An empirical comparison of surface-based and volume-based group studies in neuroimaging

Alan Tucholka, Virgile Fritsch, Jean-Baptiste Poline, Bertrand Thirion

An empirical comparison of surface-based and volume-based group studies in neuroimaging

Alan Tucholka[a,b], Virgile Fritsch[d,c], Jean-Baptiste Poline[c,d], Bertrand Thirion[d,c]

[a]*Radiology Department, Centre hospitalier de l'Université de Monréal (CHUM) - Hôpital Notre-Dame, Montreal*
[b]*Research Centre, Hôpital Sainte-Justine, Montréal*
[c]*Parietal project-team, INRIA Saclay-Île de France, 91893 Orsay, France*
[d]*CEA, DSV, I2BM, Neurospin, Bâtiment 145, 91191 Gif-sur-Yvette, France*

Abstract

Being able to detect reliably functional activity in a population of subjects is crucial in human brain mapping, both for the understanding of cognitive functions in normal subjects and for the analysis of patient data. The usual approach proceeds by normalizing brain volumes to a common three-dimensional template. However, a large part of the data acquired in fMRI aims at localizing cortical activity, and methods working on the cortical surface may provide better inter-subject registration than the standard procedures that process the data in the volume. Nevertheless, few assessments of the performance of surface-based (2D) versus volume-based (3D) procedures have been shown so far, mostly because inter-subject cortical surface maps are not easily obtained. In this paper we present a systematic comparison of 2D versus 3D group-level inference procedures, by using cluster-level and voxel-level statistics assessed by permutation, in random effects (RFX) and mixed-effects analyses (MFX). We consider different schemes to perform meaningful comparisons between thresholded statistical maps in the volume and on the cortical surface. We find that surface-based multi-subject statistical analyses are generally more sensitive than their volume-based counterpart, in the sense that they detect slightly denser networks of regions when performing peak-level detection; this effect is less clear for cluster-level inference and is reduced by smoothing. Surface-based inference also increases the reliability of the activation maps.

Keywords: Cortical surface mapping, fMRI, statistical inference, cluster-level tests, random effects analysis, mixed effects analysis, between-subject variability.

1. Introduction

Studying the localization and variability of brain activity across subjects are two crucial aspects of neuroimaging data analysis. This is important both for building models of brain activation organization and to relate inter subjects activity variations to factors of interest, such as behavioral and genetic measurements. In particular, many neuroscience results depend on the the precise localization of the Blood Oxygen-Level Dependent (BOLD) signal changes measured by functional Magnetic Resonance Imaging (fMRI). Controlling the variability of activation position across individuals is essential in order to grant enough sensitivity in activation detection, and to ensure that one-sample group-level inference, that deals with between-subject mean activation, is indeed

representative of the true population pattern. Both co-registration and activity detection are most commonly performed in 3D space, by applying linear and non-linear warps to match the anatomical and functional images to a common template, often chosen to be the MNI template [?]. Brain activations are then described through the positions of local maxima of activation maps in template space, and these locations are then interpreted in terms of brain regions using standard atlases (see e.g. [? ?]).

Activity localization. Once the data are in the common template space, the only information available on the localization of the BOLD signal are the coordinates (x,y,z) in this space. This view precludes a deep understanding of brain organization, which should consider the details of anatomical structure in order to define properly activity localization [?]; in general, functional regions should be characterized by their relative position with respect to anatomical landmarks, their extent or their connections to other regions [?]. Imperfect spatial registration procedures probably induce an additional blur that is detrimental to the accuracy of brain mapping and to sensitivity.

Activity detection. The standard approach to activation detection [?] consists in comparing the images from the different subjects on a voxel-by-voxel basis, and to compute a statistical map to test the presence of an activation in each voxel of the standard space. The ensuing multiple testing problem can be addressed directly at the voxel-level [?] or by testing the size of clusters defined above a user-chosen threshold [?].

1.1. Volume- and surface-based spatial normalization

Spatial normalization is therefore crucial to define accurately the position of functional regions, and thus to detect positive activation at the population level. Many works have been carried out to improve these volume-based co-registration procedures (see [?] for a review of most of these procedures). Yet, as a very large part of the data originates from the cortex, methods that work on the cortical surface may be more sensitive than those using the full brain volume. It is well known that volume-based normalization may introduce inaccuracies in anatomical positioning of functional data, estimated up to 1cm in several cortical regions [? ?]. For instance, it is difficult to account for the inter-subject variability of gyri size, shape or position in a 3D referential and such differences may correspond to the displacement of a functional focus to a different gyrus in some subjects.

Surface-based spatial *normalization* proceeds from a different perspective: it consists in the definition of coordinate systems on the cortical surface that match corresponding regions across individuals. This is done by using geometric descriptors of the surface, such as the main sulci [? ?] or smoothed curvature maps [?]. A review of these techniques has recently been performed in [?]. Surface-based methods do not guarantee that a perfect match will be found across individuals, in particular because sulcal patterns are known to be partly inconsistent even in standard populations [?]. However, these surface-based methods are probably more accurate for aligning cortical folds than volume-based methods, and they have been shown to correctly align cyto-architectonic boundaries [?]. Several studies have shown that a co-registration algorithm based on the cortical surface may better align the functional signal across subjects [? ? ? ?], but none of these studies used standard group-level random-effects inference procedures, such as a one-sample t-test across individuals.

More generally, besides the fact that non-cortical regions are not analyzed by a surface-based procedure, this analysis suffers from an important limitation: it requires that functional data has been accurately projected on the cortical surface, i.e. that the correspondence of functional data with anatomical data is correct. This is a real challenge in practice, as the cortical thickness, but also the distance between the banks of a given sulci or the width of a gyrus are small, in view of the typical resolution of fMRI data [**?**]. Moreover, fMRI data are artefacted by geometric distortions related to EPI acquisition. These distortion are not systematically and fully corrected during pre-processing [**?**]. In the present work, we do not focus on the evaluation of the main steps that are necessary to perform adequate surface-based analysis, such as the comparison of functional data projection techniques or EPI distortion correction, but propose different methods to compare the results of various group-level inference procedures applied on the cortical surface and in the volume. For these comparisons, we use state-of-the-art group analysis statistical approaches. A procedure has recently been proposed in [**?**] to assess the quality of the co-registration between a cortical mesh and fMRI data, and is expected to optimize the spatial match between anatomical structures and functional information. However, to the best of our knowledge, there has been no recent comprehensive investigation on the impact of 2D versus 3D activation detection on fMRI group analysis.

1.2. An empirical comparison of volume- and surface-based alignment for functional brain mapping

In this paper, we investigate whether surface-based approaches, that rely on a cortical surface referential, provide better constraints about the position of functional activity, and more precisely, whether this is reflected in state-of-the-art inter-subject statistical procedures. Following [**?**], we perform functional analysis on the cortical surface for a group of 25 subjects. The inter-subjects analysis relies on matching the subjects cortical surface [**?** **?**]. Then we systematically compare the results of surface- and volume-based statistical analysis and provide results on the difference in sensitivity of the two approaches for different tests, for a given control of the type I error. More specifically, we use for the comparison mixed- and random-effects inference at the voxel and at the cluster level [**?**] and assess their reproducibility with bootstrap.

This raises a difficult question: how should we compare group results obtained in the volume to those obtained on the cortical surface? In the present work we explore and detail three possibilities, that are based either on the relative activated volume, or the choice of surface- or volume- based spatial referential to compare the results of either kind of analysis. We also address the impact of choosing a surface- or a volume-based smoothing kernel on the results, as this is known to have an important effect on detection and localization in group analysis. We consider an fMRI protocol designed to map quickly several cognitive functions [**?**] from which we use several contrasts to compare the surface and volume analyses, and provide to cognitive neuroscientists comprehensive information on the sensitivity of those analyses.

The remainder of this paper is organized as follows: in section 2 we first present the data used in our experiments, and then the processing and inference procedures tested; we describe the results of these tests in section 3 and discuss our results in section 4.

2. Materials and methods

2.1. Dataset used in our experiments

Data were acquired from 25 subjects who performed a *functional localizer* protocol as described in [?]. This protocol is intended to activate multiple brain regions in a relatively short time (128 brain volumes acquired in 5 minutes) with ten experimental conditions, allowing the computation of many different functional contrasts: left and right button presses after auditory or visual instruction, mental computation after auditory or visual instruction, sentence listening or reading, passive viewing of horizontal and vertical checkerboards. The subjects gave informed consent and the protocol was approved by the local ethics committee.

Functional images were acquired on an 1.5T General Electric Signa System scanner (General Electric Medical Systems, Milwaukee, WI, USA) using an EPI sequence (TR = $2400ms$, TE = $60ms$, matrix size = 64×64, FOV = $24cm \times 24cm$, echo spacing = $608\mu s$). Each volume consisted of 40 $3mm$-thick axial slices without gap. A session comprised 132 EPI scans, of which the first four were discarded to allow the MR signal to reach steady state. The slices were acquired in interleaved ascending order. Anatomical fSPGR T1-weighted images were acquired on the same scanner, with a slice thickness of 1.2 mm, a field of view of 24 cm and an acquisition matrix of $256 \times 256 \times 124$ voxels, resulting in 124 contiguous double-echo slices with voxel dimensions of $(0.9375 \times 0.9375 \times 1.2)\, mm^3$.

2.2. Pre-processing

The functional data were first corrected from the EPI distortions using field maps and the SPM toolbox; the ensuing image resolution was $3.75 \times 3.75 \times 3mm^3$. Note that the fMRI were considered as initially aligned with the B0 map, and thus were not re-interpolated prior to distortion correction; the application of the correction in itself induced a re-interpolation of the data. Next, a standard pre-processing (correction of differences in slice timing, rigid-body motion correction and anatomo-functional co-registration) was performed using the SPM5 software on all subjects.

The FreeSurfer software [? ?], version 4.0.2 was used to segment and reconstruct the cortical surface from T1 MRI data of each subject, and obtained the white matter mesh for both hemispheres. This provides a common spherical coordinate system for each hemisphere in each subject. Pre-processing of the data includes *i)* segmentation of the white matter, yielding triangular meshes for grey-white and grey-csf (cerebro-spinal fluid) interfaces, *ii)* detection of the deepest sulci, *iii)* inflation of the white surface onto a sphere, *iv)* deformation to match the deepest sulci positions on the template model.

All data are then converted to the standard GIFTI format for further processing. In order to obtain a node-by-node correspondence, the mesh of each subject was then resampled : *i)* a regular sphere (icosphere) of diameter equal to the brain-sphere with a reduced number of nodes (about 40k nodes) was created, *ii)* this sphere was refolded onto the original cortical surface of each subject while preserving node-to-node correspondence with the original icosphere mesh. This corresponds to resampling the individual meshes to match a template mesh, as is usually done in volume-based normalization procedures. Downsampling to 40k nodes entails a minor loss of resolution when compared to fMRI resolution, and provides a four-fold speed gain in the ensuing statistical procedures with respect to the full mesh resolution.

An average cortical model across 25 subjects was created for visualization of the results. Each resampled mesh was embedded into the MNI space, then an average brain was obtained by

computing the mean 3D position of each node through all subjects in that space. It is important to notice that this mesh is used only for visualization purpose and not in the actual statistical analyses.

Functional images were then projected onto the resampled gray/white interface mesh of each subject using the method described in [**?**]. This projection relies on the geometry of the local anatomy: for each node of the mesh of the grey/white interface, a projection kernel is computed, so that it yields a weighted average of the neighboring voxels, where the weights decrease sharply along the normal direction outside of the cortical ribbon, and more smoothly in the tangent plane at the surface. The projection is then performed by convolution of the functional 3D map with the kernels of each node. The parameters of the decay are 5mm in the tangential direction and 2mm in the normal direction. For all subjects and hemispheres, we visually checked that there was no spatial mismatch between the brain mesh and the functional volume. A General Linear Model (GLM) analysis was applied for the volume- and surface-based data using the same analysis code, namely the nipy package `http://nipy.sourceforge.net/`. The model included the ten conditions of the experiments convolved with a standard hemodynamic filter and its time derivative, a high-pass filter (cutoff:128s) and the procedure included an estimation of the noise auto-correlation using an AR(1) model [**?**]. Activation maps were derived for the following six functional contrasts (we give short names in italic): *i) left-right*: left versus right button presses, *ii) right-left*: right versus left button presses, *iii) audio-video*: sentence listening versus sentence reading, *iv) video-audio*: sentence reading versus sentence listening, *v) computation-sentences*: computation versus sentence reading, *vi) reading-visual*: reading versus passive checkerboard viewing. These contrasts are expected to reveal different aspects of the functional organization of the brain, at different localizations, and show different sensitivity and reproducibility levels (see [**?**] for a detailed discussion).

In parallel, the fMRI data was also analyzed, using the same linear model, after non-linear spatial normalization, which was performed with the SPM5 software using the *new segment* method. This normalization was based on the coregistration with the T1 image that had been warped to the SPM T1 template. Before the linear model application, we used different levels of smoothing on the cortical surface and in the volume, corresponding to 0, 4, 8 and 12mm full width at half maximum (fwhm) for the 3D data and 0 or 8mm on the surface, in order to check the impact of smoothing kernel. We also segmented the grey matter in the individual anatomy using SPM5, and created a population-level mask of the voxels that belong to the grey matter with a probability greater than .5; this masked was resampled to the fMRI resolution.

2.3. Experiments

2.3.1. Statistical model for group analysis

In this work, $S = 25$ subjects are considered. For each subject i, and at any location of the domain under consideration (cortical mesh or brain volume), let $\hat{\beta}_i$ be the estimation of the BOLD effect related to some functional contrast of interest (for notational simplicity, we drop explicit references to the contrast under consideration). $\hat{\beta}_i$ is distributed around the true effect β_i: $\hat{\beta}_i = \beta_i + e_i$ with $e_i \sim \mathcal{N}(0, s_i^2)$ where the estimation variance s_i^2 is assumed to be known from the first-level General Linear Model (GLM)[1]. As is traditionally done in neuroimaging, we further

[1] As there are often many degrees of freedom at the first level the error made here is likely to be small.

assume that $\beta_i = \beta_G + \varepsilon_i$, where $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ and β_G is the population-level effect [?]. We thus have :

$$\hat{\beta}_i = \beta_G + \varepsilon'_i, \quad \varepsilon'_i \sim \mathcal{N}(0, \sigma^2 + s_i^2), \tag{1}$$

where σ^2 is the between-subject variance. This is a generalization of the Random Effects (RFX) model in [?] which neglects the $\hat{\beta}_i$ estimation variance, *i.e.* it assumes $s_i^2 \equiv 0$. Both β_G and σ^2 are then estimated by maximizing the log-likelihood of the model specified in Eq. (1) using the Expectation-Maximization (EM) algorithm in [?]. The following log-likelihood ratios are computed to test the positivity of β_G:

$$\mathcal{L}_{MFX} = \log \frac{\sup_{\sigma^2, \beta_G > 0} \prod_{i=1}^{S} \mathcal{N}(\hat{\beta}_i; \beta_G, \sigma^2 + s_i^2)}{\sup_{\sigma^2} \prod_{i=1}^{S} \mathcal{N}(\hat{\beta}_i; 0, \sigma^2 + s_i^2)}, \tag{2}$$

$$\mathcal{L}_{RFX} = \log \frac{\sup_{\sigma^2, \beta_G > 0} \prod_{i=1}^{S} \mathcal{N}(\hat{\beta}_i; \beta_G, \sigma^2)}{\sup_{\sigma^2} \prod_{i=1}^{S} \mathcal{N}(\hat{\beta}_i; 0, \sigma^2)} \tag{3}$$

Note that computing $\mathcal{L}_{RFX}$ is equivalent to performing a t-test.

2.3.2. Statistical calibration

The distribution of the statistics in Eq. (2-3) under the null hypothesis ($\beta_G = 0$) is unknown, but can be estimated very simply through a randomization procedure, in which the statistics are recomputed after a sign swap of the observed effects $\hat{\beta}_i$. Under the hypothesis that the distribution of the true effects is symmetric about zero everywhere in the brain –which is our null hypothesis– this procedure yields an exact (possibly conservative) specificity for the test. In order to control the family-wise error rate (FWER), *i.e.* the probability of detecting one false positive region over the search domain, we consider the distribution of the maximal statistic under the null hypothesis. For a chosen FWER α, this yields a voxel- or vertex- level corrected threshold.

A more sensitive approach to detect extended regions consists in first thresholding the statistics map at a given level (corresponding e.g. to $p < 10^{-3}$ uncorrected), and then to estimate the distribution of size (area (in mm^2) or volume (in mm^3)) of the supra-threshold clusters under the null hypothesis. To solve the multiple comparison issue, the size of the maximal cluster is tabulated under the null hypothesis. Once again, the quantile α of this simulated distribution yields a cluster-level corrected threshold.

In summary, we use the following statistics: voxel-level/vertex-level random effects test (VRFX) and mixed effects test (VMFX), cluster-level random effects test (CRFX) and mixed effects test (CMFX). We will always refer to *surface-based* or *volume-based* statistics to make the underlying spatial model explicit.

2.4. Comparison of surface-based and volume-based results

Next we discuss how to compare the difference in sensitivity between the spatial- and volume-based analyses. One first challenge is that this comparison is performed in the absence of a ground truth, so that it is difficult to characterize the correctness of surface- or volume-domain detections. However, the procedure that we use has the theoretical guarantee of providing the same rate of false detections (type I error) for each test. This holds separately for univariate tests

on the one hand, and cluster size tests on the other hand (the sensitivity of both tests are not easily compared directly as they do not have the same regional specificity). In these conditions, the fact that a spatial procedure yields the detection of *more* regions means that this procedure is better suited to detect the functional activity related to this contrast. The difficulty is thus to quantify the relative amount of detections performed in either domain, as the results are observed and characterized in two different spaces. In this work, we describe and perform three possible comparisons:

- The first one is to consider the *test sensitivity*, i.e. the relative amount of activated regions in either domain (surface or volume). This can give an indication on the performance of either domain to outline active regions, but this is confounded by the fact that 3D maps include sub-cortical regions, in which activations may (cerebellum, deep nuclei) or may not (white matter) be found. We have tried to remove this confound by running also the volume-based analysis on a mask of the grey matter.

- A second possible comparison is to *project* the result of group-level volume-based analyses onto the cortical surface of each subject, and then to report the activated surface (in mm^2) obtained, as compared to the surface declared active after full surface-based analysis. In the absence of a satisfactory population-level model of the cortical surface, we iteratively took each individual surface as reference space, and report the mean and standard deviation of the activated surface across these models. This avoids the confounds previously described, but it is difficult to decide which vertices are *active* when we recombine the projection of the group results on a given individual anatomy: in this work, we decided to declare active all nodes that reached a signal level of .5 after projection of a binary activation maps (0 for inactive, 1 for active voxels) onto the cortical surface, corresponding to a 50% confidence that the node is indeed active.

- Finally, a third solution consists in *embedding* the activated regions obtained on the cortical surface into the MNI space, and then to compare the two sets of positions in MNI space. The position comparison procedure is defined using a kernel-base metric that measures the discrepancy between different sets of positions, and is described in Section 2.4.1. Again, as the embedding depends on the chosen cortical surface, we used the grey/white interface mesh of each subject and average the result across all 25 subjects.

The *embedding* and *projection* approaches are illustrated in Fig. 1. Finally, we also considered the impact of choosing the medial cortical surface instead of the grey-white matter interface in the *projection* approach.

2.4.1. Kernel similarity between active regions

Here we define the similarity measure used to compare the coordinates obtained from surface- and volume-based analysis in the MNI space: Let $t^{surf} = (t_i^{surf})_{i=1..N}$ and $t^{vol} = (t_i^{vol})_{i=1..V}$ be the set of coordinates of supra-threshold positions obtained from surface-based and volume-based analysis respectively, N and V being the number of nodes and voxels under consideration. We define the two quantities:

$$\psi(t^{vol}; t^{surf}) = \underset{v=1..V}{\text{mean}} \max_{n=1..N} \exp\left(-\frac{\|t_v^{vol} - t_n^{surf}\|^2}{2\delta^2}\right) \tag{4}$$

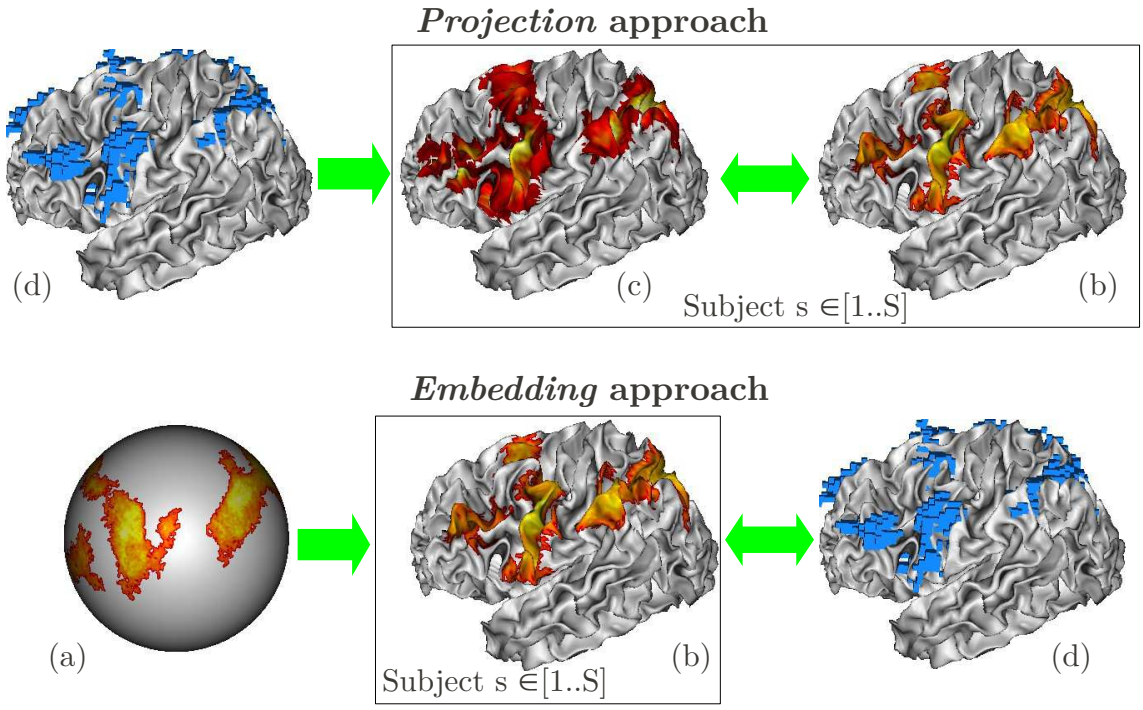

Figure 1: Illustration of the two main methodologies for the comparison of surface-based and volume-based group analysis results. In the *projection* approach (top), the activation found in MNI space (d) are projected onto the the cortical surface of each individual (c), and thus compared to the active areas found in the surface-based analysis, then represented in the cortical geometry of the same subject (b). In the *embedding* approach (bottom) the activations obtained on the surface in the common space represented by a sphere (a), are embedded in the 3D space related to the cortical geometry of each individual in the MNI space (b), and then compared to the volume activation found in MNI space (d).

and

$$\psi(t^{surf};t^{vol}) = \underset{n=1..N}{\text{mean}} \max_{v=1..V} \exp\left(-\frac{\|t_v^{vol} - t_n^{surf}\|^2}{2\delta^2}\right), \tag{5}$$

where δ is a fixed distance ($\delta = 6$mm in our experiments). $\psi(t^{surf};t^{vol})$ and $\psi(t^{vol};t^{surf})$ quantify how well the values in (t^{vol}) and (t^{surf}) correspond to each other: A high value of $\psi(t^{surf};t^{vol})$ indicates that, for each supra-threshold region on the surface, a close active region can in general be found in the volume. Reciprocally, a high value of $\psi(t^{vol};t^{surf})$ indicates that each supra-threshold voxel in the volume is closely matched by a supra-threshold node on the surface. Put differently $\frac{1}{2}\left(\psi(t^{vol};t^{surf}) + \psi(t^{surf};t^{vol})\right) \in [0,1]$ provides a concordance index of the supra-threshold region on both domains, while a positive difference $\psi(t^{vol};t^{surf}) - \psi(t^{surf};t^{vol})$ indicates that volume-based activations are more frequently found in the vicinity of surface-based activations than the converse, i.e. that surface-based analysis is more sensitive.

Regarding the choice of δ, the meaning of this parameter is that, whenever the distance between one node and it nearest voxel is larger that $2 \times \delta$, the contribution to ψ is close to 0, meaning that the node has no corresponding active voxel in the volume. δ measures the admissible discrepancy of the position of activations in MNI space, and we found $\delta = 6mm$ to be a reasonable choice. However, our evaluation metric varies smoothly with respect to this parameter, so that the results are stable with respect to this choice too.

2.5. Bootstrap reproducibility of the active regions

Finally, we performed an analysis of the reproducibility of the active regions in 2D and 3D using a bootstrap procedure: we draw S subject with replacement from the initial population B times, perform the statistical analysis, and compute a replication map R_B which yields the counts of activation of each site. A reproducibility index is derived below from this replication map. Although a measure based on a binomial mixture has been proposed previously for this purpose [**?** **?** **?**], we use here a direct estimation of the conditional probability ρ that a supra threshold voxel in one particular dataset will also be above threshold in another sample. This was simply obtained by counting the average number of co-activations across pairs of experiments, divided by the average number of activation detections across experiments. Let h be the histogram of R_B: $(h_b)_{1 \leq b \leq B}$ counts the number of times a voxel/node has been declared active among B replications; then let

$$\rho = \frac{1}{B-1} \frac{\sum_{b=1}^{B} b(b-1)h_b}{\sum_{b=1}^{B} bh_b} \tag{6}$$

$\rho \in [0,1]$ measures the probability that a given active site will in average, be reproduced in a following experiment. In our experiments, we use $B = 10$ and $S = 25$.
For visualization, we used the Anatomist software.

3. Results

A summary of the active regions activations found for the six contrasts under investigation are given in Fig. 2.

We successively report the results of the following tests for six different functional contrasts studied in the protocol:

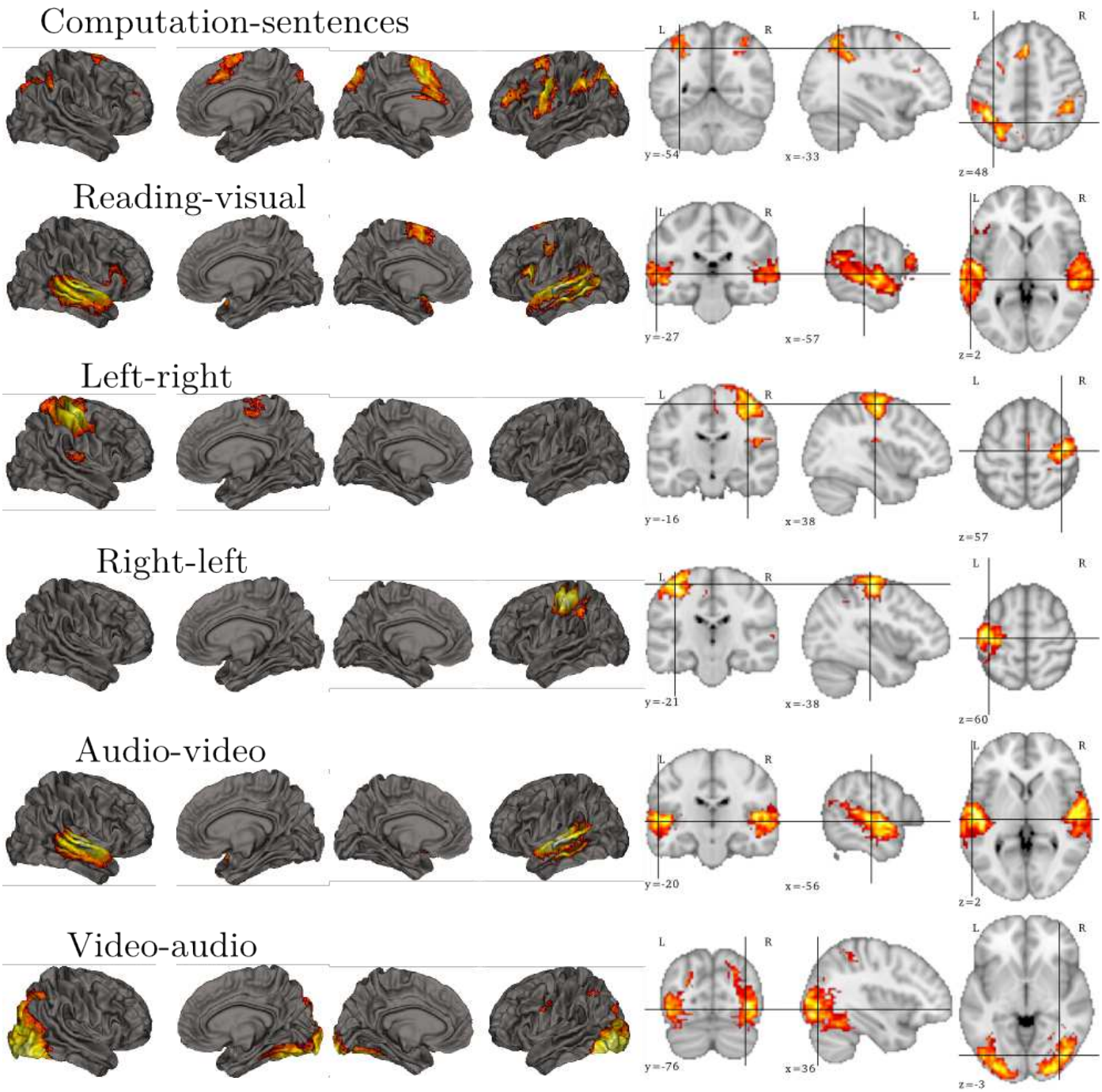

Figure 2: Activations found in the six contrasts under investigation. The results are represented for a mixed-effects model tested with cluster-level inference. The left plots show the active regions on the cortex, for the left and right hemisphere, using inner and outer views. The rights plot shows the volume maps centered at the map-level peak, and all the maps are thresholded at the $p < .05$ corrected, level.

Contrasts	computation-sentences	reading-visual	left-right	right-left	audio-video	video-audio
VMFX, volume	0.126	0.975	0.287	0.317	1.018	0.755
VMFX, surface	1.298	3.571	1.754	1.114	3.562	5.474
CMFX, volume	2.075	4.172	1.337	1.25	3.193	4.906
CMFX, surface	9.221	8.96	3.731	2.326	5.907	11.965

Table 1: Relative volume (in percent) of activated areas for different statistical inference procedures performed either on the surface or in the volume. The tests are performed for 6 different contrasts (in column), and the results are provided with voxel/vertex-level mixed effects inference (VMFX) and cluster-level mixed effects inference (CMFX). The threshold p-value is 0.05, corrected.

- Comparison of the relative amount of activated regions found in surface- and volume-based analysis.
- Comparison of surface-based and volume-based supra-threshold regions through kernel-based similarity.
- Reproducibility of the supra-threshold regions in the surface of in the volume.

The level of smoothing can have a confounding effect on the conclusions that might be drawn from these experiments. We thus systematically considered different levels of smoothing: 0, 4, 8 and 12 mm fwhm in the volume, 0 or 8 mm on the cortical surface. We report all or part of these results in the following experiments. Surface-based smoothing was performed on the mesh, using a tool from the Brainvisa toolbox `http://brainvisa.info/index.html`.

3.1. Relative sensitivity

3.1.1. Rate of significantly active regions

The proportion of supra-threshold surface or volume is given in Table 1 for mixed-effects inference, and in supplementary material for random-effects inference (see Table A.3). We observe that: *i)* cluster-level statistics systematically yield more extended supra-threshold regions than voxel- or node-level statistics; *ii)* mixed-effects statistics are more sensitive than random-effects statistics; *iii)* surface-based analysis systematically yields a larger relative portion of activated regions than volume-based analysis. Results *i)* and *ii)* are common findings, while *iii)* provides a first empirical confirmation of the advantage of surface-based analyses.

In the sequel, we no longer report results with VRFX/CRFX, since they are qualitatively similar to those obtained with VMFX and CMFX respectively, with a general tendency toward less sensitivity and reproducibility.

Next, we consider the impact of the level of smoothing on these quantities; to simplify the presentation, we only consider here a smoothing kernel of 8mm fwhm, applied on the surface or in the volume. The results are given in Table 2 using mixed-effects inference at the voxel/node (VMFX) level only. We also include the result in the volume, but limited to the grey matter, without smoothing. This should be compared with the results provided in Table 1, that gives corresponding results without smoothing.

We observe that the tendency observed previously still holds after smoothing the data: the proportion of activations is larger on the surface, and that smoothing increases this proportion.

Contrasts	computation-sentences	reading-visual	left-right	right-left	audio-video	video-audio
volume, 8mm smoothing	1.312	3.887	1.133	1.287	3.409	3.846
surface, 8mm smoothing	5.802	6.966	2.9	1.951	5.687	9.602
GM volume, no smoothing	0.204	1.516	0.447	0.483	1.634	1.166

Table 2: Relative volume (in percent) of activated areas using voxel/vertex-level mixed effects inference (VMFX), on the surface and in the volume, for a smoothing kernel of 8mm fwhm. The tests are performed for 6 different contrasts (in column). The last row yields results on volume data, restricted to the grey matter and without smoothing.

Restricting the volume analysis to grey matter increases the values, but it still remains below surface-based results.

3.1.2. Spatial extent of significantly active regions on the cortical surface (projection approach)

The spatial extent of the supra-threshold regions after surface-based analysis is compared with the extent of the 3D active regions of the volume-based group map, reprojected onto each individual surface. The average activated region is reported, as explained in section 2.4. The results are shown in Fig. 3 for the six contrasts under study, using the voxel-level mixed effects statistics (VMFX). We also consider the impact of smoothing the data in the volume or on the surface. We also ran the experiment on a mask of the grey matter in the volume, but did not find any difference with respect to a full-volume study.

As expected, we clearly see that surface-based analysis yields a larger portion of the cortical surface, and that smoothing the data tends to increase the relative size of active regions. It is also important to note that the amount of activated surface after re-projection depends only weakly on the individual anatomy used for projection, as illustrated by the relatively low variability of the area of the projection.

3.2. Kernel-based sensitivity analysis (embedding approach)

Next, we investigate the similarity of the surface and volume-based patterns by considering the natural embedding the cortical surface in MNI space, and then assessing the similarity of the positions of active regions using the metric detailed in section 2.4.1. The results are shown in Fig. 4 for six contrasts, using CMFX and VMFX inference, and with two levels of smoothing: 0mm and 8mm.

We observe a very high similarity overall, with similarity values close to .8, meaning that on average, an active voxel can be found within $d = 4mm$ ($(\exp -\frac{d^2}{2\delta^2} = .8)$ yields $d \simeq 4mm$, given that $\delta = 6mm$) of an active node, and conversely. However, some differences can be noted, and in particular between voxel-based and cluster-based inference: without additional smoothing $\psi(t^{surf}; t^{vol})$ is similar to $\psi(t^{vol}; t^{surf})$ in cluster-level inference, while $\psi(t^{vol}; t^{surf}) > \psi(t^{surf}; t^{vol})$ in voxel/vertex-level inference; when an 8mm smoothing kernel is applied to volume and surface data, $\psi(t^{surf}; t^{vol}) > \psi(t^{vol}; t^{surf})$ in cluster-level inference, while $\psi(t^{vol}; t^{surf}) \simeq \psi(t^{surf}; t^{vol})$ in voxel/vertex-level inference. This readily means that, in peak-level inference, the surface-based pattern seems to reveal more details (activation peaks) than voxel-based analysis. This effect vanishes in cluster-level inference, or when the level of smoothing is increased. The combination

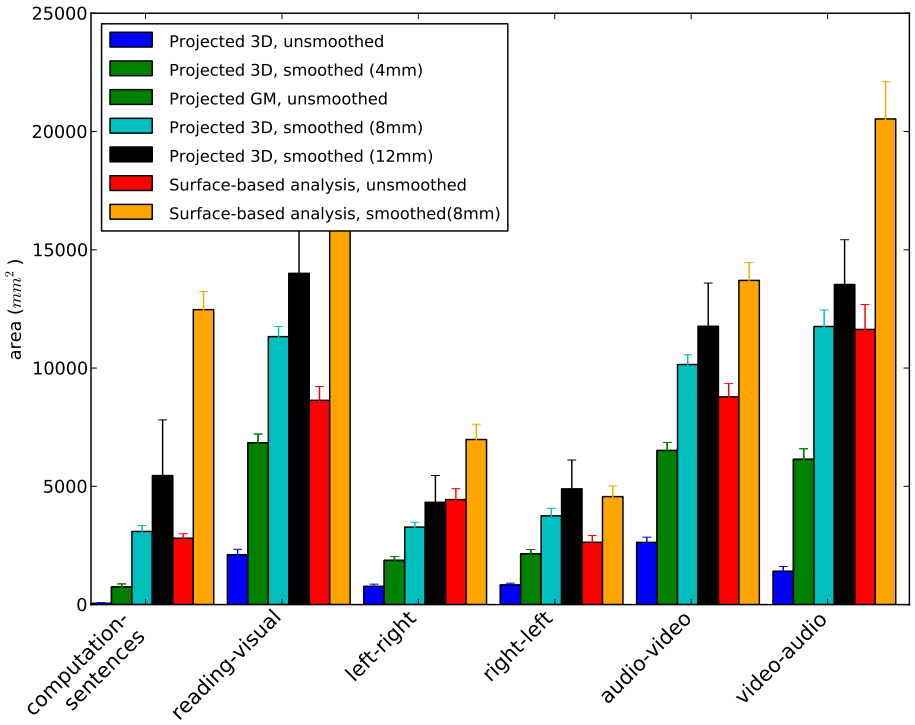

Figure 3: Comparison of the extent (in mm^2) of supra-threshold regions on the surface after full analysis on the surface or after re-projection of the 3D results onto the surface. The results are shown using the VMFX statistics on the six contrasts and for various smoothing kernel sizes, on the surface and in the volume. Similar trends are observed using other statistics, cluster-level inference, and stronger smoothing. Error bars represent the variability associated with the choice of the individual cortical surface as projection space.

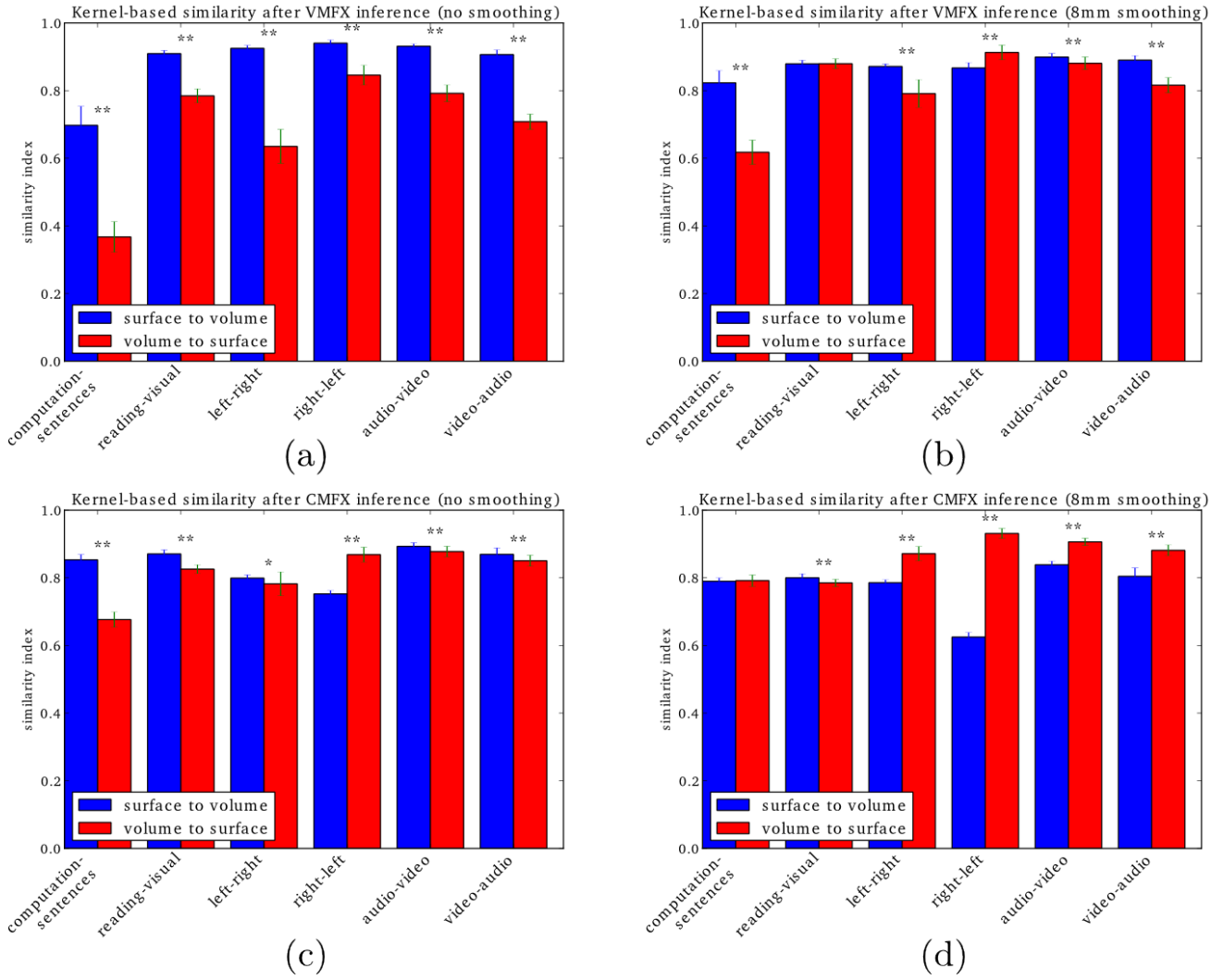

Figure 4: Kernel-based similarity of surface-based and volume-based activated regions. The bars in blue indicate how close the position of supra-threshold voxels is to nearby supra-threshold cortical nodes in average; reciprocally, the bars in red indicate how close supra-threshold nodes are to nearby supra-threshold voxels on average. The results are shown for six contrasts, using voxel-level analysis (a, b) and cluster-level (c, d) inference with the mixed-effect model, without smoothing (a, c) or after 8mm smoothing (b, d). Differences at the $p < 0.05$ uncorrected level, are highlighted with one star, and results significant at $p < 0.001$ level are highlighted by two stars. The error bars correspond to the choice of the individual surface chosen to embed active regions into MNI space.

of large smoothing and cluster-level inference shows regions in the volume that are not matched with surface-based analysis.

Note that we also investigated the use of the medial cortical surface of the grey-white matter interface when embedding the surface activation into individual space, but the difference with respect to the use of the grey-white matter interface is barely noticeable, as the distance between the two surfaces is much smaller than δ. We also considered using the activations found in grey-matter volume only, but the effect is also negligible.

3.3. Qualitative comparison

In order to compare more qualitatively the output of surface-based versus volume-based inference, we present both types of results on the same figure in Fig. 5. We present results for mixed effects analysis, given that random-effects yield very similar results, albeit with less sensitivity. We present volume-based and surface-based inference results for two contrasts, *computation-sentences* and *reading-visual*.

With the *computation-sentences* contrast, some active surface portions are relatively far from active voxels (but not the converse) in voxel-level inference (Fig. 5, rows 1 and 3), but not in cluster-level inference (Fig. 5, rows 2 and 4). This is consistent with the fact that $\psi(t^{vol}; t^{surf}) > \psi(t^{surf}; t^{vol})$ in peak-level inference. This is one of the cases, where surface-based analysis can be viewed as *more sensitive* than volume-based analysis. This effect is less obvious with the *reading-visual* contrast.

In cluster level inference, we obtain a greater consistency, i.e. all the supra-threshold locations in either domain correspond to supra-threshold location in the other domain. The most striking result is that, in the case of *reading-visual* contrast, volume-based analysis yields a huge cluster that gathers the superior temporal, Broca's and pre-central gyri, while these appear as distinct clusters in surface-based analysis. This effect can be considered as problematic, as the huge volume-based region clearly results from the agglomeration of spatially distant active foci observed in different subjects. Our interpretation is that the spatial variability of these foci creates a very large region where the average signal across subjects is positive; when many subjects are observed (here $S = 25$), the mean deviation to zero becomes statistically significant, and a very large supra-threshold cluster appears. Surface-based analysis nicely avoids this effect, as it produced spatially distant active regions on the cortical surface.

3.4. Reproducibility analysis

Finally the reproducibility statistic has been computed, based on a bootstrap procedure described in section 2.5. The results are given in Fig. 6 in the case of cluster-level or voxel-level inference, using the MFX statistic, for the six contrasts of interest.

First of all, the reproducibility values are in the [.4, 1] range, reflecting that there is a relatively high probability of reproducing the results at the voxel/vertex level when sampling a new group in the same population; this overall reproducibility level depends on the contrast under investigation, as motor and auditory activations yield the most reproducible activity patterns. As could be anticipated, smoothing increases the level of reproducibility. It can be seen that in all cases, surface-based analysis yields higher reproducibility scores than volume-based inference. There is a simple geometric reason for this: projecting the data from 3D to 2D removes one dimension of data variability (the dimension that is perpendicular to the surface), so that the apparent

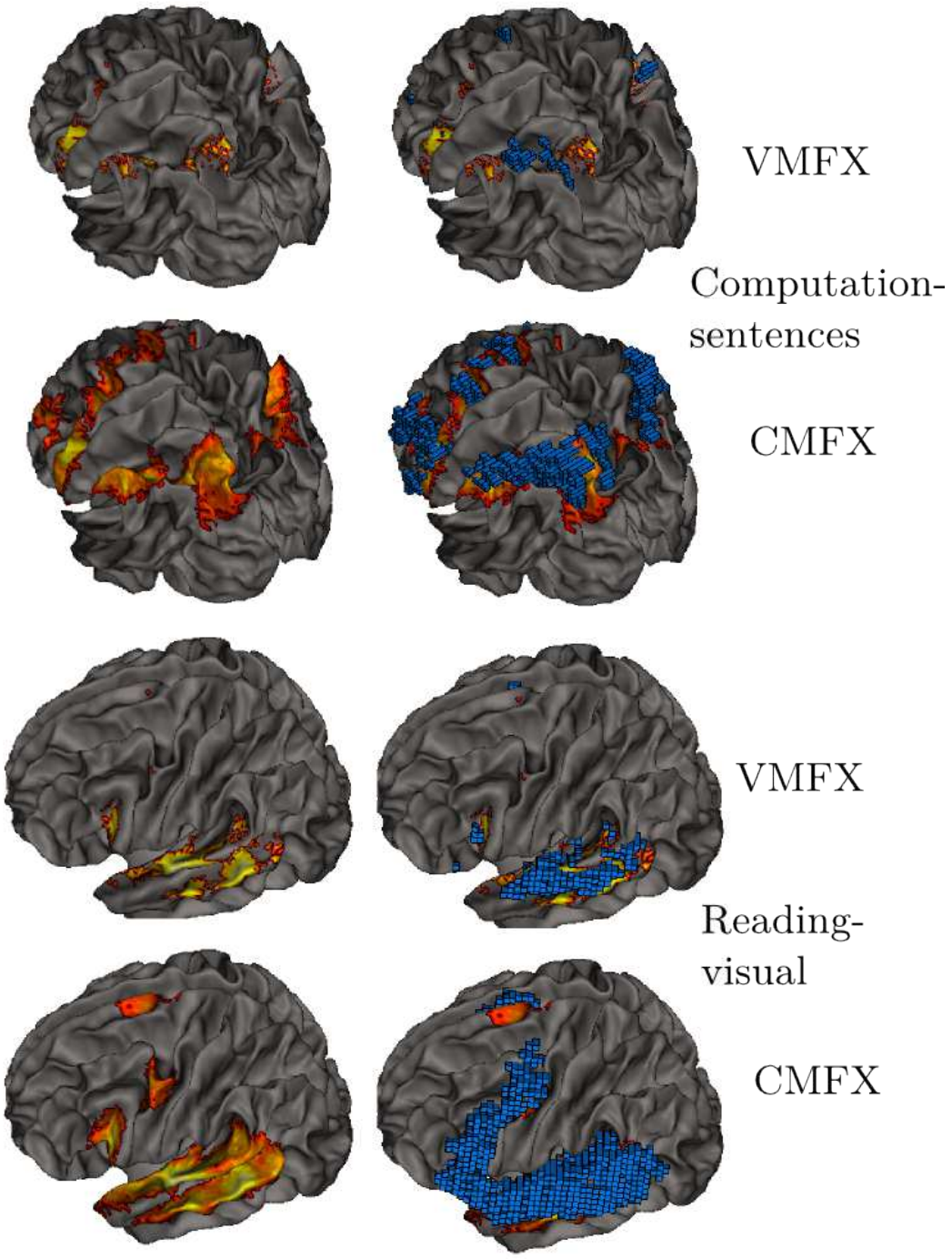

Figure 5: Volume-based versus surface-based results comparisons for the *computation-sentences* (rows 1 and 2) and the *reading-visual contrasts* (rows 3 and 4), using voxel-level (rows 1 and 3) or cluster-level (rows 2 and 4) inference. Surface activations are plotted as a red-yellow texture on the mesh and active voxels are printed in blue. Note the overall correspondence of both detection procedures in each contrast, and the dissociation between peak statistics, where the surface representation yields a wider network, and cluster size inference, where the volume representation yields extended, but non-specific networks. Maps are thresholded at the $p < 0.05$, corrected level.

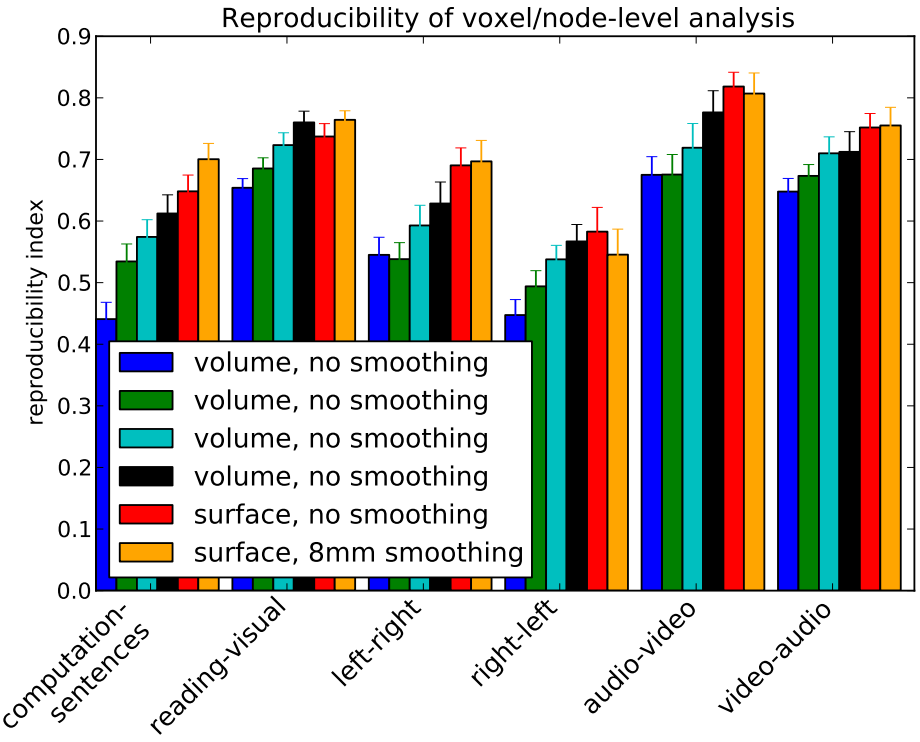

Figure 6: Reproducibility index for a voxel-/node-level inference on the surface or in the volume, after different amounts of smoothing, based on a bootstrap procedure. We can observe that surface-based inference systematically yields more reproducible results than volume-based inference for all levels of smoothing considered.

cross-subject variance is reduced after projection. Additionally, the better correspondence of cortical structures in the surface-based coordinate system should further decrease the variability of supra-threshold patterns.

4. Discussion

The results obtained can be summarized as follows. In all our experiments, we observed that:

- *Cluster-level statistics systematically yield more extended supra-threshold regions than voxel- or node-level statistics.* This was expected, as cluster-level statistics that test the size of supra-threshold regions take advantage of the intrinsic smoothness of the data that is ignored by peak statistics [**?**]. This advantage of cluster level inference is mitigated by two important drawbacks: *i)* it relies on an arbitrary cluster-forming threshold for which there is no straightforward or simple optimal choice (see [**?**] for a more complete discussion on this topic); *ii)* it provides a weak control on false detections, i.e. no guarantee on the true status (active or inactive) of each voxel or vertex within supra-threshold regions.

- *Mixed-effects statistics are more sensitive than random-effects statistics.* Although this increased sensitivity is not guaranteed in theory, it is often observed that mixed effects statistics, that take into account both intra-subject and one inter-subject variance terms, use more information from the input data, and thus allow more sensitive inference [**?**]. In particular, these approaches down-weight the observations that are less reliable (i.e. with high first-level variance), and are thus less sensitive to artefactual values.

- *Surface-based analysis systematically yields a larger portion of supra-threshold regions than volume-based analysis.* This effect was expected, as surface based analysis focuses by definition on the cortical grey matter, and thus discards many regions where BOLD activity is not expected (but also some regions where BOLD activity might occur). But, as we still obtain a significant difference in favor of surface based analysis when restricting the volume to the grey matter, the difference should rather be attributed to better between-subject correspondences in the surface space.

- *An embedding of the cortical surface into each subject's space reveals a dissociation: peak-level inference shows more details on the surface than in the volume, while cluster-size inference displays larger regions in the volume.* This dissociation can be explained as follows: in peak-level inference, surface based may be more sensitive because the search domain does not include white matter and should yield better inter-subject correspondence. Reciprocally, cluster-level inference in the volume shows very large clusters above a given threshold, because limited spatial accuracy in volume normalization tends to merge a network of regions into a large supra-threshold region: this can be viewed as a *spatial jitter* effect. This is clearly seen in Fig 5.

- *Supra-threshold regions are more stable on the surface than in the volume.* This represents another advantage of surface-based inference: the geometric variability of active regions is simply shrunk by dimension reduction, but also by better registration between subjects.

To summarize, the overall greater sensitivity of surface-based analysis is explained by two factors:

- The search domain is smaller (*i.e.* it consists in the cortex only), while the volume-based approach tests also white matter and cerebro-spinal fluid regions. The price to pay is that sub-cortical structures (thalamus, basal ganglia, cerebellum etc.) are not considered in the surface-based representations. However, our experiments show that this does not explain all the observed differences.

- The co-registration of the data is certainly more accurate, as often discussed in the literature [? ?], but hard to prove in practical settings. Note that this better co-registration could have the effect to reduce the spread of significant regions by removing spatial uncertainty on their position.

When considering cluster-level statistics, it can be expected that the finer co-registration related to the surface-based approach has a mitigated effect on the detection of large clusters: volume-based approaches tend to join smaller clusters which are close in Euclidean space, but relatively distant on the geodesic (surface-based) representation. It should also be reminded that cluster-level analyses allow only a weak control of false detections (see e.g. [?]): a supra-threshold cluster should contain at least one voxel for which the null hypothesis can be rejected, but it cannot be concluded that all the voxels or nodes in that cluster are indeed active.

However, in surface-based approaches, if the resolution is not fine enough, as this is the case in our experiments, or if EPI distortions are not perfectly corrected, the functional signal may be diluted to neighboring gyri, creating false supplementary clusters (as we can see on Fig. 7). The effect might be consistent across subjects and thus results in a spurious activation focus in a group analysis. Note that volume-based analyses suffer from the same inaccuracy, but that it appears only when activations are displayed on the cortical surface.

It is important to notice that we did not do any systematic study of different pipelines, and that part of the results described here would be slightly altered when using different pipelines and pre-processing tools, e.g. using only FreeSurfer pipeline, and that constant improvement of these tools also continuously changes the picture. In particular, the three following points should be considered carefully: *i)* the number of data resamplings performed along the analysis pipeline (three in the present case: the first one during distortion correction, a second one during motion correction and the third one during the projection onto the surface), *ii)* the potential mismatch remaining between EPI and T1 data, in particular when the acquisitions do not entirely cover the brain, e.g. if the tip of the motor cortex is not within the acquired volume, as it was the case in some subjects in our dataset and *iii)* the choice and parametrization of the projection method, which has to trade off sensitivity, e.g. integrating over a larger domain to avoid missing some signal, against specificity, avoiding bleeding across regions or tissues. We defer a more exhaustive study of these different processing steps to future work. Still, we believe that the main effects described here would carry on to many surface-based studies. It is hard to predict whether stronger fMRI contrasts would alter the conclusions: on the one hand, the decrease of false negatives may benefit surface-based analyses more than volume-based analyses; on the other hand, the sensitivity may saturate.

An interesting question that can still be addressed is how a surface co-registration procedure based on geometric features such as FreeSurfer would compare to representations of the surface that take

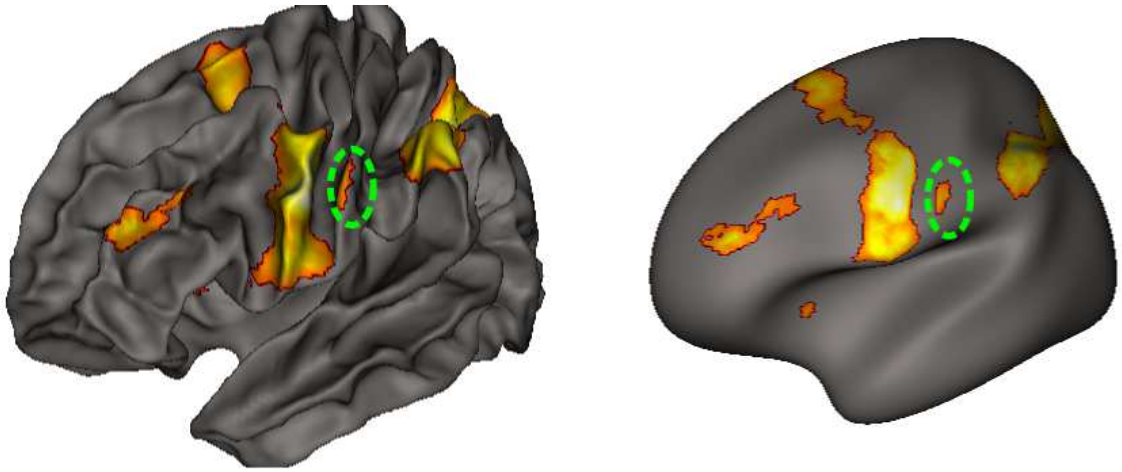

Figure 7: Potential issues with fMRI data projection to the cortical surface. (top) The activity on the pre-central gyrus for a computation task has also been projected on the neighboring gyrus (post-central gyrus) in a fraction of the subjects, and the ensuing effect can be detected in a group study ($p < 0.05$ corrected, 25 subjects, VMFX statistic). It is presented on the average cortical surface (left) and on an inflated representation (right). This kind of effect can be due to poor initial fMRI data resolution, to the limitations of EPI distortion correction procedures, and to the smoothing induced by successive interpolation steps on the BOLD data. This map was obtained using the projection method in [?]. (bottom) The limited resolution of standard fMRI data $\sim (3mm)^3$ makes this kind of effect hardly avoidable. In this schematic case, the volume-based approach would detect only one large cluster (in blue), while a surface-based approach would project the activated region on both the 2 gyri and would detect 2 different clusters.

into account sulcus labelling. Some experiments have indeed shown that sulcus-based coordinate systems tend to stabilize the position of some functional landmarks [?]. The combination of shape information exacted from gyral curves, cortical surfaces together with intensity images [? ?] is a promising approach in that respect. Some algorithms also include include information from diffusion MRI, such as the position of the main fiber bundles [?].

Conclusion. Performing fMRI group analysis on the cortical surface instead of the brain volume may benefit the detection of some foci of activity, and thus yield more specific, and possibly more sensitive analysis than standard volume-based approaches. It may reveal sharper contrasts in the functional data and thus provide more reliable markers of brain functional anatomy. A cortical surface analysis is quite costly in term of data processing, but it detects at least as many cortical regions as volume-based methods, with better reproducibility, and the additional guarantee of providing more spatially specific clusters. While it seems beneficial to group analysis, it is limited by the issues related to potentially incorrect projection of fMRI signal onto the cortical surface, which appears to be the most risky step for surface-based inference. Future work may therefore be directed at providing quality check measures of the accuracy of the projection.

Acknowledgments. The authors acknowledge support from the ANR grant ViMAGINE ANR-08-BLAN-0250-02, IRMGroup ANR-10-BLAN-0126-02 and from the Digiteo DIM grant HiDiNim. We would also like to thank Philippe Pinel for providing the data we used in this experiment.

AppendixA. Supplementary materials

Contrasts	computation-sentences	reading-visual	left-right	right-left	audio-video	video-audio
VRFX, volume	0.035	0.614	0.119	0.135	0.641	0.258
VRFX, surface	0.637	2.661	0.957	0.479	2.791	3.82
CRFX, volume	1.631	3.697	1.119	1.122	2.934	3.665
CRFX, surface	7.934	7.882	3.39	1.637	5.833	10.289

Table A.3: Relative volume (in percent) of activated areas for different statistical inference procedures performed either on the surface or in the volume. The tests are performed for 6 different contrasts (in column), and the results are provided with voxel/vertex-level random effects inference (VRFX) and cluster-level random effects inference (CRFX). The threshold p-value is 0.05, corrected.