Toblerone: surface-based partial volume estimation

Journal:	*Transactions on Medical Imaging*
Manuscript ID	Draft
Manuscript Type:	Full Paper
Date Submitted by the Author:	n/a
Complete List of Authors:	Kirk, Thomas; Institute of Biomedical Engineering, Department of Engineering Science; Wellcome Centre for Integrative Neuroimaging, Nuffield Department of Clinical Neurosciences, University of Oxford Coalson, Timothy; WUSTL Department of Neuroscience, Washington University School of Medicine Craig, Martin; Institute of Biomedical Engineering, Department of Engineering Science; Wellcome Centre for Integrative Neuroimaging, Nuffield Department of Clinical Neurosciences, University of Oxford Chappell, Michael; Institute of Biomedical Engineering, Department of Engineering Science; Wellcome Centre for Integrative Neuroimaging, Nuffield Department of Clinical Neurosciences, University of Oxford
Keywords:	Functional imaging (e.g. fMRI) < Imaging modalities, Segmentation < General methodology, Image enhancement/restoration(noise and artifact reduction) < General methodology
Specialty/Area of Expertise:	partial volume effect

Thomas Kirk
Institute of Biomedical Engineering
Old Road Campus Research Building
University of Oxford
Oxford
OX7 3DQ

To whom it may concern,

I enclose for your consideration an article entitled ‘Toblerone: surface-based partial volume estimation’. This article has been revised in response to feedback previously received from IEEE-TMI, with the existing submission number TMI-2019-0508.

I hope you will agree that we have made substantial changes in response to reviewer feedback. In particular, we now apply our method across the brain, not just in the cortex as previously; and have both added an entirely new set of experiments and expanded the scope of the existing set. As a result, the evaluation of our method is now founded upon results from pure simulations, simulated T1-weighted MRI images based on a single subject’s anatomy, and finally forty-five subjects with test-retest data. This has allowed us to more thoroughly investigate the robustness of our proposed method to random image noise and scanner field non-uniformity; and to assess the inter-session repeatability of our results, a metric clearly of interest for longitudinal functional imaging studies. The results of these new analyses are favourable for our method.

Both a clean and marked-up copy of the article are enclosed, alongside a detailed response to individual reviewer comments. A single page of supplementary material is also enclosed, containing the full versions of figures that were omitted from the main body of the paper for clarity.

I have not put down any suggested reviewers on the understanding from your email of the 15th June that manuscripts revised in good time will be handled by the same reviewers.

I hope the aforementioned will be both to your and the reviewer’s satisfaction.

Yours faithfully,

Thomas Kirk

Toblerone: surface-based partial volume estimation

Thomas F. Kirk, Timothy S. Coalson, Martin S. Craig and Michael A. Chappell

Abstract—**Partial volume effects (PVE) present a source of confound for the analysis of functional imaging data. Correction for PVE requires estimates of the partial volumes (PVs) present in an image. Conventionally these estimates are obtained via volumetric segmentation, but such an approach may not be accurate for complex structures such as the cortex. An alternative is to use surface-based segmentation, which is well-established within the literature. Toblerone is a new method for estimating PVs using such surfaces. It uses a purely geometric approach that considers the intersection between a surface and the voxels of an image. In contrast to existing surface-based techniques, Toblerone is not restricted to use with any particular structure or modality. Evaluation in a neuroimaging context has been performed on simulated surfaces, simulated T1-weighted MRI images and finally a Human Connectome Project test-retest dataset. A comparison has been made to two existing surface-based methods; in all analyses Toblerone's performance either matched or surpassed the comparator methods. Evaluation results also show that compared to an existing volumetric method (FSL FAST), a surface-based approach with Toblerone offers improved robustness to scanner noise and field non-uniformity, and better inter-session repeatability in brain volume. A surface-based approach negates the need to perform resampling (in contrast to volumetric methods) which is particularly advantageous for low-quality data.**

Index Terms—**functional imaging, partial volume effect, partial volume correction, segmentation, surface**

I. Introduction

PARTIAL volume effects (PVE) arise when an imaging matrix has low spatial resolution in relation to the structures of interest within the image, as is commonly the case for the functional imaging techniques, such as positron emission tomography (PET), blood oxygen-level dependent fMRI (BOLD) and arterial spin labelling (ASL). For example, ASL voxels typically have side lengths of 3-4mm whereas the mean thickness of the adult cortex is 2.5mm [1]. As such, voxels around the cortex will contain a mixture of cortical and non-cortical tissues, the proportions of which are termed *partial volumes* (PVs). PVE present a source of confound for functional imaging: whilst the objective is to obtain a measurement of function across some particular structure, the signal actually measured in each voxel is a sum, weighted by the partial volumes, of function both within and without said structure. This is a mixed-source problem in which the multiple tissues in each voxel constitute the sources. Partial volume correction (PVEc) uses voxel-wise estimates of PVs to separate out the signals arising from each tissue. Various PVEc methods have been developed, usually with a specific modality in mind (for example, Muller-Gartner for PET [2] and linear regression [3] or spatially-regularised variational Bayes for ASL [4]).

Estimation of PVs bears considerable similarity to volumetric segmentation and the two are typically performed concurrently on a structural image, as is demonstrated in [5]. In order to estimate PVs within the voxel grid of a functional image, it is then necessary to transform the results from the structural voxel grid to the functional. As each functional voxel corresponds to multiple smaller voxels on the structural image, the PVs of the former can be estimated using the results from the latter. The efficacy of this approach is limited by the accuracy of the volumetric segmentation approach used. For complex geometries, such as the thin and highly folded structure of the cerebral cortex, the alternative of surface-based segmentation has gained widespread support (notably through FreeSurfer [6]). The advantage of such a segmentation method is twofold. Firstly, whereas volumetric segmentation is necessarily a discrete operation in terms of voxels, a surface approach is somewhat continuous as the surface vertices are placed with subvoxel precision. Secondly, anatomically-informed constraints can be enforced anisotropically when surfaces are used: for example, the constraint that tissues should be homogenous along a surface but heterogeneous across it. This is in contrast to a volumetric tool such as FSL FAST [7] which does enforce a similar tissue continuity constraint via the use of Markov random fields but only isotropically in the neighbourhood of each voxel. In principle, it should be possible to estimate PVs by considering the geometry of intersection between the individual voxels of an image and the surface segmentations of individual structures. Being a purely geometric construct, namely, *given a surface that intersects a voxel, what is the volume within the voxel bounded by the surface,* this is a fundamentally different approach to existing methods and it is expected this will be reflected in the estimates produced.

Although surface-based PV estimation tools exist in the literature, past efforts have usually been designed with a specific modality in mind. Two notable examples for neuroimaging are the ribbon-constrained (RC) method used within the Human Connectome Project's (HCP) *fMRISurface* pipeline [8] and PETSurfer [9], [10], a variant of FreeSurfer. The former is designed for use with BOLD and so distinguishes only between cortex and otherwise, not the grey matter (GM), white matter (WM) and non-brain required for ASL and PET; whereas the latter is both PET-specific and tightly integrated into FreeSurfer such that it is hard to use independently of that

T. F. Kirk (thomas.kirk@eng.ox.ac.uk), M. S. Craig (martin.craig@eng.ox.ac.uk) and M. A. Chappell (michael.chappell@eng.ox.ac.uk) are at the Institute of Biomedical Engineering, Department of Engineering Science, and the Wellcome Centre for Integrative Neuroimaging, Nuffield Department of Clinical Neurosciences, both at the University of Oxford, UK.

T. S. Coalson (tsc5yc@mst.edu) is at the Department of Neuroscience, Washington University School of Medicine, St Louis, USA. Funding was provided by the EPSRC (EP/P012361/1) and (for T. F. Kirk) the Bellhouse scholarship at Magdalen College, Oxford. T. S. Coalson was funded by NIH Grant R01 MH-060974 (to D. Van Essen). We intend to provide access to the data underlying the figures within this paper via a DOI (as per the EPSRC).

workflow. Furthermore, both methods deal exclusively with surfaces representing the cortex. The objective of this work was to develop an algorithm, named Toblerone[1], to estimate partial volumes for both cortical and subcortical structures (where such surfaces are available, for example via FSL FIRST [11]) for neuroimaging applications. The end result is highly general and could be used with images from multiple modalities and/or in other parts of the body.

II. Theory

Voxelisation is the process of quantifying the volume contained within a surface and many algorithmic methods are given in the computer graphics literature. The key operational step for this is determining if a point lies interior or exterior to a given surface; by repeating this test entire volumes can be built up. The ray intersection test outlined by Nooruddin and Turk [12] is widely used and requires only that the surfaces be contiguous (water-tight). The test is performed by projecting an infinite ray in any direction from the point under test and counting the number of intersections made with the surface. A ray from an interior point will make an odd number of intersections as it exits the surface (including folds within the surface, there will be one more point of exit than entry); conversely an exterior point will make an even number of intersections (balanced entries and exits), if at all. This test scales badly with increasing spatial resolution: for a linear resolution of n samples per unit distance, n^3 tests per unit volume are required. Furthermore, as each ray must be tested against each surface element, the test also scales with surface complexity (linearly for a naïve implementation). For a typical functional image of 10^5 voxels and 2.5 x 10^5 surface elements in a FreeSurfer cortical surface, this is prohibitively computationally intensive.

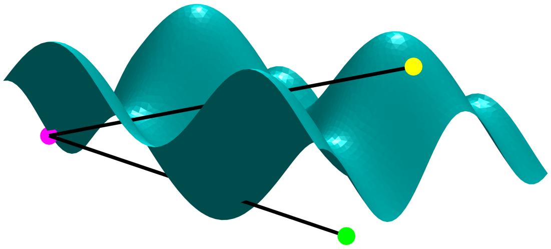


Fig. 1 Reduced ray intersection test for non-contiguous surfaces. The root point (interior) is shown in magenta. A ray from an interior point (green) makes two intersections due to the presence of a fold; from an exterior point (yellow) there is one intersection.

The method adopted in this work is to only use the portion of surface that actually intersects a given voxel (termed the 'local patch') for ray intersection testing. The local patch is defined as all triangles that intersect the voxel or, equivalently, the minimal set of triangles that unambiguously divides the voxel into two regions. This patch is by definition non-contiguous, so it is necessary to modify the ray intersection test accordingly; the modified form is referred to as the 'reduced' test in contrast to the 'classical' test. Within each voxel, a 'root point' that is known to lie within the surface is identified via the classical ray test. Any other point within the voxel may then be tested by projecting the finite line segment $\mathbf{r} = \mathbf{p_t} + \lambda(\mathbf{p_r} - \mathbf{p_t})$, where $\mathbf{p_t}$ is the point under test, $\mathbf{p_r}$ is the root point and $0 \leq \lambda \leq 1$ is a distance multiplier along the line. A parity test is then applied to the number of intersections identified between the root and test points. The fact that the line terminates at a point interior to the surface means that exterior points will lead to one more point of entry than exit; conversely interior points will lead to either zero or an even number of intersections. It is not necessary to test surface elements outside the voxel as the finite length of the line segment means it can never leave the voxel. Fig. 1 provides an illustration of the test in practice.

In order to minimise the number of tests required per voxel, convex hulls (defined as the smallest possible region enclosing a set of points within which any two points can be connected without leaving the region) are used to estimate partial volumes wherever possible. The rationale for this is that if the extrema points of a region can be classified as interior/exterior to a surface then, to an approximation, all points lying within the convex hull of these points will share the same classification.

III. Algorithm

The following section addresses PV estimation for structures within the brain, for which the tissue classes of interest are GM, WM and non-brain. The same principles would apply to structures in other areas of the body, though the interpretation of tissue classes would differ.

A. *Estimation for a single surface*

The core algorithm within Toblerone estimates the voxel-wise interior/exterior PVs arising from the intersection of a single surface with an arbitrary voxel grid. Toblerone assumes cuboid voxels with a 'boxcar' point-spread function (PSF), which is to say that it does not allow for any mixing of signal between voxels. In reality, different modalities have differing PSFs and such effects may be separately accounted for via a convolution operation. The first step is to identify and record the local patches of surface intersecting each voxel of the reference image via Moller's triangle-box overlap test [13].

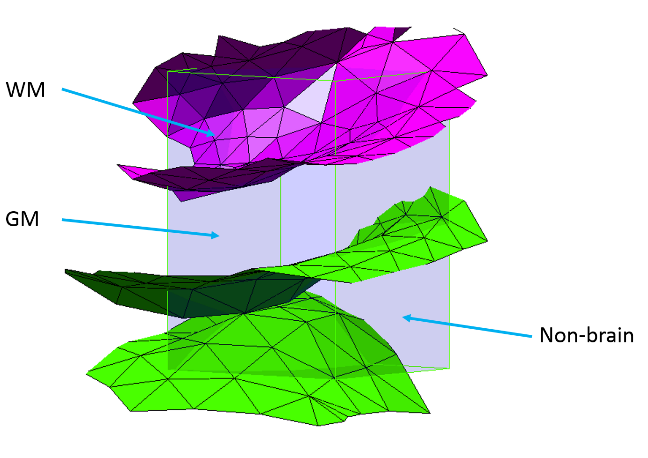


Fig. 2 Intersection of inner (magenta) and outer (green) surfaces of the cortex with a voxel. The outer surface intersects twice with distinct patches of surface; this is likely due to the presence of a sulcus. Tissue PVs are labelled.

[1] So-named because an early implementation constructed triangular prisms.

The geometry of a surface within a voxel can frequently be complex: using a sulcus of the cortex as an example, the surface may intersect the voxel multiple times, with the opposite banks of the sulcus appearing as two unconnected patches of surface, illustrated in fig. 2. Accounting for the many possible surface/voxel configurations requires numerous specific tests that rapidly become excessively complex, so the approach taken in Toblerone is to divide and conquer each voxel as required. As the length scale of a voxel decreases, the complexity of the local surface configuration within the voxel will also decrease (for example, a sulcus is less likely to intersect the voxel multiple times). Each voxel of the reference image is therefore divided into a number of subvoxels, each of which is processed individually. In the neuroimaging context of this work, the subdivision factor was set empirically as ceil$(\mathbf{v}/0.75)$ where $\mathbf{v}$ is the vector of voxel dimensions and 0.75 represents the lower limit of feature size found in the brain (in other contexts this parameter could be varied). Note that this subdivision factor transforms anisotropic voxels into approximately isotropic subvoxels. Subvoxels are then processed according to the following framework:

- If the subvoxel does not intersect the surface, it is assigned a single-class volume according to an interior/exterior classification of its centre. This is illustrated in fig. 3a.
- If the subvoxel intersects the surface, then it contains interior and exterior PVs. One of these will be estimated using a convex hull (via the Qhull implementation [14]) if the geometry of the surface is favourable, as follows:
    - If the surface intersects entirely through one face of the subvoxel, then it encloses a highly convex volume that may be reliably estimated. The other partial volume is calculated by subtraction from the total subvoxel volume. This is illustrated in fig. 3b.
    - If the surface is folded within the subvoxel (identified by multiple intersection of the surface along an edge or face diagonal of the subvoxel) then the subvoxel is subdivided a second time. This is because it is difficult to reliably identify which volume is interior or exterior in such a situation. This is illustrated in fig. 3c/d.
    - In all other cases, convex hulls are again used. In order to minimise the potential error associated with estimation of a non-convex volume via convex hulls, it is important to identify which of the two PVs within the subvoxel is closer to being convex than the other. The proxy measure used in this work is the number of subvoxel vertices lying on either side of the surface: the side with fewer vertices is assumed to enclose a more convex (and at any rate smaller) volume than the other. This is illustrated in fig. 3e.
- If the surface intersects the subvoxel multiple times (identified by the successful separation of surface nodes lying within the subvoxel into unconnected groups) then the voxel is subdivided a second time. This situation occurs for example when the opposite banks of a sulcus pass through a voxel. Although the reduced ray intersection test is accurate in such a situation, forming convex hulls is not, so subdivision is the safer option. This is illustrated in fig. 3f.

The second subdivision is performed at a constant factor of 5 to yield sub-subvoxels of approximately 0.1 to 0.2mm side length isotropic. These are always assigned a single-class volume based on a classification of their centre points as their small size

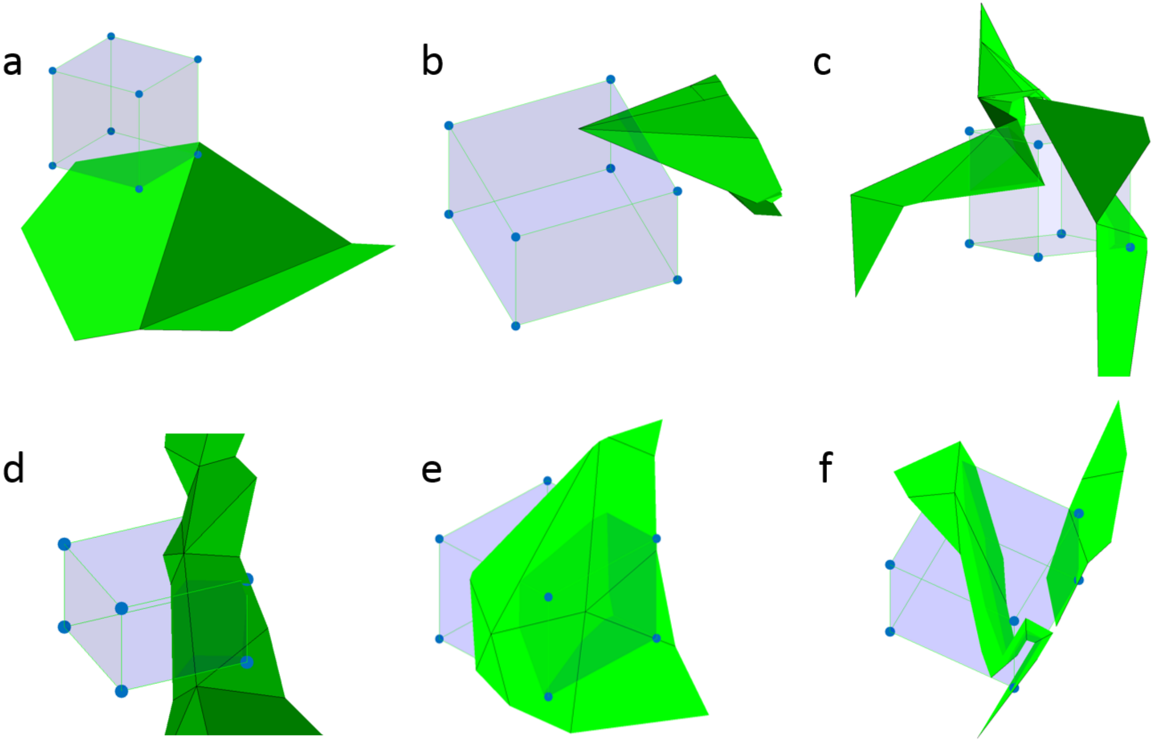


Fig. 3 Various subvoxel/surface configurations. a) no intersection: whole-volume assignment; b) single intersection through one face: a small convex hull will be formed; c/d) two examples of single intersection, folded surface: further subdivision will be used; e) single intersection through multiple faces: a convex hull will be formed; f) multiple surface intersection (unconnected patches of surface, likely a sulcus): further subdivision will be used.

means that any PVE will be negligible. Finally, voxels that do not intersect the surface (fully interior or exterior) are given single-class volumes according to tests of their centre points. Structures defined by a single surface (e.g. the thalamus) require no further processing: the estimates produced by the aforementioned steps may be used directly for PVEc.

B. *Multiple-surface structures*

Structures that are defined by multiple surfaces require further processing to yield PV estimates for all tissues of interest. With specific reference to the cortex, PVs within each hemisphere are obtained with the relations:

$$\begin{aligned} PV_{WM} &= P_{inner} \\ PV_{GM} &= \max(0, P_{outer} - P_{inner}) \\ PV_{NB} &= 1 - (PV_{WM} + PV_{GM}) \end{aligned}$$

where P_{inner} and P_{outer} denote the interior/exterior PV fractions associated with the inner and outer surfaces of the cortex respectively and PV_{WM}, PV_{GM} and PV_{NB} denote the PV estimates for WM, GM and non-brain tissue (the latter including cerebrospinal fluid, CSF). These equations are structured to account for a potential surface defect whereby the surfaces of the cortex swap relative position (the inner lying exterior to the outer) around the corpus collosum. The structure of the above relations (*N* surfaces leading to *N*+1 tissue classes) could easily be generalised to structures defined by more than two surfaces (for example, sublayers of the cortex, as used in laminar fMRI). A similar set of equations is used to merge hemisphere-specific results to cover the whole cortex, accounting for voxels lying on the mid-sagittal plane that intersect both hemispheres.

C. *Whole-brain PV estimation*

Toblerone, as outlined above, operates on a structure-by-structure basis in which the output tissue types are dependent on the structure in question. A number of methods utilising the core algorithm were implemented:

1) *estimate_structure:* estimate the inner and outer PVs associated with a structure defined by a single surface
2) *estimate_cortex:* estimate the GM, WM and non-brain PVs associated with the four surfaces of the cortex (l/r white/pial in the FreeSurfer terminology)
3) *estimate_all:* a combination of the *structure* and *cortex* methods above, this estimates PVs for the cortex and all subcortical structures identified by FIRST and combines them (with the exception of the brain stem) into a single set of GM, WM and non-brain PV estimates. The run-time for a typical subject was around 25 minutes.

The combination of FreeSurfer/FIRST and *estimate_all* provides a complete pipeline for obtaining whole-brain PV estimates in an arbitrary reference voxel grid from a single T1 structural image that may be used as a replacement for existing volumetric tools such as FAST. There is however a key conceptual difference between surface and volumetric methods concerning their interpretation of subcortical structures. Due to differences in tissue composition around the brain, cortical and subcortical GM have different intensities on a normal T1 image and are accordingly assigned different GM PVs by volumetric tools such as FAST (whereby cortical GM is seen as more 'grey' than subcortical, as illustrated in fig. 12). Surface based methods, by contrast, do not take a view on what tissue lies within the surface other than simply asserting that it is different to that which lies without. When combining the PVs of individual structures in Toblerone's *estimate_all* function, all tissue within the cortex and subcortical structures is interpreted as pure GM. The practical implication of this is that Toblerone's estimates for subcortical GM are higher than those produced by FAST. For this reason, the conventional GM/WM/CSF tissue classes used by volumetric tools may be better thought of within Toblerone's framework as *tissue of interest*, *other tissues* and *non-brain,* though for the purposes of this article the familiar names GM and WM shall be used alongside non-brain. The inherent ambiguity in determining which tissues lie outside subcortical structures, which could be either WM or CSF depending on their location within the brain, was resolved using FAST's segmentation results[2].

IV. Evaluation

Three datasets and three comparator methods were used, as summarised in Table I. The two surface-based comparator methods were restricted to use in the cortex only. By contrast, Toblerone was run on both cortical and subcortical surfaces where appropriate to provide whole-brain PV estimates.

TABLE I
DATASETS & METHODS USED

name	Simulated surfaces	BrainWeb	HCP test-retest
type	S	V + S	V + S
resolution	-	1mm iso.	0.7mm iso.
size	1 cortical hemisphere	18 simulated T1 images	45 subjects, 2 sessions each
ground truth	numerical method	volumetric segmentation*	N/A
comparator methods	NeuroPVE (S) RC (S)	RC** (S) FAST (V)	RC* (S) FAST (V)

S *surface,* V *volumetric,* RC *ribbon-constrained method*
* *established via automatic segmentation with manual intervention*
** *RC can only be run on the cortex for these datasets*

A. *Comparator methods*

The first surface-based comparator method, the ribbon-constrained (RC) algorithm, was developed for use with BOLD data in the HCP's *fMRISurface* pipeline [8] and is restricted to the cortex only. The method assumes vertex correspondence between the two surfaces of the cortex and works as follows. For each vertex in turn, the outermost edges of the triangles that surround said vertex are connected between the two surfaces to form a 3D polyhedron representing a small region of cortex.

[2] As it is ambiguous as to what tissue lies outside a given subcortical structure given only its surface, FAST's results for the same voxel are used as an estimate for the local ratio of WM and CSF. The actual quantity of non-GM tissue is still calculated from the surface estimate as the remainder 1 – GM, which is then shared between the other two classes in this ratio.

Nearby voxels are subdivided and the subvoxels centres tested to determine if they lie interior to the polyhedron. The subdivision factor used in this work was the higher value of either $\mathrm{ceil}(\max(\mathbf{v})\ /\ 0.4)$ or 4, where $\mathbf{v}$ is the vector of voxel dimensions. The fraction of subvoxel centres lying within any cortical polyhedron gives the cortical GM PV, which, as the BOLD signal is predominantly cortical in origin, is the quantity of interest for this modality. In order to obtain WM and non-brain PVs, the following post-processing steps were used. Firstly, the unassigned PV of each voxel was calculated as $1 - PV_{GM}$, which was subsequently labelled as either WM or non-brain according to a signed-distance test of the voxel centre in comparison to the cortical mid-surface: for a voxel with centre point outside the mid-surface, the unassigned PV was labelled as non-brain. A weakness of this approach is that it is unable to faithfully capture a voxel in which all three tissues are present; only the combinations WM/GM or GM/non-brain are permitted. As voxel size increases, the probability of voxels containing multiple tissues also increases; testing on a brain image of 3mm isotropic resolution showed that around 30% of voxels intersecting the cortical ribbon contain three tissues. Resampling can be used to mitigate this effect so two variants of this method were tested: 'RC', direct estimation at each resolution, and 'RC2', estimation at 1mm followed by resampling to other resolutions via the process in section IV.B. The run-time for a typical subject was around 15 minutes.

This second surface method, NeuroPVE [15], uses a voxelisation method based on the work of [9,12], applied in a brain-specific context and again restricted to the cortex only. Multiple expanded and contracted copies of each surface are created and the ratio of expanded to contracted surfaces intersecting a given voxel is used as a first approximation for partial volumes. This ratio is then mapped, along with surface orientation information, via trigonometric relations on the unit cube into a PV estimate. The estimates produced take discrete values according to the number of surfaces used (in this work the default of 5). The intended use of this tool was PV estimation at structural, not functional, resolution, so two variants were tested: 'Neuro', direct estimation at arbitrary resolutions, and 'Neuro2', estimation at structural resolution followed by resampling to other resolutions via the process in section IV.B. On the basis of NeuroPVE's results on the simulated surfaces, it was excluded from further analysis. As the process of surface inflation is slow, the run-time for a typical subject was around 12 hours.

Finally, FSL's FAST [7] is an established whole-brain volumetric segmentation tool that was used as a comparator for the surface methods. On both the BrainWeb and HCP test-retest datasets, FAST was run on the brain-extracted images at structural resolution (1mm and 0.7mm iso. respectively). PVs were then obtained at other resolutions via the resampling method detailed in section IV.B. The run-time for a typical subject was around 5 minutes.

B. *Resampling*

Resampling is an interpolation operation that is used to transform volumetric data between voxel grids (in this context, from structural to functional resolution). FSL's *applywarp* tool was used with the *-super* flag for all resampling operations. This works by creating an up-sampled copy of the target voxel grid onto which values from the input image are sampled. The average is then taken across the voxel neighbourhoods of the high-resolution grid (sized according to the up-sampling factor) to obtain the result in the target voxel grid. Such an approach is appropriate when moving from fine to coarse as each output voxel corresponds to multiple input voxels, the individual contributions of which should be accounted for to preserve overall tissue proportions. When using *applywarp* a transformation matrix between the input and output voxel grids must be given as the *-premat* argument; to denote identity for the purposes of this work, the output of the HCP *wb_command –convert-affine –from-world –to-flirt* tool operating on $\mathbf{I}_4$ was used as the *-premat* to correct for a subvoxel shift that arises due to FSL coordinate system conventions. Note that for perfectly aligned voxel grids with an integer ratio of voxel sizes, such as a 1mm and 2mm isotropic grid, this process is equivalent to averaging across blocks of the smaller grid (sized 2x2x2 in this case).

C. *Simulated surfaces*

A pair of concentric surfaces, illustrated fig. 4, were designed to capture geometric features relevant to the anatomy of a cortical hemisphere. These were produced by modulating the radius of a sphere as a function of azimuth θ and elevation ϕ to produce sulci and gyri-like features. The radius of the inner surface was defined as

$$r_{in} = 60(1 - 0.1 \max(\sin^{20} 5u\,, \sin^{20} 5v))$$

where 60 is the unmodulated radius of the sphere, 0.1 fixes the relative depth of sulci, the max function prevents sulci from constructively interfering to produce deep wells at points of intersection, the power of 20 produces broad gyri and narrow sulci, and the substitutions $u = \phi + \theta$, $v = \phi - \theta$ cause the sulci to spiral around the sphere in opposite directions. Modulation was restricted to the range $-2\pi/5 \leq \theta \leq 2\pi/5$ to leave the poles smooth and suppress unrealistic features. The outer radius was set at $r_{out} = 1.05 \cdot r_{in}$, leading to a peak radial distance between surfaces of 3mm. The outermost region was taken to represent non-brain tissue, the innermost WM and the region in between GM. The use of analytic functions to define the surfaces permitted ground truth maps to be calculated using a numerical method. Voxels were sampled at 4,096 elements per mm^3 and the positions of these sample points expressed in spherical polar coordinates. By comparing the actual radius of each point to the calculated radius of the surface boundaries for the same azimuth and elevation, the tissue type of the sample

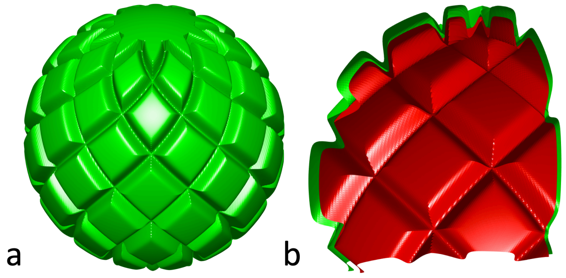


Fig. 4 a) Simulated surfaces; b) cutaway showing inner (red) and outer (green) surfaces. Peak radial distance between the two was 3mm.

point within the structure could be determined, and from there PVs obtained by aggregating results within voxels. This is referred to as the 'numerical solution' in the results section. Mean surface node spacing was set at 0.85mm, similar to that of native FreeSurfer output. Toblerone's *estimate_cortex*, NeuroPVE and the RC method were used on this dataset. PVs were obtained at voxel sizes of 1 to 3mm in steps of 0.2mm isotropic.

D. BrainWeb simulated T1 images

BrainWeb [16], [17] simulates whole-head T1 images at 1mm isotropic resolution with specified levels of random noise and field non-uniformity (NU). Eighteen images were produced to cover the available parameter space of noise levels {0, 1, 3, 5, 7, 9} and NU levels {0, 20, 40} (both quantities in percent). These were run through FAST, FIRST and FreeSurfer, after which Toblerone's *estimate_all* and the RC method (cortex only) were used on the output. FAST's output was also used to enable a comparison between surface and volumetric methods. PVs were obtained at voxel sizes of 1 to 4mm in steps of 1mm isotropic. Although ground truth PV maps exist for this dataset (produced by automatic volumetric segmentation of T1 images with manual correction [16]), both surface and volumetric methods returned significantly different results to these, raising the complicated question of determining which set of results is correct. In order to avoid making this judgement, each method was instead referenced to its own results on the ideal T1 image (0% noise 0% NU) in the 1mm isotropic voxel grid of the structural images. The voxel grids associated with each voxel size were aligned such that results at 1mm could be used to calculate a reference at other sizes (for example, summing across 3x3x3 blocks to get a 3mm reference).

E. Human Connectome Project test-retest data

This dataset comprises 45 subjects from the main HCP cohort who underwent two separate structural scan sessions (mean age 30.2 years, mean time between sessions 4.8 months). Each session was processed using the pipeline in [8] to obtain cortical surfaces via FreeSurfer. Separately, the distortion-corrected T1 images were fed into FAST (brain-extracted) and FIRST (whole-head) to produce volumetric segmentations and subcortical surfaces. Toblerone's *estimate_all* and the RC method (for the cortex only) were used on this dataset, as well as FAST for a comparison between surface and volumetric methods. PVs were obtained at voxel sizes of 1 to 3.8mm in steps of 0.4mm isotropic, as well as the native 0.7mm isotropic voxel grid of the structural images. Although a ground truth is not defined for this dataset, each method's results from the first session were used as a reference for the second session.

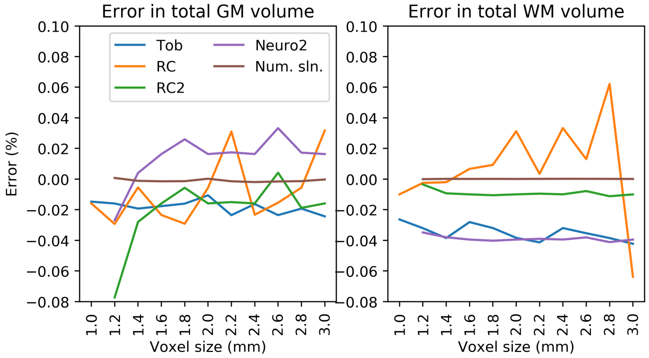
F. Evaluation metrics

Errors were measured in both a per-voxel (root-mean-square, RMS, of individual voxel errors) and aggregate (total tissue volume) sense. The former basis is important as PVEc is locally sensitive to the PV estimates [18]; the latter basis reflects systematic bias at the aggregate level. All error quantities are expressed in percent and map directly to PV estimates without scaling: for example, a PV estimate of 0.5 against a reference value of 0.55 corresponds to an error of -0.05 or -5%.

A further analysis of voxel-wise differences between Toblerone and FAST was performed on the HCP dataset at multiple voxel sizes by sorting voxels into 5% width bins according to their Toblerone GM PV estimate. The difference (Toblerone – FAST) was calculated for each voxel and the mean taken across each bin. This quantity was then averaged across subjects and sessions (weighted to respect differences in brain volume).

V. Results

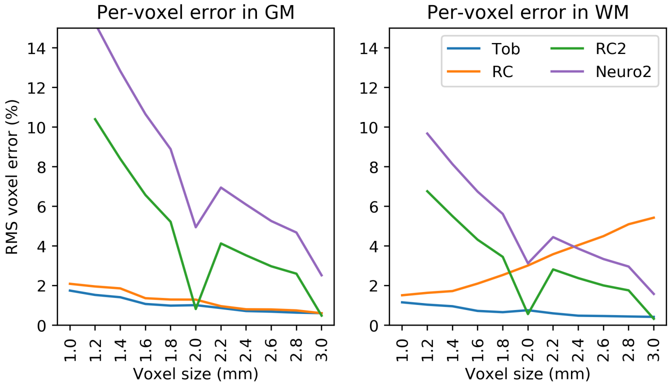
A. Simulated surfaces


Error in total GM volume

Error in total WM volume

Fig. 5 Simulated surfaces: error in total tissue volume. Toblerone showed consistency, though with small bias, for both GM and WM. RC1 errors were lower for GM than WM. Resampling-based methods (RC2, Neuro2) showed particular consistency in WM. [Full results in supplementary, fig. s5]

Fig. 5 shows the error in total tissue volume for the simulated surfaces. The numerical solution at 1mm was used as the reference. Toblerone showed consistency across voxel sizes, though with a small negative bias in both tissues. RC estimates showed variation in both. The resampling-based methods RC2 and Neuro2 showed high consistency in WM but less so in GM. The numerical solution was stable across voxel sizes. Neuro's results are excluded from this and subsequent graphs for clarity; the full results are given in the supplementary material (figs. s5 and s6).


Per-voxel error in GM

Per-voxel error in WM

Fig. 6 Simulated surfaces: per-voxel error. Toblerone and RC produced the lowest errors in GM; in WM there was a clear difference to Toblerone. RC2 and Neuro2's errors both decreased with increasing voxel size, with a characteristic notch observed at 2mm. [Full results in supplementary, fig. s6]

Fig. 6 shows per-voxel error for the simulated surfaces. Results were masked to consider voxels intersecting either surface of the cortex as only these contain PVs. Toblerone and RC produced the lowest errors at all voxel sizes in GM; in WM only Toblerone retained this behaviour. Both resampling-based methods (RC2, Neuro2) produced lower errors as voxel size increased, and a characteristic notch in their results was observed at 2mm. Although RC initially performed better than RC2 in WM, the inverse was true above 2mm voxel size.

B. BrainWeb simulated T1 images

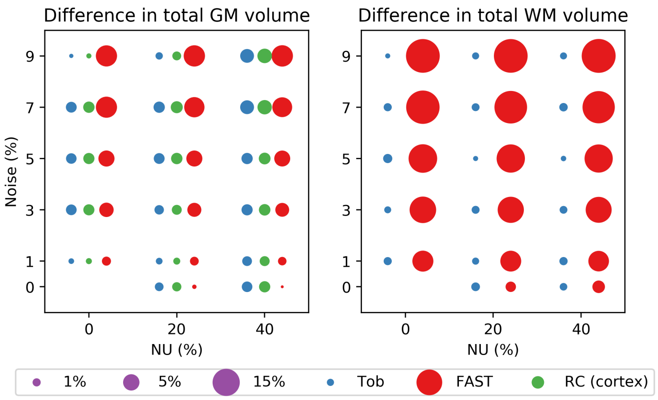


Fig. 7 BrainWeb: difference in total tissue volume referenced to each method's 0% noise 0% NU result. Surface-based methods were more consistent at almost all noise and NU levels; FAST was more consistent in GM than WM.

Fig. 7 shows the difference in total tissue volume across the brain as a function of noise and NU levels, referenced to each method's results at 0% noise and 0% NU. PV estimates at 1mm isotropic voxel size were used for this analysis. RC's GM result was for the cortex only as it cannot process subcortical structures. In general, the surface-based methods showed more consistency in their estimates across all levels of noise and NU, with the notable exception of GM at 40% NU. FAST's consistency was notably better in GM than WM.

Fig. 8 shows the RMS per-voxel difference in PV estimates at 3mm voxel size as a function of noise and NU. Each method's 1mm results at 0% noise 0% NU were used as the reference. Toblerone returned lower RMS voxel differences in both GM and WM at all levels of NU and noise except 0% noise 0% NU; a pattern that was repeated at other voxel sizes (these are shown in supplementary fig. s8).

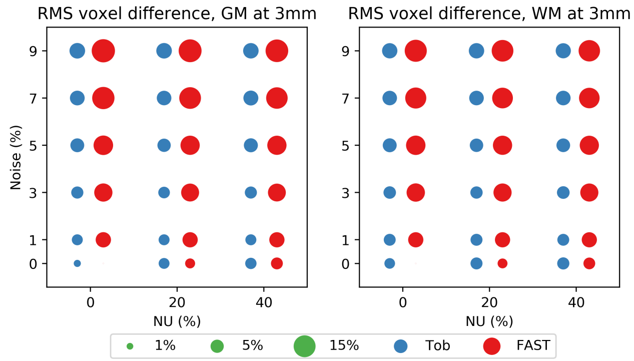


Fig. 8 BrainWeb: RMS per-voxel differences at 3mm voxel size, referenced to each method's 1mm 0% noise 0% NU results. Toblerone's differences were smaller at almost all levels of noise and NU, as was also the case at other voxel sizes. [Results for other voxel sizes are given in supplementary fig. s8]

C. HCP test-retest subjects

Fig. 9 shows violin plots of inter-session difference (retest minus test) in tissue volume across the 45 subjects of the HCP dataset. PV estimates at 0.7mm isotropic voxel size were used for this analysis. RC's GM result was for the cortex only. Both surface methods gave a tighter distribution than FAST, suggesting greater repeatability between sessions. All methods showed greater variability in GM than WM.

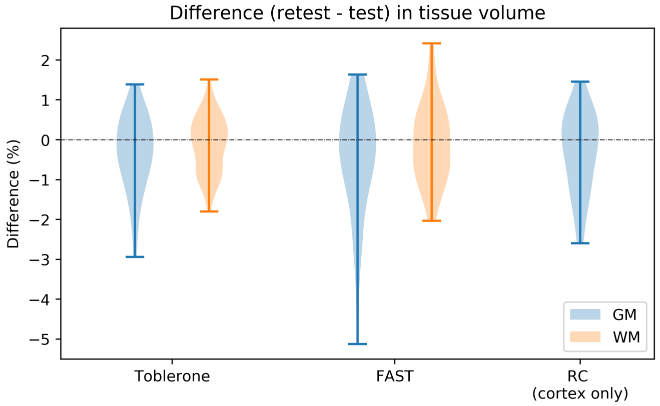


Fig. 9 HCP test-retest: inter-session (retest minus test) difference in total tissue volume. PVs were estimated in the native 0.7mm isotropic space of the structural images. RC's result is for the cortex only. Both surface methods show a tighter distribution than FAST.

Fig. 10 shows the mean per-voxel difference between Toblerone and FAST's GM PV estimates as a function of Toblerone's GM PV estimate. Excepting the 0.7mm result, the positive slope of each line shows that in voxels with a low Toblerone GM PV estimate, FAST was more likely to assign a higher value, and vice-versa at high Toblerone GM PV estimates. The strength of this relationship decreased with increasing voxel size. It should be noted that the 0.7mm result is the only one *not* to make use of resampling (for all others, FAST's 0.7mm estimates were resampled onto the target voxel grid).

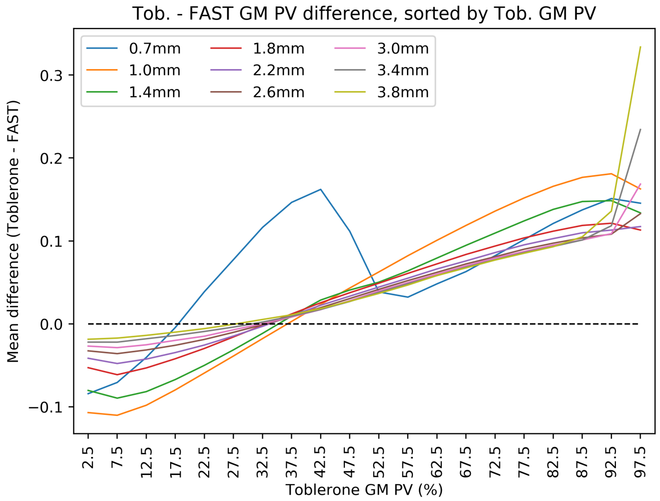


Fig. 10 HCP test-retest: mean difference between Toblerone and FAST GM PVs, sorted into 5% width bins according to Toblerone's GM PV. As Toblerone's GM PV estimate in a given voxel increases, FAST is more likely to assign a smaller value, and vice-versa. The strength of this relationship decreases with increasing voxel size. An inverse, but weaker, effect is seen for WM (supplementary fig s10).

VI. Discussion

Results from the simulated surfaces showed that Toblerone produced estimates with a comparatively low and consistent error. Although the RC method was able to perform similarly for GM, there was a clear advantage for Toblerone in WM. Results from the BrainWeb images suggested that a surface-based approach (the combination of FreeSurfer/FIRST and Toblerone) is more robust to noise and NU than FAST's volumetric approach. Further analysis suggested it is the consistency of FAST's WM estimates that suffers in the presence of these scanner imperfections. Finally, results from the HCP test-retest dataset showed that the surface-based approach provide better inter-session repeatability in total tissue volume.

The use of resampling – unavoidable for all volumetric methods that must transform PV estimates from a structural voxel grid to a functional voxel grid – degrades data quality in an unpredictable and highly localised manner. This chiefly arises due to so-called *subvoxel effects*, which may be illustrated via the following 1D example. Consider a row of voxels of size 1mm that are to be resampled onto 1.4mm voxels. A larger voxel overlaps evenly onto two smaller voxels, covering 0.7mm of each. The resampled value will be the mean of the two smaller, on the implicit and unlikely assumption that the tissues within each are evenly distributed. Next, consider a row of 1mm voxels that are to be resampled onto 3.4mm voxels, whereby a larger voxel overlaps by 0.2, 1, 1, 1 and 0.2mm onto smaller voxels. Again, the resampled value will be a weighted mean of the smaller voxels, but as the central three voxels are included wholly in the new voxel, the spatial distribution of tissues within these voxels is irrelevant and the assumption of even distribution can safely be made. As the ratio of output voxel size to input voxel size increases, the significance of subvoxel effects are therefore reduced.

It is extremely difficult to quantitatively measure the impact of resampling, particularly on non-simulated data. To do so would require the ability to express some volumetric reference data in an arbitrary voxel grid *without* making use of resampling, otherwise a trap of circular reasoning results. Nevertheless, such an analysis can be performed using the simulated surfaces presented earlier. The key conceptual difference is that the ground truth for this dataset is defined by a surface and can therefore calculated in any voxel grid without resampling. Fig. 11 shows the results of resampling ground truth results from the numerical method at each resolution to all other resolutions above the one in question (for example, the 1.4mm truth was resampled to 1.6, 1.8, … etc). At each voxel size, the resampled results can be compared to a ground truth that has been calculated without the use of resampling. RMS per-voxel error was measured using the same mask as before, namely all voxels intersecting either surface of the cortex, as only these contain PVs. Multiple trends can be seen: firstly, as the input voxel size increases, error at all output voxel sizes increases. Secondly, as the ratio of output to input voxel size increases, the error decreases. Finally, the error falls to zero when this ratio takes an integer value. This is due to the use of perfectly aligned voxel grids in this work (which would not be the case with patient data) and is discussed in section IV.B. This likely explains the interesting behaviour observed in various analyses, namely: the notches seen in fig. 6, as well as supplementary figs. s5 and s6 (perfect voxel correspondence means no subvoxel effects); the 0.7mm result in fig. 10 (for all other sizes, resampling by a non-integer ratio of voxel size blurs the FAST results, reducing image contrast and the number of high GM PV voxels); and the lack of error observed in FAST's GM and WM results at 2, 3 and 4mm voxel size, 0% noise 0% NU in figs. 8 and s8 (again, perfect voxel correspondence with the reference set of 1mm estimates). These considerations do not apply to surface-based methods as they do not make use of resampling.

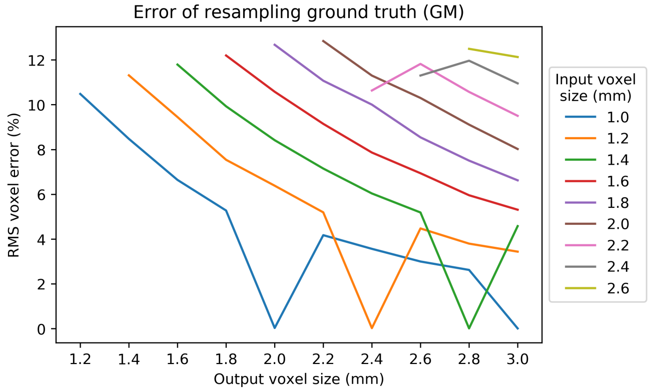


Fig. 11 Simulated surfaces: error induced by resampling the ground truth GM PV map, masked to voxels intersecting either surface of the cortex. As the input voxel size increases, error increases, but as the ratio output / input voxel size increases, error falls. Finally, error falls to zero when the ratio takes an integer value.

A further advantage of surface-based methods concerns their application of transformations. Notwithstanding the fact that volumetric methods require resampling to transform data from one resolution to another, they also require it to apply a registration transformation between the structural voxel grid in which PVs are estimated and the functional voxel grid in which PVEc is to be performed. Once again, the impact of this upon data quality is highly localised and difficult to measure quantitatively. It can however be illustrated via the following experiment, illustrated in fig. 12. GM PV maps for the 0% noise 0% NU BrainWeb image were translated by 0.5mm in each of the *x,y,z* axes. For FAST, this translation took the form of an affine transformation applied during a resampling operation. Significant blurring is seen, particularly around the edges of structures where there was previously good edge definition. As these edge voxels by definition contain PVs this is a particularly undesirable outcome. By contrast, blurring within a structure is of little consequence as the tissue is already homogenous. For Toblerone, this translation was performed by shifting the *surfaces* into the new reference voxel grid represented by the translation and then estimating the PVs afresh with no noticeable reduction in edge definition.

In its native form, the RC method is unable to correctly handle voxels in which all three tissue types are present (due to the fact that it estimates GM first and then assigns the remainder to either WM or non-brain). The impact of this is seen in the positive relationship between per-voxel error in WM and voxel size in fig. 6. Resampling can help to minimise this error: at

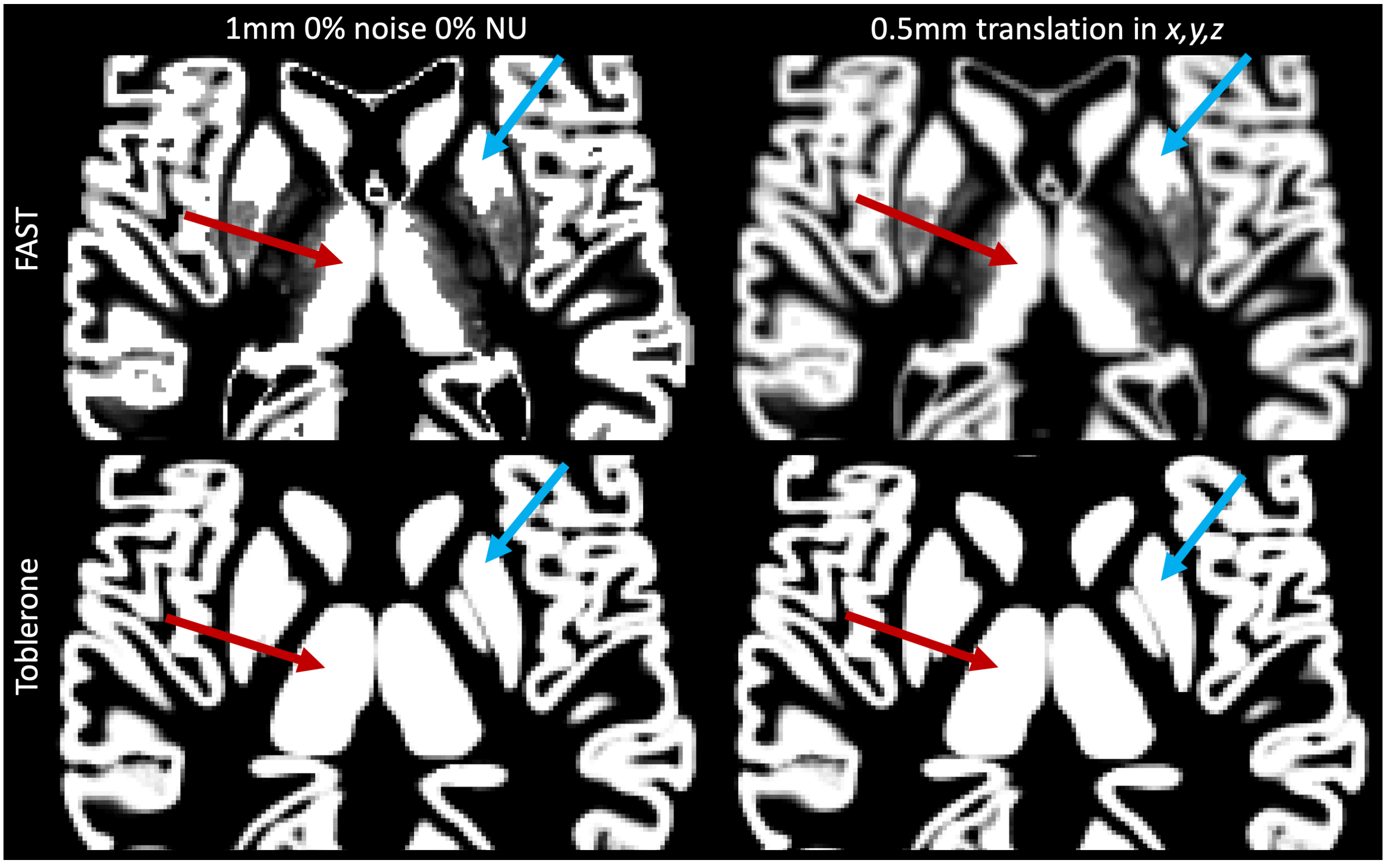


Fig. 12 Illustration of resampling-induced blurring on the 1mm isotropic GM PV map from the 0% noise 0% NU BrainWeb image. The left column shows the original estimates produced by FAST and Toblerone, the right shows the result of a 0.5mm translation along each axis. The left thalamus (red) and right putamen (blue) are highlighted in each, showing how surface and volumetric methods differ markedly in their interpretation of subcortical structures (FAST does not regard them as pure GM, whereas Toblerone does for the analyses presented in this work).

small voxel sizes, the probability of voxels containing three tissue types is smaller, so the error is minimised, but this does not hold true at larger voxel sizes. Accordingly, as voxel size increases, it is increasingly beneficial to obtain PV estimates by resampling those from a smaller voxel size. Set against this, however, are the aforementioned problems introduced by resampling: when the ratio of output to input voxel size is small, subvoxel effects are significant and high per-voxel errors result (as shown in fig. 6). A threshold voxel value above which resampling is beneficial therefore results (at around 2mm in the figure). The exact value of this threshold would be difficult to predict in the general case (in particular, the use of aligned voxel grids in this work is both highly significant and extremely unrealistic). By contrast, Toblerone is able to produce consistent estimates in all tissue classes at arbitrary voxel sizes without the use of resampling.

We were unable to further analyse the HCP test-retest dataset in order to establish where in the brain the differences between Toblerone's and FAST's estimates arise. As this would require extensive use of non-linear registrations and resampling to transform all subjects onto a common template, it is likely that the artefacts imposed by this process would obscure the true methodological differences of interest. Furthermore, an analysis on the BrainWeb database would be of limited use as this only represents the cortical anatomy of a single subject and would therefore ignore population variability.

VII. Conclusion

Toblerone is a new method for estimating PVs using surface segmentations. Unlike existing surface-based tools, it is not closely tied to any specific modality or structure and can therefore be adapted to multiple use cases (notably, providing PV estimates for the whole brain). It is able to operate at arbitrary resolutions without recourse to resampling, thereby avoiding the highly localised degradation of image quality that this process entails. Three datasets have been used to evaluate the algorithm. Results from simulated surfaces show consistently low errors at both the voxel and aggregate level, either matching or surpassing other surface-based methods. Results on simulated T1 images from the BrainWeb database show that a FreeSurfer/FIRST/Toblerone surface-based pipeline used as an alternative to FAST is more robust in the presence of random noise and field non-uniformity. Finally, results from the HCP test-retest dataset of 45 subjects show that the surface-based pipeline produces a tighter distribution of inter-session tissue volumes than FAST, suggesting the surface approach has greater repeatability. The magnitude of methodological differences observed in this work, and related conceptual questions concerning the interpretation of subcortical tissue between surface and volumetric methods, will have implications for the wider process of PVEc.

Acknowledgement

The authors would like to thank Gabriel Mangeat of Polytechnique Montréal for assistance with the NeuroPVE method. The University of Oxford Advanced Research Computing (ARC) facility was used in preparing this work (http://dx.doi.org/10.5281/zenodo.22558). Data was provided by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University.

References

[1] B. Fischl and A. M. Dale, “Measuring the thickness of the human cerebral cortex from magnetic resonance images,” *Proc. Natl. Acad. Sci.*, vol. 97, no. 20, pp. 11050 LP – 11055, Sep. 2000.

[2] H. W. Müller-Gärtner, J. M. Links, J. L. Prince, R. N. Bryan, E. McVeigh, J. P. Leal, C. Davatzikos, and J. J. Frost, “Measurement of Radiotracer Concentration in Brain Gray Matter Using Positron Emission Tomography: MRI-Based Correction for Partial Volume Effects,” *J. Cereb. Blood Flow Metab.*, vol. 12, no. 4, pp. 571–583, Jul. 1992.

[3] I. Asllani, A. Borogovac, and T. R. Brown, “Regression algorithm correcting for partial volume effects in arterial spin labeling MRI,” *Magn. Reson. Med.*, vol. 60, no. 6, pp. 1362–1371, Sep. 2008.

[4] M. A. Chappell, A. R. Groves, B. J. MacIntosh, M. J. Donahue, P. Jezzard, and M. W. Woolrich, “Partial volume correction of multiple inversion time arterial spin labeling MRI data,” *Magn. Reson. Med.*, vol. 65, no. 4, pp. 1173–1183, 2011.

[5] D. W. Shattuck, S. R. Sandor-Leahy, K. A. Schaper, D. A. Rottenberg, and R. M. Leahy, “Magnetic resonance image tissue classification using a partial volume model,” *Neuroimage*, vol. 13, no. 5, pp. 856–876, 2001.

[6] B. Fischl, “FreeSurfer,” *Neuroimage*, vol. 62, no. 2, pp. 774–781, 2012.

[7] Y. Zhang, M. Brady, and S. Smith, “Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm,” *IEEE Trans. Med. Imaging*, vol. 20, no. 1, pp. 45–57, 2001.

[8] M. F. Glasser, S. N. Sotiropoulos, J. A. Wilson, T. S. Coalson, B. Fischl, J. L. Andersson, J. Xu, S. Jbabdi, M. Webster, J. R. Polimeni, D. C. Van Essen, and M. Jenkinson, “The minimal preprocessing pipelines for the Human Connectome Project,” *Neuroimage*, vol. 80, pp. 105–124, 2013.

[9] D. N. Greve, C. Svarer, P. M. Fisher, L. Feng, A. E. Hansen, W. Baare, B. Rosen, B. Fischl, and G. M. Knudsen, “Cortical surface-based analysis reduces bias and variance in kinetic modeling of brain PET data,” *Neuroimage*, vol. 92, pp. 225–236, May 2014.

[10] D. N. Greve, D. H. Salat, S. L. Bowen, D. Izquierdo-Garcia, A. P. Schultz, C. Catana, J. A. Becker, C. Svarer, G. M. Knudsen, R. A. Sperling, and K. A. Johnson, “Different partial volume correction methods lead to different conclusions: An (18)F-FDG-PET study of aging,” *Neuroimage*, vol. 132, pp. 334–343, May 2016.

[11] B. Patenaude, S. M. Smith, D. N. Kennedy, and M. Jenkinson, “A Bayesian model of shape and appearance for subcortical brain segmentation,” *Neuroimage*, vol. 56, no. 3, pp. 907–922, Jun. 2011.

[12] F. S. Nooruddin and G. Turk, “Simplification and repair of polygonal models using volumetric techniques,” *IEEE Trans. Vis. Comput. Graph.*, vol. 9, no. 2, pp. 191–205, 2003.

[13] T. Akenine-Möller, “Fast 3D Triangle-Box Overlap Testing,” *J. Graph. Tools*, vol. 6, no. 1, pp. 29–33, 2001.

[14] C. B. Barber, D. P. Dobkin, and H. Huhdanpaa, “The Quickhull Algorithm for Convex Hulls,” *ACM Trans. Math. Softw.*, vol. 22, no. 4, pp. 469–483, Dec. 1996.

[15] C. Van Assel, G. Mangeat, B. De Leener, N. Stikov, C. Mainero, and J. Cohen, “Partial volume effect correction for surface-based cortical mapping,” *Proc. Int. Soc. Magn. Reson. Med. Annu. Meet. Paris*, 2017.

[16] D. L. Collins, A. P. Zijdenbos, V. Kollokian, J. G. Sled, N. J. Kabani, C. J. Holmes, and A. C. Evans, “Design and Construction of a Realistic Digital Brain Phantom,” *IEEE Trans. Med. Imaging*, vol. 17, no. 3, pp. 463–468, 1998.

[17] C. Cocosco, V. Kollokian, R. K. Kwan, G. B. Pike, and A. C. Evans, “BrainWeb : Online Interface to a 3D MRI Simulated Brain Database,” *3-rd Int. Conf. Funct. Mapp. Hum. Brain*, vol. 5, no. 4, p. S425, 1997.

[18] M. Y. Zhao, M. Mezue, A. R. Segerdahl, T. W. Okell, I. Tracey, Y. Xiao, and M. A. Chappell, “A systematic study of the sensitivity of partial volume correction methods for the quantification of perfusion from pseudo-continuous arterial spin labeling MRI,” *Neuroimage*, 2017.

SUPPLEMENTARY MATERIAL

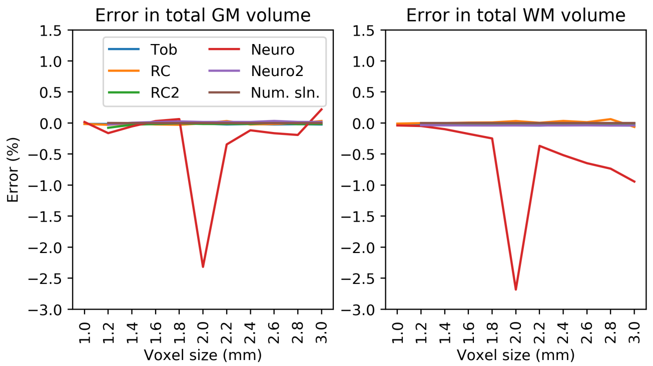


Fig. s5 Simulated surfaces: error in total tissue volume, all methods. The notch in the Neuro method at 2mm may arise due to an interplay between the number of expanded surfaces created (5) and the voxel size.

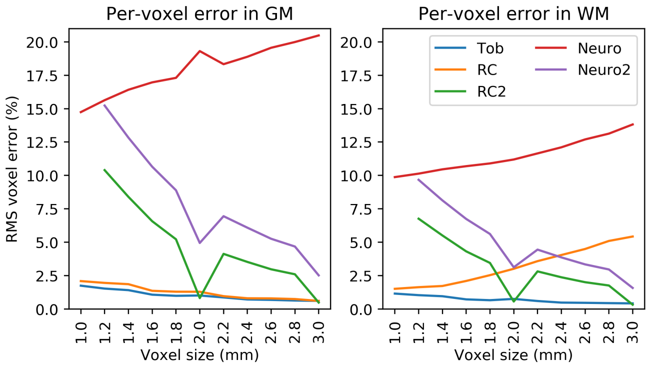


Fig. s6 Simulated surfaces: RMS per-voxel error. Neuro's results are significantly worse than all other methods at all other resolutions, though the resampled version (Neuro2) performs better.

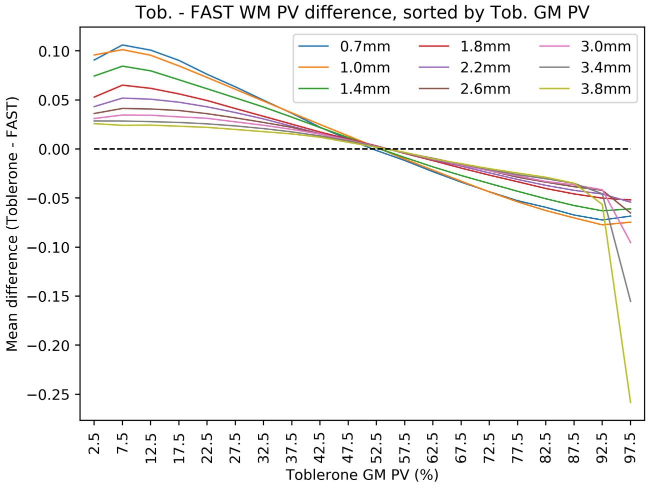


Fig. s10 HCP test-retest: mean difference between Toblerone and FAST WM PVs, sorted into 5% width bins according to Toblerone's GM PV. This is the analogue of fig. 10, showing a weaker and inverse relationship.

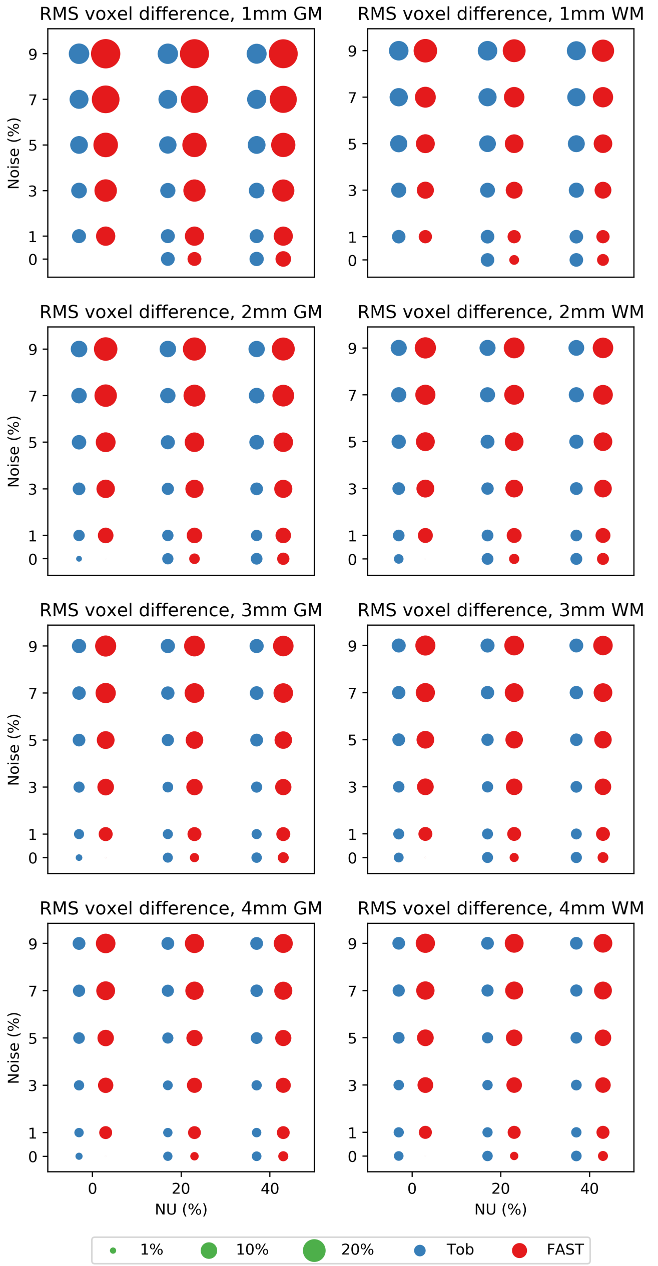


Fig. s8 BrainWeb: RMS per-voxel differences at voxel sizes of 1 to 4mm isotropic, referenced to each method's 1mm 0% noise 0% NU results. Toblerone's differences were smaller at almost all levels of noise and NU, the exception being 0% noise 0% NU.

All of the below responses are from the lead author (Thomas Kirk).

Reviewers' Comments:
Reviewer: 1

Comments to the Author
The article deals with the issue of estimating the partial-volume priors necessary for a PV-correction. A novel method that estimates the PV maps from a surface-based segmentation instead of by using volumetric segmentation is presented. This an important topic, as PV-maps from volumetric segmentation can contain errors which are then propagated in the partial volume correction. The article is well written, clear, and easy to understand in general. The method seems to perform better than the reference surface-based segmentations. Intuitively, the surface based segmentation can offer superior results also to the volumetric segmentation, but this still needs to be very clearly demonstrated. However, there are certain things unclear to me in the validation and results sections. These need to be better explained or fixed before the publication of the manuscript (see specific comments below):

MAJOR:
1. I understand that for Toblerone, boxcar point-spread function is assumed when calculating PV at different resolutions. In Section IV.D, applywarp is used to resample images to a coarser resolution. What point-spread-function is assumed? Boxcar as well to make it comparable to Toblerone, or a more realistic PSF like sinc/Guassian etc.? In case of a difference it would be good to discus if this could influence the results. Also, ASL, BOLD, and PET data all have slightly different PSF. Can you, please, clarify what voxel shape is assumed for Toblerone, and then mention this in the discussion?

We assume cuboid voxels (they do not need to be isotropic) but do not account for modality specific PSF or mixing of signal between voxels. This is because such effects can be accounted for by a separate convolution operation, in a manner appropriate to the modality in question. This also applies to applywarp. Please see the first paragraph of section III A and section IV B.

2. Methods IV.E. How exactly was the ground-truth established using the surfaces for the simulations? Please, provide details for the numerical method.

The numerical solution is discussed in more detail in section IV C.

3. Figures 5 and 6 show a voxel-wise error and a total error. From the text and figures, I understand that the mean is shown for the total error, and STD for the voxel-wise error. It would help to show either the mean voxel-wise error, or Jaccard or Tanimoto's coefficient to show the voxel-wise performance of the methods. Or the graph gives mean voxel-wise error and the bars give standard deviation? Then the description in the text, in the figure caption, and in the graph axes needs to be improved.

The voxel-wise metric has been changed to RMS. Axes labels have been updated as appropriate.

4. Large voxels are claimed to be problematic for RC1 and this is addressed with RC2. RC2 looks worse on Figure 5 than RC1 and in the discussion, it is claimed that this is due to the resampling issues with non-match voxel-sizes. I understand that this could be a problem with e.g. 1mm->1.5mm resampling. But I don't see why there should be a resampling error between 1mm->3mm - at least not if the matrix are similarly positioned. Also, I don't understand why the RC method cannot be calculated on sub-divided voxels with size ~0.1-0.2mm as is done for Toblerone. Can you comment on the effect of matrix positioning, error on resampling for multiples like 1->2mm or 3mm, and on calculating RC on 0.1mm -> otherwise it seems that this is the only advantage of using Toblerone.

The discussion section now addresses in resampling in some detail. With regards to the figure referred to by the reviewer, the corresponding figure in the revised paper hopefully better illustrates the behaviour mentioned (fig 6). The RC method does indeed make use of voxel subdivision and this fact is now mentioned explicitly (section IV A). Where image matrices are perfectly aligned this is also mentioned explicitly.

5. You claim that for Figure 6: "The results for Neuro2 and RC2 are not shown for clarity as they are resampling methods; excepting for rounding errors, total class volume will not change between resolutions.". However, for figure 8, the FAST results are different across voxel-sizes. Despite that both FAST and RC2 were upsampled in the same way using the approach in IV.D. It appears to me as inconsistency both in results and in reasoning. Can you, please, comment on that?

Fig 5 is the corresponding figure in the revised paper to which the reviewer refers. Resampling methods are now shown in this and it can be seen there is some small variation between voxel sizes, which logically can only be attributed to numerical errors in the resampling process (rounding etc). The greater variability in GM is perhaps explained by the fact that there is less GM present in the simulated surfaces, whereas the comparatively larger mass of WM – most of which do not have PVs – presumably acts to stabilise the result against numerical errors.

MINOR:

6. Page 2 - "...non-brain required for ASL..." - you might add PET here as all three modalities (PET, BOLD, ASL) were mentioned in the first paragraph.

Done

7. Page 3. You mention a second subdivision and a further subdivision. First subdivision would give voxels of around 0.5-1mm, the second division would thus give something between 0.1-0.2mm if I understand everything correctly. Can you, please, clarify if 'further subdivison' gives you 0.1-0.2mm and also specify if there are 2 or 3 subdivisions (avoiding the term 'further subdivision').

The paragraph (end of section II A) has been revised to make it clear that there are at most two levels of subdivision. The second level does lead to voxels around 0.1 – 0.2 mm isotropic.

8. "Toblerone" - it might be useful to explain why this name/acronym was chosen. Is it only per shape resemblance to Fig 1?

My mistake! A footnote has been added at the end of section I

9. "Corpus collosum" -> "corpus callosum"

Thank you

10. Methods IV.C. - You mention that FAST has to be restricted to cortex as the surface-based methods can only segment cortex. In III.B, thalamus and sub-cortex is mentioned for Toblerone. Can you, please, specify that the cortex limitations apply only to "RC" and "Neuro" methods, or clarify this otherwise.

This is addressed via the expanded scope of the paper: as we now use Toblerone across the full brain, masking FAST for individual structures is no longer an issue, which enables a much fairer comparison to be made between the two.

11. Please, include also a figure with at least a single axial slice of the resulting segmented maps to allow visual assessment of the differences.

See fig 12.

12. "The significance of this _is_ increases with voxel size."

Thank you

Reviewer: 2

Comments to the Author
The authors presented a surface-based method for quantifying partial voluming effect in neuroimaging studies. The method adopts graphics algorithms to compute the portion of a voxel intersected by a local surface patch. The method was compared with variants of two existing surface-based method and a traditional volume-based method in simulated and HCP experiments. Results showed indication of the superiority of the proposed method.

As techniques used in Toblerone for estimating partial voluming are standard, focus should be placed on the validation, whether the method has the potential to make real impact on current practices. Regarding this, the manuscript could benefit from clarifying the following questions.

1. I believe the authors wanted to plot something like 'mean squared error' rather than std of error. Std of error means first taking the average of the error and then computing root of mean square modulated by that average.

We now use the metric of RMS voxel error

2. While methods based on resampling (e.g. Neuro2 an RC2) produce constant metrics, authors could still display them in the plots as flat lines, because those metrics still indicate the accuracy of those approaches (being invariant to resolution is presumably a better property).

These methods are now displayed in fig 5.

3. It doesn't really justify to omit Neuro1 and Neuro2 in followup analyses as simulation and real data are totally different monsters. Furthermore, if the accuracy results (pattern of the curves) of Neuro1,2 are similar to the simulation study, it further validates the correctness of Toblerone.

Although it is undoubtedly the case that simulation data and real data are very different, I believe that the difference referred to by the reviewer here concerns the question of ‘how well does the simulation represent the features and challenges found in real surfaces’. The decision for omitting Neuro from further analysis was made on the basis that it was unable to produce estimates for the simulated surfaces as accurately as the other methods, regardless of whether those simulated surfaces are a realistic representation of anatomy or not. A robust surface estimation method should be able to operate well on any surface, including those that are not necessarily anatomical. Good performance on such a surface is a prerequisite for good performance on anatomical surfaces. Neuro was least able to meet this first criterion.

4. As also suggested by the authors, the comparison between Toblerone and FAST is ambiguous because the two methods start from different stages. I understand this comparison is challenging in general. In the simulation study, is it a good idea to start the comparison from simulated MRI images? That is, simulate 3D volume images (with typical noise and artifacts) from ground-truth cortical surfaces, then perform 1) FAST; 2) FIRST+Toblerone. This would support the claim that Toblerone works better in real life.

This is something that has been considered at *great* length. Ultimately, it was decided that simulating T1 images from cortical surfaces would be too methodologically challenging: moving from surface to volume (the inverse of the FreeSurfer process) would be complicated and would involve some sort of PV estimation tool – exactly what we are trying to evaluate here. We have performed a very similar experiment using the BrainWeb MRI simulator which provides us with a direct comparison between FAST and FreeSurfer/FIRST + Toblerone in the presence of scanner imperfections (section IV D)

5. What is the 'numerical method' that the authors mentioned to compute ground truth in the simulation? If it is so accurate, why not use it as the final PV estimation tool?

The numerical method is now discussed in section IV C. It relies upon the fact that the simulated surfaces are defined by analytic functions so we could not use it for real surfaces!

6. Computation resources should be given with "FSL FIRST is around 20 minutes"

Run time is now included.

7. The validation on the HCP gives really indirect evidence. Is it possible to use phantom studies or clinical repeatability studies as other validation approaches?

We have made two major changes to address this point: the use of simulated T1 MRI images and the use of the HCP test-retest data to investigate repeatability (sections IV D and IV E).

Typo: IV C "...surround said vertex are connected..."

Thank you

Reviewer: 3

Comments to the Author
The paper proposes a novel method for estimation of partial volumes: Toblerone which is a new method for estimating PVs using surface segmentation. The problem of accurate tissue classification is of wide interest and therefore the paper could be a positive contribution. This is an interesting work and the idea is interesting.

Having said that, I think significantly more effort needs to be spend before the paper becomes journal paper. Some, of the issues are pointed below.

* Some important citations are missing. e.g. DW Shattuck - 2001 .

Included

* My main issue with this paper is that all the methods for surface extraction that I know, first use tissue classification, partial volume estimation and then surface generation. These methods have been validated and well established and validated. Is there any need for another method? Moreover, the problem of errors in tissue classification is generally bias field. The bias field in MRI images tends to have a much significant impact than anything else as far as tissue classification in concerned. It tends to be severe in 3T and 7T images. The proposed method should be evaluated against conventional methods for different field strength images to show that the proposed method has some advantages.

It is important to note that the PV estimation referred to here is within the space of the structural image being used for surface generation, which is not necessarily the space in which we would like to perform PVEc. Furthermore, even if surface generation methods saved these PV estimates as an extra output, the fact that they are intrinsically tied to the volumetric grid of the structural image means that one would still need to resample them to another space to perform PVEc which entails a loss of accuracy. By contrast, Toblerone always estimates in the space in which estimates are required. Regarding field non-uniformity, we hope that the BrainWeb (section IV D) experiments will satisfy this question.

* The marching cube algorithm for surface extraction uses surface constraints similar to the ones used in the paper.

Many surface generation methods do indeed make use of marching cubes. This is to perform the operation of volume data -> surface segmentations, whereas what we aim to do here is the inverse: surface -> volume (though in a different space to the original volumetric data). In conjunction with the previous point, marching cubes operates within the voxel grid of the structural image, whereas in this work we seek to express PVs in arbitrary grids, not necessarily the original grid.

* One way to validate the proposed method is to check intersession /scanner test -retest studies. For example, acquire a high resolution scan, compute PVC. Then downsample the scan, or change bias field or acquire scan with different field strength and then show consistency. The existing simulation results show promise but need more work to show that this will be useful in practice.

The HCP test-retest data has provided us with a much better analysis than what we had before.

* Another issue could be that the assumption that a sharp inner cortex and pial cortex exists may not be valid in the first place,. Some citations should be added about this.

This is a very valid point but is more appropriately directed at the entire class of surface segmentation methods. This work seeks to answer the question of ‘can PVs be estimated from surface segmentations, and are there any benefits to doing so’? It is therefore beyond the scope of this work to second-guess the quality of the surfaces, though it is undoubtedly the case that any criticism that may be levelled at surface segmentation methods may also be levelled at this work.

Toblerone: surface-based partial volume estimation

Thomas F. Kirk, Timothy S. Coalson, Martin S. Craig and Michael A. Chappell

***Abstract*—Partial volume effects (PVE) present a source of confound for the analysis of functional imaging data. Correction for PVE requires estimates of the partial volumes (PVs) present in an image. Conventionally these estimates are obtained via volumetric segmentation, but such an approach may not be accurate for complex structures such as the cortex. An alternative is to use surface-based segmentation, which is well-established within the literature. Toblerone is a new method for estimating PVs using such surfaces. It uses a purely geometric approach that considers the intersection between a surface and the voxels of an image. In contrast to existing surface-based techniques, Toblerone is not restricted to use with any particular structure or modality. Evaluation in a neuroimaging context has been performed on simulated surfaces, simulated T1-weighted MRI images and finally a Human Connectome Project test-retest dataset. A comparison has been made to two existing surface-based methods; in all analyses Toblerone's performance either matched or surpassed the comparator methods. Evaluation results also show that compared to an existing volumetric method (FSL FAST), a surface-based approach with Toblerone offers improved robustness to scanner noise and field non-uniformity, and better inter-session repeatability in brain volume. A surface-based approach negates the need to perform resampling (in contrast to volumetric methods) which is particularly advantageous for low-quality data.**

Commented [TK1]: Now covering the whole brain, instead of just the cortex as previously.

Commented [TK2]: This is an addition (BrainWeb)

Commented [TK3]: Previously we used single session data

Commented [TK4]: Shown using the BrainWeb and HCP test-retest data, both of which are new to this paper

Commented [TK5]: This forms a greater part of the discussion than previously.

***Index Terms*—functional imaging, partial volume effect, partial volume correction, segmentation, surface**

I. Introduction

PARTIAL volume effects (PVE) arise when an imaging matrix has low spatial resolution in relation to the structures of interest within the image, as is commonly the case for the functional imaging techniques, such as positron emission tomography (PET), blood oxygen-level dependent fMRI (BOLD) and arterial spin labelling (ASL). For example, ASL voxels typically have side lengths of 3-4mm whereas the mean thickness of the adult cortex is 2.5mm [1]. As such, voxels around the cortex will contain a mixture of cortical and non-cortical tissues, the proportions of which are termed *partial volumes* (PVs). PVE present a source of confound for functional imaging: whilst the objective is to obtain a measurement of function across some particular structure, the signal actually measured in each voxel is a sum, weighted by the partial volumes, of function both within and without said structure. This is a mixed-source problem in which the multiple tissues in each voxel constitute the sources. Partial volume correction (PVEc) uses voxel-wise estimates of PVs to separate out the signals arising from each tissue. Various PVEc methods have been developed, usually with a specific modality in mind (for example, Muller-Gartner for PET [2] and linear regression [3] or spatially-regularised variational Bayes for ASL [4]).

Estimation of PVs bears considerable similarity to volumetric segmentation and the two are typically performed concurrently on a structural image, as is demonstrated in [5]. In order to estimate PVs within the voxel grid of a functional image, it is then necessary to transform the results from the structural voxel grid to the functional. As each functional voxel corresponds to multiple smaller voxels on the structural image, the PVs of the former can be estimated using the results from the latter. The efficacy of this approach is limited by the accuracy of the volumetric segmentation approach used. For complex geometries, such as the thin and highly folded structure of the cerebral cortex, the alternative of surface-based segmentation has gained widespread support (notably through FreeSurfer [6]). The advantage of such a segmentation method is twofold. Firstly, whereas volumetric segmentation is necessarily a discrete operation in terms of voxels, a surface approach is somewhat continuous as the surface vertices are placed with subvoxel precision. Secondly, anatomically-informed constraints can be enforced anisotropically when surfaces are used: for example, the constraint that tissues should be homogenous along a surface but heterogeneous across it. This is in contrast to a volumetric tool such as FSL FAST [7] which does enforce a similar tissue continuity constraint via the use of Markov random fields but only isotropically in the neighbourhood of each voxel. In principle, it should be possible to estimate PVs by considering the geometry of intersection between the individual voxels of an image and the surface segmentations of individual structures. Being a purely geometric construct, namely, *given a surface that intersects a voxel, what is the volume within the voxel bounded by the surface*, this is a fundamentally different approach to existing methods and it is expected this will be reflected in the estimates produced.

Although surface-based PV estimation tools exist in the literature, past efforts have usually been designed with a specific modality in mind. Two notable examples for neuroimaging are the ribbon-constrained (RC) method used within the Human Connectome Project's (HCP) *fMRISurface* pipeline [8] and PETSurfer [9], [10], a variant of FreeSurfer. The former is designed for use with BOLD and so distinguishes only between cortex and otherwise, not the grey matter (GM), white matter (WM) and non-brain required for ASL and PET; whereas the latter is both PET-specific and tightly integrated into FreeSurfer such that it is hard to use independently of that

T. F. Kirk (thomas.kirk@eng.ox.ac.uk), M. S. Craig (martin.craig@eng.ox.ac.uk) and M. A. Chappell (michael.chappell@eng.ox.ac.uk) are at the Institute of Biomedical Engineering, Department of Engineering Science, and the Wellcome Centre for Integrative Neuroimaging, Nuffield Department of Clinical Neurosciences, both at the University of Oxford, UK.

T. S. Coalson (tsc5yc@mst.edu) is at the Department of Neuroscience, Washington University School of Medicine, St Louis, USA. Funding was provided by the EPSRC (EP/P012361/1) and (for T. F. Kirk) the Bellhouse scholarship at Magdalen College, Oxford. T. S. Coalson was funded by NIH Grant R01 MH-060974 (to D. Van Essen). We intend to provide access to the data underlying the figures within this paper via a DOI (as per the EPSRC).

workflow. Furthermore, both methods deal exclusively with surfaces representing the cortex. The objective of this work was to develop an algorithm, named Toblerone[1], to estimate partial volumes for both cortical and subcortical structures (where such surfaces are available, for example via FSL FIRST [11]) for neuroimaging applications. The end result is highly general and could be used with images from multiple modalities and/or in other parts of the body.

Commented [TK6]: In response to reviewer

II. Theory

Voxelisation is the process of quantifying the volume contained within a surface and many algorithmic methods are given in the computer graphics literature. The key operational step for this is determining if a point lies interior or exterior to a given surface; by repeating this test entire volumes can be built up. The ray intersection test outlined by Nooruddin and Turk [12] is widely used and requires only that the surfaces be contiguous (water-tight). The test is performed by projecting an infinite ray in any direction from the point under test and counting the number of intersections made with the surface. A ray from an interior point will make an odd number of intersections as it exits the surface (including folds within the surface, there will be one more point of exit than entry); conversely an exterior point will make an even number of intersections (balanced entries and exits), if at all. This test scales badly with increasing spatial resolution: for a linear resolution of n samples per unit distance, n^3 tests per unit volume are required. Furthermore, as each ray must be tested against each surface element, the test also scales with surface complexity (linearly for a naïve implementation). For a typical functional image of 10^5 voxels and 2.5 x 10^5 surface elements in a FreeSurfer cortical surface, this is prohibitively computationally intensive.

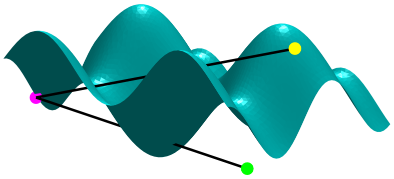


Fig. 1 Reduced ray intersection test for non-contiguous surfaces. The root point (interior) is shown in magenta. A ray from an interior point (green) makes two intersections due to the presence of a fold; from an exterior point (yellow) there is one intersection.

The method adopted in this work is to only use the portion of surface that actually intersects a given voxel (termed the 'local patch') for ray intersection testing. The local patch is defined as all triangles that intersect the voxel or, equivalently, the minimal set of triangles that unambiguously divides the voxel into two regions. This patch is by definition non-contiguous, so it is necessary to modify the ray intersection test accordingly; the modified form is referred to as the 'reduced' test in contrast to the 'classical' test. Within each voxel, a 'root point' that is known to lie within the surface is identified via the classical ray test. Any other point within the voxel may then be tested by projecting the finite line segment $\mathbf{r} = \mathbf{p_t} + \lambda(\mathbf{p_r} - \mathbf{p_t})$, where $\mathbf{p_t}$ is the point under test, $\mathbf{p_r}$ is the root point and $0 \leq \lambda \leq 1$ is a distance multiplier along the line. A parity test is then applied to the number of intersections identified between the root and test points. The fact that the line terminates at a point interior to the surface means that exterior points will lead to one more point of entry than exit; conversely interior points will lead to either zero or an even number of intersections. It is not necessary to test surface elements outside the voxel as the finite length of the line segment means it can never leave the voxel. Fig. 1 provides an illustration of the test in practice.

In order to minimise the number of tests required per voxel, convex hulls (defined as the smallest possible region enclosing a set of points within which any two points can be connected without leaving the region) are used to estimate partial volumes wherever possible. The rationale for this is that if the extrema points of a region can be classified as interior/exterior to a surface then, to an approximation, all points lying within the convex hull of these points will share the same classification.

III. Algorithm

The following section addresses PV estimation for structures within the brain, for which the tissue classes of interest are GM, WM and non-brain. The same principles would apply to structures in other areas of the body, though the interpretation of tissue classes would differ.

Commented [TK7]: An explicit statement of the fact that the focus of this work is a neuroimaging context but the principles could be adapted for other uses.

A. Estimation for a single surface

The core algorithm within Toblerone estimates the voxel-wise interior/exterior PVs arising from the intersection of a single surface with an arbitrary voxel grid. Toblerone assumes cuboid voxels with a 'boxcar' point-spread function (PSF), which is to say that it does not allow for any mixing of signal between voxels. In reality, different modalities have differing PSFs and such effects may be separately accounted for via a convolution operation. The first step is to identify and record the local patches of surface intersecting each voxel of the reference image via Moller's triangle-box overlap test [13].

Commented [TK8]: In response to reviewer feedback

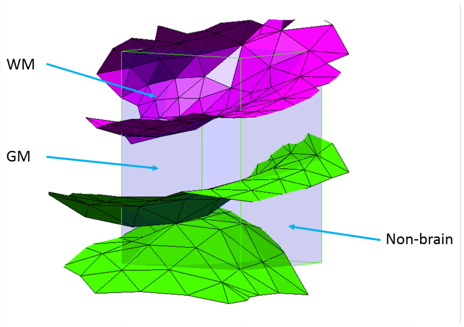


Fig. 2 Intersection of inner (magenta) and outer (green) surfaces of the cortex with a voxel. The outer surface intersects twice with distinct patches of surface; this is likely due to the presence of a sulcus. Tissue PVs are labelled.

[1] So-named because an early implementation constructed triangular prisms.

The geometry of a surface within a voxel can frequently be complex: using a sulcus of the cortex as an example, the surface may intersect the voxel multiple times, with the opposite banks of the sulcus appearing as two unconnected patches of surface, illustrated in fig. 2. Accounting for the many possible surface/voxel configurations requires numerous specific tests that rapidly become excessively complex, so the approach taken in Toblerone is to divide and conquer each voxel as required. As the length scale of a voxel decreases, the complexity of the local surface configuration within the voxel will also decrease (for example, a sulcus is less likely to intersect the voxel multiple times). Each voxel of the reference image is therefore divided into a number of subvoxels, each of which is processed individually. In the neuroimaging context of this work, the subdivision factor was set empirically as $\text{ceil}(\mathbf{v}/0.75)$ where $\mathbf{v}$ is the vector of voxel dimensions and 0.75 represents the lower limit of feature size found in the brain (in other contexts this parameter could be varied). Note that this subdivision factor transforms anisotropic voxels into approximately isotropic subvoxels. Subvoxels are then processed according to the following framework:

Commented [TK9]: Changed value in relation to the last submission

- If the subvoxel does not intersect the surface, it is assigned a single-class volume according to an interior/exterior classification of its centre. This is illustrated in fig. 3a.
- If the subvoxel intersects the surface, then it contains interior and exterior PVs. One of these will be estimated using a convex hull (via the Qhull implementation [14]) if the geometry of the surface is favourable, as follows:
  - ○ If the surface intersects entirely through one face of the subvoxel, then it encloses a highly convex volume that may be reliably estimated. The other partial volume is calculated by subtraction from the total subvoxel volume. This is illustrated in fig. 3b.
  - ○ If the surface is folded within the subvoxel (identified by multiple intersection of the surface along an edge or face diagonal of the subvoxel) then the subvoxel is subdivided a second time. This is because it is difficult to reliably identify which volume is interior or exterior in such a situation. This is illustrated in fig. 3c/d.
  - ○ In all other cases, convex hulls are again used. In order to minimise the potential error associated with estimation of a non-convex volume via convex hulls, it is important to identify which of the two PVs within the subvoxel is closer to being convex than the other. The proxy measure used in this work is the number of subvoxel vertices lying on either side of the surface: the side with fewer vertices is assumed to enclose a more convex (and at any rate smaller) volume than the other. This is illustrated in fig. 3e.
- If the surface intersects the subvoxel multiple times (identified by the successful separation of surface nodes lying within the subvoxel into unconnected groups) then the voxel is subdivided a second time. This situation occurs for example when the opposite banks of a sulcus pass through a voxel. Although the reduced ray intersection test is accurate in such a situation, forming convex hulls is not, so subdivision is the safer option. This is illustrated in fig. 3f.

The second subdivision is performed at a constant factor of 5 to yield sub-subvoxels of approximately 0.1 to 0.2mm side length isotropic. These are always assigned a single-class volume based on a classification of their centre points as their small size

Commented [TK10]: Hopefully less ambiguous than before!

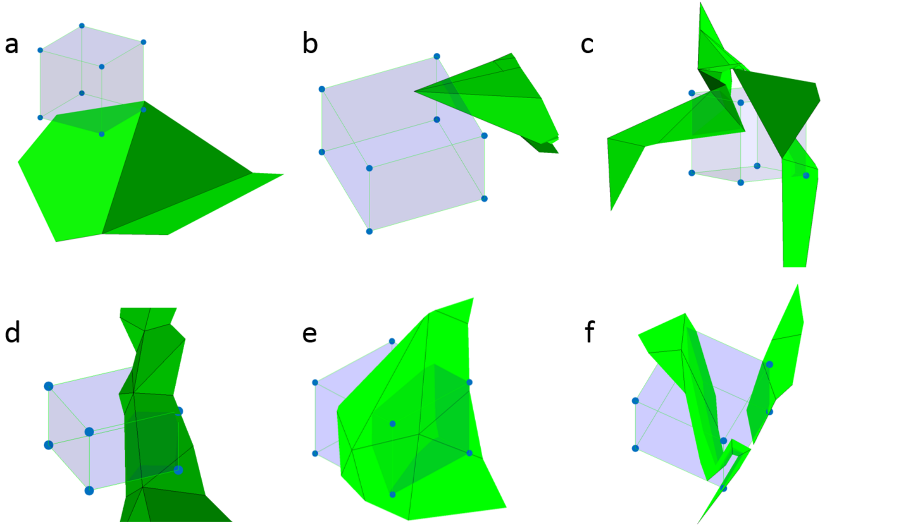


Fig. 3 Various subvoxel/surface configurations. a) no intersection: whole-volume assignment; b) single intersection through one face: a small convex hull will be formed; c/d) two examples of single intersection, folded surface: further subdivision will be used; e) single intersection through multiple faces: a convex hull will be formed; f) multiple surface intersection (unconnected patches of surface, likely a sulcus): further subdivision will be used.

means that any PVE will be negligible. Finally, voxels that do not intersect the surface (fully interior or exterior) are given single-class volumes according to tests of their centre points. Structures defined by a single surface (e.g. the thalamus) require no further processing: the estimates produced by the aforementioned steps may be used directly for PVEc.

B. Multiple-surface structures

Structures that are defined by multiple surfaces require further processing to yield PV estimates for all tissues of interest. With specific reference to the cortex, PVs within each hemisphere are obtained with the relations:

$$PV_{WM} = P_{inner}$$
$$PV_{GM} = \max(0, P_{outer} - P_{inner})$$
$$PV_{NB} = 1 - (PV_{WM} + PV_{GM})$$

where P_{inner} and P_{outer} denote the interior/exterior PV fractions associated with the inner and outer surfaces of the cortex respectively and PV_{WM}, PV_{GM} and PV_{NB} denote the PV estimates for WM, GM and non-brain tissue (the latter including cerebrospinal fluid, CSF). These equations are structured to account for a potential surface defect whereby the surfaces of the cortex swap relative position (the inner lying exterior to the outer) around the corpus collosum. The structure of the above relations (*N* surfaces leading to *N*+1 tissue classes) could easily be generalised to structures defined by more than two surfaces (for example, sublayers of the cortex, as used in laminar fMRI). A similar set of equations is used to merge hemisphere-specific results to cover the whole cortex, accounting for voxels lying on the mid-sagittal plane that intersect both hemispheres.

C. Whole-brain PV estimation

Toblerone, as outlined above, operates on a structure-by-structure basis in which the output tissue types are dependent on the structure in question. A number of methods utilising the core algorithm were implemented:

1) *estimate_structure:* estimate the inner and outer PVs associated with a structure defined by a single surface

2) *estimate_cortex:* estimate the GM, WM and non-brain PVs associated with the four surfaces of the cortex (l/r white/pial in the FreeSurfer terminology)

3) *estimate_all:* a combination of the *structure* and *cortex* methods above, this estimates PVs for the cortex and all subcortical structures identified by FIRST and combines them (with the exception of the brain stem) into a single set of GM, WM and non-brain PV estimates. The run-time for a typical subject was around 25 minutes.

The combination of FreeSurfer/FIRST and *estimate_all* provides a complete pipeline for obtaining whole-brain PV estimates in an arbitrary reference voxel grid from a single T1 structural image that may be used as a replacement for existing volumetric tools such as FAST. There is however a key conceptual difference between surface and volumetric methods concerning their interpretation of subcortical structures. Due to differences in tissue composition around the brain, cortical and subcortical GM have different intensities on a normal T1 image and are accordingly assigned different GM PVs by volumetric tools such as FAST (whereby cortical GM is seen as more ‘grey’ than subcortical, as illustrated in fig. 12). Surface based methods, by contrast, do not take a view on what tissue lies within the surface other than simply asserting that it is different to that which lies without. When combining the PVs of individual structures in Toblerone’s *estimate_all* function, all tissue within the cortex and subcortical structures is interpreted as pure GM. The practical implication of this is that Toblerone’s estimates for subcortical GM are higher than those produced by FAST. For this reason, the conventional GM/WM/CSF tissue classes used by volumetric tools may be better thought of within Toblerone’s framework as *tissue of interest*, *other tissues* and *non-brain,* though for the purposes of this article the familiar names GM and WM shall be used alongside non-brain. The inherent ambiguity in determining which tissues lie outside subcortical structures, which could be either WM or CSF depending on their location within the brain, was resolved using FAST’s segmentation results[2].

IV. Evaluation

Three datasets and three comparator methods were used, as summarised in Table I. The two surface-based comparator methods were restricted to use in the cortex only. By contrast, Toblerone was run on both cortical and subcortical surfaces where appropriate to provide whole-brain PV estimates.

TABLE I
DATASETS & METHODS USED

name	Simulated surfaces	BrainWeb	HCP test-retest
type	S	V + S	V + S
resolution	-	1mm iso.	0.7mm iso.
size	1 cortical hemisphere	18 simulated T1 images	45 subjects, 2 sessions each
ground truth	numerical method	volumetric segmentation*	N/A
comparator methods	NeuroPVE (S) RC (S)	RC** (S) FAST (V)	RC* (S) FAST (V)

S *surface,* V *volumetric,* RC *ribbon-constrained method*
* *established via automatic segmentation with manual intervention*
** *RC can only be run on the cortex for these datasets*

A. Comparator methods

The first surface-based comparator method, the ribbon-constrained (RC) algorithm, was developed for use with BOLD data in the HCP’s *fMRISurface* pipeline [8] and is restricted to the cortex only. The method assumes vertex correspondence between the two surfaces of the cortex and works as follows. For each vertex in turn, the outermost edges of the triangles that surround said vertex are connected between the two surfaces to form a 3D polyhedron representing a small region of cortex.

[2] As it is ambiguous as to what tissue lies outside a given subcortical structure given only its surface, FAST’s results for the same voxel are used as an estimate for the local ratio of WM and CSF. The actual quantity of non-GM tissue is still calculated from the surface estimate as the remainder 1 – GM, which is then shared between the other two classes in this ratio.

Commented [TK13]: This is entirely new and should be read in conjunction with the comment on section C(3). We are now able to run FreeSurfer and FIRST output through and get a single set of PV estimates for the whole brain in much the same way FAST operates. We cannot do the same with RC and Neuro as they are cortex only methods

Commented [TK14]: An overview of all datasets and methods used.

Commented [TK11]: This is a new section

Commented [TK12]: With the addition of the BrainWeb and HCP test-retest datasets (as discussed later), we were able to expand the scope of our experiments to do a direct comparison of Toblerone and FAST on whole-brain PV estimation, whereas previously we were restricted to the cortex only

Nearby voxels are subdivided and the subvoxels centres tested to determine if they lie interior to the polyhedron. The subdivision factor used in this work was the higher value of either ceil$(\max(\mathbf{v}) / 0.4)$ or 4, where $\mathbf{v}$ is the vector of voxel dimensions. The fraction of subvoxel centres lying within any cortical polyhedron gives the cortical GM PV, which, as the BOLD signal is predominantly cortical in origin, is the quantity of interest for this modality. In order to obtain WM and non-brain PVs, the following post-processing steps were used. Firstly, the unassigned PV of each voxel was calculated as $1 - PV_{GM}$, which was subsequently labelled as either WM or non-brain according to a signed-distance test of the voxel centre in comparison to the cortical mid-surface: for a voxel with centre point outside the mid-surface, the unassigned PV was labelled as non-brain. A weakness of this approach is that it is unable to faithfully capture a voxel in which all three tissues are present; only the combinations WM/GM or GM/non-brain are permitted. As voxel size increases, the probability of voxels containing multiple tissues also increases; testing on a brain image of 3mm isotropic resolution showed that around 30% of voxels intersecting the cortical ribbon contain three tissues. Resampling can be used to mitigate this effect so two variants of this method were tested: 'RC', direct estimation at each resolution, and 'RC2', estimation at 1mm followed by resampling to other resolutions via the process in section IV.B. The run-time for a typical subject was around 15 minutes.

This second surface method, NeuroPVE [15], uses a voxelisation method based on the work of [9,12], applied in a brain-specific context and again restricted to the cortex only. Multiple expanded and contracted copies of each surface are created and the ratio of expanded to contracted surfaces intersecting a given voxel is used as a first approximation for partial volumes. This ratio is then mapped, along with surface orientation information, via trigonometric relations on the unit cube into a PV estimate. The estimates produced take discrete values according to the number of surfaces used (in this work the default of 5). The intended use of this tool was PV estimation at structural, not functional, resolution, so two variants were tested: 'Neuro', direct estimation at arbitrary resolutions, and 'Neuro2', estimation at structural resolution followed by resampling to other resolutions via the process in section IV.B. On the basis of NeuroPVE's results on the simulated surfaces, it was excluded from further analysis. As the process of surface inflation is slow, the run-time for a typical subject was around 12 hours.

Finally, FSL's FAST [7] is an established whole-brain volumetric segmentation tool that was used as a comparator for the surface methods. On both the BrainWeb and HCP test-retest datasets, FAST was run on the brain-extracted images at structural resolution (1mm and 0.7mm iso. respectively). PVs were then obtained at other resolutions via the resampling method detailed in section IV.B. The run-time for a typical subject was around 5 minutes.

B. Resampling

Resampling is an interpolation operation that is used to transform volumetric data between voxel grids (in this context, from structural to functional resolution). FSL's *applywarp* tool was used with the *-super* flag for all resampling operations. This works by creating an up-sampled copy of the target voxel grid onto which values from the input image are sampled. The average is then taken across the voxel neighbourhoods of the high-resolution grid (sized according to the up-sampling factor) to obtain the result in the target voxel grid. Such an approach is appropriate when moving from fine to coarse as each output voxel corresponds to multiple input voxels, the individual contributions of which should be accounted for to preserve overall tissue proportions. When using *applywarp* a transformation matrix between the input and output voxel grids must be given as the *-premat* argument; to denote identity for the purposes of this work, the output of the HCP *wb_command –convert-affine –from-world –to-flirt* tool operating on $\mathbf{I}_4$ was used as the *-premat* to correct for a subvoxel shift that arises due to FSL coordinate system conventions. Note that for perfectly aligned voxel grids with an integer ratio of voxel sizes, such as a 1mm and 2mm isotropic grid, this process is equivalent to averaging across blocks of the smaller grid (sized 2x2x2 in this case).

C. Simulated surfaces

A pair of concentric surfaces, illustrated fig. 4, were designed to capture geometric features relevant to the anatomy of a cortical hemisphere. These were produced by modulating the radius of a sphere as a function of azimuth θ and elevation ϕ to produce sulci and gyri-like features. The radius of the inner surface was defined as

$$r_{in} = 60\left(1 - 0.1\max\left(\sin^{20} 5u\,, \sin^{20} 5v\right)\right)$$

where 60 is the unmodulated radius of the sphere, 0.1 fixes the relative depth of sulci, the max function prevents sulci from constructively interfering to produce deep wells at points of intersection, the power of 20 produces broad gyri and narrow sulci, and the substitutions $u = \phi + \theta$, $v = \phi - \theta$ cause the sulci to spiral around the sphere in opposite directions. Modulation was restricted to the range $-2\pi/5 \leq \theta \leq 2\pi/5$ to leave the poles smooth and suppress unrealistic features. The outer radius was set at $r_{out} = 1.05 \cdot r_{in}$, leading to a peak radial distance between surfaces of 3mm. The outermost region was taken to represent non-brain tissue, the innermost WM and the region in between GM. The use of analytic functions to define the surfaces permitted ground truth maps to be calculated using a numerical method. Voxels were sampled at 4,096 elements per mm^3 and the positions of these sample points expressed in spherical polar coordinates. By comparing the actual radius of each point to the calculated radius of the surface boundaries for the same azimuth and elevation, the tissue type of the sample

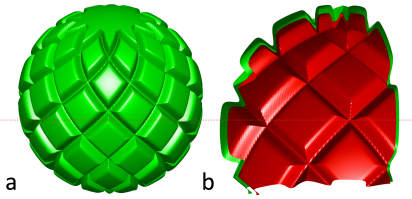


Fig. 4 a) Simulated surfaces; b) cutaway showing inner (red) and outer (green) surfaces. Peak radial distance between the two was 3mm.

Commented [TK15]: In response to reviewer

Commented [TK19]: Re-phrased compared to previously.

Commented [TK20]: Re-phrased

Commented [TK21]: An explicit statement of how resampling on perfectly aligned voxel grids

Commented [TK16]: In response to reviewer.

Commented [TK22]: Explicit statement of what 'tissue' we refer to

Commented [TK17]: In response to reviewer.

Commented [TK18]: In response to reviewer point within the structure could be determined, and from there PVs obtained by aggregating results within voxels. This is referred to as the 'numerical solution' in the results section. Mean surface node spacing was set at 0.85mm, similar to that of native FreeSurfer output. Toblerone's *estimate_cortex*, NeuroPVE and the RC method were used on this dataset. PVs were obtained at voxel sizes of 1 to 3mm in steps of 0.2mm isotropic.

D. BrainWeb simulated T1 images

BrainWeb [16], [17] simulates whole-head T1 images at 1mm isotropic resolution with specified levels of random noise and field non-uniformity (NU). Eighteen images were produced to cover the available parameter space of noise levels {0, 1, 3, 5, 7, 9} and NU levels {0, 20, 40} (both quantities in percent). These were run through FAST, FIRST and FreeSurfer, after which Toblerone's *estimate_all* and the RC method (cortex only) were used on the output. FAST's output was also used to enable a comparison between surface and volumetric methods. PVs were obtained at voxel sizes of 1 to 4mm in steps of 1mm isotropic. Although ground truth PV maps exist for this dataset (produced by automatic volumetric segmentation of T1 images with manual correction [16]), both surface and volumetric methods returned significantly different results to these, raising the complicated question of determining which set of results is correct. In order to avoid making this judgement, each method was instead referenced to its own results on the ideal T1 image (0% noise 0% NU) in the 1mm isotropic voxel grid of the structural images. The voxel grids associated with each voxel size were aligned such that results at 1mm could be used to calculate a reference at other sizes (for example, summing across 3x3x3 blocks to get a 3mm reference).

E. Human Connectome Project test-retest data

This dataset comprises 45 subjects from the main HCP cohort who underwent two separate structural scan sessions (mean age 30.2 years, mean time between sessions 4.8 months). Each session was processed using the pipeline in [8] to obtain cortical surfaces via FreeSurfer. Separately, the distortion-corrected T1 images were fed into FAST (brain-extracted) and FIRST (whole-head) to produce volumetric segmentations and subcortical surfaces. Toblerone's *estimate_all* and the RC method (for the cortex only) were used on this dataset, as well as FAST for a comparison between surface and volumetric methods. PVs were obtained at voxel sizes of 1 to 3.8mm in steps of 0.4mm isotropic, as well as the native 0.7mm isotropic voxel grid of the structural images. Although a ground truth is not defined for this dataset, each method's results from the first session were used as a reference for the second session.

F. Evaluation metrics

Errors were measured in both a per-voxel (root-mean-square, RMS, of individual voxel errors) and aggregate (total tissue volume) sense. The former basis is important as PVEc is locally sensitive to the PV estimates [18]; the latter basis reflects systematic bias at the aggregate level. All error quantities are expressed in percent and map directly to PV estimates without scaling: for example, a PV estimate of 0.5 against a reference value of 0.55 corresponds to an error of -0.05 or -5%.

A further analysis of voxel-wise differences between Toblerone and FAST was performed on the HCP dataset at multiple voxel sizes by sorting voxels into 5% width bins according to their Toblerone GM PV estimate. The difference (Toblerone – FAST) was calculated for each voxel and the mean taken across each bin. This quantity was then averaged across subjects and sessions (weighted to respect differences in brain volume).

V. Results

A. Simulated surfaces

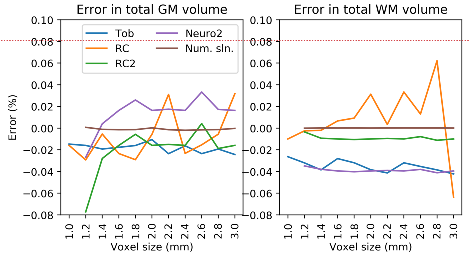


Fig. 5 Simulated surfaces: error in total tissue volume. Toblerone showed consistency, though with small bias, for both GM and WM. RC1 errors were lower for GM than WM. Resampling-based methods (RC2, Neuro2) showed particular consistency in WM. [Full results in supplementary, fig. s5]

Fig. 5 shows the error in total tissue volume for the simulated surfaces. The numerical solution at 1mm was used as the reference. Toblerone showed consistency across voxel sizes, though with a small negative bias in both tissues. RC estimates showed variation in both. The resampling-based methods RC2 and Neuro2 showed high consistency in WM but less so in GM. The numerical solution was stable across voxel sizes. Neuro's results are excluded from this and subsequent graphs for clarity; the full results are given in the supplementary material (figs. s5 and s6).

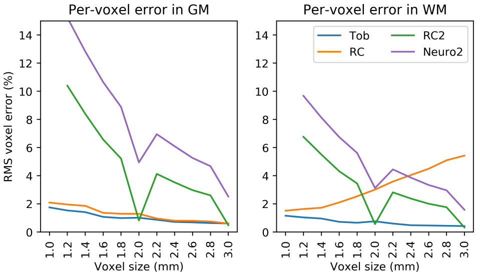


Fig. 6 Simulated surfaces: per-voxel error. Toblerone and RC produced the lowest errors in GM; in WM there was a clear difference to Toblerone. RC2 and Neuro2's errors both decreased with increasing voxel size, with a characteristic notch observed at 2mm. [Full results in supplementary, fig. s6]

Fig. 6 shows per-voxel error for the simulated surfaces. Results were masked to consider voxels intersecting either surface of the cortex as only these contain PVs. Toblerone and RC produced the lowest errors at all voxel sizes in GM; in WM only Toblerone retained this behaviour. Both resampling-based

Commented [TK23]: The numerical method multiple reviewers enquired about.

Commented [TK28]: This analysis was also performed previously but we now discuss the methodology here.

Commented [TK24]: A more thorough investigation than before, in order to tease out the effects of resampling (discussed later)

Commented [TK25]: This forms one of the two major changes.
We are investigating the robustness of a surface based approach in the presence of 'real' scanner artefacts and imperfections, through the use of a T1 MRI simulator.

Commented [TK29]: We now include the numerical result at sizes other 1mm to show its stability.

Commented [TK30]: As previously: the magnitude of errors for Neuro is so high that it dwarfs other methods. As requested by reviewers, we do include resampling based methods on this graph, although theoretically they should be very stable between voxel sizes.

Commented [TK26]: This forms the second of two major changes.
The test-retest data provides us with a reference that was not previously available when using the 100 HCP subjects. Although we still do not have ground truth partial volume maps, we do at least have a subject-specific reference (the session 1 data) to which we can compare session 2 for repeatability.

Commented [TK27]: New per-voxel error metric methods (RC2, Neuro2) produced lower errors as voxel size increased, and a characteristic notch in their results was observed at 2mm. Although RC initially performed better than RC2 in WM, the inverse was true above 2mm voxel size.

B. BrainWeb simulated T1 images

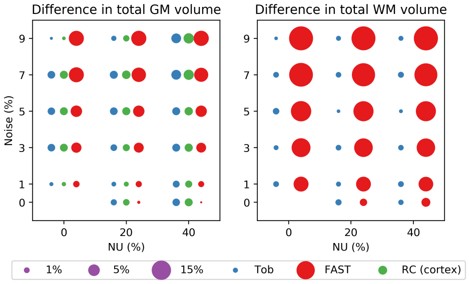


Fig. 7 BrainWeb: difference in total tissue volume referenced to each method's 0% noise 0% NU result. Surface-based methods were more consistent at almost all noise and NU levels; FAST was more consistent in GM than WM.

Fig. 7 shows the difference in total tissue volume across the brain as a function of noise and NU levels, referenced to each method's results at 0% noise and 0% NU. PV estimates at 1mm isotropic voxel size were used for this analysis. RC's GM result was for the cortex only as it cannot process subcortical structures. In general, the surface-based methods showed more consistency in their estimates across all levels of noise and NU, with the notable exception of GM at 40% NU. FAST's consistency was notably better in GM than WM.

Fig. 8 shows the RMS per-voxel difference in PV estimates at 3mm voxel size as a function of noise and NU. Each method's 1mm results at 0% noise 0% NU were used as the reference. Toblerone returned lower RMS voxel differences in both GM and WM at all levels of NU and noise except 0% noise 0% NU; a pattern that was repeated at other voxel sizes (these are shown in supplementary fig. s8).

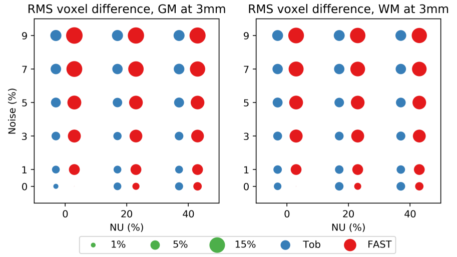


Fig. 8 BrainWeb: RMS per-voxel differences at 3mm voxel size, referenced to each method's 1mm 0% noise 0% NU results. Toblerone's differences were smaller at almost all levels of noise and NU, as was also the case at other voxel sizes. [Results for other voxel sizes are given in supplementary fig. s8]

C. HCP test-retest subjects

Fig. 9 shows violin plots of inter-session difference (retest minus test) in tissue volume across the 45 subjects of the HCP dataset. PV estimates at 0.7mm isotropic voxel size were used for this analysis. RC's GM result was for the cortex only. Both surface methods gave a tighter distribution than FAST, suggesting greater repeatability between sessions. All methods showed greater variability in GM than WM.

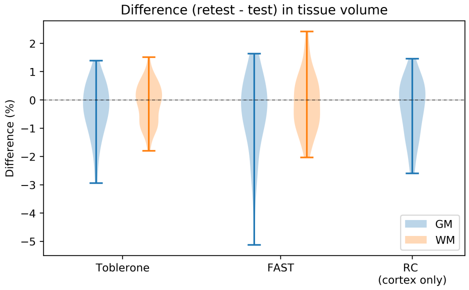


Fig. 9 HCP test-retest: inter-session (retest minus test) difference in total tissue volume. PVs were estimated in the native 0.7mm isotropic space of the structural images. RC's result is for the cortex only. Both surface methods show a tighter distribution than FAST.

Fig. 10 shows the mean per-voxel difference between Toblerone and FAST's GM PV estimates as a function of Toblerone's GM PV estimate. Excepting the 0.7mm result, the positive slope of each line shows that in voxels with a low Toblerone GM PV estimate, FAST was more likely to assign a higher value, and vice-versa at high Toblerone GM PV estimates. The strength of this relationship decreased with increasing voxel size. It should be noted that the 0.7mm result is the only one *not* to make use of resampling (for all others, FAST's 0.7mm estimates were resampled onto the target voxel grid).

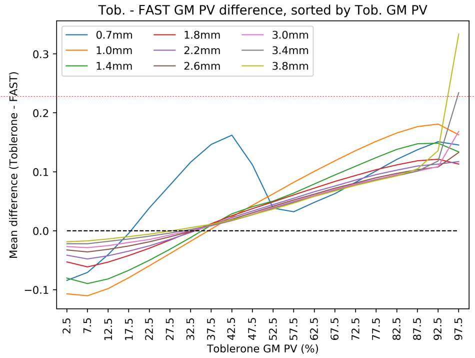


Fig. 10 HCP test-retest: mean difference between Toblerone and FAST GM PVs, sorted into 5% width bins according to Toblerone's GM PV. As Toblerone's GM PV estimate in a given voxel increases, FAST is more likely to assign a smaller value, and vice-versa. The strength of this relationship decreases with increasing voxel size. An inverse, but weaker, effect is seen for WM (supplementary fig s10).

Commented [TK31]: This effect is discussed later and is more readily visible within the range of voxel sizes now investigated. This hopefully clears up a question the reviewers asked of us.

Commented [TK34]: Entirely new analysis

Commented [TK32]: Entirely new analysis

Commented [TK35]: An analysis that was previously included, but looks different due to the differing resolutions used. In the last submission, 1mm was used as the base voxel size, which meant that the base set of PVs used for all FAST results were in fact resampled from 0.7mm. In this submission, we used the 0.7mm space of the HCP structural images as the base, from which all other results were resampled. In summary, this graph now looks different because the input data used to produce FASTs results at various resolutions has changed.

Commented [TK33]: Entirely new analysis

VI. Discussion

Results from the simulated surfaces showed that Toblerone produced estimates with a comparatively low and consistent error. Although the RC method was able to perform similarly for GM, there was a clear advantage for Toblerone in WM. Results from the BrainWeb images suggested that a surface-based approach (the combination of FreeSurfer/FIRST and Toblerone) is more robust to noise and NU than FAST's volumetric approach. Further analysis suggested it is the consistency of FAST's WM estimates that suffers in the presence of these scanner imperfections. Finally, results from the HCP test-retest dataset showed that the surface-based approach provide better inter-session repeatability in total tissue volume.

The use of resampling – unavoidable for all volumetric methods that must transform PV estimates from a structural voxel grid to a functional voxel grid – degrades data quality in an unpredictable and highly localised manner. This chiefly arises due to so-called *subvoxel effects*, which may be illustrated via the following 1D example. Consider a row of voxels of size 1mm that are to be resampled onto 1.4mm voxels. A larger voxel overlaps evenly onto two smaller voxels, covering 0.7mm of each. The resampled value will be the mean of the two smaller, on the implicit and unlikely assumption that the tissues within each are evenly distributed. Next, consider a row of 1mm voxels that are to be resampled onto 3.4mm voxels, whereby a larger voxel overlaps by 0.2, 1, 1, 1 and 0.2mm onto smaller voxels. Again, the resampled value will be a weighted mean of the smaller voxels, but as the central three voxels are included wholly in the new voxel, the spatial distribution of tissues within these voxels is irrelevant and the assumption of even distribution can safely be made. As the ratio of output voxel size to input voxel size increases, the significance of subvoxel effects are therefore reduced.

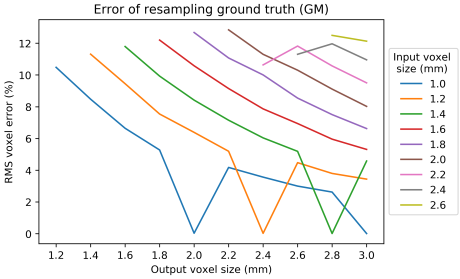


Fig. 11 Simulated surfaces: error induced by resampling the ground truth GM PV map, masked to voxels intersecting either surface of the cortex. As the input voxel size increases, error increases, but as the ratio output / input voxel size increases, error falls. Finally, error falls to zero when the ratio takes an integer value.

It is extremely difficult to quantitatively measure the impact of resampling, particularly on non-simulated data. To do so would require the ability to express some volumetric reference data in an arbitrary voxel grid *without* making use of resampling, otherwise a trap of circular reasoning results. Nevertheless, such an analysis can be performed using the simulated surfaces presented earlier. The key conceptual difference is that the ground truth for this dataset is defined by a surface and can therefore calculated in any voxel grid without resampling. Fig. 11 shows the results of resampling ground truth results from the numerical method at each resolution to all other resolutions above the one in question (for example, the 1.4mm truth was resampled to 1.6, 1.8, … etc). At each voxel size, the resampled results can be compared to a ground truth that has been calculated without the use of resampling. RMS per-voxel error was measured using the same mask as before, namely all voxels intersecting either surface of the cortex, as only these contain PVs. Multiple trends can be seen: firstly, as the input voxel size increases, error at all output voxel sizes increases. Secondly, as the ratio of output to input voxel size increases, the error decreases. Finally, the error falls to zero when this ratio takes an integer value. This is due to the use of perfectly aligned voxel grids in this work (which would not be the case with patient data) and is discussed in section IV.B. This likely explains the interesting behaviour observed in various analyses, namely: the notches seen in fig. 6, as well as supplementary figs. s5 and s6 (perfect voxel correspondence means no subvoxel effects); the 0.7mm result in fig. 10 (for all other sizes, resampling by a non-integer ratio of voxel size blurs the FAST results, reducing image contrast and the number of high GM PV voxels); and the lack of error observed in FAST's GM and WM results at 2, 3 and 4mm voxel size, 0% noise 0% NU in figs. 8 and s8 (again, perfect voxel correspondence with the reference set of 1mm estimates). These considerations do not apply to surface-based methods as they do not make use of resampling.

A further advantage of surface-based methods concerns their application of transformations. Notwithstanding the fact that volumetric methods require resampling to transform data from one resolution to another, they also require it to apply a registration transformation between the structural voxel grid in which PVs are estimated and the functional voxel grid in which PVEc is to be performed. Once again, the impact of this upon data quality is highly localised and difficult to measure quantitatively. It can however be illustrated via the following experiment, illustrated in fig. 12. GM PV maps for the 0% noise 0% NU BrainWeb image were translated by 0.5mm in each of the *x,y,z* axes. For FAST, this translation took the form of an affine transformation applied during a resampling operation. Significant blurring is seen, particularly around the edges of structures where there was previously good edge definition. As these edge voxels by definition contain PVs this is a particularly undesirable outcome. By contrast, blurring within a structure is of little consequence as the tissue is already homogenous. For Toblerone, this translation was performed by shifting the *surfaces* into the new reference voxel grid represented by the translation and then estimating the PVs afresh with no noticeable reduction in edge definition.

In its native form, the RC method is unable to correctly handle voxels in which all three tissue types are present (due to the fact that it estimates GM first and then assigns the remainder to either WM or non-brain). The impact of this is seen in the positive relationship between per-voxel error in WM and voxel size in fig. 6. Resampling can help to minimise this error: at

Commented [TK37]: In response to various reviewer questions

Commented [TK38]: New discussion

Commented [TK39]: New figure

Commented [TK36]: New discussion. Ideally we would have investigated resampling effects on 'real' data but cannot find a way to do so that does not require resampling to produce a reference. The simulated surfaces are the next best thing.

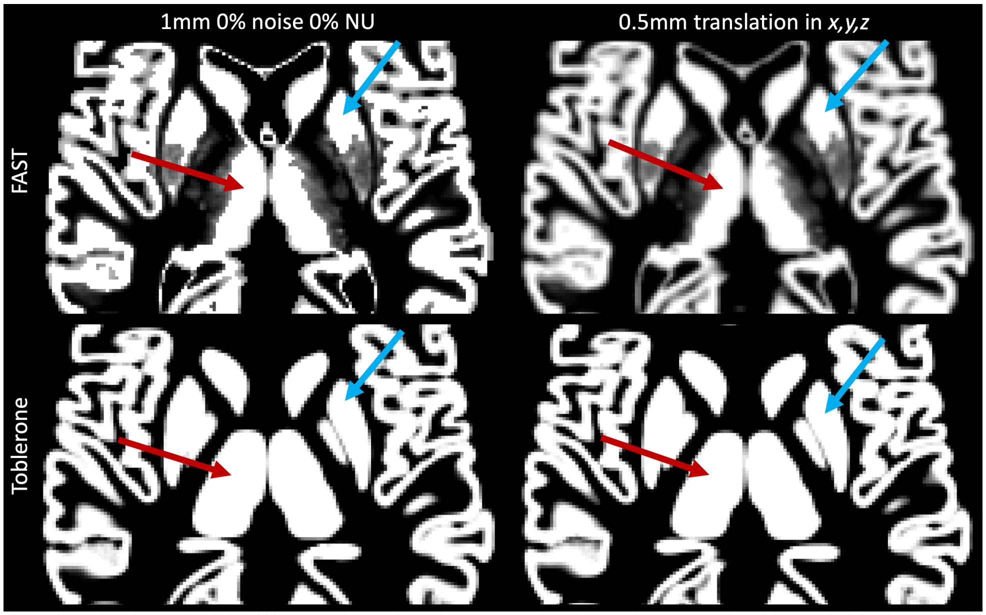


Fig. 12 Illustration of resampling-induced blurring on the 1mm isotropic GM PV map from the 0% noise 0% NU BrainWeb image. The left column shows the original estimates produced by FAST and Toblerone, the right shows the result of a 0.5mm translation along each axis. The left thalamus (red) and right putamen (blue) are highlighted in each, showing how surface and volumetric methods differ markedly in their interpretation of subcortical structures (FAST does not regard them as pure GM, whereas Toblerone does for the analyses presented in this work).

small voxel sizes, the probability of voxels containing three tissue types is smaller, so the error is minimised, but this does not hold true at larger voxel sizes. Accordingly, as voxel size increases, it is increasingly beneficial to obtain PV estimates by resampling those from a smaller voxel size. Set against this, however, are the aforementioned problems introduced by resampling: when the ratio of output to input voxel size is small, subvoxel effects are significant and high per-voxel errors result (as shown in fig. 6). A threshold voxel value above which resampling is beneficial therefore results (at around 2mm in the figure). The exact value of this threshold would be difficult to predict in the general case (in particular, the use of aligned voxel grids in this work is both highly significant and extremely unrealistic). By contrast, Toblerone is able to produce consistent estimates in all tissue classes at arbitrary voxel sizes without the use of resampling.

We were unable to further analyse the HCP test-retest dataset in order to establish where in the brain the differences between Toblerone's and FAST's estimates arise. As this would require extensive use of non-linear registrations and resampling to transform all subjects onto a common template, it is likely that the artefacts imposed by this process would obscure the true methodological differences of interest. Furthermore, an analysis on the BrainWeb database would be of limited use as this only represents the cortical anatomy of a single subject and would therefore ignore population variability.

VII. Conclusion

Toblerone is a new method for estimating PVs using surface segmentations. Unlike existing surface-based tools, it is not closely tied to any specific modality or structure and can therefore be adapted to multiple use cases (notably, providing PV estimates for the whole brain). It is able to operate at arbitrary resolutions without recourse to resampling, thereby avoiding the highly localised degradation of image quality that this process entails. Three datasets have been used to evaluate the algorithm. Results from simulated surfaces show consistently low errors at both the voxel and aggregate level, either matching or surpassing other surface-based methods. Results on simulated T1 images from the BrainWeb database show that a FreeSurfer/FIRST/Toblerone surface-based pipeline used as an alternative to FAST is more robust in the presence of random noise and field non-uniformity. Finally, results from the HCP test-retest dataset of 45 subjects show that the surface-based pipeline produces a tighter distribution of inter-session tissue volumes than FAST, suggesting the surface approach has greater repeatability. The magnitude of methodological differences observed in this work, and related conceptual questions concerning the interpretation of subcortical tissue between surface and volumetric methods, will have implications for the wider process of PVEc.

Commented [TK40]: A more thorough discussion of the resampling tradeoff (referred to as the 'cross-over point' previously)

Commented [TK42]: New conclusions in light of the additional analyses performed here.

Commented [TK41]: In response to a question asked by many: where in the brain do the differences arise? The methodological implications of performing such an analysis in volume space across subjects are complex given natural population variability.

Commented [TK43]: The focus of future work, but it is reasonable to expect the implications will be modality-specific, in contrast to the thrust of this paper which aims to demonstrate a modality-independent approach.

Acknowledgement

The authors would like to thank Gabriel Mangeat of Polytechnique Montréal for assistance with the NeuroPVE method. The University of Oxford Advanced Research Computing (ARC) facility was used in preparing this work (http://dx.doi.org/10.5281/zenodo.22558). Data was provided by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University.

References

[1] B. Fischl and A. M. Dale, "Measuring the thickness of the human cerebral cortex from magnetic resonance images," *Proc. Natl. Acad. Sci.*, vol. 97, no. 20, pp. 11050 LP – 11055, Sep. 2000.

[2] H. W. Müller-Gärtner, J. M. Links, J. L. Prince, R. N. Bryan, E. McVeigh, J. P. Leal, C. Davatzikos, and J. J. Frost, "Measurement of Radiotracer Concentration in Brain Gray Matter Using Positron Emission Tomography: MRI-Based Correction for Partial Volume Effects," *J. Cereb. Blood Flow Metab.*, vol. 12, no. 4, pp. 571–583, Jul. 1992.

[3] I. Asllani, A. Borogovac, and T. R. Brown, "Regression algorithm correcting for partial volume effects in arterial spin labeling MRI," *Magn. Reson. Med.*, vol. 60, no. 6, pp. 1362–1371, Sep. 2008.

[4] M. A. Chappell, A. R. Groves, B. J. MacIntosh, M. J. Donahue, P. Jezzard, and M. W. Woolrich, "Partial volume correction of multiple inversion time arterial spin labeling MRI data," *Magn. Reson. Med.*, vol. 65, no. 4, pp. 1173–1183, 2011.

[5] D. W. Shattuck, S. R. Sandor-Leahy, K. A. Schaper, D. A. Rottenberg, and R. M. Leahy, "Magnetic resonance image tissue classification using a partial volume model," *Neuroimage*, vol. 13, no. 5, pp. 856–876, 2001.

[6] B. Fischl, "FreeSurfer," *Neuroimage*, vol. 62, no. 2, pp. 774–781, 2012.

[7] Y. Zhang, M. Brady, and S. Smith, "Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm," *IEEE Trans. Med. Imaging*, vol. 20, no. 1, pp. 45–57, 2001.

[8] M. F. Glasser, S. N. Sotiropoulos, J. A. Wilson, T. S. Coalson, B. Fischl, J. L. Andersson, J. Xu, S. Jbabdi, M. Webster, J. R. Polimeni, D. C. Van Essen, and M. Jenkinson, "The minimal preprocessing pipelines for the Human Connectome Project," *Neuroimage*, vol. 80, pp. 105–124, 2013.

[9] D. N. Greve, C. Svarer, P. M. Fisher, L. Feng, A. E. Hansen, W. Baare, B. Rosen, B. Fischl, and G. M. Knudsen, "Cortical surface-based analysis reduces bias and variance in kinetic modeling of brain PET data," *Neuroimage*, vol. 92, pp. 225–236, May 2014.

[10] D. N. Greve, D. H. Salat, S. L. Bowen, D. Izquierdo-Garcia, A. P. Schultz, C. Catana, J. A. Becker, C. Svarer, G. M. Knudsen, R. A. Sperling, and K. A. Johnson, "Different partial volume correction methods lead to different conclusions: An (18)F-FDG-PET study of aging," *Neuroimage*, vol. 132, pp. 334–343, May 2016.

[11] B. Patenaude, S. M. Smith, D. N. Kennedy, and M. Jenkinson, "A Bayesian model of shape and appearance for subcortical brain segmentation," *Neuroimage*, vol. 56, no. 3, pp. 907–922, Jun. 2011.

[12] F. S. Nooruddin and G. Turk, "Simplification and repair of polygonal models using volumetric techniques," *IEEE Trans. Vis. Comput. Graph.*, vol. 9, no. 2, pp. 191–205, 2003.

[13] T. Akenine-Möller, "Fast 3D Triangle-Box Overlap Testing," *J. Graph. Tools*, vol. 6, no. 1, pp. 29–33, 2001.

[14] C. B. Barber, D. P. Dobkin, and H. Huhdanpaa, "The Quickhull Algorithm for Convex Hulls," *ACM Trans. Math. Softw.*, vol. 22, no. 4, pp. 469–483, Dec. 1996.

[15] C. Van Assel, G. Mangeat, B. De Leener, N. Stikov, C. Mainero, and J. Cohen, "Partial volume effect correction for surface-based cortical mapping," *Proc. Int. Soc. Magn. Reson. Med. Annu. Meet. Paris*, 2017.

[16] D. L. Collins, A. P. Zijdenbos, V. Kollokian, J. G. Sled, N. J. Kabani, C. J. Holmes, and A. C. Evans, "Design and Construction of a Realistic Digital Brain Phantom," *IEEE Trans. Med. Imaging*, vol. 17, no. 3, pp. 463–468, 1998.

[17] C. Cocosco, V. Kollokian, R. K. Kwan, G. B. Pike, and A. C. Evans, "BrainWeb : Online Interface to a 3D MRI Simulated Brain Database," *3-rd Int. Conf. Funct. Mapp. Hum. Brain*, vol. 5, no. 4, p. S425, 1997.

[18] M. Y. Zhao, M. Mezue, A. R. Segerdahl, T. W. Okell, I. Tracey, Y. Xiao, and M. A. Chappell, "A systematic study of the sensitivity of partial volume correction methods for the quantification of perfusion from pseudo-continuous arterial spin labeling MRI," *Neuroimage*, 2017.

SUPPLEMENTARY MATERIAL

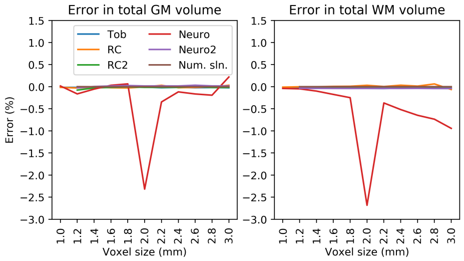


Fig. s5 Simulated surfaces: error in total tissue volume, all methods. The notch in the Neuro method at 2mm may arise due to an interplay between the number of expanded surfaces created (5) and the voxel size.

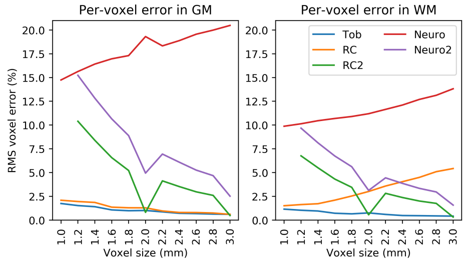


Fig. s6 Simulated surfaces: RMS per-voxel error. Neuro's results are significantly worse than all other methods at all other resolutions, though the resampled version (Neuro2) performs better.

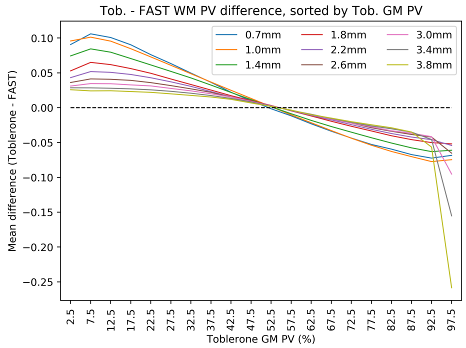


Fig. s10 HCP test-retest: mean difference between Toblerone and FAST WM PVs, sorted into 5% width bins according to Toblerone's GM PV. This is the analogue of fig. 10, showing a weaker and inverse relationship.

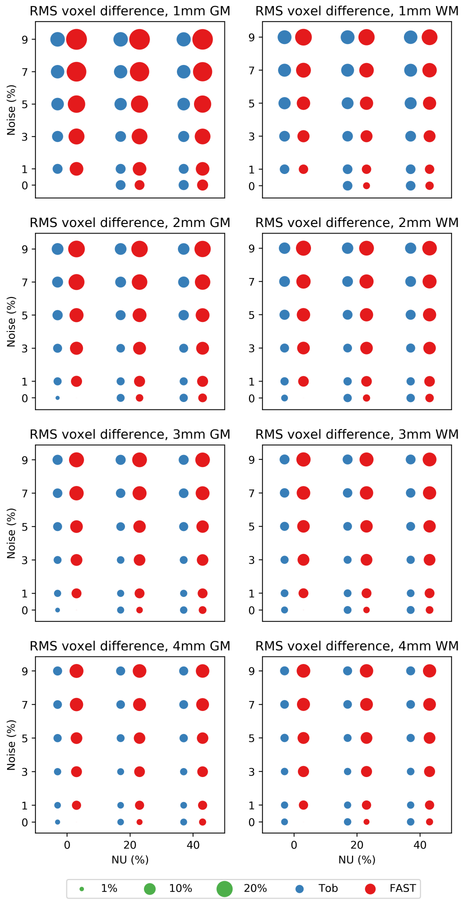


Fig. s8 BrainWeb: RMS per-voxel differences at voxel sizes of 1 to 4mm isotropic, referenced to each method's 1mm 0% noise 0% NU results. Toblerone's differences were smaller at almost all levels of noise and NU, the exception being 0% noise 0% NU.

Commented [TK44]: I appreciate this is a complex figure: the high dimensionality of the parameter space (noise, non-uniformity, method, tissue type and voxel size) does not lend itself well to a 2D representation! It is included for completeness with the full awareness that reviewers may object to it in the current form. Suggestions gratefully received.