemented in an attempt to minimize the residual variability in anatomy and functional organization of individual brain in stereotaxic space.

In order to limit the computational burden of the bootstrap analysis, a region-growing algorithm (Bellec et al., 2006) was applied to the concatenated time series of every subjects in order to derive a common segmentation of the brain into small functionally homogeneous regions. To limit the memory demand, the region growing was applied independently in each of the 116 areas of the AAL template (Tzourio-Mazoyer et al., 2002). The resulting regions were spatially connected with roughly equal size, which was set to the smallest value that would lead to amenable computations, i.e. a size of 800 mm³ translating into 1191 regions covering the grey matter. The average time series within each region was derived after correction to a zero temporal mean and unit variance, and then concatenated across all the fMRI datasets for each subject. This process resulted into individual time series with $T=405$ time points¹⁰ and $R=1191$ regions.

Choice of BASC parameters

The block length for CBB of individual time series was selected randomly at each bootstrap sample in the interval $([10,30])$. BASC was

⁶ http://www.nitrc.org/projects/fcon_1000/.

⁷ <http://wiki.bic.mni.mcgill.ca/index.php/NiakFmriPreprocessing>.

⁸ <http://wiki.bic.mni.mcgill.ca/index.php/MinctraccManPage>.

⁹ <http://wiki.bic.mni.mcgill.ca/index.php/CIVET>.

¹⁰ (138 volumes – 3 dummy scans) \times 3 runs = 405 time points per subject.

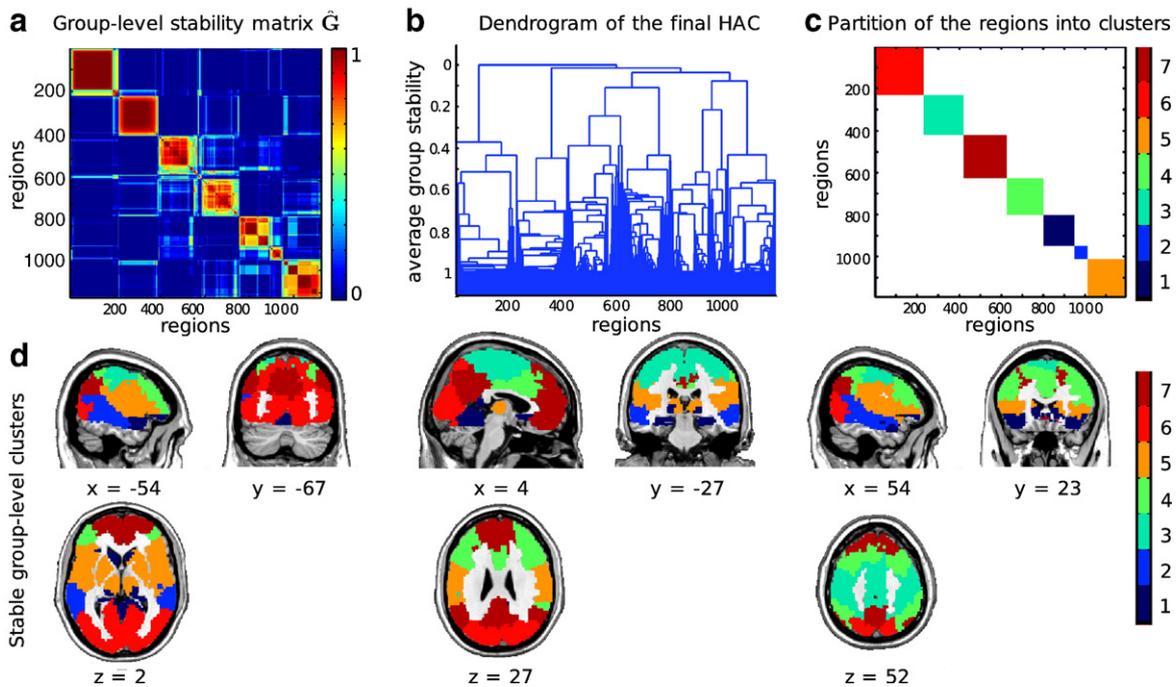


Fig. 7. Partition of the brain into stable group-level networks. The group-level stability matrix (a) underwent a hierarchical clustering which was represented as a dendrogram (b) and was used to generate a partition into $M=7$ clusters. These clusters were color-coded in matrix form (c), see text for details, and the same color code was applied to represent the brain regions associated with each cluster, superimposed with the MNI152 non-linear template (d). Some representative sagittal, coronal and axial slices were selected manually.

first applied in order to select the number of clusters with a small number $B=30$ of bootstrap samples at the individual level and $C=250$ bootstrap samples at the group level. Fig. 6 presented the stability contrast σ as a function of the number of final clusters M , on a grid of 5 to 30, with different choices of (K,L) . The σ curves were smooth and the global maximum was achieved for M close to the employed K . The locally maximal contrast σ_{\max} was also smooth, but did not present a clear global maximum. Instead, the first local maximum at $M=7$ was followed by a plateau in the range 12 to 21. This result demonstrated that more than one stable clustering structure could be identified in resting-state fMRI data. The purpose of this article being to illustrate the mechanics of the multi-level BASIC rather than investigating such multi-scale architecture, we reported the results for a single M corresponding to the first local maximum.

Group-level stable clusters

To derive stability maps with high accuracy, the multi-level BASIC was applied a second time with $B=200$ bootstrap samples at the individual level and $C=1000$ at the group level for the following number of clusters ($K=8, L=7, M=7$). The group-level stability matrix exhibited a very high stability contrast ($\sigma_{\max}(7)=0.59$), which graphically translated into yellow–red diagonal squares (within-clusters stability) on a deep blue background, see Fig. 7a. The dendrogram representation of the HAC applied on the group-level stability matrix showed that most of the merging between regions occurred at a very high level of stability, above 0.9, see Fig. 7b. A matrix representation of the RSNs was derived by coding every pairs of regions within a given cluster with a specific color (Fig. 7c). Each RSN of Fig. 7c was associated with a group of brain regions represented in Fig. 7d. Some RSNs were strikingly similar to those reported previously using group-level ICA (Damoiseaux et al., 2006): a sensorimotor network (RSN3), a visual network (RSN6); the default-mode network (RSN7). The bilateral temporal RSN2 resembled a component from (Damoiseaux et al., 2006), yet the correspondence was not clear due to differences in the employed field of views. The fronto-parietal network (RSN4) included regions involved in

working-memory tasks, and has often been reported splitted into two left and right subnetworks (Damoiseaux et al., 2006). The RSN5, comprising ventro-frontal cortex, the thalami and the striatum, also appeared like a merging of two components reported by (Damoiseaux et al., 2008). RSN1 included some ventro-medial cortex and anterior caudate, yet it also included regions around the cerebellar tentorium indicating that it might have been corrupted by residual physiological noise.

Stability maps

For each stable group RSN, the average group stability matrix was translated into a stability map, as illustrated for RSN6 in Fig. 8a. The same process was applied with the average individual stability matrix (Fig. 8b) and each of the 43 individual stability matrices (Fig. 8c). For a region i , the associated value of the stability map was the average of all columns $j \neq i$ of row i in the stability matrix, where j belonged to the target cluster. Note that the stability score was represented only for the regions inside the cluster in Fig. 8, but could actually be derived for every brain region.

The group and average individual stability maps for the 7 RSNs were presented in Fig. 9. The maps were derived on the whole brain, yet only the values higher than 0.2 were represented and superimposed to an average structural scan to facilitate the identification of anatomical landmarks. The percentiles of the distribution of these stability maps within each RSN were reported in Table 1. At the group level, the two most stable RSNs were the visual and sensorimotor (respective median 0.91 and 0.94). The remaining RSNs had comparable levels of group stability (median ranging from 0.71 to 0.77). The observed average individual stability was markedly smaller, with a median of 0.49 and 0.53 for the two most stables, and a median ranging from 0.3 to 0.36 for the other RSNs. It was interesting to note that a relatively low level of average individual stability could result in a high level of group stability, e.g. in the striatum Fig. 9e. The gradient observed on the individual average stability of the sensorimotor also resulted in a plateau of high stability at the group level 9c. This result illustrated the fact that the group-level clustering was based on

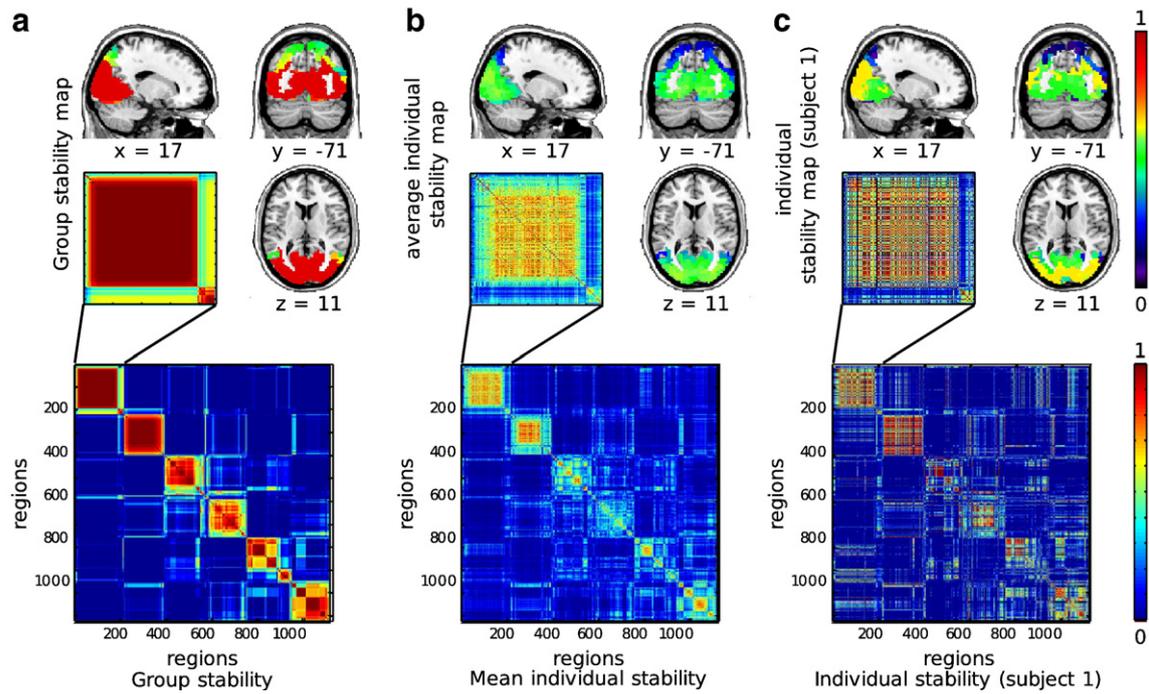


Fig. 8. Multi-level stability maps. The group (a), average individual (b) and individual (c) stability matrices were represented in panels alongside with the associated stability maps for one of the cluster (RSN6, visual).

the balance of within-cluster and between-cluster average individual stability, regardless of the absolute value. As long as the number of subjects is high enough to ensure the stability of the structure of the average individual stability matrix, the group-level stability may achieve high values.

Ideally, a group-level analysis would achieve a good trade-off between the establishment of clear correspondences across subjects and the ability to identify subject-specific features. A standard deviation map was thus generated to represent the spatial variability of the individual maps associated with each group stable cluster. The distance toolbox¹¹ was also used to generate a multi-dimensional scaling (MDS) representation of the 43 individual stability maps, and the maps corresponding to the two subjects most distant on the MDS first axis were selected. The average and standard deviation of individual stability maps, as well as the two selected individual maps were represented for four networks¹² (RSN3, RSN4, RSN6, and RSN7) in Fig. 10. A first general qualitative comment was that individual RSN stability maps did match well with the group-level stable RSNs, even though the maps were derived on the full brain, which indicated that the group-level clustering process was efficient at picking up stable features of the individual clusterings. Another general remark was that there was a widespread distribution of the median values of stability across subjects. For example for RSN3, the 5% upper subjects had a stability median above 0.68 while for the 5% lower subjects it was below 0.36 (see Table 1), which was apparent with the two selected subjects in Fig. 10b. This may have reflected the absence or presence of a particular RSN for some subject or a large deviation from the normal group pattern for some subjects. Such global variation will contribute to large standard deviation colocalized with large average individual stability, as could be observed for RSN4 (Fig. 10b) and the frontal part of RSN7 (Fig. 10d). The individual stability maps also evidenced differences in the edges of the RSNs across subjects. For example, the

RSN3 had more anterior components for subject 36 than for subject 40, see Fig. 10a; the frontal regions were much smaller for subject 22 than for subject 23 in RSN4, see Fig. 10b; the posterior cingulate had less anterior extension for subject 31 than subject 42 in RSN7, see Fig. 10d. These variations of edge position were clearly visible in the standard deviation maps as well, see for example the sensorimotor network (Fig. 10a) or the posterior cingulate area of the default-mode network (Fig. 10d). Some regions were also present or absent from RSNs on different subjects. For example, subject 23 had subparts of the caudate nuclei associated with its RSN4, which were absent in subject 22, see Fig. 10b, coronal slice; subject 42 had the thalami associated with RSN6, which were absent in subject 17, see Fig. 10c, axial slice. The standard deviation maps could identify areas where such change occurred frequently, such as the medial part of the occipital cortex in RSN6 (Fig. 10c). Taken together, these results demonstrated that the multi-level BASC was able to identify some subject-specific features associated with each of the group stable RSNs.

Discussion

We developed a general method called bootstrap analysis of stable clusters (BASC) to extract the stable features of a random clustering process. This method was applied to fMRI datasets, both at the individual and group levels. A measure called stability contrast was also designed to assess the quality of the clustering and was used to select the parameters of the clustering.

BASC adds to the quickly growing set of techniques available to identify RSNs. There is however an important distinction to be made between the algorithm which is used to identify RSNs and the mechanics employed to assess their stability. The BASC framework belongs to the second category and is not limited to the *k*-means algorithm. It would be applicable to any other clustering algorithm and also to component analysis, if the spatial components were transformed into clusters by application of a winner-take-all principle where each voxel is associated to the component with maximal contribution at this location. BASC therefore does not stand as an alternative to ICA, but rather as a generic technique to investigate

¹¹ <http://groups.google.com/group/distance-toolbox>.

¹² These networks were selected because they minimized or maximized the 5% upper percentile across subjects of the median individual stability within the group RSN, see Table 1.

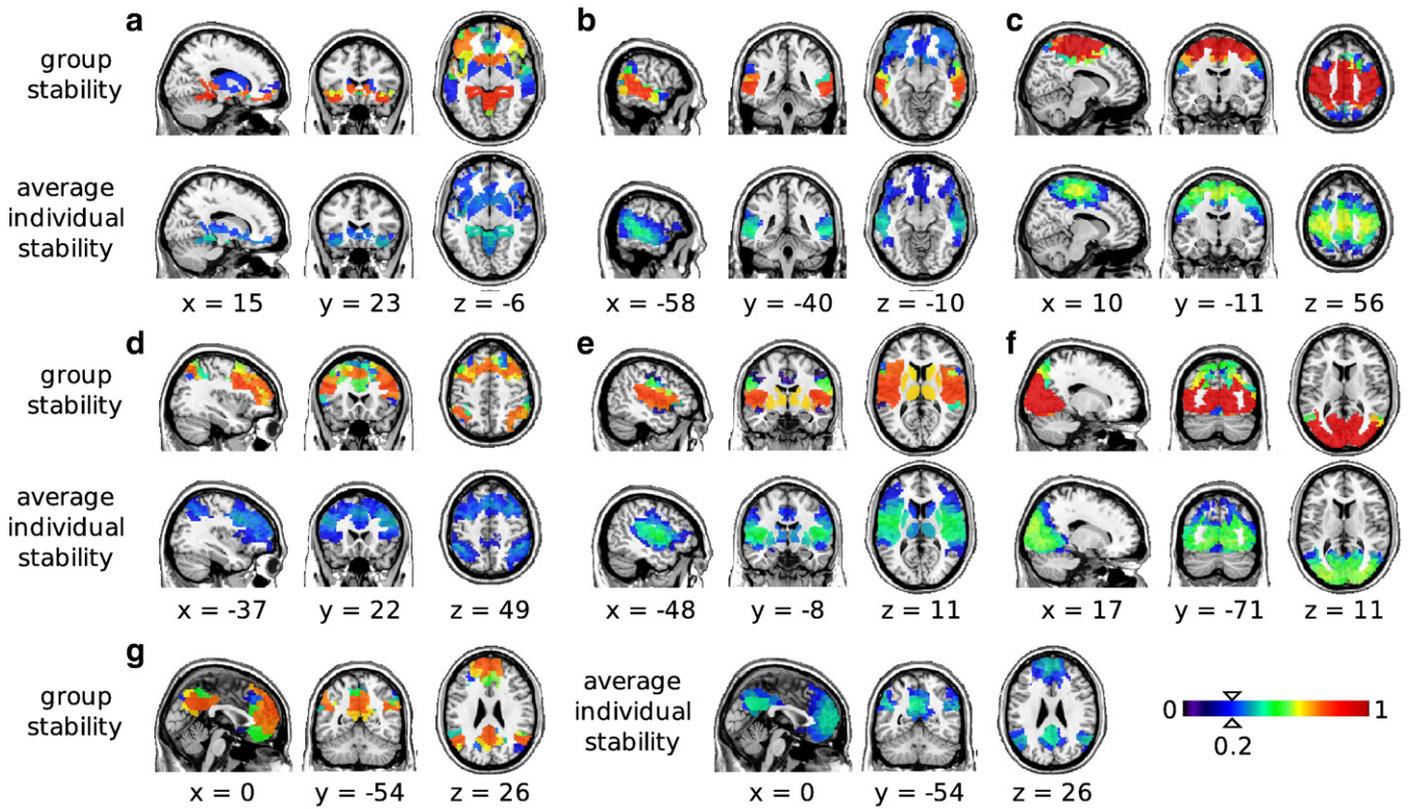


Fig. 9. Group and average individual stability maps. The group and average individual stability maps were generated for each of the 7 group stable clusters and superimposed on the MNI152 non-linear template. The stability score were derived in the whole brain, yet a threshold of 0.2 was applied. Some representative sagittal, coronal and axial slices were selected manually.

Table 1
Percentiles of the stability for each RSN, at different levels. Percentiles of the distribution of the stability for each RSN at the group, average individual and individual levels. Only the regions within a RSN were considered to derive the percentiles of the distribution. For the individual stability and for each stability percentile, the 5% lower and upper percentiles were derived across the 43 subjects.

Percentile	Stability type	RSNs						
		1	2	3	4	5	6	7
0.01	Group	0.54	0.40	0.45	0.41	0.43	0.41	0.07
	Average individual	0.20	0.22	0.20	0.21	0.21	0.17	0.13
	Individual (5% percentile)	0.03	0.04	0.02	0.03	0.03	0.02	0.02
	Individual (95% percentile)	0.13	0.16	0.13	0.12	0.15	0.11	0.09
0.25	Group	0.72	0.65	0.86	0.62	0.72	0.88	0.55
	Average individual	0.30	0.33	0.39	0.27	0.33	0.41	0.27
	Individual (5% percentile)	0.14	0.15	0.19	0.16	0.20	0.14	0.14
	Individual (95% percentile)	0.42	0.55	0.51	0.28	0.40	0.58	0.30
0.5	Group	0.77	0.76	0.94	0.72	0.77	0.91	0.71
	Average individual	0.33	0.36	0.49	0.30	0.39	0.53	0.33
	Individual (5% percentile)	0.20	0.23	0.36	0.22	0.29	0.32	0.23
	Individual (95% percentile)	0.57	0.65	0.68	0.37	0.58	0.72	0.47
0.75	Group	0.81	0.80	0.94	0.75	0.82	0.91	0.74
	Average individual	0.38	0.40	0.57	0.33	0.43	0.56	0.38
	Individual (5% percentile)	0.24	0.28	0.47	0.25	0.31	0.40	0.28
	Individual (95% percentile)	0.63	0.67	0.71	0.50	0.64	0.73	0.53
0.99	Group	0.81	0.80	0.94	0.79	0.83	0.91	0.76
	Average individual	0.45	0.45	0.62	0.38	0.48	0.60	0.44
	Individual (5% percentile)	0.29	0.30	0.50	0.29	0.34	0.43	0.32
	Individual (95% percentile)	0.64	0.70	0.71	0.52	0.65	0.74	0.55

the individual and group stability of any type of random clustering process.

There was a good agreement between the RSNs found in this paper and those previously reported in the literature, despite a large methodological variability. The default-mode network is already known to be robustly identified by different techniques (Long et al., 2008). Some networks reported in (Damoiseaux et al., 2006) were still not found here, such as the executive control network. Note that there are actually discrepancies in the ICA studies themselves, e.g. (Damoiseaux et al., 2006, 2008; Perlberg et al., 2008; Smith et al., 2009). Such discrepancies might be explained by many factors, including differences in preprocessing strategies, differences between the *k*-means and ICA, and also differences between the inference framework employed to derive the group-level networks. Moreover, the differences may also be related to the number of clusters or components that were selected. ICA and clustering are indeed expected to behave quite differently regarding this parameter: by construction of the linear model assumption in ICA, the maximal possible number of components equals the number of time points. By contrast, there is no constraint on the number of clusters in a dataset even in dimension 1. These questions need to be investigated in the future. In this context, BASC may prove useful by providing a single inference framework able to deal with all the types of clustering algorithms. The stability contrast could be used to perform model selection in this context, i.e. to quantify which technique performs best.

The multi-level BASC has a substantial computational cost, as it involves thousands of replications of a clustering process in high dimension. The region-growing stage of the preprocessing took 30 h, the first pass of BASC which consisted of testing multiple numbers of clusters took 66 h. The second pass to generate the stability measures with high accuracy took 8 h. This computation time is still amenable on a single workstation, but markedly slower than a standard general

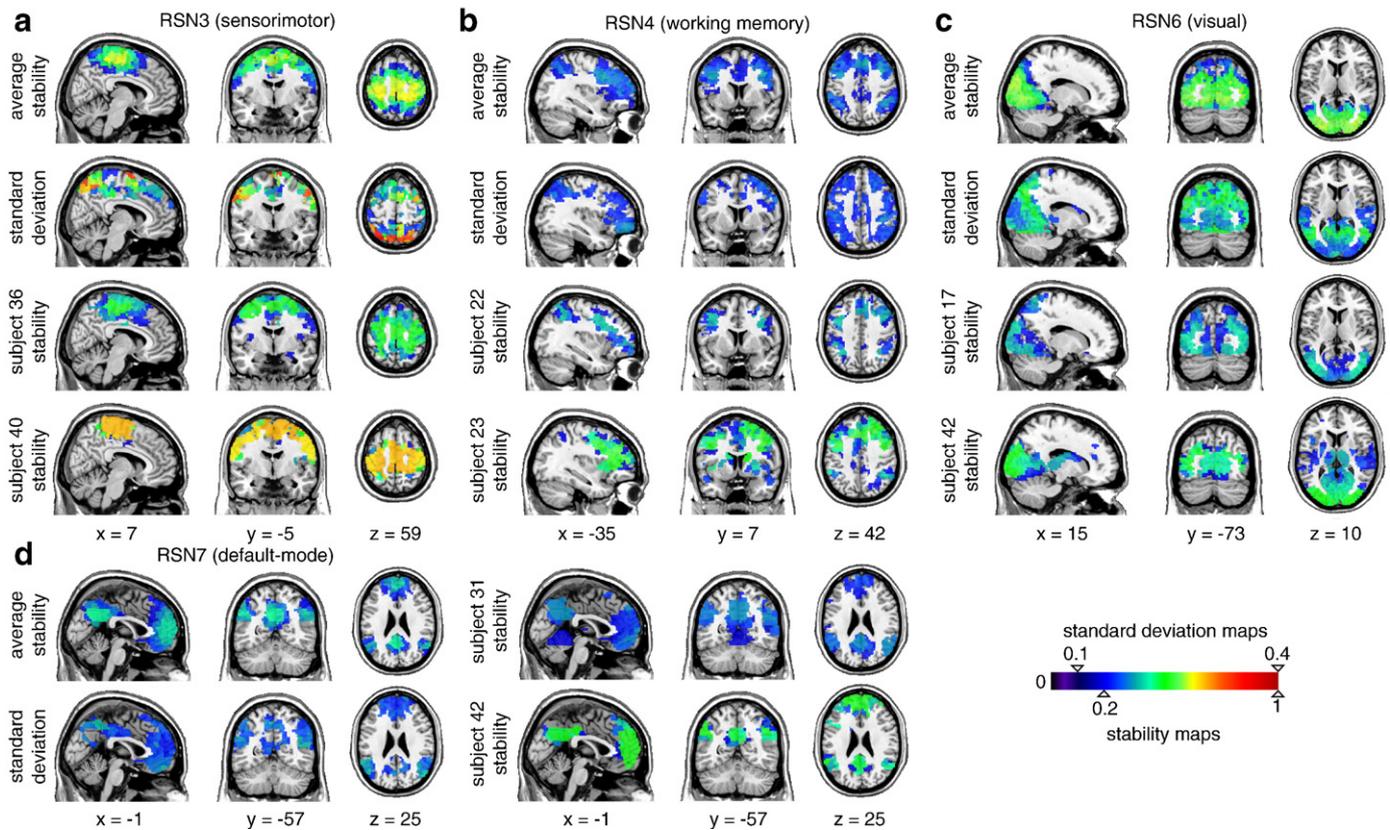


Fig. 10. Individual stability maps. The two most stable RSNs, i.e. RSN3 (a), RSN6 (c), and the two least stable RSNs, RSN4 (b) RSN7 (d), were selected to illustrate the variability of individual stability maps associated with stable group clusters. For each RSN, the average and standard deviation of individual stability maps as well as two individual maps were selected (see text for details) and superimposed on the MNI152 non-linear template. Some representative sagittal, coronal and axial slices were selected manually.

linear model analysis. The implementation of BASC takes advantage of parallel computation on multiple computational cores¹³ and the actual computation time for the whole BASC analysis was 3 h 15 min. This computation time is still reasonable considering that a model selection was performed over a large grid of clustering parameters. The multi-level BASC for a fixed number of clusters actually ran in a decent time (8 h in our case). Computational cost is one common limitation of resampling-based method, but it is the price to pay for deriving non-parametric statistics on a complicated stochastic process.

An important contribution of this work was to provide a probabilistic measure of the clustering stability. This formulation did not rely on component matching, as is commonly done in ICA (Esposito et al., 2005; Damoiseaux et al., 2006; Perlberg et al., 2008). While the latter measures are known to exhibit an upward statistical bias (Langers, 2010), i.e. the stability is overestimated, the BASC approximation of clustering stability exhibited only little bias in finite samples with synthetic data. For stability measures in the range 0.5–1, the bias was actually found to be downward, i.e. the stability was underestimated. Some extensions of the bootstrap, e.g. (Shimodaira, 2004), would allow to further reduce the bias as is done in the package pvclust (Suzuki and Shimodaira, 2006). Bootstrap bias-correction has unfortunately a large computational cost which makes its application to massive datasets such as resting-state fMRI challenging. In the perspective of further validation of the bootstrap approximation of the stability, the ideal strategy would be to compare the bootstrap estimate to measures derived using real replications of the experiment. This can be done at the individual level by scanning multiple times the same population, e.g. (Zuo et al., 2010). Validation of BASC at the group level would require a large number of independent

groups of subjects. Such ambitious validation strategy is now feasible thanks to the recent release of extensive public databases of resting-state fMRI, both for test–retest analysis¹⁴ and for multi-group analysis¹⁵ (over 1000 subjects).

The idea of building the group-level RSNs by performing a clustering on a similarity matrix averaged across subjects has been proposed previously for various similarity measures: Salvador et al. (2005) used partial correlation, while van den Heuvel et al. (2008) considered directly the adjacency matrix of an individual clustering. The choice of the individual stability matrix as a measure of similarity in the group clustering was motivated primarily by the purpose of the multi-level BASC: the group clustering was designed to reflect the stable features of the individual clustering. In the pure perspective of group clustering, it would still be possible to employ the group-level BASC with any type of individual similarity measure. The stability contrast could again serve as a measure of the clustering quality to compare approaches.

The preprocessing strategy implemented in this work included a physiological noise correction and a region-growing algorithm to reduce the spatial dimension. Data-driven noise correction has been reported to improve the detection of RSNs (Giove et al., 2009; Weissenbacher et al., 2009), yet the procedure that we used Perlberg et al. (2007) has not been evaluated in this context. Spatial dimension reduction has been employed quite often in the past, e.g. (Salvador et al., 2005), but using anatomically defined brain areas such as the AAL template (Tzourio-Mazoyer et al., 2002). A couple of groups recently applied some data-driven techniques to derive a large number of brain regions and conduct a region-based analysis (Bellec et al., 2006; Thirion et al., 2006; Meunier et al., 2009).

¹³ The processing was performed on SUN Dual-Dual Opteron 875 nodes with a total of 90 cores available.

¹⁴ http://www.nitrc.org/projects/nyu_trt.

¹⁵ http://www.nitrc.org/projects/fcon_1000/.

Because of its computational cost, the preprocessing of fMRI time series was not included in the bootstrap replication of the clustering and its impact on stability was therefore not assessed by BASC. As a preliminary experiment to test the robustness of the BASC results to different choices of preprocessing strategies, the identification of the group stable RSNs reported in this manuscript was replicated with two other preprocessing strategies, see [Supplementary Material C](#). Our conclusions were that excluding the physiological noise correction (CORSICA) and the spatial smoothing from the preprocessing had only a marginal influence on the stable group clusters. The choice of the template of brain areas used to constrain the region growing had a more pronounced effect, notably through the segmentation of the grey matter, yet most features of the clustering were retained. This suggested that the BASC group results were reasonably robust to the choice of the preprocessing strategy. Assessing the impact of various preprocessing strategies on the multi-level BASC results more thoroughly will be an important area for future work.

An interesting finding that was not further investigated in this work was the existence of stable clusters at multiple spatial scales, i.e. with different numbers of clusters. A recent article ([Smith et al., 2009](#)) has reported about 45 RSNs with consistent ICA maps at the group-level. Our results suggest that there may be many numbers of clusters where local maxima of clustering stability can be identified. While the current view of RSNs is focussed on a few widespread networks, the organization of RSNs at finer spatial scales may also inform us on critical aspects of the brain functional architecture.

The BASC method naturally extends to perform group comparison. There are currently only a few techniques available to compare the RSNs patterns across group, e.g. ([Calhoun et al., 2004](#); [Sui et al., 2009](#)). The present work concentrated on a single group, but the difference between the average individual-level stability matrices derived on two distinct groups of subjects would allow to test for differences in the underlying clustering structures. The group-level clustering would then search for groups of brain regions with a large difference in stability between the two groups.

The ability of the multi-level BASC to provide individual stability maps associated with group-level RSNs is a key feature for clinical applications. For example, it has been shown that the overlap between some individual default-mode network maps and a group template allowed to differentiate between patients with Alzheimer's disease and healthy controls ([Greicius et al., 2004](#)). This result however required to select a default-mode network component for each subject on the basis of an a priori template. The BASC method provides a fully automated alternative which applies on an arbitrary large number of networks. As was illustrated in the present work, the individual stability maps derived from group RSNs provide a well-defined correspondence of RSNs across subjects while allowing to extract some substantial subject-specific features.

Acknowledgments

Part of this work was presented at the 15th annual meeting of the organization for human brain mapping ([Bellec et al., 2009](#)). The authors would like to thank Samir Das for editing an earlier version of this manuscript and to acknowledge the work of the International Consortium for Brain Mapping¹⁶ (ICBM) fMRI community in creating the resting-state database. The acquisition of the ICBM resting-state database was supported by NIH grant 9P01EB001955-11. The computational resources used to perform the data analysis were provided by CLUMEQ¹⁷, which is funded in part by NSERC (MRS), FQRNT, and McGill University.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at [doi:10.1016/j.neuroimage.2010.02.082](https://doi.org/10.1016/j.neuroimage.2010.02.082).

References

- Baumgartner, R., Windischberger, C., Moser, E., 1998. Quantification in functional magnetic resonance imaging: fuzzy clustering vs. correlation analysis. *Magn. Reson. Imaging* 16 (2), 115–125.
- Beckmann, C.F., Smith, S.M., 2005. Tensorial extensions of independent component analysis for multisubject fMRI analysis. *NeuroImage* 25 (1), 294–311.
- Bellec, P., Marrelec, G., Benali, H., 2008. A bootstrap test to investigate changes in brain connectivity for functional MRI. *Stat. Sin.* 18, 1253–1268.
- Bellec, P., Perlberg, V., Jbabdi, S., Pélégrini-Issac, M., Anton, J., Doyon, J., Benali, H., 2006. Identification of large-scale networks in the brain using fMRI. *NeuroImage* 29 (4), 1231–1243.
- Bellec, P., Rosa-Neto, P., Benali, H., Evans, A.C., 2009. Multi-level bootstrap analysis of stable clusters (BASC) in resting-state fMRI. *Proceedings of the Human Brain Mapping 2009 Annual Meeting*, San Francisco: *NeuroImage*, vol. 47, p. S123.
- Ben-David, S., Pál, D., Simon, H., 2007. Stability of *k*-means clustering. 20th Annual Conference on Learning Theory, COLT 2007, San Diego, CA, USA; June 13–15, 2007, vol. 4539/2007. Springer, Berlin/Heidelberg, pp. 20–34.
- Ben-Hur, A., Elisseeff, A., Guyon, L., 2002. A stability based method for discovering structure in clustered data. *Pacific Symposium on Biocomputing*, pp. 6–17.
- Biswal, B., Yetkin, F.Z., Haughton, V.M., Hyde, J.S., 1995. Functional connectivity in the motor cortex of resting human brain using echo-planar MRI. *Magn. Reson. Med.* 34 (4), 537–541.
- Brodmann, K., 1909. *Vergleichende Lokalisationslehre der Grosshirnrinde in ihren Prinzipien dargestellt auf Grund des Zellenbaues*. Johann Ambrosius Barth Verlag, Leipzig.
- Broyd, S.J., Demanuele, C., Debener, S., Helps, S.K., James, C.J., Sonuga-Barke, E.J., 2009. Default-mode brain dysfunction in mental disorders: a systematic review. *Neurosci. Biobehav. Rev.* 33 (3), 279–296.
- Bullmore, E., 2000. How good is good enough in path analysis of fMRI data? *NeuroImage* 11 (4), 289–301.
- Calhoun, V.D., Adali, T., Pekar, J.J., 2004. A method for comparing group fMRI data using independent component analysis: application to visual, motor and visumotor tasks. *Magn. Reson. Imaging* 22 (9), 1181–1191.
- Calhoun, V.D., Liu, J., Adal, T., 2009. A review of group ICA for fMRI data and ICA for joint inference of imaging, genetic, and ERP data. *NeuroImage* 45 (1), S163–S172.
- Chen, S., Ross, T.J., Zhan, W., Myers, C.S., Chuang, K.-S.S., Heishman, S.J., Stein, E.A., Yang, Y., 2008. Group independent component analysis reveals consistent resting-state networks across multiple sessions. *Brain Res.* 1239, 141–151.
- Collins, D.L., Neelini, P., Peters, T.M., Evans, A.C., 1994. Automatic 3D intersubject registration of MR volumetric data in standardized talairach space. *J. Comput. Assist. Tomogr.* 18 (2), 192–205.
- Cordes, D., Haughton, V., Carew, J.D., Arfanakis, K., Maravilla, K., 2002. Hierarchical clustering to measure connectivity in fMRI resting-state data. *Magn. Reson. Imaging* 20 (4), 305–317.
- Damoiseaux, J.S., Beckmann, C.F., Arigita, S.J., Barkhof, F., Scheltens, P., Stam, C.J., Smith, S.M., Rombouts, S.A., 2008. Reduced resting-state brain activity in the “default network” in normal aging. *Cereb. Cortex* 18 (8), 1856–1864 (New York, N.Y.: 1991).
- Damoiseaux, J.S., Greicius, M.D., 2009. Greater than the sum of its parts: a review of studies combining structural connectivity and resting-state functional connectivity. *Brain Struct. Funct.* 213 (6), 525–533.
- Damoiseaux, J.S., Rombouts, S.A., Barkhof, F., Scheltens, P., Stam, C.J., Smith, S.M., Beckmann, C.F., 2006. Consistent resting-state networks across healthy subjects. *Proc. Natl Acad. Sci. USA* 103 (37), 13848–13853.
- Day, W.H., Edelsbrunner, H., 1984. Efficient algorithms for agglomerative hierarchical clustering methods. *J. Classif.* 1 (1), 7–24.
- De Luca, M., Beckmann, C.F., De Stefano, N., Matthews, P.M., Smith, S.M., 2006. fMRI resting state networks define distinct modes of long-distance interactions in the human brain. *NeuroImage* 29 (4), 1359–1367.
- Duda, R.O., Hart, P.E., Stork, D.G., 2000. *Pattern Classification*, 2nd Edition. Wiley-Interscience.
- Efron, B., Tibshirani, R.J., 1994. *An Introduction to the Bootstrap*. Chapman and Hall/CRC.
- Esposito, F., Scarabino, T., Hyvarinen, A., Himberg, J., Formisano, E., Comani, S., Tedeschi, G., Goebel, R., Seifritz, E., Disalle, F., 2005. Independent component analysis of fMRI group studies by self-organizing clustering. *NeuroImage* 25 (1), 193–205.
- Field, C.A., Welsh, A.H., 2007. Bootstrapping clustered data. *J. R. Stat. Soc. B Methodol.* 69 (3), 369–390.
- Filippini, N., MacIntosh, B.J., Hough, M.G., Goodwin, G.M., Frisoni, G.B., Smith, S.M., Matthews, P.M., Beckmann, C.F., Mackay, C.E., 2009. Distinct patterns of brain activity in young carriers of the apoe-epsilon4 allele. *Proc. Natl Acad. Sci. USA* 106 (17), 7209–7214.
- Fox, M.D., Raichle, M.E., 2007. Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. *Nat. Rev. Neurosci.* 8 (9), 700–711.
- Fred, A., Lourenço, A., 2008. Cluster ensemble methods: from single clusterings to combined solutions. *Supervised and Unsupervised Ensemble Methods and their Applications*, pp. 3–30.
- Fred, A.L.N., Jain, A.K., 2005. Combining multiple clusterings using evidence accumulation. *IEEE Trans. Pattern Anal. Mach. Intell.* 27 (6), 835–850.

¹⁶ <http://www.loni.ucla.edu/ICBM/>.

¹⁷ <http://www.clumeq.mcgill.ca/>.

- Giove, F., Gili, T., Iacovella, V., Macaluso, E., Maraviglia, B., 2009. Images-based suppression of unwanted global signals in resting-state functional connectivity studies. *Magn. Reson. Imaging* 27 (8), 1058–1064.
- Greicius, M., 2008. Resting-state functional connectivity in neuropsychiatric disorders. *Curr. Opin. Neurol.* 24 (4), 424–430.
- Greicius, M.D., Srivastava, G., Reiss, A.L., Menon, V., 2004. Default-mode network activity distinguishes Alzheimer's disease from healthy aging: evidence from functional MRI. *Proc. Natl Acad. Sci. USA* 101 (13), 4637–4642.
- Himberg, J., Hyvarinen, A., Esposito, F., 2004. Validating the independent components of neuroimaging time series via clustering and visualization. *NeuroImage* 22 (3), 1214–1222.
- Jain, A., 1987. Bootstrap technique in cluster analysis. *Pattern Recogn.* 20 (5), 547–568.
- Jain, A.K., 2010. Data clustering: 50 years beyond *k*-means. *Pattern Recognition Letters* 31, 651–666.
- Kerr, M.K., Churchill, G.A., 2001. Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. *Proc. Natl Acad. Sci. USA* 98 (16), 8961–8965.
- Lahiri, S.N., 2003. *Resampling Methods for Dependent Data*, 1st Edition. Springer.
- Langers, D.R.M., 2010. Unbiased group-level statistical assessment of independent component maps by means of automated retrospective matching. *Hum. Brain Mapp.* 31 (5), 727–742.
- Laufs, H., Krakow, K., Sterzer, P., Eger, E., Beyerle, A., Salek-Haddadi, A., Kleinschmidt, A., 2003. Electroencephalographic signatures of attentional and cognitive default modes in spontaneous brain activity fluctuations at rest. *Proc. Natl Acad. Sci. USA* 100 (19), 11053–11058.
- Li, J., Tai, B.C., Nott, D.J., 2009. Confidence interval for the bootstrap P-value and sample size calculation of the bootstrap test. *J. Nonparametric Stat.* 21 (5), 649–661.
- Long, X.Y., Zuo, X.N., Kiviniemi, V., Yang, Y., Zou, Q.H., Zhu, C.Z., Jiang, T.Z., Yang, H., Gong, Q.Y., Wang, L., Li, K.C., Xie, S., Zang, Y.F., 2008. Default mode network as revealed with multiple methods for resting-state functional MRI analysis. *J. Neurosci. Methods* 171 (2), 349–355.
- Margulies, D.S., Vincent, J.L., Kelly, C., Lohmann, G., Uddin, L.Q., Biswal, B.B., Villringer, A., Castellanos, F.X., Milham, M.P., Petrides, M., 2009. Precuneus shares intrinsic functional architecture in humans and monkeys. *Proc. Natl Acad. Sci. USA* 106 (47), 20069–20074.
- McKeown, M.J., Makeig, S., Brown, G.G., Jung, T.P., Kindermann, S.S., Bell, A.J., Sejnowski, T.J., 1998. Analysis of fMRI data by blind separation into independent spatial components. *Hum. Brain Mapp.* 6 (3), 160–188.
- Meunier, D., Lambiotte, R., Fornito, A., Ersche, K.D., Bullmore, E.T., 2009. Hierarchical modularity in human brain functional networks. *Front. Neuroinformatics* 3.
- Meyer-Baeae, A., Wismueller, A., Lange, O., 2004. Comparison of two exploratory data analysis methods for fMRI: unsupervised clustering versus independent component analysis. *IEEE transactions on information technology in biomedicine: a publication of the IEEE Engineering in Medicine and Biology Society*, 8(3), pp. 387–398.
- Nigam, A.K., Rao, J.N.K., 1996. On balanced bootstrap for stratified multistage samples. *Stat. Sin.* 6, 199–214.
- Perlberg, V., Bellec, P., Anton, J.-L., Pélégrini-Issac, M., Doyon, J., Benali, H., 2007. CORSICA: correction of structured noise in fMRI by automatic identification of ICA components. *Magn. Reson. Imaging* 25 (1), 35–46.
- Perlberg, V., Marrelec, G., 2008. Contribution of exploratory methods to the investigation of extended large-scale brain networks in functional MRI: methodologies, results, and challenges. *Int. J. Biomed. Imaging* 218519. doi:10.1155/2008/218519.
- Perlberg, V., Marrelec, G., Doyon, J., Pélégrini-Issac, M., Lehericy, S., Benali, H., 2008. NEDICA: detection of group functional networks in fMRI using spatial independent component analysis. 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro, 2008. ISBI 2008, pp. 1247–1250.
- Raghavan, V.V., 1982. Approaches for measuring the stability of clustering methods. *SIGIR Forum* 17 (1), 6–20.
- Rousseeuw, P., 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* 20 (1), 53–65.
- Salvador, R., Suckling, J., Coleman, M.R., Pickard, J.D., Menon, D., Bullmore, E., 2005. Neurophysiological architecture of functional magnetic resonance images of human brain. *Cereb. Cortex* 15 (9), 1332–1342.
- Shao, J., Tu, D., 1995. *The Jackknife and Bootstrap*. Springer.
- Shehzad, Z., Kelly, A.M.C., Reiss, P.T., Gee, D.G., Gotimer, K., Uddin, L.Q., Lee, S.H., Margulies, D.S., Roy, A.K., Biswal, B.B., Petkova, E., Castellanos, F.X., Milham, M.P., 2009. The resting brain: unconstrained yet reliable. *Cereb. Cortex* 19 (10), 2209–2229.
- Shimodaira, H., 2004. Approximately unbiased tests of regions using multistep-multiscale bootstrap resampling. *Ann. Stat.* 32 (6), 2616–2641.
- Shmuel, A., Leopold, D.A., 2008. Neuronal correlates of spontaneous fluctuations in fMRI signals in monkey visual cortex: implications for functional connectivity at rest. *Hum. Brain Mapp.* 29 (7), 751–761.
- Sitter, R.R., 1992. A resampling procedure for complex survey data. *J. Am. Stat. Assoc.* 87 (419), 755–765.
- Smith, S.M., Fox, P.T., Miller, K.L., Glahn, D.C., Fox, P.M., Mackay, C.E., Filippini, N., Watkins, K.E., Toro, R., Laird, A.R., Beckmann, C.F., 2009. Correspondence of the brain's functional architecture during activation and rest. *Proc. Natl Acad. Sci. USA* 106 (31), 13040–13045.
- Smith, S.P., Dubes, R., 1980. Stability of a hierarchical clustering. *Pattern Recogn.* 12 (3), 177–187.
- Sorenson, J.A., Wang, X., 1996. ROC methods for evaluation of fMRI techniques. *Magn. Reson. Med.* 36 (5), 737–744.
- Steinley, D., 2008. Stability analysis in *k*-means clustering. *Br. J. Math. Stat. Psychol.* 61, 255–273.
- Sui, J., Adali, T., Pearlson, G.D., Clark, V.P., Calhoun, V.D., 2009. A method for accurate group difference detection by constraining the mixing coefficients in an ICA framework. *Hum. Brain Mapp.* 30 (9), 2953–2970.
- Suzuki, R., Shimodaira, H., 2006. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* 22 (12), 1540–1542.
- Thirion, B., Flandin, G., Pinel, P., Roche, A., Ciuciu, P., Poline, J.B., 2006. Dealing with the shortcomings of spatial normalization: multi-subject parcellation of fMRI datasets. *Hum. Brain Mapp.* 27 (8), 678–693.
- Toro, R., Fox, P.T., Paus, T., 2008. Functional coactivation map of the human brain. *Cereb. Cortex* 18 (11), 2553–2559.
- Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., Joliot, M., 2002. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage* 15 (1), 273–289.
- van den Heuvel, M., Mandl, R., Hulshoff, 2008. Normalized cut group clustering of resting-state fMRI data. *PLoS ONE* 3 (4), e2001+.
- Vincent, J.L., Patel, G.H., Fox, M.D., Snyder, A.Z., Baker, J.T., Van Essen, D.C., Zempel, J.M., Snyder, L.H., Corbetta, M., Raichle, M.E., 2007. Intrinsic functional architecture in the anaesthetized monkey brain. *Nature* 447 (7140), 83–86.
- Weissenbacher, A., Kasess, C., Gerstl, F., Lanzenberger, R., Moser, E., Windischberger, C., 2009. Correlations and anticorrelations in resting-state functional connectivity MRI: a quantitative comparison of preprocessing strategies. *NeuroImage* 47 (4), 1408–1416.
- Woolrich, M.W., Behrens, T.E., Beckmann, C.F., Jenkinson, M., Smith, S.M., 2004. Multilevel linear modelling for fMRI group analysis using Bayesian inference. *NeuroImage* 21 (4), 1732–1747.
- Zhong, Y., Wang, H., Lu, G., Zhang, Z., Jiao, Q., Liu, Y., 2009. Detecting functional connectivity in fMRI using PCA and regression analysis. *Brain Topogr.* 22 (2), 134–144.
- Zijdenbos, A.P., Forghani, R., Evans, A.C., 2002. Automatic "pipeline" analysis of 3-D MRI data for clinical trials: application to multiple sclerosis. *IEEE Trans. Med. Imaging* 21 (10), 1280–1291.
- Zuo, X.-N., Kelly, C., Adelstein, J.S., Klein, D.F., Castellanos, F.X., Milham, M.P., 2010. Reliable intrinsic connectivity networks: test-retest evaluation using ICA and dual regression approach. *NeuroImage* 49 (3), 2163–2177.